Proceedings of the 35th WIC Symposium on Information Theory in the Benelux

and

The 4th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux

Eindhoven, the Netherlands
May 12–13, 2014

Organized by TU Eindhoven
Previous symposia

1. 1980 Zoetermeer, The Netherlands, Delft University of Technology
2. 1981 Zoetermeer, The Netherlands, Delft University of Technology
3. 1982 Zoetermeer, The Netherlands, Delft University of Technology
15. 1994 Louvain-la-Neuve, Belgium ISBN 90-71048-10-1
20. 1999 Haasrode, Belgium ISBN 90-71048-14-4
23. 2002 Louvain-la-Neuve, Belgium ISBN 90-71048-16-0

Proceedings

Edited by Boris Škorić and Tanya Ignatenko.
A catalogue record is available from the Eindhoven University of Technology Library.
The 35th Symposium on Information Theory in the Benelux and The 4th JointWIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux have been organized by

Signal Processing Systems group, Dept. of Electrical Engineering, TU Eindhoven
and
SEC group, Dept. of Mathematics and Computer Science, TU Eindhoven

on behalf of the

Werkgemeenschap voor Informatie- en Communicatietheorie, the IEEE Benelux Information Theory Chapter and the IEEE Benelux Signal Processing Chapter.

Organizing committee
Tanya Ignatenko (TU Eindhoven)
Boris Škorić (TU Eindhoven)

The organizing committee gratefully acknowledges the financial support of the Gauss Foundation for the “Best Student Paper Award”, of the IEEE Benelux Information Theory Chapter, and of the WIC for the “Best Student Presentation Award”.
Preface

The Werkgemeenschap voor Informatie- en Communicatietheorie (WIC) has organized the annual Symposium on Information Theory in the Benelux (SITB) since 1980. This year’s symposium, the 35th in the series, takes place in Eindhoven, the Netherlands. For the fourth time, it is organized jointly with the IEEE Benelux Signal Processing Chapter. The symposium is organized by TU Eindhoven. These proceedings contain the papers which were presented during the symposium. We are grateful to the authors for sharing their latest research with us.

This year we were very fortunate to have two IEEE SPS Distinguished Lecturers as keynote speakers: prof. Bhaskar Rao (University of California, San Diego) and prof. Alex Kot (Nanyang Technological University).

We gratefully acknowledge the sponsorship provided by the Gauss Foundation (presenting the Best Student Paper Award) and the IEEE Benelux Information Theory Chapter. Also, the WIC sponsored the Best Student Presentation Award.

We also express our sincere thanks to Diana Heijnerman and Jolande Matthijssse for their great help in the organization and hosting of the symposium.

We hope that this symposium offers a good opportunity to exchange knowledge and improve personal contacts among the participants, and results in novel scientific, cultural and personal discoveries.

Eindhoven, May 2014,
Tanya Ignatenko and Boris Škorić
(Symposium Organizers)
SITB 2014 program

Monday May 12

09:15   Registration

10:00   Opening

SESSION 1: WIRELESS NETWORKS
10:10   M. Mitici (UT)
        Energy-delay trade-off of wireless data collection in the plane

10:30   A. Tsiatmas (TUE, Philips)
        Maximal-ratio combining in optical wireless communications

10:50   T. Vermeulen (KUL)
        Energy-delay analysis of full duplex wireless communication for sensor networks

break

11:35   F. Rottenberg (UCL)
        Dominant eigenmode transmission in MIMO FBMC for frequency selective channels

11:55   Z. Xu (IMEC)
        Digital intensive architecture exploration for low-power 60 GHz polar transmitter

12:15   LUNCH

13:45   Invited speaker Bhaskar Rao
        Bayesian methods for sparse signal recovery

break

SESSION 2: CODING & CRYPTO
15:00   G. op 't Veld (EPFL)
        Successive refinement of Gaussian projections

15:20   Z. Ren (TUD)
        On the energy benefit of compute-and-forward on the hexagonal lattice

15:40   F. Rosas (KU Leuven)
        Increasing the energy-efficiency of point-to-point low-power wireless communications

break

16:20   L. Tolhuizen (Philips Research)
        HIMMO: a collusion-resistant, identity-based symmetric key establishment scheme

16:40   T. Laarhoven (TU/e)
        Asymptotics of fingerprinting and group testing

17:00-18:00   General Assembly
18:30-23:00   Social event: DAF Museum visit, including dinner
Tuesday May 13

09:00 Invited speaker Alex Kot
Is your biometric data safe?

10:00 break

SESSION 3: BIOMETRICS AND SURVEILLANCE
10:20 C. Zeinstra (UT)
Towards the automation of forensic facial individualisation: Comparing forensic to non forensic eyebrow features

10:40 D. Kostadinov (U. Geneva)
Visual information encoding for face recognition: sparse coding vs vector quantization

11:00 A. Dutta (UT)
Automatic eye detection error as predictor of face recognition performance

break

11:35 S. Javanbakhti (TU/e)
Automatic generic Region-Of-Interest selection for video surveillance applications

11:55 N. Jovanović (TU/e)
In-plane user positioning indoors

12:15 LUNCH

13:45 POSTER SESSION

SESSION 4: CLASSIFICATION
14:55 S. Ferdowski (U. Geneva)
Content identification: machine learning meets coding

15:15 N. Bouhouch (TU/e, AMC)
Information-theoretic measures for angiogenesis imaging by contrast-enhanced ultrasound

15:35 A. Jalalirad (TU/e)
Maximum-likelihood feature model selection

15:55 F. Farhadzadeh (TU/e)
Information theoretical analysis of identification based on active content fingerprinting

16:15 Awards and closing

16:30 Drinks
# Contents

## Invited presentations

Bhaskar D. Rao  
*Bayesian methods for sparse signal recovery* ................................................................. 1

Alex C. Kot  
*Is your biometric data safe?* .................................................................................................. 2

## Session 1: Wireless networks

M. Mitici, J. Goseling, M. de Graaf and R.J. Boucherie  
*Energy-delay trade-off of wireless data collection in the plane* ........................................... 3

A. Tsiatmas, C.P.M.J. Baggen and F.M.J. Willems  
*Diversity combining techniques for visible light communications* ........................................ 11

T. Vermeulen and S. Pollin  
*Energy-delay analysis of full duplex wireless communication for sensor networks* ... 19

F. Rottenberg and J. Louveaux  
*Dominant eigenmode transmission in MIMO FBMC for frequency selective channels* 27

Z. Xu, C. Li, A. Bourdoux, K. Khalaf, S. Pollin and L. Van der Perre  
*Digital intensive architecture exploration for low-power 60GHz polar transmitter* ... 35

## Session 2: Coding and Crypto

G.J. Op ‘t Veld and M.C. Gastpar  
*Successive refinement of Gaussian projections* ................................................................. 42

Z. Ren, J. Goseling, J. Weber and M.C. Gastpar  
*On the energy benefit of compute-and-forward on the hexagonal lattice* ....................... 50

F. Rosas, R. Demo Souza and C. Oberli  
*Comparative analysis of the optimization of error correcting codes and MIMO SVD schemes for increasing the energy efficiency of communications* ......................................................... 58

O. Garcia-Morchon, R. Rietman, L. Tolhuizen, D. Gomez-Perez and J. Gutierrez  
*HIMMO: a collusion-resistant, non-interactive, identity-based symmetric key establishment scheme* ................................................................................................................................. 66

T. Laarhoven  
*Asymptotics of fingerprinting and group testing (Extended abstract)* ............................... 70
Session 3: Biometrics and Surveillance

C.G. Zeinstra, R.N.J. Veldhuis and L.J. Spreeuwers
Towards the automation of forensic facial individualisation: Comparing forensic to non forensic eyebrow features ................................................................. 73

D. Kostadinov, S. Voloshynovskiy and M. Diephuis
Visual information encoding for face recognition: sparse coding vs vector quantization .................................................................................................................. 81

A. Dutta, R.N.J. Veldhuis and L.J. Spreeuwers
Automatic Eye Detection Error as a Predictor of Face Recognition Performance ...... 89

S. Javanbakhti, X. Bao, S. Zinger and P.H.N. de With
Automatic generic Region-Of-Interest selection for video surveillance applications .. 97

N. Jovanovic, T. Ozcelebi and J.J Lukkien
In-plane user positioning indoors ......................................................................... 105

Session 4: Classification

S. Ferdowsi and S. Voloshynovskiy
Content identification: machine learning meets coding ........................................... 113

N. Bouhouch, L. Demi, M.P.J. Kuenen, H. Wijkstra, T.J. Tjalkens and M. Mischi
Information-theoretic measures for angiogenesis imaging by contrast-enhanced ultrasound ........................................................................................................ 121

A. Jalalirad and T.J Tjalkens
Maximum-likelihood feature model selection ....................................................... 126

F. Farhadzadeh, F.M.J. Willems and S. Voloshinovskiy
Information theoretical analysis of identification based on active content fingerprinting ........................................................................................................ 134

Poster session

C. Papatsimpa, J.A. de Groot and J.-P. Linnartz
Error checking and near matching in helper data systems for biometric authentication ........................................................................................................... 140

M.L.L. Rompen, W.P. Sanberg, G. Dubbelman and P.H.N. de With
Online self-supervised learning for road detection ............................................... 148
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.I. Koutrouvelis, R. Heusdens and N.D. Gaubitch</td>
<td>Comparison of linear prediction methods in terms of sparsity, stability and robustness to reverberation</td>
<td>156</td>
</tr>
<tr>
<td>D. Vukobratovic and A. Pizurica</td>
<td>Compressed sensing using adaptive sparse measurements</td>
<td>164</td>
</tr>
<tr>
<td>C. Li, I. Creusen, L. Hazelhoff and P.H.N. de With</td>
<td>Recognition of lane markings in panoramic images</td>
<td>172</td>
</tr>
<tr>
<td>A. Callejon, M. Li, Y. Huang, S. Pollin and L. Van der Perre</td>
<td>Gbps throughput architecture for turbo decoder</td>
<td>178</td>
</tr>
<tr>
<td>E. De Brouwer and J. Louveaux</td>
<td>Comparisons of pre-coders under power normalization in massive-MIMO systems</td>
<td>186</td>
</tr>
</tbody>
</table>
Bayesian methods for sparse signal recovery

Bhaskar D. Rao
University of California, San Diego

Abstract

Compressive sensing (CS) as an approach for data acquisition has recently received much attention. In CS, the signal recovery problem from the observed data requires the solution of a sparse vector from an underdetermined system of equations. The underlying sparse signal recovery problem is quite general with many applications and is the focus of this talk. The main emphasis will be on Bayesian approaches for sparse signal recovery. We will examine sparse priors such as the super-Gaussian and student-t priors and appropriate MAP estimation methods. In particular, re-weighted $l_2$ and re-weighted $l_1$ methods developed to solve the optimization problem will be discussed. The talk will also examine a hierarchical Bayesian framework and then study in detail an empirical Bayesian method, the Sparse Bayesian Learning (SBL) method. If time permits, we will also discuss Bayesian methods for sparse recovery problems with structure; Intra-vector correlation in the context of the block sparse model and inter-vector correlation in the context of the multiple measurement vector problem.

Biography

Bhaskar D. Rao received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively. Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering. He is the holder of the Ericsson Endowed Chair in Wireless Access Networks and was the Director of the Center for Wireless Communications (2008–2011).

Prof. Rao was elected IEEE Fellow in 2000 for his contributions to the statistical analysis of subspace algorithms for harmonic retrieval. His work has received several paper awards; Best Paper Award (2013) for the paper “Multicell Random Beamforming with CDF-based Scheduling: Exact Rate and Scaling Laws”; SPS Best Paper Award (2012) for the paper “An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem”; Stephen O. Rice Prize Paper Award in the Field of Communication Systems (2008) for the paper “Network Duality for Multiuser MIMO Beamforming Networks and Applications”; Best Paper Award (2000) for the paper “PDF Optimized Parametric Vector Quantization of Speech Line Spectral Frequencies”.

Prof. Rao’s interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and biomedical signal processing.
Is your biometric data safe?

Alex C. Kot
Nanyang Technological University

Abstract

Nowadays, biometrics is widely used in authentication systems. In general, biometrics needs to be stored in a database for subsequent authentication. However, templates stored in the database are at the risk of being stolen or modified. Once the template is stolen, it is difficult to be replaced like passwords and the private user information associated with the stolen template would also be exposed. Thus, biometrics templates should be stored in the database such that both the security of the template and the privacy of the user are not compromised under various attacks. This talk will cover some existing techniques in dealing with biometrics data protection. New schemes in creating new fingerprint based on two sets of minutiae from two different fingers will be presented and a novel data hiding scheme is also proposed for the thinned fingerprint template.

Biography

Alex C. Kot received his BS in electrical engineering and MBA degrees both from the University of Rochester. He obtained his MS and PhD in electrical engineering from the University of Rhode Island. Prof. Kot has been with the Nanyang Technological University, Singapore since 1991. He headed the Division of Information Engineering with more than 40 faculty members at the School of Electrical and Electronic Engineering for eight years and served as Associate Chair/ Research and Vice Dean Research for the School of Electrical and Electronic Engineering. He is currently Professor and Associate Dean for the College of Engineering. He is the Director of Rapid-Rich Object SEarch (ROSE) Lab, partnering with Peking University.

Dr. Kot has served the IEEE SP Society in various capacities such as General Co-Chair, ICIP 2004 and Chairman, SPS Chapter Chairs. He served as Member, IEEE Fellow Evaluation Committee; Vice-President Finance, IEEE Signal Processing Society (2013–2014); Member, SPS Conference Board (2013–2014); and Member, SPS Publications Board (2013–2014). He received the Best Teacher of the Year Award and is a co-author for several Best Paper Awards including ICPR, IEEE WIFS, ICEC and IWDW. He was the IEEE CASS Distinguished Lecturer in 2005 and 2006 and is a Fellow of IES, a Fellow of IEEE, and a Fellow of Academy of Engineering, Singapore.

Dr. Kot has published extensively in the areas of signal processing for communication, biometrics, data-hiding, image forensics, information security. His new research area is in the domain object search and recognition.
Energy-Delay Trade-off of Wireless Data Collection in the Plane

Mihaela Mitici\textsuperscript{*}, Jasper Goseling\textsuperscript{†}, Maurits de Graaf\textsuperscript{‡} and Richard J. Boucherie\textsuperscript{*}

\textsuperscript{*}Stochastic Operations Research, University of Twente, The Netherlands
\textsuperscript{†} Department of Intelligent Systems, Delft University of Technology, The Netherlands
\textsuperscript{‡} Thales BV, The Netherlands

\{m.a.mitici, j.goseling, m.degraaf, r.j.boucherie\}@utwente.nl

Abstract

We analyze the Pareto front of the delay of collecting data from wireless devices located in the plane according to a Poisson process and the energy needed by the devices to transmit their observations. Fundamental bounds on the energy-delay trade-off over the space of all achievable scheduling strategies are provided.

1 Introduction

We consider wireless devices placed in the plane according to a homogeneous Poisson process. The devices have noisy observations of an attribute, e.g. temperature. A collector, positioned at a random location in the plane, is interested in a reliable estimate of this attribute, which is obtained by combining a subset of available observations from the devices. We assume that time is slotted and that the devices are awake independently at random every slot. Our interest is in which wireless devices should be scheduled for transmission and at which time slot such that the collector obtains this reliable estimate.

We focus on two performance measures: i) the delay experienced by the collector until retrieving a reliable estimate of the attribute and ii) the transmission energy used by the wireless devices to transmit their observations, which is assumed to be an increasing function of the distance between the devices and the collector. One could reduce the energy used for transmission by waiting longer until a device located closer to the collector is available for transmission. Alternatively, one could select in every time slot an available device that is closest to the location of the collector, keeping the time needed to retrieve a reliable estimate of the attribute from the devices minimal. In this case, the transmission energy is expected to increase since the devices that are selected can be arbitrarily far from the location of the collector. A trade-off arises between the energy used for data transmission and the time needed for the collector to retrieve a reliable estimate of the attribute.

The problem of scheduling wireless devices located in the plane has been studied in [1–7]. In [1], data from a network of sensors is aggregated at a hub node, which transmits to a base station. Since the base station is located at a considerable distance from the network of sensors, the sensors take turns in serving as a hub node. This minimizes the transmission energy. Simultaneous sensor transmission are possible, provided that the sensors are sufficiently separated in space. In [2], energy minimization within a wireless sensor network is considered such that all the sensors in the network transmit their data within a time threshold. The authors determine the Pareto front of the transmission energy and delay. In our setting, only a subset of the wireless devices is necessary to transmit its observations. Moreover, we consider single-hop transmissions, whereas in [1,2] multi-hop transmission is considered. Energy-latency trade-offs are analyzed in [3] using a data aggregation tree for wireless sensors.

The problem of collecting observations from a network of wireless devices is closely related to the problem of collecting fragments of a file from a wireless distributed storage network. The energy-delay trade-off for caches in the plane in which data is
stored according to partitioning or coding strategies is analyzed in [4,5]. In [6], energy-
delay and storage-delay trade-offs are analyzed. In [7], the effect of the topology and
density of a network of wireless devices on the trade-off between the energy and the
delay associated with the data collection, is analyzed. Closed form expressions for the
energy used by a specific set of nodes to transmit to a client are provided in [8].

The contributions of this paper are:

- An inner bound on the Pareto front of the energy-delay tradeoff under a maximum
delay constraint.

- Closed form expressions for the performance of a Greedy schedule that minizes
the delay. It is shown that performance can expressed in terms of Mahonian
numbers.

- Closed form expressions for the performance of a schedule that achieves minimum
possible energy consumption under an expected delay constraint.

- We propose a scheduling strategy that achieves expected delay and outperforms
any schedule that satisfies a maximum delay constraint.

The remainder of this paper is organizes as follows. In Section 2 we provide an exact
problem statement. In Section 3 we present our inner bound. The Greedy schedule is
analyzed in Section 4. Finally, we consider expected delay constraints in Section 5.

2 Problem Statement

Consider wireless devices located in the plane according to a homogeneous Poisson
process with intensity $\lambda$. Let $x$ denote the device that is the $x$–th closest neighbor
of the collector in the plane. Let $\delta_x$ denote the distance between the collector and the
device $x$.

Each wireless device makes an independent observation on an attribute $\theta$. The
observations are subject to independent and identically distributed additive Gaussian
noise with variance $\sigma^2$, i.e. $X_i \sim \mathcal{N}(\theta, \sigma^2)$. A collector, located at a random location
in the plane, is interested in retrieving a reliable estimate $\hat{X}$ with a variance that is
below a threshold $T$. This estimate can be obtained by retrieving an arbitrary set
of $s$ observations from the wireless devices such that $\text{Var}(\hat{X}) = \text{Var}(\frac{1}{s} \sum_{i=1}^{s} X_i) =
\frac{1}{s^2} \sum_{i=1}^{s} \text{Var}(X_i) = \frac{s}{s} < T$.

Every time slot, the wireless devices are awake with probability $p$, $0 < p < 1$ and
asleep with probability $q = 1 - p$. The probability of being awake is independent over
time and across the devices. We say that a wireless device is eligible to transmit if
the device has not transmitted its observation to the collector in previous time slots.
Transmission eligibility prevents the collector from receiving the same observation mul-
tiple times. A device may be scheduled for transmission only if it is awake and eligible.
No restrictions on the maximal transmission range of the wireless devices are imposed,
i.e a device at any random location in the plane can transmit its observation to the
receiver at a corresponding energy.

A centralized scheduler has knowledge about the position of the devices in the
plane, their eligibility and whether they are awake at a specific time slot. Our interest
is the fundamental performance limits that can be achieved by any possible schedule.
The performance measures the we consider are:

1. The energy, denoted by $P_x$ needed by the devices to transmit to the collector.
   Let the energy used by device $x$ to transmit to the collector be $\delta^a_x$, where $a \geq 1$ is
   a fixed parameter. The form of the energy function is motivated by, for instance,
the minimum power $P = (e^{2R} - 1)\delta^a$ required to transmit at a fixed rate $R$ over a distance $\delta$ given that the capacity of a AWGN channel is $1/2 \log(1 + P\delta^{-a})$.

Let $S = \{x_1, ..., x_s\}$ be the set of devices that transmitted their observations to the collector according to a centralized schedule, where the transmission order of the s devices can be any of the $s!$ possible sequences.

The total energy used for transmission by a set $S = \{x_1, ..., x_s\}$ of devices such that a reliable estimate is achieved at the collector is:

$$P_s(\delta x_1, ..., \delta x_s) = \sum_{i=1}^{s} \delta x_i^a. \quad (1)$$

We are interested in the expected energy $\mathbb{E}[P_s]$ used by the wireless devices to transmit $s$ observations to the collector, where the expectation is taken over the randomness in the Poisson process according to which the devices are located in the plane and the randomness in the awake/asleep status of the wireless devices.

2. The retrieval delay $W_s$. The retrieval delay $W_s$ is the time until the collector retrieves $s$ observations based on which a reliable estimate is retrieved. Initially we consider a maximum delay within which a collector must have retrieved a reliable estimate. Later, we relax this constraint and consider also the expected delay.

Consider the multi-objective optimization problem which aims at minimizing the energy and the retrieval delay under all possible centralized schedules. In general, a single point that simultaneously minimizes both performance measures does not exist, in which case we do not have a unique optimal solution. Therefore, we characterize the Pareto front [9] of the energy and delay performance measures.

We will make use of the Gamma function, which for $x > 0$ can be represented as $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$.

### 3 Inner bound under a maximum delay constraint

In this section we consider a maximum retrieval delay constraint. We provide a lower bound on the expected energy that is required to meet a maximum delay constraint. As a result we provide an inner bound on the energy-delay Pareto front under a maximum delay constraint.

The lower bound is constructed by considering the best possible schedule that could be obtained if the awake/asleep status of the devices would be known ahead of time. Such a schedule corresponds to a minimum matching in a bipartite graph that we describe next.

We consider time slots $1, 2, ..., t$, where $t$ is the maximal delay allowed, and the devices that are awake and eligible during those time slots. A matching is an assignment of devices to time slots. In each time slot of the matching, the devices transmit their observations to the collector. Let the weight of a matching be the energy used for transmission according to this matching. We are interested in a matching of size $s$ of minimal weight. This matching provides a transmission schedule for the devices such that the collector retrieves a reliable estimate within a delay of at most $t$ slots.

More precisely, let $G((T, N), E)$ be a bipartite graph with left vertices $T = \{1, \ldots, t\}$ and right vertices $N = \mathbb{N}$, edges $E \subset T \times N$ and weight function $w : E \to \mathbb{R}$. Let $T$ denote the time slots in which the observations are gathered and $N$ the devices in the plane. Devices $N$ are sorted according to their distance to the collector. An edge $(i, j)$ exists if node $j$ is awake in time slot $i$, i.e. it exists with probability $p$. Let the weight of an edge $w_e, e = (i, j)$, be the energy needed for device $i$ to transmit to the collector.
In order to compute the actual inner bound, we need to find on \( G((T, N), E) \), a minimum matching \( A \) of size \( s \). We proceed in two steps. First, we compute a minimum matching \( A' \) on graph \( G((T, N), E) \) using the Hungarian algorithm [10]. Clearly, the size of \( A' \) is \( t \). Next, we determine a matching \( A \) of minimum weight of size \( s \) starting from the matching \( A' \). More precisely, we consider the matching \( A' \) and eliminate the \( t - s \) largest weight edges. We are left with a matching \( A \) of minimum weight and of size \( s \). We refer to this two step algorithm as the modified Hungarian algorithm. The optimality of the modified Hungarian algorithm follows from the next result, which we state without proof due to space constraints.

**Lemma 1.** A minimum weight matching \( A \) of size \( s \) on \( G((T, N), E) \) can be found by eliminating the \( t - s \) largest weight edges from the matching given by the Hungarian algorithm applied to \( G((T, N), E) \).

For a given placement of nodes the modified Hungarian algorithm provides the minimum energy at a given delay constraint. The average expected energy follows from taking an average over the node placement. We illustrate the result in Figure 1 in Section 5.

### 4 Energy at minimum delay

In the previous section we provided a bound on the energy consumption at any given maximum delay constraint \( t \geq s \). In this section we provide an exact closed form result for the minimum energy consumption that can be obtained given delay constraint \( t = s \).

In particular, we consider a Greedy schedule, which schedules every time slot the closest awake and eligible device. Clearly, the delay in this case is minimal, i.e. \( W_s = s \). Also, it is clear that given the delay constraint one cannot achieve lower energy consumption.

Our analysis of the Greedy schedule is based on analyzing the probability that a given set \( M = \{x_1, ..., x_s\} \), \( x_i \in \mathbb{N}, i = \{1, ..., s\} \) of devices is scheduled for transmission, where \( x_i \) is the \( x_i \)-th closest device and \( x_1 < x_2 < .. < x_s \).

To illustrate the main idea of our result, consider the case that \( s = 3 \) and that these devices transmit in order \( x_1, x_2, x_3 \). Device \( x_1 \) transmits in the first slot if no other device is awake and closer than \( x_1^{ST} \) device. Also, \( x_1 \) has to be awake in the first slot. Thus, \( x_1 \) transmits in the first slot with probability \( q^{x_1^{ST}}p \). Similarly, in the second slot, \( x_2 \) transmits if \( x_2 \) is awake and all \( x_2 - 1 \) closer devices to the collector than \( x_2 \) are not awake. Also, recall that \( x_1 \) is closer than \( x_2 \) with respect to the location of the collector and that \( x_1 \) has transmitted in the first slot. Thus, \( x_2 \) transmits in the second slot with probability \( q^{x_2^{ST} - 1}p \). Lastly, \( x_3 \) transmits in the third slot if \( x_3 \) is awake and all \( x_3 - 1 \) nodes closer to the collector are not awake. Again, recall that both \( x_1 \) and \( x_2 \) have transmitted previously. Hence, \( x_3 \) transmits in the third slot with probability \( q^{x_3 - 1}p \). Therefore, the probability that sequence \( (x_1, x_2, x_3) \) transmits under Greedy is:

\[
p^3 q^{(x_1+x_2+x_3-1-2-3)} = p^3 q^{(x_1+x_2+x_3-6)}.\]

The general result depends on the order in which the devices transmit. Therefore, we need to notion of the disorder of a permutation.

**Definition 1.** The disorder of a permutation is the number of pairs of entries that appear in reversed order. Formally, given a sequence \((x_1, x_2, ..., x_n)\), with \( x_1 < x_2 < .. < x_n \), a permutation \((x_1, x_2, ..., x_n)\) has disorder \( \sum_{i=1}^{n} \sum_{j>i} \mathbb{I}(x_i > x_j) \).

For example, the disorder of permutation \((1, 2, 4, 3, 5)\) is one since elements 3 and 4 are in disorder. The disorder of \((2, 1, 4, 3, 5)\) is two since elements 1 and 2 are in disorder, as well as elements 3 and 4.
To illustrate the relevance of a disorder of a transmission sequence, we now consider the transmission sequence \((x_1, x_3, x_2)\), which has disorder one. Again, device \(x_1\) transmits in the first slot with probability \(q^{x_1-1}p\) and device \(x_2\) transmits in the third slot with probability \(q^{x_2-1}p\). The only change is in the probability of device \(x_3\) transmitting in the second slot. Since \(x_3\) transmits before \(x_2\) and \(x_2\) is closer to the collector, the probability of \(x_3\) transmitting in the second slot depends on the awake/asleep status of \(x_2\). More precisely, for \(x_3\) to transmit in the second slot, \(x_2\) has to be asleep in this slot. Thus, \(x_3\) transmits in the second slot with probability \(q^{x_2-1-2+1}p\). Therefore, the probability that sequence \((x_1, x_3, x_2)\) transmits under Greedy is:

\[
p^3q^{(x_1+x_2+x_3-6+1)}.
\]

Observe that in the above expression the term +1 corresponds exactly to the disorder of the transmission sequence. Note also that there are 2 sequences with disorder 1, 2 sequences with disorder 2 and one sequence with disorder 3. By generalizing the above derivations we obtain the following expression for \(p_{x_1,x_2,...,x_s}\), the probability that \(x_1, x_2\) and \(x_3\) are transmitting:

\[
p_{x_1,x_2,x_3} = p^3q^{\sum_{i=1}^{\infty} x_i-6} + 2p^3q^{\sum_{i=1}^{\infty} x_i-5} + 2p^3q^{\sum_{i=1}^{\infty} x_i-4} + p^3q^{\sum_{i=1}^{\infty} x_i-3}
\]

\[
= p^3q^{\sum_{i=1}^{\infty} x_i-6}(1+q)(1+q+q^2).
\]

In general, the probability that a sequence of \(s\) devices transmits under Greedy can be expressed in terms of the disorder of that sequence. In addition, the number of sequences that have a given disorder is given by the Mahonian numbers, which are the coefficients in the expansion of \(\Pi_{m=0}^{s-1} (1 + x + ... + x^m)\) [11]. This leads to the following theorem, which again we state without proof.

**Theorem 1.** The probability that set \((x_1, x_2, ..., x_s)\) of devices transmits under Greedy scheduling is:

\[
p_{x_1,x_2,...,x_s} = C_s \cdot q^{(x_1+x_2+...+x_s)},
\]

where \(C_s = p^s q^{-s(s+1)/2} \Pi_{m=0}^{s-1}(1 + q + ... + q^m)\).

We can now determine the expected energy when Greedy scheduling is used.

**Theorem 2.** The expected energy needed to transmit \(s\) observations to the collector when Greedy scheduling is used, is:

\[
\mathbb{E}[P_s] = \frac{C_s}{(s-1)!} \sum_{x_1=1}^{\infty} \sum_{x_2=1}^{\infty} \cdots \sum_{x_s=1}^{\infty} \sum_{x_s \neq x_{s-1}} \cdots \sum_{x_2 \neq x_1} \Gamma(\frac{a}{2} + x_1) \frac{(\lambda \pi)^{a/4}}{2 \Gamma(\frac{a}{2})} q^{x_1+...+x_s},
\]

where \(C_s = p^s q^{-s(s+1)/2} \Pi_{m=0}^{s-1}(1 + q + ... + q^m)\).

**Proof.** Let \(e_{x_i}\) denote the expected energy needed for the \(x_i^{th}\) closest device to transmit its observations to the collector. From [8] we have

\[
e_{x_i} = \frac{\Gamma(a/2 + i)}{(\lambda \pi)^{a/4} \Gamma(i)}.
\]
Let $\beta_{x_i}$ is the probability that device $x_i$ transmits under Greedy schedule. Then, the expected energy used for transmission under Greedy can be written as

$$
\mathbb{E}[P_s] = \sum_{x_i=1}^{\infty} e_{x_i} \beta_{x_i} = \frac{1}{(s-1)!} \sum_{x_1=1}^{\infty} \cdots \sum_{x_s \neq x_{s-1}}^{\infty} e_{x_1} p_{x_1,x_2,\ldots,x_s}.
$$

The last equality in the above expression is obtained by summing over all sets that include $x_i$. The result follows from (3) and Theorem 1.

The expression in Theorem 2 provides a convenient means of computing $\mathbb{E}[P_s]$. As an example, consider $s = 3$ and $a = 2$:

$$
\mathbb{E}[P_3] = \frac{C_3}{2!} \sum_{x_1=1}^{\infty} \sum_{x_2=1}^{\infty} \sum_{x_3=1}^{\infty} \sum_{x_2 \neq x_1} \sum_{x_3 \neq x_2} \sum_{x_3 \neq x_1} \frac{x_1}{\lambda \pi} q^{x_1+x_2+x_3}
$$

$$
= \frac{C_3}{2\lambda \pi} \sum_{x_1=1}^{\infty} x_1 q^{x_1} \sum_{x_2=1}^{\infty} q^{x_2} \left( \sum_{x_3=1}^{\infty} q^{x_3} - q^{x_2} - q^{x_1} \right)
$$

$$
= \frac{C_3}{2\lambda \pi} q^3 \left( \frac{1}{p^4} - \frac{2}{p(1-q^2)^2} - \frac{1}{p^2(1-q^2)} + \frac{2}{(1-q^3)^2} \right).
$$

5 The benefit of an expected delay constraint

Observe that the minimum expected energy consumption is achieved by scheduling only the $s$ closest devices. Since for any fixed number of time slots there is a positive probability that at least one of these devices is asleep in all these time slots, it is not possible to meet a maximum delay constraint. In this section we relax the delay constraint and analyze the performance of schedules that satisfy an expected delay constraint.

First, we determine the point on the Pareto front given by the minimal expected energy under all achievable schedules and the corresponding expected delay. As indicated above, the minimal expected energy under all achievable schedules is the energy needed for the closest $s$ devices, with respect to the location of the collector in the plane, to transmit their observations.

**Theorem 3.** The point on the Pareto front given by the minimal expected energy $\mathbb{E}[P_{s_{\min}}]$ and its corresponding expected delay $\mathbb{E}[W_s]$ under all achievable schedules:

$$
\left\{ \begin{array}{c}
(x, y) = \left( \frac{\Gamma(s + \frac{a}{2} + 1)}{(\lambda \pi)^{a/2} \cdot (\frac{a}{2} + 1) \Gamma(s)} \cdot \sum_{i=1}^{s} \frac{1}{1 - q^{s-i+1}} \right), x = \mathbb{E}[P_{s_{\min}}] \text{and } y = \mathbb{E}[W_s] \\
\end{array} \right\}.
$$

**Proof.** Using the expected energy needed for the $x_i^{th}$ neighbor of the collector in the plane, as in (3), and summing over all $s$ expected energy transmissions [8]:

$$
\mathbb{E}[P_{s_{\min}}] = \sum_{i=1}^{s} \frac{\Gamma(a/2 + i)}{(\lambda \pi)^{a/2} \Gamma(i)} = \frac{\Gamma(s + \frac{a}{2} + 1)}{(\lambda \pi)^{a/2} \cdot (\frac{a}{2} + 1) \Gamma(s)}.
$$

(4)
The corresponding expected delay is computed as follows. The scheduling scheme is independent and identical over time. Thus, the time until the collector retrieves the $i^{th}$ observation, given that it already received $i - 1$ observations, can be viewed as the time until a first success in a Bernoulli trial is achieved, where the success probability $T_i$ is the probability that at least one eligible device is awake. Thus, the probability of success is $T_i = 1 - q^{s-i+1}$. Then, the expected delay is:

$$E[W_s] = \sum_{i=1}^{s} \frac{1}{T_i} = \sum_{i=1}^{s} \frac{1}{1 - q^{s-i+1}}.$$ (5)

Next, we determine the point on the Pareto front given by the minimal delay and the corresponding minimal expected energy under any achievable schedule. The following result directly follows from Theorem 2.

**Corollary 1.** The point on the Pareto front given by the minimal delay and its corresponding minimal expected energy $E[P_s]$ under all achievable schedules:

$$\begin{cases} (x, y) = \left( \frac{C_s}{(s-1)!} \sum_{x_1=1}^{\infty} ... \sum_{x_s=1}^{\infty} \sum_{x_s \neq x_1} \frac{\Gamma(s + x_1)}{(\lambda q)^{s/2} \Gamma(x_1)} q^{x_1+...+x_s}, s \right), x = E[P_s], y = W_s^{\text{min}} \right) \end{cases}$$

where $C_s = p^s q^{-s(s+1)/2} \prod_{m=0}^{s-1} (1 + x + ... + x^m)$.

Finally, we consider a class of expected delay strategies that interpolate between Theorem 3 and Corollary 1. The strategy considers the $N \geq s$ closest neighbors. In each time slot the closest eligible and awake device among these $N$ devices is scheduled to transmit. We denote this strategy as Modified Greedy. Note that for $N = s$, we have the case of the minimal expected energy and the corresponding expected delay of Theorem 3. For $\lim N \to \infty$ we approach the minimal delay and the corresponding expected energy of Corollary 1.
Figure 1 shows the Pareto front of the expected energy and delay and the inner bound of the Pareto front, as described in Section 3. An important observation from Figure 1 is that for small energy consumption, schedules that satisfy only the relaxed, expected delay constraint, can outperform all schedules that satisfy a maximum delay constraint.

Acknowledgments

This work was performed within the project RRR (Realisation of Reliable and Secure Residential Sensor Platforms) of the Dutch program IOP Generieke Communicatie, number IGC1020, supported by the Subsidiering Sterktes in Innovatie, and is partly supported by the Netherlands Organisation for Scientific Research (NWO), grant 612.001.107.

References


Diversity Combining Techniques for Visible Light Communications

Anagnostis Tsiatmas\textsuperscript{1,2}, Constant P. M. J. Baggen\textsuperscript{2}, and Frans M. J. Willems\textsuperscript{1}

\textsuperscript{1}Department of Electrical Engineering, TU/e, The Netherlands
\textsuperscript{2}Philips Research Eindhoven, High Tech Campus 34, The Netherlands
emails: a.tsiatmas@tue.nl, stan.baggen@philips.com, f.m.j.willems@tue.nl

Abstract

Diversity transmission is often encountered in Visible Light Communications (VLC) either in the form of wavelength diversity when a white LED source is used as a transmitter or in the form of spatial diversity when simultaneous transmission from multiple luminaires takes place. In this paper, we examine optimal diversity combining techniques for VLC systems in the presence of strong ambient light. We show that the signals from the different channels can be either combined in the optical domain with the use of special optics such as filters and prisms when all the channels are detected by a single photodetector or in the electrical domain when the light of each separate channel is collected by a separate photodetector. In the case of optical combining, we prove that the total signal-to-noise ratio (SNR) at the receiver is maximised by a fundamentally different solution compared to the well-known maximal-ratio combining (MRC) in radio wireless communications (RWC). However, in the case of electrical combining, MRC solution still achieves the optimum. Furthermore, in the case of photon noise only, we prove that combining the channels in the electrical domain is never worse than combining them in the optical domain. Finally, we provide a physical interpretation for our results using the compound Poisson model.

1 Introduction

Diversity techniques have been examined extensively as a means of improving the received signal-to-noise ratio (SNR) and combating fading in wireless radio communications, already, since the 50’s [1]. In diversity systems, communication is provided through multiple channels. The main idea is that when different channels are separated enough in space, time, or frequency, their behaviour (e.g., fading) can be considered more or less independent. Thus, each component channel yields a different SNR at the receiver and the probability of a message’s successful reception is increased.

Various techniques of different complexity and performance exist on how to combine the received signals from the available channels of a diversity system. For example, with selection combining, which is the simplest of all, the receiver chooses always the channel with the highest SNR. Differently, the receiver can apply a weighting factor to the signal received from each of the channels and then combine them all together. Applying the proper weighting coefficients the receiver’s SNR is maximised. The technique that achieves the latter is known as Maximal-Ratio Combining (MRC).

Recently, the idea of wireless data transmission from infrared- and visible-light sources is proposed for a variety of indoor applications such as device interconnection, internet access, positioning, and navigation [2]. Optical wireless (OW) systems gather a number of attractive qualities, such as low-cost components and a license-free electromagnetic spectral range [3]. However, the commercialisation of OW systems was hindered by the high propagation losses and the low-power LED sources used as
transmitters. The aforementioned situation has changed with the advent of the high-brilliance LED (HB-LED) source [5]. In particular, white HB-LEDs offer the potential for simultaneous illumination and communication functionality and are commonly encountered in Visible Light Communication (VLC) systems developed nowadays [4].

Most VLC systems that use HB-LEDs as transmitters employ intensity-modulation/direct-detection (IM/DD), since LED sources do not permit control of the phase of the optical radiation. IM/DD channels are best modelled as Poisson channels. The latter means that only the expected and not the actual number of generated electrons at the output of the photodetector (PD) is known [5]. The variance in the number of generated electrons depends on the expected number of photon arrivals and results in a signal-dependent photon-noise component. At the same time, photons from other interfering light sources reach the surface of the PD and cause both a positive DC offset in the expected number of generated electrons and an additional shot-noise component.

The white HB-LED source often encountered in VLC systems provides inherent wavelength diversity by encoding the same information in different visible wavelengths. Note that different light sources have different optical spectra, meaning that the number of emitted photons in each wavelength depends on the optical properties of the light source. Thus, the received visible wavelengths are equivalent to channels that can differ in their SNRs, since in each wavelength the detected number of signal and non-signal photons can be different. A similar situation occurs when simultaneous transmission from spatially separated light sources is considered. The SNR of received signals from signal sources near ambient-light (non-signal) sources is smaller compared to the SNR of received signals from signal-sources that are not affected by other non-signal light sources. To maximise the total SNR at the receiver, the different wavelength and/or spatial channels should be combined optimally.

The signals received from the different channels can be either processed in the optical or in the electrical domain. The former happens with the use of special optics such as optical filters (e.g., absorptive or resonant filters) and a single PD to detect the photons from all channels, while the latter is the case when a separate PD is used to detect the signal in each separate channel. In this paper, we show that contrary to radio wireless systems, in the case of OW communications, the performance of the OW diversity system depends on the domain in which combining occurs. The reader can find a first reason for this, on the stochastic (quantum) interaction of photons with optical components in the optical domain. In the following sections, we examine VLC systems operating in the presence of strong ambient light and we show that combining in the optical domain results in a fundamentally different solution compared to the MRC. In our analysis, in the case of photon noise only, we prove that combining the channels in the electrical domain outperforms combining them in the optical domain. Finally, we explain the presented counter-intuitive results regarding the combining of the channels in the optical domain with the use of the compound Poisson model.

2 Diversity Combining in Radio Communications

We start by presenting the well-known case of optimal combining in RWC diversity systems. We assume a system with $K$ separate channels. The gain of the transmitted signal in channel $i$ is $g_i$, while the received signal $y_i$ is affected by additive white Gaussian noise (AWGN). The equation of channel $i$ is:

$$y_i = g_i X + n_i,$$

where the r.v. $X \in \mathbb{R}$ is the transmitted signal and the independent r.v.'s $n_i \sim \mathcal{N}(0, \sigma_{n_i}^2)$. Assuming that $E[X] = 0$ and $\text{Var}[X] = 1$, the SNR in channel $i$ is $\text{SNR}_i = (g_i^2 \text{Var}[X]) / \sigma_{n_i}^2 = g_i^2 / \sigma_{n_i}^2$. 

12
In the case of MRC, the signals received from the different channels are brought to the same phase and weighted with coefficients $w_i$ before they are linearly combined. The total received signal $y_{tot}$ is written as:

$$y_{tot} = \sum_{i=1}^{K} w_i y_i,$$

and achieves an SNR equal to:

$$\text{SNR}_{tot} = \frac{\left(\sum_{i=1}^{K} w_i g_i\right)^2}{\sum_{i=1}^{K} |w_i|^2 \sigma_{Ni}^2}.$$  \hspace{1cm} (3)

With the aid of the Cauchy-Schwarz inequality, we calculate the maximum possible SNR attained at the receiver:

$$\text{SNR}_{tot} = \frac{\left(\sum_{i=1}^{K} w_i g_i\right)^2}{\sum_{i=1}^{K} |w_i|^2 \sigma_{Ni}^2} \leq \frac{\left(\sum_{i=1}^{K} |w_i|^2 \sigma_{Ni}^2\right) \left(\sum_{i=1}^{K} |g_i|^2 / \sigma_{Ni}^2\right)}{\sum_{i=1}^{K} |w_i|^2 \sigma_{Ni}^2} \leq \sum_{i=1}^{K} \left|\frac{g_i}{\sigma_{Ni}}\right|^2. \hspace{1cm} (5)$$

It is easy to verify that the upper bound on the $\text{SNR}_{tot}$ is achieved by taking $w_i = k g_i^* / \sigma_{Ni}^2$, where $k \in \mathbb{R}$ is a proportionality constant.

**Comment:** The MRC solution reveals that in order to maximise the total SNR at the receiver, the signals from all the channels should be taken into account by the combining method.

### 3 The Optical Wireless (OW) Channel

While OW systems employing IM/DD are best modelled as Poisson channels \cite{6}, the non-additive behaviour of the noise in the Poisson model complicates the expression between the channel input and output and does not allow us to use well-established concepts such as the SNR and the Euclidean distance between signals, which are common in AWGN channels. Thus, using the Poisson channel model presents difficulties in the study of optimum methods of diversity combining in OW communications. In \cite{6}, we introduced the Square Root (SR) channel model, which is in the form of an AWGN channel model and simplifies the analysis of OWC diversity systems.

#### 3.1 Channel Model

In \cite{6} we showed that the SR channel model can be used to analyse an OW IM/DD channel after the application of a square-root instantaneous nonlinearity at the output of the PD. The SR channel model accounts both for the signal-dependent noise due to the inherent Poisson nature of the optical channels and for the thermal and the ambient-light noise contributions existing in OW channels. The advantage of the SR model is that the only noise source in the channel is additive Gaussian with constant variance $\sigma_N^2 = 1/4$, while the signal degradation due to thermal and ambient light noise is expressed through a constant parameter $\lambda_0 \geq 0$. In particular, with the r.v. $\Lambda \geq 0$

\footnote{A possible thermal noise contribution is modelled as an equivalent shot-noise component caused by a virtual number of photon arrivals that results in the same signal degradation.}
expressing the channel input (i.e., the expected number of received photons within a specific time-interval), the SR channel output is:

\[ Y = \sqrt{\Lambda + \lambda_0} + N, \]  

where the r.v. \( N \sim \mathcal{N}(0,1/4) \). In [6] we proved that as \( \Lambda \) increases the statistics of the output \( Y \) of the SR channel model match the statistics of the square-rooted output of a Poisson channel for the same input \( \Lambda \).

In this paper, we are interested in the practical scenario of communication in the presence of strong ambient-light illumination. We assume that the r.v. \( \Lambda \) takes values on the finite set \( \mathcal{L} \) and that the non-signal photon arrivals dominate over the signal photon arrivals. Thus, we have that \( \lambda_0 \gg \Lambda, \forall \Lambda \in \mathcal{L} \). Using a Taylor expansion for the square-root term on the righthand side of Eq. (6), we obtain:

\[ Y = \sqrt{\Lambda + \lambda_0} + N = \sqrt{\lambda_0} \sqrt{1 + \frac{\Lambda}{\lambda_0}} + N = \sqrt{\lambda_0} + \frac{1}{2} \frac{\Lambda}{\sqrt{\lambda_0}} - \frac{1}{8} \frac{\Lambda^2}{\lambda_0^{3/2}} + \ldots + N. \]  

We assume that the receiver knows the level of ambient-light illumination. Thus, any DC offset is known to the receiver and can be removed. Furthermore, since \( \lambda_0 \gg \Lambda, \forall \Lambda \in \mathcal{L} \), we can approximate the equivalent channel equation as:

\[ Y' = \frac{1}{2} \sqrt{\lambda_0} + N. \]  

### 3.2 Signal-to-Noise Ratio in OW channels

**Definition 1:** We define as Signal-to-Noise Ratio of the SR channel the quantity \( \text{SNR}_{\text{SR}} = \text{Var}(\Lambda)/(4\lambda_0\text{Var}(N)) = \text{Var}(\Lambda)/\lambda_0. \)

In the following of this section we show that increasing \( \text{SNR}_{\text{SR}} \) increases the maximum information that can be transmitted through an OW channel for any given input distribution. In addition, increasing \( \text{SNR}_{\text{SR}} \) minimises the minimum mean-square error (mmse) in the estimation of the channel input given the channel output and at the same time minimises the symbol-error-rate (SER). In this sense, **Definition 1** for the SNR in OW systems is justified and, at the same time, it becomes clear that an optimum diversity combining method should maximise the \( \text{SNR}_{\text{SR}} \).

**Proposition 1:** The mutual information \( \text{I}(\Lambda';Y') \) between the normalised input \( \Lambda' = (\Lambda - E[\Lambda]) / \text{Var}(\Lambda) \) and output \( Y' \) of the SR channel is a monotonically increasing function of \( \text{SNR}_{\text{SR}} \).

**Proof:** By dividing both sides of Eq. (8) with \( \sigma_N \), we obtain:

\[ Y'' = \frac{1}{\sigma_N} Y' = \sqrt{\text{SNR}_{\text{SR}}} \Lambda' + N', \]

where \( N' \sim \mathcal{N}(0,1) \). From **Theorem 1** in [7], we get:

\[ \frac{d}{d\text{SNR}_{\text{SR}}} \text{I}(\Lambda';Y'') = \frac{1}{2} \text{mmse}(\Lambda'|Y''), \]

where \( \text{mmse}(\Lambda'|Y'') \) is the mmse in the estimation of \( \Lambda' \) given \( Y'' \). The proposition is proven by noticing that always \( \text{mmse}(\Lambda'|Y'') \geq 0 \) and that \( \text{I}(\Lambda';Y'') = \text{I}(\Lambda';Y') \).

**Corollary 1:** The \( \text{mmse}(\Lambda'|Y'') \) is a monotonically decreasing function of the \( \text{SNR}_{\text{SR}} \). This follows from **Corollary 5** in [8].
Corollary 2: From Eq. (9) it becomes clear that increasing SNR$_{SR}$ increases the Euclidean distance between the input signals $\Lambda'$. Thus, increasing SNR$_{SR}$ results in a lower SER.

In the rest of this paper, we focus on the maximisation of SNR$_{SR}$ as the criterion that achieves optimal combining of the different channels in Optical Wireless Communications (OWC).

4 Diversity Combining in OWC Systems

As we explained in the Introduction, in OWC systems that provide transmit diversity, the number of signal photons as well as non-signal photons reaching the receiver through each separate channel can differ. We assume that there are $K$ channels available, with the components $a_i, i = 1, \ldots, K$ of the $K$-dimensional vector $a$ representing the received signal photons in each channel $i$ and the components $b_i, i = 1, \ldots, K$ of the $K$-dimensional vector $b$ representing the received non-signal photons in each channel $i$. We also assume that there is thermal noise present in the channel, which is expressed by the parameter $\lambda_{th} = \sigma^2_{\text{thermal}}$. We examine both the cases that the different channels are combined in the optical domain and in the electrical domain.

4.1 Diversity Combining in the Optical Domain

We start by examining how optimum combining should occur in the optical domain. OWC systems can either exploit tuneable optical filters or change the field-of-view (FOV) of the receiver in order to control the amount of optical intensity (i.e., the number of photon arrivals during a specific time-interval) detected from each of the $K$ channels by the single PD of the receiver. Let the portion of the optical intensity that reaches the PD from channel $i$ be $f_i ^\dagger$. It is clear that $0 \leq f_i \leq 1$. Therefore, the PD detects $a_i f_i$ signal photons and $b_i f_i$ non-signal photons from channel $i$. Of course, $\lambda_{th}$ is not affected by the $f_i$ coefficients. Using the SR model formulation, we can write $\Lambda = (a \cdot f)U$, with the r.v. $U \geq 0$ expressing the signal that modulates the photons and takes values on the finite set $\mathcal{U}$, and $\lambda_0 = b \cdot f + \lambda_{th}$. Thus, the expression of Definition 1 for the SNR in the SR channel, SNR$_{SR}$, now becomes:

$$\text{SNR}_{SR} = \frac{(a \cdot f)^2}{b \cdot f + \lambda_{th}} \text{Var}(U) = G(f)\text{Var}(U),$$

where $G(f) = (a \cdot f)^2 / (b \cdot f + \lambda_{th})$. In order to maximise SNR$_{SR}$, we have to determine the vector $f$ that maximises $G(f)$, since $\text{Var}(U)$ is independent of $f$.

\footnote{The quantity $f_i$ may correspond to the transmittance of an optical filter or the reduction in the number of the collected photons because of the PD’s FOV.}

Figure 1: (a) Diversity combining in the optical domain. (b) Diversity combining in the electrical domain.
Proposition 2: The function $G(f)$ is quasi-convex on $L = [0, 1]^K - \{0\}$.

Proof: The numerator of $G(f)$ is a convex function of $f$ on $L$ and the denominator of $G(f)$ is a concave function of $f$ on $L$. Based on the Example 3.38 in [9] we know that a convex-over-concave function is a quasi-convex function. Thus, $G(f)$ is a quasi-convex function of $f$ on $L$.

Theorem 1. $\text{SNR}_\text{SR}^\text{e} = \max_f \text{SNR}_{\text{SR}}(f)$ is always achieved by a vector $f'$ with components $f'_i$ equal to either 0 or 1.

Proof: Since $G(f)$ is a quasi-convex function of $f$ on $L$, its global maximum is always attained on one of the extreme points of the $K$-dimensional hypercube defined by $L$, as stated in Proposition 14 in [10]. Thus, $\max_f G(f)$ is achieved by a vector $f'$ with components 0 and 1. As $\text{Var}(U)$ does not depend on $f$, $\text{SNR}_{\text{SR}}$ is maximised by the vector $f'$.

Interpretation: The result of Theorem 1 implies that in order to maximise the $\text{SNR}_{\text{SR}}$, we should allow all the photons from $N \leq K$ channels to reach the surface of the PD, while the rest of the channels should be blocked completely. This is in contrast to the MRC solution, which would imply that photons from all channels should reach the PD with the intensity in each of the component channels weighted with a proper complex coefficient. We call the derived solution, which maximises the $\text{SNR}_{\text{SR}}$ of an OWC system, Optical Diversity Combining (ODC).

While Theorem 1 states that the optimum vector $f'$ is an extreme point of the hypercube defined by $L$, it does not provide any information on which of the extreme points achieves the global maximum. In general, an algorithm is necessary to find the exact value of $f'$ that achieves $\text{SNR}_{\text{SR}}^\text{e}$.

4.2 Diversity Combining in the Electrical Domain

We now focus on the case of combining the signals from the $K$ component channels in the electrical domain. We assume that the receiver has $K$ PDs, each one detecting photons only from one of the $K$ channels. This can happen, e.g., either by using a prism to decompose the white light into different wavelengths and using a dedicated PD for every wavelength channel or by using PDs oriented towards each of the different directions in the case of OW systems with spatial diversity. At the output of each of the PDs there is a signal-processing unit that applies the square-root nonlinearity and subtracts the mean value in each of the channels. By linearly combining the channel outputs $Y''$, as given by Eq. (9), and by using the notation presented in Subsection 4.1, we obtain:

$$Y''_{\text{tot}} = \sum_{i=1}^{K} c_i Y''_i = \sum_{i=1}^{K} c_i \sqrt{\frac{a_i^2}{b_i + \lambda_{\text{th},i}}} U + c_i N''_i,$$

where $c_i$ is the weighting factor applied to the output of the square-root nonlinearity by the signal processing unit of branch $i$ of the receiver. In Eq. (12), $\lambda_{\text{th},i}$ is a constant that refers to the thermal noise component in channel $i$. From the form of Eq. (12) it becomes clear that the MRC is the optimum diversity combining technique in the electrical domain. Since the noise variance, $\sigma_{N''}^2 = 1$, is the same in every channel, the values $c_i$ that lead to the maximum $\text{SNR}_{\text{SR}}$ for the combined signal are $c'_i = \sqrt{\frac{a_i^2}{b_i + \lambda_{\text{th},i}}}$ [1]. Substituting $c_i$ values with $c'_i$, for $i = 1, \ldots, K$, in Eq. (12), we get:

$$\text{SNR}_{\text{SR}}^\text{e} = \sum_{i=1}^{K} \frac{a_i^2}{b_i + \lambda_{\text{th},i}} \text{Var}(U) = \sum_{i=1}^{K} \text{SNR}_{\text{SR},i}.$$

We now examine the case that the photon noise in the OWC system is very strong compared to the thermal noise (i.e., we take $\lambda_{\text{th},i} = 0$ for $i = 1, \ldots, K$).
Theorem 2. When only photon noise exists, the maximum SNR achieved in the electrical domain SNR$^e_{SR}$ exceeds the maximum SNR achieved in the optical domain SNR$^o_{SR}$.

Proof: Assume that $f' = \arg \max_f \text{SNR}_{SR}(f)$. In the derivation of Theorem 1, we showed that $f'$ maximises $G(f)$. Thus, by applying the Cauchy-Schwartz inequality to $G(f')$, we obtain:

$$G(f') = \left( \frac{\sum_{i=1}^{K} f'_i a_i}{\sum_{i=1}^{K} f'_i b_i} \right)^2 = \left( \frac{\sum_{i=1}^{K} (b_i f'_i)^{1/2} a_i}{b_i^{1/2} \sum_{i=1}^{K} f'_i b_i} \right)^2 \leq \frac{\sum_{i=1}^{K} (f'_i a_i^2)}{\sum_{i=1}^{K} f'_i b_i} \leq \frac{\sum_{i=1}^{K} a_i^2}{b_i}.$$ (14)

By multiplying all terms in Ineq. (14) with Var$(U)$, we derive the following inequality:

$$\text{SNR}^e_{SR} \leq \sum_{i=1}^{K} f'_i \frac{a_i^2}{b_i} \text{Var}(U) \leq \text{SNR}^e_{SR}.$$ (15)

5 Discussion

In the previous sections we presented how optimum combining of OWC channels is performed in the electrical and in the optical domain. The counter-intuitive result of ODC is better understood when we consider the stochastic nature of optical radiation and the way optical radiation interacts with matter (e.g., with filters, detectors, etc.). To explain this fundamental difference of OWC systems from RWC systems, we have to examine the RWC systems first. The main noise source in a RWC system is the thermal noise induced by the electronics in the receiver. The thermal-noise component is independent from the signal level. Thus, weighting the received signal from each component channel in order to combine the different channels, as MRC implies, does not change the SNR of each component channel.

In the case of OWC systems, the generation and detection of photons are Poisson processes. The latter implies that the photon-noise level depends on the expected number of photon arrivals. At the same time, the operation of optical filters that achieve the weighting of the intensities of the component channels in the optical domain is best modelled as a Bernoulli process. In particular, the transmittance $t$ of an optical filter expresses the probability of a photon propagating through the filter to appear at the filter’s output.

In order to understand why the MRC method is not optimum, we have to explain the optical radiation-filter interaction process. We assume that the photon arrivals at the input of an optical filter follow a Poisson distribution with parameter $\lambda$. Thus, the SNR at the input of an optical filter is $\text{SNR}_{IN} = \lambda$. Since the interaction of a particular photon with the filter is a Bernoulli process, we can use the compound Poisson model to describe the distribution of the r.v. $X$ representing the photons at the output of the filter [11]. We assume that the behaviour of the filter is expressed through the r.v. $K \sim \text{Bernoulli}(t)$, for which $E[K] = t$, $\text{Var}[K] = t(1 - t)$ and $E[K^2] = t$. Thus, the compound Poisson second-order statistics are $E[X] = \lambda t$ and $\text{Var}[X] = \lambda t$, leading to an SNR at the output of the filter equal to $\text{SNR}_{OUT} = (E[X])^2/\text{Var}[X] = \lambda t$. Since photon interactions with optical components affect the SNR of a channel in optical communication systems, MRC is not an optimal combining method for these systems.
Note that in the case of cascaded optical filters the enhancement of the power of the noise is proportional to the product of the transmittances of the filters. This is not the case in radio systems, where the noise enhancement is expressed by the Friis equation and is typically dominated by the first amplifier.

6 Conclusion

In this paper, we examined optimal combining methods for OWC systems in the presence of strong ambient light and we showed that optimal combining of the different channels in the optical domain is achieved by a fundamentally different technique compared to the MRC technique due to the quantum nature of light. Furthermore, we demonstrated that in the case of electrical combining, MRC remains the optimum combining technique. Finally, in the case of photon-noise only, we proved that combining the channels in the electrical domain is never worse than combining them in the optical domain.

References

Energy-Delay Analysis of Full Duplex Wireless Communication for Sensor Networks

Tom Vermeulen  
KU Leuven  
Electrical Engineering, Telemic  
Kasteelpark Arenberg 10 - bus 2444, Leuven  
tom.vermeulen@esat.kuleuven.be  
Sofie Pollin  
KU Leuven  
Electrical Engineering, Telemic  
Kasteelpark Arenberg 10 - bus 2444, Leuven  
sofie.pollin@esat.kuleuven.be

Abstract

Full duplex wireless communication is a promising new technique that enables simultaneous transmission and reception of a packet on the same frequency. So far most research focused on proving the feasibility of full duplex Wi-Fi systems, focusing on PHY analysis of the BER. In this paper, the use of full duplex in wireless sensor networks is analysed, and it is shown that this can enable significant energy and delay gains, even when considering a realistic MAC protocol. This paper presents a novel full duplex energy model and MAC protocol for wireless sensor networks, compatible with off-the-shelf 802.15.4 chips. We show the potential of full duplex sensor networks, both for networks with low and high loads. Especially for high loads, full duplex enables a promising collision detection, avoiding wasting scarce resources in long packet collisions. Full duplex nodes outperform half duplex nodes both in terms of energy as in terms of delay, even in case of asymmetric traffic conditions. In addition, several advantages exist in terms of fairness of downlink traffic towards uplink traffic.

1 Introduction

Wireless sensor networks (WSNs) are becoming more prevalent in our daily lives. From smart wristwatches to smart thermostats, everything is connected and measuring our surroundings. From a communications point-of-view, this so called Internet-of-Things has two main challenges, (1) how can we allow all these nodes to send and receive data with a reasonable delay and (2) how to do all this with a minimum of energy. The first challenge is mostly overlooked as the general focus of wireless communication is often throughput or spectral efficiency. However, in the coming years delay will become more important, and important physical layer (PHY) innovations need to be evaluated at medium access control (MAC) layer. New applications where tactile feedback is necessary, like exoskeletons for elderly people and self-driving cars all need low delay communication, beyond what is currently guaranteed with most communication standards [1]. The second challenge is inversely proportional to the first one, meaning that typically, when a network is optimised for energy, the delay increases. Lowering the energy consumption of sensor nodes is necessary because these nodes should have a lifetime of 10 years. We show that both challenges can be tackled by using full duplex wireless communication, i.e., full duplex promises an improved uplink/downlink delay at MAC layer without an additional energy penalty.

By using full duplex it is possible to send and receive data in the same time and frequency slot. To do this the self-transmitted signal needs to be removed, ideally below the noise floor. This should be simple because the self-transmitted signal is a known signal and can thus be subtracted from the received signal. In practice however it is more complicated due to non-idealities in the front-end [2]. In most full duplex designs [3], [4], the self-transmitted signal is first cancelled in the analog domain to
prevent the analog-to-digital convertor from saturating and then the residual interference is cancelled in the digital domain [2], [3], [4]. Various research groups have shown that it is possible to cancel the self-transmitted signal to enable full duplex. Recently, this has been proven to be feasible using commodity hardware [2], however thus far the focus has always been on achieving a higher throughput in Wi-Fi systems.

WSNs rarely use Wi-Fi because of the high power consumption, instead they use the IEEE 802.15.4 standard [5]. In this paper we analyse the effect of using full duplex in WSNs connected in a star topology, and prove that the energy consumption is not increased by doing so, making it a viable candidate for next generation low-delay WSNs. We modify the 802.15.4 PHY and MAC layers to accommodate for full duplex and run simulations in MATLAB. The most commonly used mode of this MAC is the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) mode, which we will model here. We will show that even this MAC model, although not designed for low delay, can benefit greatly from the introduction of full duplex. The output of these simulations is then linked with a novel full duplex energy model.

Full duplex can solve one of the biggest problems that make QoS guarantees for wireless communication challenging, i.e., collisions. Every time a collision takes place, the medium will be busy and other nodes will have to defer their transmissions. The colliding nodes will have to try again after the collision. Numerous attempts have been made to solve the collision problem, e.g., CSMA/CA tries to avoid collisions by randomly backing off, more on this in Section 2. However CSMA/CA does not work for hidden terminals because the nodes cannot sense each other. Further, CSMA/CA doesn’t solve the issue for a large number of nodes as the probability to send successfully decreases drastically with the number of nodes [6].

Full duplex easily expands to CSMA with collision detection (CSMA/CD) because whenever a node is transmitting a packet to the coordinator, it can use the downlink slot to sense for collisions. However this collision detection only works when both colliding nodes are in range of each other, because otherwise the nodes will not be able to detect each others transmission and the hidden terminal problem occurs. Our full duplex design enables collision detection while also allowing to solve the hidden terminal problem or increase throughput when uplink/downlink traffic is balanced. This allows the coordinator to transmit a downlink packet without affecting the network capacity. We will also show that full duplex lowers the energy consumption and allows more nodes to be active in the network.

This paper is constructed as follows, we first briefly explain the IEEE 802.15.4 standard. Next we will show how WSNs can benefit from full duplex. In section 4 we explain the simulation model and in Section 5 we explain our full duplex energy model and finally we look at our simulation results.

2 IEEE 802.15.4 slotted CSMA/CA

In this section we explain the IEEE 802.15.4 slotted CSMA/CA mode. The communication in this mode is build around a superframe which starts with a beacon sent by the coordinator. All nodes synchronise to the superframe using this beacon. The superframe is divided in multiple backoff slots, the beacon is always sent in the first slot, then the active portion of the superframe follows where nodes can contend for the medium in the contention access period or if they have a slot given to them, they can send in the contention free period and then an inactive portion can follow.

Whenever a node wants to transmit a packet it will need to follow the steps defined in Figure 1. The node will first locate the backoff period boundary to synchronise itself with the superframe slots. Next it will wait for a random time, this random backoff assures that it is unlikely that multiple nodes try to send simultaneously. Next the nodes perform clear channel assessment (CCA) for two backoff slots. If both times the medium is free they will send their packet. If the channel is not free, the nodes will
increase their backoff exponent to backoff even further. If the node has reached its maximum number of backoffs, the transmission has failed.

3 Benefits of full duplex for WSNs

WSNs can benefit from full duplex in a number of ways. First the node is able to transmit and receive at the same time, meaning that it can receive a downlink packet from the coordinator while transmitting an uplink packet. Secondly a full duplex node can use the full duplex downlink slot for collision detection. We will now explain both benefits, and show how they can be implemented in enhanced 802.15.4 protocols.

3.1 Full duplex transmissions

Full duplex sensor nodes have two slots simultaneously available, one for downlink and one for uplink. Figure 2a shows the message exchanges for full duplex, it is very similar to the exchange for downlink packets. First the coordinator will announce in the beacon that it has a packet available for the node. If the node has an uplink packet waiting it will, when it has successfully acquired the medium, send a data request packet to announce a full duplex opportunity. The coordinator will acknowledge and then both will send data in full duplex followed by a full duplex acknowledgement (ACK).

In terms of transmission delay, these full duplex transmissions take 1.92ms longer than normal uplink transmissions because of the data request and ACK. Compared to
downlink transmissions, there is no difference in transmission delay. For low throughput networks this would increase the delay, however in this regime most transmissions will be half duplex because both coordinator and node will not have a packet ready at the same time. In high throughput networks the delay is mainly dominated by collisions, here full duplex transmissions solve some of the congestion problems.

3.2 Full duplex collision detection

With full duplex it is possible to implement collision detection, however because of the limited range of the sensor nodes this algorithm only works for non-hidden nodes. Our solution solves both the collision detection and hidden terminal problem and is shown in Figure 2b. To address the hidden terminal problem, the coordinator will as soon as it senses a packet, transmit a dummy packet to let all other nodes know that the channel is not idle. If it then discovers there has been a collision, it will stop transmitting this dummy packet. The transmitting nodes will listen on the downlink channel for this dummy packet and will stop transmission as soon as they no longer receive this dummy packet. In a way the dummy packet acts as an instantaneous acknowledgement of the uplink packet or can be used to transfer other control information.

Whenever there is no collision this scheme will introduce some overhead in terms of power consumption because each transmitting node will have to listen the whole time it is transmitting. However in the CSMA/CA protocol (Figure 1) a node needs to sense the channel idle (CCA) for two consecutive timeslots before it can transmit anything. Therefore, because of the dummy packet, it is only possible that a collision occurs in the first timeslot. This collision will occur when the delay between the reception of the uplink packet and the transmission of the dummy packet is larger than the CCA length of the other nodes. The other nodes will then assess the medium free and start transmission in the next timeslot. A collision will thus take place in the next timeslot and the coordinator will react to this by stopping the transmission of the dummy packet. This means that it is sufficient that the transmitting node only listens to the dummy packet for two timeslots, after that it can turn of its receiver to conserve energy.

Both changes are fully backwards compatible with current 802.15.4 networks, furthermore full duplex nodes can coexist with half duplex nodes as the packet and contention structure is basically kept standard compliant, additionally the dummy packet also solves the hidden terminal problem for legacy nodes as well. To accommodate for all this we’ve added an extra full duplex state to an existing energy model as will be explained in Section 5. In the next section we will first explain our simulation model.

4 Simulation model

To compare our novel full duplex MAC protocol, we have built a simulator which is based on [6]. The pseudocode of our simulator is shown in Algorithm 1. The variables in italic are used for the energy calculations in the next section. Each iteration of the for loop depicts a backoff slot of the MAC protocol. The simulator keeps also track of the state of the different nodes, to know the current backoff slot’s state. These tracking variables are not shown in Algorithm 1.

The simulator first checks if the current slot is a beacon slot, if so we check for downlink traffic and update the necessary variables. If it is not a beacon slot, the simulator checks if the channel is idle and which nodes are ready to transmit. If there is a collision and collision detection is enabled, the number of transmission slots is set to two, as explained in the previous section. Otherwise the nodes will transmit for the full packet length. If there is no collision, the simulator checks the mode of the transmitting node and increments all the necessary variables for the energy calculation. The variable ‘packetDelay’ is used to calculate the delay, it is the time between the
arrival of the packet and the reception of the acknowledgement after the transmission. If the channel is busy, we follow the regular CSMA/CA algorithm. In the end we update the packet arrivals and keep track of the arrival time.

Algorithm 1 Pseudocode of the simulator

1: initNodes(1:N) = uniformly distributed;
2: upPacketArrival(1:N) = poisson dist.;
3: downPacketArrival(1:N) = poisson dist.;
4: mode(1:N): %uplink, downlink, full duplex
5: for slot=0:nbSlots do
6: if beaconSlot then
7: check nodes i for downlink;
8: update delay(i), mode(i);
9: increment N_B;
10: else
11: check idle nodes i;
12: increment N_I(i);
13: check nodes i performing CCA;
14: increment N_CCA(i);
15: if channel idle then
16: check nodes j from i ready to TX;
17: if enough slots left for TX then
18: if more than one node j then %collision
19: txSlots(j) = 2;
20: increment N_TXcd(j);
21: if collision detection then
22: increment N_TXcd(j);
23: else
24: txSlots(j) = packetLength;
25: increment N_TXncd(j);
26: end if
27: else
28: switch (mode(j))
29: case uplink:
30: increment N_TXcd or N_TXncd;
31: increment N_ACKr;
32: case downlink:
33: increment N_TXcd or N_TXncd;
34: increment N_RX, N_ACKt, N_ACKr;
35: case full duplex:
36: increment N_TXcd, N_ACKr;
37: increment N_RXTX, N_ACKr;
38: end switch
39: end if
40: defer nodes j;
41: end if
42: %channel is busy
43: update i according to CSMA/CA;
44: end if
45: end if
46: end if
47: update packet arrivals;
48: end for

5 Energy model

Currently there is no off-the-shelf support for full duplex. Therefore we developed an energy model based on a popular 802.15.4 chipset, the TI CC2420 [7]. We start from the energy model from [8], which consists of four states: Shutdown (clock is turned off), Idle (clock is turned on), transmit (TX) and receive (RX) and add a fifth state: full duplex (RXTX) (both RX and TX are active). A node is in the full duplex state whenever it uses collision detection or transmits and receives a packet in full duplex.

5.1 Full duplex energy

To describe the energy model of a full duplex wireless transceiver we need the power each state consumes and the transition energy between states. Figure 3 shows the five different states and state transitions. The four basic states are unchanged with respect to [8], only the full duplex state is added. In this state both the transmitter and receiver chain will be active, both operate on the same frequency so only one Phase Locked Loop (PLL) is needed. From a similar chipset [9] we identified $P_{PLL}$ to be around 9mW. Table 1 gives an overview of the power consumption of the different states. The power consumption of the full duplex state is $P_{RXTX} = P_{RX} + P_{TX} - P_{PLL}$. 
Figure 3: Transition diagram with transition energies and times

Table 1: Power consumption of the different states

<table>
<thead>
<tr>
<th>State</th>
<th>Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shutdown</td>
<td>144mW</td>
</tr>
<tr>
<td>Idle</td>
<td>712µW</td>
</tr>
<tr>
<td>RX</td>
<td>35.28mW</td>
</tr>
<tr>
<td>TX</td>
<td>30.67mW</td>
</tr>
<tr>
<td>RXTX</td>
<td>56.95mW</td>
</tr>
</tbody>
</table>

Figure 3 also shows the transition energies and times. The transition from Idle to RXTX is calculated using [8]'s rule of thumb ($E = TIV_{DD}$). The transition time is the same as the one for RX and TX because most of the time is lost in the PLL. From our MAC model, we can count states and state transitions, to determine energy cost.

5.2 Average power consumption

The average power consumption of the sensor node is given in Eq. (1), which consists of the sum of all state energies with $s \in \{\text{shutdown, idle, RX, TX, RXTX}\}$ and all state transitions with $t \in \{\text{SI, IR, IT, IRT}\}$, representing a transition from Shutdown to Idle (SI), from Idle to RX (IR), from Idle to TX (IT) and from Idle to RXTX (IRT), respectively, divided by the total simulation time. The number of slots is equal to the number of simulated timeslots and the length of a timeslot is equal to 0.32ms.

$$P_{\text{avg}} = \frac{\sum_s P_s T_s + \sum_t P_t T_t}{N_{\text{slot}}T_{\text{slot}}}$$

We can now link the simulations with the energy model by using the variables from Algorithm 1, which counted the number of times a certain action is performed. Table 2 gives the subtimings and if not defined they are equal to the slot time (0.32ms).

Several protocol states map to the RX power state, as a node is in receive mode during CCA, packet reception, ACK reception and beacon reception. The total time in the receive state is $T_{RX} = \sum_s N_s T_s$, with $s \in \{\text{CCA, RX, B, ACKr}\}$.

The transmit state is mainly used in half duplex mode without collision detection (ncd), but remember from Section 3, a node will only detect collisions during the first two timeslots so for the remaining time he will be in the TX state hence the final term of $T_{RX} = \sum_s N_s T_s + N_{TXcd}(T_{\text{packet}} - T_{TXcd})$, with $s \in \{\text{TXncd, ACKt}\}$.

Table 2: Subtimings

<table>
<thead>
<tr>
<th>T_{CCA}</th>
<th>T_{TXcd}</th>
<th>T_{SI}</th>
<th>T_{IR}</th>
<th>T_{IT}</th>
<th>T_{IRT}</th>
</tr>
</thead>
<tbody>
<tr>
<td>128µs</td>
<td>640µs</td>
<td>970µs</td>
<td>194µs</td>
<td>194µs</td>
<td>194µs</td>
</tr>
</tbody>
</table>
Similarly, a node is in the full duplex state the first two timeslots when transmitting with collision detection and when he’s transmitting a packet or ACK in full duplex. The total time is equal to \( T_{RXTX} = \sum_s N_s T_s \), with \( s \in \{TXcd, RXTX, ACKrt\} \).

Nodes are in the idle state when they are backing off and for the remainder of the slot when they perform CCA. This gives a total time of \( T_{idle} = N_I T_I + N_{CCA}(T_{Slot} - T_{CCA}) \).

Finally, nodes are in the shutdown state when they are not in the other states.

6 Simulation Results

In this section we analyse the delay and energy performance for a range of relevant network configurations. We compare the results of our proposed MAC enhancement with the standard half duplex CSMA/CA protocol. First in Figure 4a, we look at uplink/downlink throughput fairness in our network. We ran simulations with 25 nodes, the total uplink network traffic is fixed at 3 packets/s of 100 bytes. The throughput in the half duplex case starts to decrease starting from 20 kbits/s of downlink traffic, this is mainly due to downlink prioritization and collisions. In the full duplex case, uplink is not affected by the downlink traffic because they can be transmitted simultaneously, showing the effect of full duplex transmissions.

(a) Effect of downlink traffic on uplink(b) Energy per bit using a fixed throughput of 3 packets/s with 10% of the packets downlink.

(c) Delay analysis using a fixed throughput of 3 packets/s with 10% of the packets downlink.

Figure 4: Simulation results

In most WSNs there is an asymmetry between up and downlink, therefore in the following results, only 10% of all traffic in the network is downlink. Figure 4b shows the energy per bit. The nodes in this figure transmit packets with a constant throughput of 3 packets/s of 100 bytes each. We see that full duplex starts saturating much later than
half duplex. When the network is saturated, the energy consumption with full duplex is lower because of the collision detection. The nodes are less in the high-energy RX or TX states. In the non-saturated region, full duplex performs slightly worse because of the increased energy consumption of the collision detection. In this region, there aren’t much collisions so collision detection is not necessary here. Overall, we can conclude that the energy penalty is low for full duplex. More importantly, nodes could easily learn when to do full duplex collision detection as function of networking conditions.

Figure 4c shows the delay results, it uses the same parameters as in the previous figure. Figure 4c shows us that it takes full duplex almost double the amount of active network nodes before the network starts to saturate. The saturation is caused by collisions in the network, and full duplex collision detection has the potential to delay network collapse with 50%. If we compare half duplex and full duplex in non-saturated regions, full duplex is only slightly superior than half duplex in terms of delay.

7 Conclusions

In this paper, a novel full duplex MAC protocol and energy model for full duplex WSNs is presented. The MAC protocol implements a collision detection scheme using an immediate acknowledgement in the form of a dummy packet. We have shown that using full duplex, downlink traffic is almost free and it does not decrease the uplink traffic. We also showed that full duplex only starts to saturate when almost double the nodes are active compared to half duplex. Our novel energy model uses an extra full duplex state. Simulations have shown that when the network is saturated, it is better to switch to full duplex. In a non-saturated network, full duplex is only slightly worse.

References

Dominant eigenmode transmission in MIMO FBMC for frequency selective channels

François Rottenberg  Jérôme Louveaux
ICTEAM
Université catholique de Louvain, Place du Levant 2, Louvain-la-Neuve, Belgium
francois.rottenberg@student.uclouvain.be  jerome.louveaux@uclouvain.be

Abstract
In this paper, we consider a MIMO transmission using FBMC (filter-bank based multicarrier) modulation where channel state information (CSI) is assumed to be available at the transmitter. The so-called dominant eigenmode transmission (DET) scheme is investigated. This technique uses a joint beamforming scheme to transmit a single stream of data and can extract both full spatial diversity and a high array gain. It has already been shown that it can be applied to the FBMC modulation if the channel is not too selective in frequency. The paper then aims at comparing the performance of the DET technique with FBMC (more precisely the OFDM/OQAM version) and OFDM respectively, in a variety of channels including highly frequency selective channels. It is known that FBMC outperforms OFDM in mildly selective channels thanks to its better energy-efficiency. The paper investigates the effect of higher frequency selectivity on the DET technique both in the case of FBMC and OFDM. It is shown that FBMC slightly suffers from inter-carrier and inter-symbol interference while OFDM robustness heavily depends on the cyclic prefix length. To improve the FBMC behavior against channel selectivity, a low complexity per-subcarrier equalizer is designed based on the frequency sampling (FS) technique. The paper analyzes the trade-off between energy-efficiency and robustness against frequency selective channels that the two modulation techniques offer.

1 Introduction

Most multicarrier systems are using orthogonal frequency division multiplexing (OFDM) as a modulation scheme nowadays. This scheme has many advantages including a low complexity implementation based on the fast Fourier transform and relying on the transmission of the cyclic prefix (CP) to fight against channel selectivity. However, the rectangular pulse shaping around each subcarrier induces very high spectral leakage, which significantly decreases the system flexibility, especially for user synchronization or cognitive radio issues. Therefore, filter-bank based multicarrier (through its OFDM/OQAM version) is becoming more and more interesting. This scheme doesn’t require any CP leading to a larger spectral efficiency. At the cost of a higher complexity implementation, the pulse shape around each subcarrier is made much more frequency selective and is spread only up to the adjacent subcarriers.

Furthermore, the interest in multiple input multiple output (MIMO) systems has grown very rapidly this last years. More specifically, the focus in this article is to look at spatial diversity schemes. It has been shown in [3] that dominant eigenmode transmission (DET) is applicable to FBMC in the frequency flat case. Supposing channel state information (CSI) at transmitter and receiver sides, this technique uses a joint beamforming scheme to transmit a single stream of data and can extract both full spatial diversity and a high array gain [2].
This paper has two main objectives. Firstly, it investigates how DET applied to FBMC will perform under channel selectivity. A comparison will be made with OFDM under well defined conventions. The DET implementation will be made at the symbol rate and not at the frequency sampling rate as in [3], making it more efficient. Secondly, this paper investigates if it is possible to extend efficiently the frequency sampling-based SISO equalizer derived in [4] to the MIMO case and more specifically to the DET and in an alternative way than [5].

2 System Model

The system considered here is depicted in Figure 2. A MIMO transmission scheme using FBMC (filter-bank based multicarrier) modulation is investigated. For the sake of clarity, we first review the basic description of FBMC transmission in a SISO scenario and then extend it to the MIMO case.

2.1 SISO system model

The block diagram of a SISO FBMC transmission scheme is shown in Figure 1.

The purely real PAM symbols entering the OQAM pre-processing stage are denoted by \( d_{k,n} \) for subcarrier \( k \) at instant \( n \). The symbol duration is given by \( T_0/2 \) and \( \Delta f = 1/T_0 \) is the subcarrier spacing. After multiplication by \( \theta_{k,n} = j^{k+n} \), the symbols \( d_{k,n} \) that were purely real are now alternatively purely real and purely imaginary giving the OQAM pattern.

In the synthesis filterbank (SFB), the symbols are first upsampled and next filtered by the synthesis filters \( g_k(m) \). The indexes \( n \) and \( m \) refer to the time instants at the low rate \( 2/T_0 \) and at the sampling rate \( 1/MT_0 \) respectively. The output of the SFB is obtained by the sum of the signals coming from each subcarrier branch at the rate \( 1/MT_0 = 1/T_0 \). At the analysis filterbank (AFB), the received signal is first filtered by the analysis filters \( f_k(m) \) and then downsampled. The subchannel filters are obtained by an exponential modulation of a real-valued linear-phase FIR prototype filter \( p[m] \). The prototype filter used in this paper is defined as in [1] with overlapping factor \( K = 4 \) and roll-off factor \( \rho = 1 \) which means that \( p[m] \) is spread over 4 symbols in the time domain and its bandwidth extends up to the two neighboring subcarriers.

The signal transmitted can be written as the sum of the signals coming from each branch of the SFB

\[
s[m] = \sum_{n=-\infty}^{+\infty} \sum_{k=0}^{M-1} d_{k,n} \theta_{k,n} g_k[m - nM/2]
\]  

(1)
and the received signal is given by

\[ r[m] = s[m] * h[m] + w_m, \]  

where \( * \) denotes the convolution operator and where \( w_m \) is a zero mean circularly symmetric complex white Gaussian noise. The received signal is then processed by the AFB filters and downsampled by a factor \( M/2 \). The expression of the AFB output \( y_{k,n} \) can be greatly simplified by assuming that the channel is flat at the subcarrier level and on the adjacent subcarriers. However, this assumption is not always true and will be revisited later on. The AFB outputs are given by

\[
y_{k,n} = H_k \sum_{k'=-K}^{K} \sum_{n'=-k+1}^{k+1} d_{k',n'} \theta_{k',n'} t_{k-k',n-n'} + w_{k,n} \tag{3}
\]

where \( w_{k,n} \) denotes the noise \( w_m \) filtered and downsampled at subcarrier \( k \) and where

\[
y_{k,n} = H_k \theta_{k,n} (d_{k,n} + j u_{k,n}) + w_{k,n}, \tag{4}
\]

Table 1: Transmultiplexer response - Overlapping factor \( K = 4 \)

<table>
<thead>
<tr>
<th>subcarrier/time</th>
<th>n-4</th>
<th>n-3</th>
<th>n-2</th>
<th>n-1</th>
<th>n</th>
<th>n+1</th>
<th>n+2</th>
<th>n+3</th>
<th>n+4</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-1</td>
<td>0.0054</td>
<td>0.0429j</td>
<td>-0.1250</td>
<td>-0.2058j</td>
<td>0.2393</td>
<td>0.2058j</td>
<td>-0.1250</td>
<td>-0.0429j</td>
<td>0.0054</td>
</tr>
<tr>
<td>k</td>
<td>0</td>
<td>-0.0668</td>
<td>0.0002</td>
<td>0.5644</td>
<td>1</td>
<td>0.5644</td>
<td>0.0002</td>
<td>-0.0668</td>
<td>0</td>
</tr>
<tr>
<td>k+1</td>
<td>0.0054</td>
<td>-0.0429j</td>
<td>-0.1250</td>
<td>0.2058j</td>
<td>0.2393</td>
<td>-0.2058j</td>
<td>-0.1250</td>
<td>0.0429j</td>
<td>0.0054</td>
</tr>
</tbody>
</table>

the transmultiplexer response \( t_{k-k',n-n'} \) is given in Table 1 and corresponds to the contribution to the AFB output on subcarrier \( k \) at time instant \( n \) due to an impulse \( \delta_{k',n'} \) at the SFB input. Due to the roll-off factor of 1, the contributions from non-adjacent subcarriers are negligible and are thus not shown in Table 1. Remembering that \( \theta_{k,n} = j^{k+n} \), one can see the advantage of OQAM modulation. Due to the OQAM pattern of real and imaginary symbols, the useful symbol \( d_{k,n} \) is not mixed with its neighboring elements except the very small interference due to \( t_{k,n+2} \) because of the NPR filter design. \( j u_{k,n} \) is called the intrinsic interference ans is purely imaginary. Finally, channel equalization and OQAM demodulation is performed by

\[
\tilde{d}_{k,n} = \Re \{ \theta_{k,n}^* H_k^* y_{k,n} \} \tag{5}
\]

\[
\tilde{d}_{k,n} = |H_k|^2 d_{k,n} + H_k^* w_{k,n} \tag{6}
\]

2.2 MIMO system model

![MIMO filter-bank multicarrier system](image-url)
We now extend the previous SISO model to the MIMO environment. Figure 2 illustrates the FBMC MIMO system studied in this paper. The transmission of one stream of data is considered. The symbols that are entering the SFB are precoded by the vector $a_k \in \mathbb{C}^{N_T \times 1}$. If we make the assumption of approximately flat channel inside each subcarrier, the output of the AFB can be written as

$$z_{k,n} = \sum_{n'=-K}^{K} \sum_{k'=k-1}^{k+1} H_{k'}a_k d_{k',n'}\theta_{k',n'} t_{k-k',n-n'} + w_{k,n} \quad (7)$$

If we also assume that the channel is frequency flat on the adjacent subcarriers, that is $H_k = H_{k+1}$, it simplifies to

$$z_{k,n} = H_k a_k \sum_{n'=-K}^{K} \sum_{k'=k-1}^{k+1} d_{k',n'}\theta_{k',n'} t_{k-k',n-n'} + w_{k,n} \quad (8)$$

$$= H_k a_k (d_{k,n} + j u_{k,n})\theta_{k,n} + w_{k,n} \quad (9)$$

$$H_k = \begin{pmatrix} H_{k1}^{11} & \cdots & H_{k1}^{1N_T} \\ \vdots & \ddots & \vdots \\ H_{kN_T}^{N_T1} & \cdots & H_{kN_T}^{N_TN_T} \end{pmatrix} \quad (10)$$

where $H_{kij}$ is the channel coefficient between transmit and receive antenna $i$ and $j$ at subcarrier $k$. The decoding vector $g_k^H \in \mathbb{C}^{N_R \times 1}$ is then applied

$$y_{k,n} = g_k^H z_{k,n} \quad (11)$$

followed by the OQAM demodulation $\tilde{d}_{k,n} = \Re\{y_{k,n}\theta_{k,n}^*\}$ giving

$$y_{k,n} = g_k^H H_k a_k (d_{k,n} + j u_{k,n})\theta_{k,n} + g_k^H w_{k,n} \quad (12)$$

$$\tilde{d}_{k,n} = \Re\{g_k^H H_k a_k (d_{k,n} + j u_{k,n})\} + \Re\{g_k^H w_{k,n}\theta_{k,n}^*\} \quad (13)$$

In [3], the precoding and decoding stages were performed at the rate $M/T_0$ while it can be checked in Figure 2 that it is done here at the low rate $2/T_0$. This is a major difference causing the proposed scheme to have a much lower complexity.

3 Dominant eigenmode transmission under channel selectivity

This section explains how to choose the precoding and decoding vector $a_k$ and $g_k$ through the dominant eigenmode transmission. However, this technique is optimized for frequency flat channels and not for selective channels. To counterbalance the effect of channel selectivity, the choice being made here is to keep the same precoding and decoding vectors as in the frequency flat case and add a per-subcarrier equalizer at the AFB output. Its design is based on the frequency sampling technique.

3.1 Dominant eigenmode transmission

If the channel is frequency flat at the subcarrier level, it can be shown that the optimal beamforming scheme is the dominant eigenmode transmission. In practice, $a_k$ and $g_k$
are chosen as the input and output singular vectors respectively corresponding to the maximum singular value $\sigma_{k,max}$ of $H_k$, the channel matrix, giving

$$\hat{d}_{k,n} = \sigma_{k,max} d_{k,n} + \Re \left\{ g_k^H w_k d_{k,n} \right\}. \quad (14)$$

In [2], it is proven that this scheme reaches a spatial diversity gain of $N_T N_R$ and an array gain of $\sigma_{k,max}^2$. However, this expression relies strongly on the assumption about the channel flatness. As the channel will get more frequency selective, the performance will degrade progressively.

### 3.2 Frequency sampling-based equalizer design

We now look at the case where the channel is becoming frequency selective. We know that dominant eigenmode transmission is only optimal for frequency flat channel. Nevertheless, we will still use the same precoding and decoding vectors optimized for the frequency flat case to avoid a more complex optimization. However, to counterbalance the performance degradation, we add after the decoding vector a per-subcarrier equalizer $C_k(z)$ as shown in Figure 3.

![Receiver block diagram for one particular subcarrier](image)

**Figure 3:** Receiver block diagram for one particular subcarrier

In [4], the authors derive a low complexity per-subcarrier equalizer based on the frequency sampling technique for the SISO case. This technique is here extended to the dominant eigenmode transmission. The principle of the technique is to choose certain values for the equalizer coefficients so that its frequency response is forced to pass through some target frequency points in the subchannel of interest. The target frequency points are chosen thanks to a ZF or MMSE criteria. If the channel is not too selective in frequency, we can expect the equalizer response between the target frequency points to approximate the optimal ZF or MMSE response. The choice of a three coefficient per-subcarrier equalizer has been made here.

The signal received at the AFB output $y_{k,n}$ can be seen as having passed through an equivalent SISO channel $h_k[n] = g_k^H h[n] \alpha_k, n = [0, ..., L - 1]$ with

$$h[n] = \begin{pmatrix} h^{11}[n] & ... & h^{1N_T}[n] \\ ... & ... & ... \\ h^{N_R1}[n] & ... & h^{N_RN_T}[n] \end{pmatrix} \quad (15)$$

where $h^{ij}[n]$ denotes the $n^{th}$ tap coefficient of the channel impulse response between transmit antenna $i$ and receive antenna $j$. Taking the Fourier transform of $h_k[n]$ and evaluating at three evenly-spaced target frequency points in subchannel $k$, that is $H_k^i$ at frequency $\Omega_i = \frac{2\pi(2k+i)}{2M}$ for $i = [-1, 0, 1]$. The target frequency points are obtained depending on the ZF or MMSE criteria as

$$G_{k,ZF}^i = \frac{1}{H_k^i} \quad (16)$$

$$G_{k,MMSE}^i = \frac{(H_k^i)^*}{|H_k^i|^2 + \eta} \quad (17)$$
where $\eta$ denotes the noise to signal ratio. The expression of the 3-taps complex FIR non causal equalizer can be written as:

$$C_k(z) = c_{k,-1}z + c_{k,0} + c_{k,1}z^{-1}$$

(18)

Forcing the equalizer to pass through the three target frequency points gives a system of three equations and three unknowns that can be solved easily and finally gives the equalizer coefficients.

$$C_k(e^{-j\pi/2}) = -jc_{k,-1} + c_{k,0} + jc_{k,1} = G^{-1}_{k,ZF/MMSE}$$

(19)

$$C_k(e^0) = c_{k,-1} + c_{k,0} + c_{k,1} = G^0_{k,ZF/MMSE}$$

(20)

$$C_k(e^{j\pi/2}) = jc_{k,-1} + c_{k,0} - jc_{k,1} = G^{+1}_{k,ZF/MMSE}$$

(21)

### 4 Simulation results

All presented simulations are performed in quasi-static conditions, that is each channel realization is considered constant over the transmitted frame. The different taps of the channel are independent zero mean circularly symmetric complex Gaussian random variables. The subcarrier spacing is $\Delta f = 1/T_0 = 15kHz$ with $M = 128$ subcarriers leading to $T_s = T_0/M = 520ns$. An exponentially decaying Power Delay Profile (PDP) is considered*. The prototype filter has a roll-off factor $\rho = 1$ and an overlapping factor $K = 4$. QPSK symbols are sent. Perfect channel knowledge is assumed at transmitter and receiver sides ($N_R = N_T = 2$). Various frequency selective channels are considered by changing the number of taps from 1 to 60.

![Symbol Error Rate vs. SNR](image1.png)

(a) Effect of channel frequency selectivity on FBMC

![Symbol Error Rate vs. SNR](image2.png)

(b) Difference between ZF/MMSE criteria on the equalizer

Figure 4: Effect of frequency selectivity channel - Performance of the ZF/MMSE equalizer

Figure 4(a) illustrates the effect of increasing channel selectivity on FBMC. We see that as the length of the channel increases, the assumption about its flatness is not holding anymore and interference appears, leading to saturation of the BER at high

*The last tap of the channel has a power of $-20dB$ relatively to the first tap. All channels are normalized so that the total channel power averaged over all channel realizations is equal to 1.
SNR. In other words, the interference caused by inter-antenna, inter-subcarrier and inter-symbol interference is becoming bigger than the noise and imposes a SER floor.

Figure 4(b) shows the performance gain due to the designed equalizer. We see that the performance gain is more significant for highly selective channels. Moreover, the performance difference induced by using the target frequency points based on the ZF or MMSE criteria is negligible and can only be seen for very highly selective channel in the low SNR region. This means that the equalizer design doesn’t even require the working SNR, which lowers its complexity.

Figure 5 compares OFDM and FBMC as a function of the channel frequency selectivity for different $\frac{E_s}{N_0}$. However, obtaining a perfectly fair comparison OFDM/FBMC is very difficult. The two modulation schemes are depending on different parameters and have different characteristics. The choice being made here is to normalize each system for the same $\frac{E_s}{N_0}$ so that the same total energy is sent over a frame transmission. This means that the throughput rate of CP-OFDM is smaller than that of FBMC/OQAM. Another way of comparing the two schemes would have been to use a redundancy code of rate $\frac{M}{M+CP}$ so that the throughput would be equal. It would result in better FBMC performance, as shown in [4]. Different length of CP are considered. For the sake of comparison, the CP lengths in LTE are : $T_{CP,normal} = 4.7\mu s$ or $T_{CP,extended} = 16.7\mu s$ for $T_0 = 66, 7\mu s$ leading to $CP_{normal} \approx M/16$ and $CP_{extended} \approx M/5$.

![Figure 5: Performance comparison OFDM/FBMC over frequency selective channels for different $\frac{E_s}{N_0}$](image)

(a) $\frac{E_s}{N_0} = 4.5dB$, $\frac{E_s}{N_0} = 9dB$

(b) $\frac{E_s}{N_0} = 13.5dB$

In Figure 5(a), we can see that FBMC outperforms OFDM for very short channels. This is due to the lack of energy-efficiency of OFDM transmitting the CP. As expected, this lack of energy-efficiency is bigger for a larger CP. At $\frac{E_s}{N_0} = 4.5dB$, the noise power is predominant and FBMC performance only slightly degrades for increasing selective channels, this allows FBMC to be more energy-efficient.

For bigger $\frac{E_s}{N_0}$, OFDM shows its very good robustness against channel selectivity. OFDM is not affected as far as the channel length is smaller than the CP. Once the channel length gets bigger, OFDM degrades quicker than FBMC. This can be explained by the fact that only adjacent subcarriers are interfering in FBMC while in OFDM, once the CP gets passed, all subcarriers are interfering with each other causing in the end a bigger total interference. The gap between FBMC and OFDM will thus directly depend on the CP length. For a small CP ($CP = M/32$ in Figure 5(b)), FBMC regains its gap against OFDM very quickly as the channel gets longer. For larger CP
OFDM is very robust and the gap OFDM/FBMC is only reduced for very long channels.

5 Conclusion

The behavior of DET, which is computed based on a flat channel assumption, has been investigated for varying selective channels. The DET implementation is made more efficient than in [3] by performing the precoding and decoding stage at the low rate $2/T_0$ instead of the sampling rate $T_0/M$. The FS-based equalizer derived in [4] has been extended to the DET transmission and results in significant performance improvement in highly selective channels. From the comparison OFDM/FBMC, it appears clearly that OFDM is very robust against channel selectivity. However, for very short or very long channels, FBMC takes the lead thanks to its better energy-efficiency and its well chosen time-frequency location. The gap between the two modulation schemes will directly depend on the length of the CP.

References


\( CP = M/4 \) in Figure 5(b),
Digital Intensive Architecture Exploration for Low-power 60GHz Polar Transmitter

Zhijie Xu, Chunshu Li, Andre Bourdoux, Khaled Khalaf, Sofie Pollin, Liesbet Van der Perre
Interuniversity Micro-Electronics Center (IMEC) vzw
Department of Circuits and Systems
Kapeldreef-75, Leuven, B-3001, Belgium
first name.last name@imec.be

Abstract

Wireless communication continues to evolve, which is enabled by technology scaling. However, the advantages of scaling for digital circuits can not be projected directly to analog circuits. Actually, analog circuits suffer from scaling, e.g. loss of signal strength, increasing variance, which makes single chip integration challenging with new technology node. Digital intensive and digitally assisted transceivers are proposed to achieve an optimal solution. This paper presents an analysis of 60GHz digital polar transmitter. The 60GHz communication system is modelled in Matlab to evaluate system performance based on error vector magnitude (EVM) and power spectrum density (PSD). Also, a power consumption analysis is provided to prove polar transmitter a low-power solution.

Keywords: Digital Intensive, 60 GHz Transmitter, Rectangular to Polar Conversion

1 Introduction

Wireless communications technology continues to evolve and regularly yields new standards for different application areas. To enable seamless and transparent inter-working between these different wireless access systems, communication systems are moving towards an era where ubiquitous connectivity and growing levels of integration will be essential for most applications. In this context, the mobile devices impose a real challenge for designers since they incorporate several concurrent constraints, such as battery life, cost, performance, size and weight.

On the other hand, the CMOS technology keeps scaling according to Moore’s Law, which provides ever improving cost advantages and processing capabilities to ensure the ever improving of communication technology. Digital gate density doubles and computation power improves with every new technology node. However, analog circuit does not benefit from these advantages of scaling. The voltage resolution is substantially decreased and accurate modelling of transistors becomes impossible. In addition, variation problems become much more severe. Thermal effects, decreased reliability and aging effects are all degrading analog devices. All these disadvantages make the technology used by commercial radio frequency (RF) transceivers lag behind digital circuits. In the end, analog circuit still has to follow the scaling trend of digital design for single chip integration.

Digital intensive and digitally assisted transceivers have been proposed and experimented in the past decades to achieve the above-mentioned challenging goal. The concept of digital RF and digitally assisted RF were proposed in [1][2] and many designs with such philosophies have been presented. When targeting emerging communication standards and deeply scaled technologies, the signal processing design and implementations become much more sophisticated.

Digital intensive transmitter design, as the name also implies, moves digital processing closer to antennas. Recently, 60GHz systems (e.g., IEEE 802.11ad [3]) are
Figure 1: Block diagram of a mm-wave polar transmitter

becoming the target for such designs. A key challenge is that signal processing needs to work at very high sampling frequency which may even stay close to RF. This very high sample rate processing, combined with substantial bit width requirement creates substantial challenges.

In this work, the challenges of signal processing in 60GHz wireless communication are described. One of the solutions, polar transmitter is proposed to deal with this problem. In addition, the effect of bit resolution of signal processing in polar conversion on system performance is analyzed in section 3. In section 4 the power budget for polar transmitter is compared with non-polar solution. Section 5 draws a conclusion for this paper.

2 60GHz Digital Polar Transmitter

In 60 GHz chips, the power amplifier (PA) is usually the most power hungry block. Moreover, in order to overcome signal losses at 60 GHz, phased arrays are often employed and at least the front-ends have to be multiplied with the same number of antenna paths. This increases the PA share in the total chip power consumption. Different applications benefit from improving the PA power efficiency, such as high datarate short-range portable applications that require minimal power consumption for longer battery lifetime and high datarate backhaul systems that transmit with high output powers for longer range communication. Most 60 GHz PAs operate in class-A linear mode [4][5][6] due to the use of variable envelope modulations that are required for high datarates and high spectral efficiency. This causes the PA to work at power efficiency values of less than 5% although values up to 30% could be achieved [4]. In order to improve the PA power efficiency, the PA needs to work in its nonlinear region to utilize the peak efficiency. The polar architecture is one interesting solution that allows the PA to operate in saturation without the need for duplicating the signal path or using power combiners. The phase signal with constant envelop goes to the PA input, while the amplitude is extracted and applied to the PA through a separate modulation path, as shown in Fig.1.

Polar conversion can be done with digital signal processing to avoid the need of an RF limiter that can introduce extra nonlinearity and bandwidth limitations. The amplitude signal can then digitally modulate an RF digital-to-analog converter (DAC) working as a variable-size PA. This eliminates the need to have an additional RF amplitude detection circuit and also avoids modulating the supply.
3 Bit Resolution Analysis

There are several parameters needed to be decided for the polar architecture, including the bit resolution of rectangular signal, also amplitude and phase signal after conversion. The nonlinear polar conversion leads to an spectrum expansion. Fig.2 (a) depicts the PSD of rectangular signal, which is compliant with the spectrum mask of IEEE 802.11ad[3]. After non-linear conversion to polar signal, the spectrum of the converted signal greatly expands, as shown in Fig.2 (b). To avoid the spectrum overlap caused by non-linear conversion, the rectangular signal needs to be firstly upsampled and then digitally filtered before converting to polar signal. The first residual image due to oversampling appears at an offset equal to the sampling frequency. With a symbol rate of 1760MS/s (according to the IEEE 802.11ad standard), an OSR of 6 is normally employed to push the first residual image out of the RF band of 802.11ad standard spanning from 57 GHz to 66 GHz. However an OSR of 4 may still yield a PSD compliant with the spectral mask due to the additional filtering after upsampling, which will significantly reduce the power consumption of DSP.

Bit resolution of rectangular signal and converted polar signal are analyzed to have a better knowledge of the complexity of signal processing in this polar conversion. The complete 802.11ad transmission system with 16QAM modulation is modelled in...
Figure 4: Output signal spectrum with OSF = 6

Figure 5: Output EVM results in terms of quantization accuracies with OSF = 4

MATLAB. A system with OSR of 6 is firstly simulated with different bit resolution. The EVM results are shown in Fig.3. A notable improvement can be seen when increasing the accuracy of transmitted rectangular signal(IQ denoted in the figure) from 6 bits to 7 bits, while minor improvement with further increase. Based on a rectangular signal of 7 bits, the requirements on bit resolution for AM and PM are further investigated, as Fig.3 (b) shows. According to the IEEE802.11ad standard, the constellation error should not be worse than -21dB for 16QAM. With a design margin of 10dB, the design target is -31dB. Note that although there are multiple choices of quantization accuracies to achieve -31dB EVM, the one with minimum bits of AM signal is chosen to make layout easier when routing the digital AM bit-wires to the RFDAC. Fig.4 shows the PSD of the output signal with 7, 5 and 7 bits for rectangular signal, converted AM signal and PM signal respectively, which is compliant with the spectrum mask. The same analysis is done to verify that system with an OSR of 4 still works. Bit resolution of 7, 6 and 7 are chosen for rectangular signal, converted AM and PM signal respectfully, the results are shown in Fig.5 and Fig.6.
4 Power Consumption Analysis

Since the digital front end is tightly connected with RF, the DSP needs to operate at very high sampling frequency which imposes large difficulty on reducing power. The following power consumption calculations can be used to estimate the power budget for the DSP in polar solution. A phased array chip[4] is taken as a reference. The top-level Tx architecture with 4-antenna paths is shown in Fig.7. The PA will be replaced by a variable size RFDAC for amplitude modulation when using the same chip in polar mode. Table 1 shows the advantage of a chip used in polar mode over used in non-polar mode in output power, power consumption and efficiency. With a 5 dB Peak-to-Average Power Ratio (PAPR), the linear PA operates at 5 dB back-off from the 1 dB compression point of 10.8 dBm, which gives a PA efficiency of 4.9% compared to the maximum value of 32.2% in saturation. If the same chip is used in polar mode, the RFDAC input includes only phase information with constant envelop and is allowed to operate in the saturation region. With a PAPR of 5 dB, the amplitude will modulate the RFDAC such that the average output power is 5 dB less than the peak saturated 14dBm(see Fig.8). This causes the average RFDAC size to be $10^{(-5/20)} = 0.56 \times$ the full size, and the power consumption to reduce with the same factor. The PA operating efficiency is then 18.2% in the polar mode compared to 4.9% in the linear mode. The polar mode total Tx output power is 3 dB higher than the linear mode, and the total Tx efficiency goes to 18.94% compared to 7.25% in the linear mode. A fair comparison of power consumption between two different modes should be taken at the same Tx output power. Assuming the same peak saturated efficiency,
the total Tx power consumption in the polar mode reduces to 487.6mW compared to 724mW at the same output power. In order for the DSP to have a minor influence on the total power budget, a value of 10% of the total Tx power consumption should be considered. This concludes an average of 50mW for the extra digital processing required for the polar operation.

The above analysis puts a challenging task on the optimization of algorithms and design techniques of the additional signal processing circuitry. The 50mW budget for signal processing power consumption needs to cover 7040 (1760 \times 4) Msps I/Q to phase/amplitude transformation. We expect aggressive algorithms and circuit level optimizations to achieve this target.

5 Conclusion

In this paper, a digital intensive polar transmitter for 60GHz communication is proposed to deal with degraded analog performance at deeply scaled technology nodes. The performance of this transmitter is evaluated at system level regarding EVM and PSD with Matlab. Design parameters such as bit resolution of transmitted rectangular signal, converted amplitude and phase signal are assessed to achieve optimal performance. The power budget of this polar transmitter is also analyzed with comparison to traditional architecture.
References


Successive Refinement of Gaussian Projections

Giel J. Op ’t Veld  Michael C. Gastpar
École Polytechnique Fédérale de Lausanne
Dept. I&C, LINX
Lausanne, Switzerland
giel.optveld@epfl.ch  michael.gastpar@epfl.ch

Abstract

Successive refinement is a technique to encode a source iteratively, improving the distortion at each step. If the code rates of all steps combined can be as efficient as coding only for the final distortion in one go, then the source is said to be successively refinable. We turn our attention to vector sources and we propose to replace the high-distortion first stage by a projection to low-dimensional space. The goal remains the same: code the original, high-dimensional source afterwards by only a small addendum. We look at Gaussian sources subject to an MSE criterion as a first step in this direction. A coding scheme for this problem is developed by turning the projection into an estimator of the source. It is shown that Gaussian vector sources coded via a projection can be successively refinable. We derive a parametric description of distortion pairs that can achieve this state; this region exists for any projection matrix, but its size may vary.

1 Introduction

With the rapid growth of data, it becomes increasingly difficult to assess in advance whether a data set is going to be of value. To prevent any unnecessary processing, one can evaluate either a fragment or a lower quality thumbnail first to decide if one wants to have the entire set. Equitz and Cover pitched the idea of successive refinement as an improvement on efficiency: can one optimally code a random variable $X$ at a low distortion $D_1$ and then use the bits such that only an addendum is required to code a second representation at a better distortion $D_2 < D_1$? Moreover, can such two stages combined be as efficient as the rate-distortion optimal rate $R_X(D_2)$? The answer turned out to be affirmative, be it under strict conditions and for specific sources.

We extend this idea to involve projections. First, a user requests only a low-dimensional projection of the original data. This projection reflects his interests: he may value certain components more than others in his assessment of the data’s value. Can one still use this first low-dimensional projection to code a final high-dimensional representation by only a small addendum and if so, can one do it without the loss of rate? In a continuous setting, we adopt the mean squared error as our distortion measure. Gaussian sources, as maximizers of differential entropy, form our choice for a start to the problem.

The amount of classic results on successive refinement is vast, featuring amongst others [1, 2]. Both [3] and [4] also considered the successive coding of two correlated variables $X$ and $Y$; our problem is in principle encapsulated in their models. Both papers also go into the specific case of a pair of jointly Gaussians. We let the source $X$ and the projection $U = A^T X$ be of arbitrary dimensions, $N$ and $M$ respectively. Other work by Nayak and Tuncel involved the extension to vector sources as well, but focused on accommodating individual distortion criteria [5].

We present a sequential coding scheme for Gaussian sources with a projection on the first stage in Section 3. The strategy is to, if the projection was satisfactory, turn it into an estimator of the original source and use it as side information on the
refinement. The algebraic nature of Gaussian sources allows us to find the estimator in closed form and we show that the refinement boils down to a rank-$M$ downdate of the source statistics in Section 4. Furthermore, we will derive a lemma on the rates and comment on the successive refinability of the problem afterwards in Section 5. We find a parametric expression for a distortion level on the projection that marks a threshold in successive refinability. The value of this threshold depends on the projection $\mathbf{A}$, but it always exists, provided the required distortion on the second stage is not too large.

2 Preliminaries

Let $X$ be a random variable over some alphabet $\mathcal{X}$ and let $\hat{X}$ be its reconstruction, whose precision one evaluates by some distortion measure $d(X, \hat{X}) : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$. A classic, two stage successive refinement problem codes $X$ twice, first at a distortion $D_1$ and then at a $D_2 \leq D_1$. The scheme is such that the second stage only codes an addendum to the bits of the first so as to meet $d(X, \hat{X}_2) \leq D_2$. In particular, a source is called successively refinable if it achieves the rate distortion limit on both stages, i.e., $R_1 = R(D_1)$ and $R_1 + R_2 = R(D_2)$. Equitz and Cover showed that Gaussians subject to an MSE criterion are successively refinable [1].

This source coding problem has been extended to involve different, yet correlated sources on the first and second stage $(X_1, X_2)$ by most notably [4, 3]. One first codes $X_1$ and subsequently, $X_2$ can be described by fewer bits if one exploits the correlation with $\hat{X}_1$. Viswanathan and Berger adopt the terminology sequential coding to refer to a situation in which the first encoder only has access to the first source, while the second has access to both [4]. Nayak and Tuncel allowed both encoders access to both sources and dubbed their problem successive coding [3]. Of these two, we adopt the sequential coding approach.

In this paper, $\hat{X}$ is an $N$-dimensional vector and we replace the first stage by a projection to $M$-dimensional space for $M < N$. We label it

$$ U = \mathbf{A}^T X, \quad (1) $$

where $\mathbf{A} \in \mathbb{R}^{N \times M}$ is an arbitrary matrix with orthonormal columns. In contrast to $U$, we label the vector corresponding to the refinement in the second stage as $\hat{V}$. One could view upon the source and its projection as two correlated sources, $X$ and $U$. Their correlation is of course special, since $p(X, U) = p(X) \delta(U - \mathbf{A}^T X)$. All in all, we seek the following encoder-decoder pairs:

$$ \begin{align*}
  f_1 & : \mathcal{U}^n \to \{1, \cdots, 2^{nR_U}\} \\
  f_2 & : \mathcal{X}^n \to \{1, \cdots, 2^{nR_X}\} \\
  g_1 & : \{1, \cdots, 2^{nR_U}\} \to \hat{\mathcal{U}}^n \\
  g_2 & : \{1, \cdots, 2^{nR_U}\} \times \{1, \cdots, 2^{nR_X}\} \to \hat{\mathcal{X}}^n,
\end{align*} \quad (2)$$

where $\hat{U} = g_1(f_1(U))$ and $\hat{X} = g_2(f_1(U), f_2(X))$. A schematic is depicted in Figure 1.

We investigate the impact of these projections for a specific case: Gaussians under mean squared error distortion measure. The source thus follows $X \sim \mathcal{N}(0, \Sigma_X)$, with $\Sigma_X = \diag(\lambda_1, \lambda_2, \cdots, \lambda_N)$, ordered such that $\lambda_1 \geq \cdots \geq \lambda_N$. The same ordering and numbering applies to the eigenvalues of any matrix. We adopt scalar distortions $(D_U, D_X)$ to evaluate $\hat{U}$ and $\hat{X}$ respectively. Since we are working with vectors, an encoder will allocate rate to each vector component such that their individual distortion levels satisfy the end distortion constraint. To accommodate that procedure with terminology, $\hat{D}_{X,i}$ refers to the individual distortion of the $i$'th component of $X$. For
convenience, it is sometimes shorter to aggregate the distortion profile into a diagonal matrix $\hat{D}_X = \text{diag}(\hat{D}_{X,i})$.

The achievable rate region of our problem is contained in the results of [4]. Thus if we reformulate, the achievable rate region is already known to be:

**Theorem 1.** The rate distortion quadruple $(R_U, R_V, D_U, D_X)$ is achievable if and only if there exists a joint distribution $p(X, \hat{X}, \hat{U})$ such that

\begin{align}
I(U; \hat{U}) &\leq R_U \\
I(X; \hat{X}|\hat{U}) &\leq R_V \\
E[||U - \hat{U}||^2] &\leq D_U \\
E[||X - \hat{X}||^2] &\leq D_X
\end{align}

In the following sections, we will introduce a sequential coding scheme for Gaussian sources, whose optimality we will evaluate with respect to the following definition:

**Definition 1.** We say the source $X$ is successively refinable via a projection $U$ at $(D_U, D_V)$ if the quadruple $(R_U, R_V, D_U, D_X)$ is achievable and it holds that $R_U = R_U(D_U)$ and $R_U + R_V = R_X(D_X)$.

### 3 Coding Strategy

Let us first emphasize that Definition 1 states that successive refinability requires both stages to be rate-distortion optimal. Hence, any possible code for a Gaussian source that we wish to be fully successively refinable can be constructed by minimizing the rate of the first stage first; the second stage will follow suit. A trivial first observation is that if $X$ is Gaussian then so is $U = A^TX$. Hence, the rate $R_U \geq R_U(D_U)$ is uniquely (see [6, Theorem 12.1.1]) minimized by the straightforward procedure of reversed water filling [6, p. 314], which leads to a $\hat{U}$ that is Gaussian as well.

It is not immediately clear that the refinement, which we dubbed $V$, is best coded in a Gaussian fashion as well. After the first stage, the decoder has a copy of $\hat{U}$ at a distortion no more than $D_U$; and he may decide he is also interested in a copy of the high-dimensional $X$. Before even starting to code, the decoder already knows something on $X$, namely he can compute $E[X|\hat{U}]$ as its MMSE-estimate. Hence, the second stage only revolves around the residual, that which is not yet known on $X$:

\begin{equation}
V = X - E[X|\hat{U}],
\end{equation}

so that coding this leads to a final estimate on $X$ afterwards as:

\begin{equation}
\hat{X} = \hat{V} + E[X|\hat{U}].
\end{equation}
Conveniently, the distortion measures on $X$ and the residual $V$ match as follows:

$$
\mathbb{E}[||X - \hat{X}||^2] = \mathbb{E}[||V + \mathbb{E}[X|\hat{U}] - (\hat{V} + \mathbb{E}[X|\hat{U}]||^2] = \mathbb{E}[||V - \hat{V}||^2].
$$ (9)

In other words, coding $V$ for a distortion $D_X$ after having received $\hat{U}$ ensures one of also having a copy of $\hat{X}$ at the same distortion. In this setup, the jointly Gaussian nature of $(U, V)$ follows from the following:

$$
I(U; \hat{U}) + I(X; \hat{X}|\hat{U}) \geq R_U(D_U) + h(X|\hat{U}) - h(X|\hat{X}, \hat{U})
$$ (10)

$$
= R_U(D_U) + h(X - \mathbb{E}[X|\hat{U}]|\hat{U}) - h(X - \hat{X}|\hat{X}, \hat{U})
$$ (11)

$$
\geq R_U(D_U) + h(V) - h(V - \hat{V})
$$ (12)

$$
\geq \frac{1}{2} \log \left( \frac{\Sigma_U}{|D_U|} \right) + \frac{1}{2} \log \left( \frac{\Sigma_V}{|D_V|} \right),
$$ (13)

where $\hat{D}_U$ is the reversed water filling solution, computed as $\hat{D}_{U,i} = \min(\theta_U, \lambda_i(\Sigma_U))$ with $\theta_U$ chosen such that $\sum_{i=1}^n \hat{D}_{U,i} = D_U$ is satisfied. The same goes for $\hat{D}_V$, but with respect to end distortion constraint $D_X$. (†) Follows from the fact that via this construction, $\hat{U}$ is jointly Gaussian with $X$ and hence $X - \mathbb{E}[X|\hat{U}] \perp \hat{U}$. With this set-up in check, it remains to be found what the covariance of the residual, $\Sigma_V$, is and what impact it has on the sum-rate $R_U + R_V$ when compared to $R_X(D_X)$.

### 4 Estimator and Refinement Algebra

Thanks to the Gaussian nature of both stages, rates can be computed in closed form. To that end, we first derive the estimator $\mathbb{E}[X|\hat{U}]$ and the residual statistics $\Sigma_V$. We already concluded that $R_U$ is minimized by applying reversed water filling to the principal components of covariance $\Sigma_U$, which we find as

$$
\Sigma_U := \mathbb{E}[UU^T] = A^T \Sigma_X A.
$$ (14)

Without loss of generality, we define our projections $A$ actually as follows:

$$
A = \tilde{A} Q,
$$ (15)

in which $Q$ are the eigenvectors of $\tilde{A}^T \Sigma_X \tilde{A}$. $\tilde{A}$ can still be arbitrary, as long as its columns are orthonormal. Incorporating $Q$ inside $A$ has as an effect that $\Sigma_U$ is already diagonalized into its eigendecomposition. Consequently, also the $D_U$ found by means of (13) is diagonal. Making this diagonalization implicit is without loss of generality and allows us to construct simpler equations. Namely, we build a simple Gaussian test channel $\hat{U} = B(U + W)$ (see, e.g., [6, p. 339]) where $U \perp W$, $W \sim \mathcal{N}(0, \Sigma_W)$ and

$$
\Sigma_W = \text{diag}_{i=1,\ldots,M} \left( \frac{\lambda_i(\Sigma_U)}{\lambda_i(\Sigma_U) - \delta_{U,i}} \right) = \Sigma_U \hat{D}_U (\Sigma_U - \hat{D}_U)^{-1}
$$ (16)

$$
B = \text{diag}_{i=1,\ldots,M} \left( \frac{\lambda_i(\Sigma_U) - \delta_{U,i}}{\lambda_i(\Sigma_U)} \right) = (\Sigma_U - \hat{D}_U) \Sigma_U^{-1}.
$$ (17)

Since $X$ and $\hat{U}$ are jointly Gaussian, the estimation of $X$ follows as:

$$
\mathbb{E}[X|\hat{U}] = \mathbb{E}[X\hat{U}^T] \mathbb{E} \left[ \hat{U}\hat{U}^T \right]^{-1} \hat{U},
$$ (18)
of which the unknown expectations on the right hand side are found as:

\[ \mathbb{E} \left[ X \hat{U}^T \right] = \Sigma_X A B \]  
(19)
\[ \mathbb{E} \left[ \hat{U} \hat{U}^T \right] = B (\Sigma_U + \Sigma_W) B = \Sigma_U - \hat{D}_U. \]  
(20)

Observe that (20) confirms that the channel follows the desired distribution, i.e., \( p(\hat{U}) \sim \mathcal{N}(0, \Sigma_U - \hat{D}_U) \). This ultimately gives us the first estimation of \( X \) as

\[ \mathbb{E}[X|\hat{U}] = \Sigma_X A B \left( \Sigma_U - \hat{D}_U \right)^{-1} B (A^T X + W) \]  
(21)
\[ = \Sigma_X A \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) (A^T X + W). \]  
(22)

The only hiccup in this analysis is that \( (\Sigma_U - \hat{D}_U)^{-1} \) might not be invertible, which happens when \( \hat{D}_{U,i} = \lambda_i(\Sigma_U) \) for some \( i \). This situation occurs when the reversed water filling procedure concludes that some principal components of \( U \) do not require coding to meet the distortion constraint \( D_U \). These components do not participate in the coding and thus cannot contribute to the estimator and can therefore be safely excluded to make everything invertible. Notice, though, that the exclusion is implicit in (22).

As for the covariance of the residual \( V = X - \mathbb{E}[X|\hat{U}] \),

\[ \Sigma_V(D_U) := \mathbb{E}[VV^T] = \mathbb{E}[(X - \mathbb{E}[X|\hat{U}])(X - \mathbb{E}[X|\hat{U}])^T] \]  
(23)
\[ = \mathbb{E}[(X - \mathbb{E}[X|\hat{U}])^T X] \]  
(24)
\[ = \Sigma_X - \Sigma_X A \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) A^T \Sigma_X. \]  
(25)

One term in (23) drops out by the orthogonality principle. The statistics of this residual \( V \) are thus a perturbation of the original source statistics by subtracting a real symmetric matrix of at most rank \( M \), representing what one learned on the first stage. We denote this covariance \( \Sigma_V(D_U) \) as a function of \( D_U \) to emphasize its dependency. Sometimes the function notation is dropped if the context does not benefit. On a side note, observe that a projection back would lead one back to \( A^T \Sigma_V A = \hat{D}_U \).

**Lemma 1.** For all \( D_U \geq 0 \), the eigenvalues of \( \Sigma_V \) satisfy

\[ \lambda_i(\Sigma_V) \geq \lambda_{i+M}(\Sigma_X) \]  
for \( 1 \leq i \leq N - M \)  
(26)
\[ \lambda_i(\Sigma_V) \geq 0 \]  
for \( N - M < i \leq N. \)  
(27)

**Proof.** Observe that (25) is a subtraction of two real symmetric matrices. The subtracted matrix is of at most rank \( M < N \). Consequently, we can apply a theorem by Weyl [7, Thm 4.3.6] that for any two \( N \times N \) Hermitian matrices \( A, B \), of which \( \text{rank}(B) \leq M \), the following holds:

\[ \lambda_k(A) \geq \lambda_{k+M}(A + B) \geq \lambda_{k+2M}(A). \]  
(28)

Let \( A = \Sigma_X \) and pick for \( B \) the perturbation matrix on the right of (25), including the minus sign. Conclude that the update matrix was positive semidefinite by construction and is thus now negative semidefinite if one indeed includes the subtraction as its sign. All \( \lambda_i(\Sigma_V) \) can thus only decrease with respect to \( \lambda_i(\Sigma_X) \), but Weyl’s theorem now implies that in any case they can never drop below \( \lambda_{i+M}(\Sigma_X) \). Hence, (26) holds.

The second line of the lemma is somewhat trivial, since \( \Sigma_V \) is a covariance matrix. However, to prove that this is indeed true, one could apply an argument similar to that of [8, Thm 1]. Due to space limitations, we leave it to a reference for now. \(\square\)
5 Rates and Successive Refinability

Lemma 1 and (25) show that the eigenvalues of $\Sigma_V$ are strictly non-decreasing in $D_U$. An example is drawn in the left plot of Figure 2. A trivial observation is that if $D_U \geq \text{tr}(\Sigma_U)$ then $R_U(D_U) = 0$, implying that $\Sigma_V = \Sigma_X$; the entire first stage is skipped and the 'refinement' is the one-stage RD-coding problem. Furthermore, note that $\Sigma_X \geq \Sigma_V$ and thus by the fact that both $X$ and $V$ are Gaussian, we have

$$R_V(D_X) \leq R_X(D_X),$$

(29)

with equality if and only if $R_U = 0$. Furthermore, the two-stage refinement can never produce a sum-rate lower than the one-step optimal coding of $R_X(D_X)$. Combining the latter observation with (29), we find the following bounds:

$$\max\{R_X(D_X), R_U(D_U)\} \leq R_U(D_U) + R_V(D_X) \leq R_U(D_U) + R_X(D_X).$$

(30)

An example is shown in the center plot of Figure 2. Evaluating the sum-rate is more convenient than one might expect, as becomes clear from the following lemma:

**Lemma 2.** *For $0 \preceq \hat{D}_U \preceq \Sigma_U$, it holds that

$$\frac{1}{2} \log \left( \frac{|\Sigma_U|}{|\hat{D}_U|} \right) = \frac{1}{2} \log \left( \frac{|\Sigma_X|}{|\Sigma_V|} \right).$$

(31)

Proof. Let us start with the determinant of the residual $\Sigma_V$ by making use of (25).

$$|\Sigma_V| = |\Sigma_X - \Sigma_X A \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) A^T \Sigma_X|$$

(32)

$$= |\Sigma_X| \cdot |I_N - \Sigma_X^{1/2} A \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) A^T \Sigma_X^{1/2}|.$$  

(33)

The rest of the proof relies on Sylvester’s theorem for determinants. It states that for any $P \in \mathbb{R}^{M \times N}$ and $Q \in \mathbb{R}^{N \times M'}$ it holds that $|I_M + PQ| = |I_N + QP|$ (see, e.g., [9]). Now, continue expanding the right hand side

$$\frac{|\Sigma_V|}{|\Sigma_X|} = \left| I_M - \Sigma_X^{1/2} \left( \begin{array}{c} A \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) A^T \Sigma_X^{1/2} \\ P \end{array} \right) \right|$$

(34)

$$= \left| I_M - \Sigma_U^{-2} (\Sigma_U - \hat{D}_U) A^T \Sigma_X^{1/2} \Sigma_U^{1/2} A \Sigma_U^{-1} \right|_{=I_m}$$

(35)

$$= |\Sigma_U^{-1}| \cdot |\hat{D}_U|.$$  

(36)

Since the arguments of the logs in the lemma are equal, so are the logs.  

This Lemma turns out to be the crucial tool to comment on successive refinability. To that end, let us zoom in on a specific region of distortions, namely $D_X < N \lambda_N(\Sigma_X)$. For these $D_X$, we are ready to prove the absence of rate loss for any $A$. The case of larger $D_X$ requires some care, which we will explain after the following theorem:

---

*The authors would like to thank Stefan Apostol for his substantial contribution to Lemma 2.*

47
Theorem 2. For sufficiently small distortion, i.e., $D_X \leq N \lambda_N(\Sigma_X)$, a Gaussian source $X$ is successively refinable after a projection $A$ if and only if the projection is requested at a $D_U \geq D_{U,c}$ for a critical $D_{U,c}$ that satisfies

$$\lambda_N(\Sigma_V(D_{U,c})) = \frac{D_X}{N},$$

where $\Sigma_V(D_{U,c})$ follows (25).

Proof. The theorem says that one must pick a $D_U$ so that no eigenvalue of the residual $\Sigma_V$ drops below the distortion threshold that is set by reversed water filling for the optimal one-stage coding of $R_X(D_X) = \frac{1}{2} \log \left( \frac{\|\Sigma_X\|}{(D_X/\theta)^n} \right)$. Assume for convenience that all the following $\hat{D}$ are already fixed so that rates are minimized. Then by Lemma 2,

$$R_U(D_U) + R_V(D_X) = \frac{1}{2} \log \left( \frac{\|\Sigma_U\|}{\|D_U\|} \right) + \frac{1}{2} \log \left( \frac{\|\Sigma_V\|}{\|D_V\|} \right)$$

$$= \frac{1}{2} \log \left( \frac{\|\Sigma_X\|}{\|\Sigma_V\|} \right) + \frac{1}{2} \log \left( \frac{\|\Sigma_V\|}{\|D_V\|} \right)$$

$$= \frac{1}{2} \log \left( \frac{\|\Sigma_X\|}{\|D_V\|} \right).$$

If $\lambda_N(\Sigma_V) \geq \frac{D_X}{N}$ then $\hat{D}_{V,i} = \frac{D_X}{N}$ for all $i = 1, \ldots, N$ by the reversed water filling procedure. That means that $\hat{D}_V = \hat{D}_X$ and thus $R_U(D_U) + R_V(D_X) = R_X(D)$. For the converse assume that $D_U$ is picked so small that $\lambda_N(\Sigma_V) < \frac{D_X}{N}$. Then the reversed water filling procedure will assign individual distortion levels,

$$|\hat{D}_V| = \prod_{i=1}^N \min(\theta_V, \lambda_i(\Sigma_V)),$$

where $\theta_V$ is picked such that $\sum_{i=1}^N \min(\theta_V, \lambda_i(\Sigma_V)) = D_X$. This product is maximized by $(\frac{D_X}{n})^N$, which can only be a solution if all $\lambda_i(\Sigma_V) \geq \frac{D_X}{N}$, contradicting the assumption. We have $|\hat{D}_V| < |\hat{D}_X|$ otherwise, resulting in a sum-rate strictly higher than $R_X(D_X)$. Since $\lambda_i(\Sigma_V)$ are continuous and non-decreasing in $D_U$, successive refinability is guaranteed for all $D_U$ larger than the $D_U$ at which $\lambda_N(\Sigma_V(D_U)) = \frac{D_X}{N}$.  

Figure 2: Example for $(N, M) = (6, 3)$ for $\Sigma_X = \text{diag}(2.5, 1.8, 1.6, 1.4, 1.3, 0.9)$ and $\frac{D_X}{N} = 0.675$. Left: $\lambda_i(\Sigma_V)$ are non-decreasing in $D_U$ by Lemma 1. Middle: The sum-rate is bounded as in (30) and is equal to $R_X(D_X)$ for $D_U \geq D_{U,c}$. Right: Eigenvalue drop for $\Sigma_X \rightarrow \Sigma_V$ at the critical point $D_U = D_{U,c} \approx 2.36$. 

48
The right plot of Figure 2 shows an example of the eigenvalues at this critical point. Note that everything above the dotted line $\frac{D_X}{N}$ is what $\hat{X}$ would have been by a one-stage optimal coding via $R_X(D_X)$. A too small $D_U$ so that any $\lambda_i(\Sigma_V)$ drops below $\frac{D_X}{N}$ means that $E[X|\hat{U}]$ contains information on $X$ not present in the $\hat{X}$ that would have been the one-stage optimal solution. One can show that if $A$ consists of any set of $M$ eigenvectors of $\Sigma_X$, then $D_{U,c} = \frac{M \cdot D_X}{N}$. It appears that $D_{U,c} > \frac{M \cdot D_X}{N}$ for any other projection matrix, but we are yet to formally show this.

Low distortion, $D_X < N\lambda_N(\Sigma_X)$, ensures that all principal components of $X$ require coding in the first place. Under this condition there always exists a $D_{U,c} < \text{tr}(\Sigma_U)$ for any $A$. Were the condition not true, then $A$ could have been aligned in the direction of components that are not part of the one-stage optimal description of $X$. One could thus, for example, exclude these directions from $A$ to also achieve successive refinability for $D_X > N\lambda_N(\Sigma_X)$. In general, the critical point $D_{U,c}$ in Theorem 2 is more formally found by comparing the partaking principal components $\lambda_i(\Sigma_V(D_{U,c}))$ to $\theta_X$, the coding threshold the reversed water filling procedure uses to compute $R_X(D_X)$. All in all, the successive refinability of $X$ is not so much restricted by the dimensionality reduction imposed by $A$, it is directionality to which it is most sensitive.

References


On the Energy Benefit of Compute-and-Forward on the Hexagonal Lattice

Zhijie Ren\textsuperscript{1}, Jasper Goseling\textsuperscript{2,1}, Jos Weber\textsuperscript{1} and Michael Gastpar\textsuperscript{3,1},

\textsuperscript{1}Department of Intelligent Systems, Delft University of Technology, The Netherlands
\textsuperscript{2}Stochastic Operations Research, University of Twente, The Netherlands
\textsuperscript{3}School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

z.ren@tudelft.nl, j.goseling@utwente.nl, j.h.weber@tudelft.nl, michael.gastpar@epfl.ch

Abstract

We study the energy benefit of applying compute-and-forward on a wireless hexagonal lattice network with multiple unicast sessions with a specific session placement. Two compute-and-forward based transmission schemes are proposed, which allow the relays to exploit both the broadcast and superposition nature of the wireless network. The energy consumption of both transmission and reception of the nodes are taken into account. We show that with our schemes, the total energy consumption of the network is significantly reduced compared to traditional routing based transmission schemes. Moreover, our schemes also outperform the plain network coding based transmission scheme in terms of power saving as long as the receive energy of the devices is not negligible.

1 Introduction

Compute-and-forward [1], also known as reliable physical layer network coding, is a technique that enables to exploit the features of broadcast and superposition in wireless networks. It has been shown (e.g., [1]-[3]) that in the scenario of multiple unicast, the throughput in a network can be significantly improved by this technique, compared to traditional routing or network coding [4].

On the other hand, network coding, sometimes referred as plain network coding to distinguish from physical layer network coding, has been proven to be beneficial to a wireless network in the aspect of energy savings [5] [6]. Most of the studies consider transmit energy only, and show that energy can be saved by using plain network coding compared to traditional routing. The ratio of the energy consumption of traditional routing to advanced schemes is sometime referred as the energy benefit. In particular, on the hexagonal lattice with multiple unicast, the coding scheme proposed in [7] lower bounds this energy benefit by 2.4, and the lower bound is further improved to 3 by the scheme proposed in [8]. However, an interesting problem is studied in [9] that in the scenario that the receive energy (used for supporting circuit for receiving, for example) is not negligible, some plain network coding based scheme will have less energy benefit, or even no benefit at all.

Besides the above-mentioned studies, our study is also motivated by the classic example of the two-way relay channel shown in Fig. 1(a). As observed, the energy consumption is reduced by using compute-and-forward compared to traditional routing and plain network coding, since the relay node needs fewer transmissions and receptions, and thus consumes less energy.

In this paper, we study the energy benefit of applying compute-and-forward on a hexagonal lattice with specific session placement as in [8] and [9]. We consider both the
transmit and receive energy of the wireless nodes in the network, and design two energy efficient schemes based on compute-and-forward. It is shown that the energy benefit of the two schemes is between 1.5 and 3, for any value of transmit and receive energy of the nodes. Thus, by using compute-and-forward in this network, the energy consumption is at least reduced by a factor of 1.5. This result is essentially different from the energy benefit of the plain network coding based scheme in [9], which is severely degraded or completely gone when the receive energy is not negligible, since it reduces the number of transmissions of the network at the cost of increasing the number of receptions. Our schemes, on the other hand, show that compute-and-forward is able to save the energy by reducing the number of both transmissions and receptions in a wireless network with multiple unicast.

The remainder of the paper is organized as follows. In Section 2 we describe our model. In Section 3, we propose two compute-and-forward based coding schemes and prove their validity. In Section 4, we define the energy benefit, briefly introduce some previous schemes and compare the energy benefit of our schemes to previous schemes to show our contribution. At last, we conclude our work in Section 5.

2 Model

2.1 Hexagonal Lattice and Session Placement

We consider a subset of the hexagonal lattice network with nodes representing wireless terminals, denoted as \( \mathcal{V} = \{ v | 0 \leq v_1, v_2 \leq K, v_1 + v_2 \leq K \} \), in which \( v \) is a node defined by an index tuple \( (v_1, v_2) \) and \( K \) is a positive integer. The location of the node \( v \in \mathcal{V} \) in \( \mathbb{R}^2 \) is given by \( vG \), where \( G = \left( \begin{array}{cc} 1/2 & \sqrt{3}/2 \\ -1/2 & \sqrt{3}/2 \end{array} \right) \). We then denote the interior of the network as \( \mathring{\mathcal{V}} = \{ v \in \mathcal{V} | 0 < v_1, v_2 < K, v_1 + v_2 < K \} \), and define the boundary nodes as \( \mathcal{V} = \mathcal{V} \setminus (\mathring{\mathcal{V}} \cup \{(0,0),(0,K),(K,0)\}) \) and the three borders as \( \mathcal{V}^L = \{ v \in \mathcal{V} | v_1 = 0 \} \) for the left border, \( \mathcal{V}^R = \{ v \in \mathcal{V} | v_1 + v_2 = K \} \) for the right border and \( \mathcal{V}^B = \{ v \in \mathcal{V} | v_2 = 0 \} \) for the bottom border. We consider local interference, i.e. for any node, the transmitted signals can only be received by its neighbors which are unit distance away. More precisely, the neighbors of node \( v \) are \( \mathcal{O}(u_1, u_2) = \{(u_1 - 1, u_2 + 1), (u_1, u_2 + 1), (u_1 - 1, u_2), (u_1 + 1, u_2), (u_1, u_2 - 1), (u_1 + 1, u_2 - 1)\} \cap \mathcal{V} \).

Now we place \( 3(K-1) \) unicast sessions, denoted as \( \mathcal{M} = \{m^1(i), m^2(i), m^3(i) | i \in \{1,2, \ldots, K-1\} \} \). Sources \( s^i(i) \) and destinations \( d^i(i) \) of the sessions are positioned as follows:

\[
\begin{align*}
m^1(i) &: s^1(i) = (0, i), \quad d^1(i) = (K - i, i). \\
m^2(i) &: s^2(i) = (i, K - i), \quad d^2(i) = (i, 0). \\
m^3(i) &: s^3(i) = (K - i, 0), \quad d^3(i) = (0, K - i).
\end{align*}
\]

We assume that the source symbols for session \( m^j(i) \) are drawn from finite field \( \mathbb{F}_q \) and denote the source symbols as \( m_0^j(i), m_1^j(i), m_2^j(i), \ldots \). The hexagonal lattice with the sessions is illustrated in Fig 1(b).

2.2 Energy Model

In [9], an energy consumption model is used that includes both the energy for transmitting data and the energy for receiving data. The energy consumed when receiving consists of, for instance, the energy consumed by supporting circuitry. This model is, for instance, useful if the transmit energy is very small and reception energy cannot be neglected. In this paper, we study a similar energy model defined as follows: In each time slot, a symbol from \( \mathbb{F}_q \) transmitted by node \( v \) can be successfully received by node \( u \) if \( u \) is a neighbor of \( v \), \( v \) transmits with energy \( e_t \) and \( u \) receives with energy \( e_r \).
2.3 Compute-and-forward

Compute-and-forward [1] provides a way to exploit both the broadcast and the superposition nature in wireless networks. With compute-and-forward, a node is able to retrieve a linear sum of the symbols that are transmitted by its neighbors. It has been proved that compute-and-forward achieves a rate very close to the channel capacity on the Gaussian channel, more precisely, a rate

$$R \leq \frac{1}{2} \log_2 \left( \frac{1}{k} + \text{SNR} \right)$$

is achievable [1], where $k$ is the number of superposed symbols, which is at most 6 in our network. Comparing this rate to the capacity, only a term of $\frac{k-1}{k}$ is missing inside the logarithm, which has only a minor effect when SNR is large. Hence, in this paper, we neglect this term and assume that in a time slot, node $v$ can successfully retrieve the sum of the transmitted symbols by all of its non-silent neighbors if it receives with energy $e_r$ and the non-silent neighbors transmit with energy $e_t$.

3 Compute-and-forward Based Schemes

In [8], an energy efficient network coding scheme is designed in such a way that the number of the transmissions of the interior nodes is decreased at the cost of each of the boundary nodes transmitting extra symbols for the successful decoding at the destinations. However, the extra energy consumption at the boundaries turns out to be negligible for a network with large enough $K$.

In this section, two compute-and-forward based coding and scheduling schemes inspired by [8] are proposed. Both of them support the multiple unicast sessions $\mathcal{M}$ on the hexagonal network $\mathcal{V}$. The schemes work in rounds, in which at each destination, a new source symbol for its corresponding session is decoded after the initial startup phase. We define the notation $x_i^t(v)$ and $y_i^t(v)$, respectively, as the transmission and reception of node $v$ in time slot $i$ of round $t \in \mathbb{Z}^+$. Before starting the description of both schemes, we divide all nodes into 3 categories. We define category $i$, $i \in \mathbb{Z}_3$ as $\mathcal{V}_i = \{ (v_1, v_2) \in \mathcal{V} | v_1 \equiv v_2 + i \pmod{3} \}$, see Fig. 2(a).
3.1 Scheme 1

We consider a round of 6 time slots $i \in \{0, 1, 2, 3, 4, 5\}$, and describe the scheme by defining the transmissions of node $v = (v_1, v_2)$ at round $t$. Here, we define function $i \oplus j$ as the summation in $\mathbb{Z}_3$. This notation will be used throughout this paper for simplicity.

If node $v \in \hat{V} \cap V_i$, it receives at time slot $i \oplus 1$ and $i \oplus 2$ and transmits

$$x_t^i(v) = y_t^{i \oplus 2}(v) - y_t^{i \oplus 1}(v) + x_{t-3}^i(v).$$

at time slot $i$.

If node $v \in \hat{V} \cap V_i$, it receives 3 times at time slot $i \oplus 1$, $i \oplus 2$ and $(i \oplus 2) + 3$, and transmits twice. At time slot $i$ it transmits

$$x_t^i(v) = \begin{cases} m_1^1(v_2), & \text{if } v \in V^L, \\ m_2^2(v_1), & \text{if } v \in V^R, \\ m_3^3(K - v_1), & \text{if } v \in V^B, \end{cases}$$

and at time slot $i + 3$ it transmits

$$x_t^{i+3}(v) = \begin{cases} m_3^3(K - v_2) - m_1^1(v_2), & \text{if } v \in V^L, \\ m_1^1(v_2) - m_2^2(v_1), & \text{if } v \in V^R, \\ m_2^2(v_1) - m_3^3(K - v_1), & \text{if } v \in V^B. \end{cases}$$

Scheme 1 is illustrated in Fig. 2(b).

3.2 Scheme 2

Scheme 2 is a dual scheme of Scheme 1, in which each interior node needs to transmit twice but only receive once in each round. Similarly, we consider the transmissions of node $v = (v_1, v_2)$ in round $t$.

If node $v \in \hat{V} \cap V_i$, it receives at time slot $i$, transmits

$$x_t^{i \oplus 1}(v) = y_t^i(v) + x_t^{i \oplus 1}(v)$$

at time slot $i \oplus 1$ and transmits

$$x_t^{i \oplus 2}(v) = -y_t^{i \oplus 2}(v) + x_t^{i \oplus 2}(v)$$

at time slot $i \oplus 2$.

If node $v \in \hat{V} \cap V_i$, it receives at time slot $i$ and $(i \oplus 2) + 3$, transmits

$$x_t^{i \oplus 1}(v) = \begin{cases} m_1^1(v_2), & \text{if } v \in V^L, \\ m_2^2(v_1), & \text{if } v \in V^R, \\ m_3^3(K - v_1), & \text{if } v \in V^B, \end{cases}$$

at time slot $i \oplus 1$, transmits

$$x_t^{i \oplus 2}(v) = \begin{cases} -m_1^1(v_2), & \text{if } v \in V^L, \\ -m_2^2(v_1), & \text{if } v \in V^R, \\ -m_3^3(K - v_1), & \text{if } v \in V^B, \end{cases}$$

at time slot $i \oplus 2$, and transmits (5) at time slot $i + 3$.

Scheme 2 is illustrated in Fig. 2(c).
3.3 Validity of the Schemes

Firstly, we denote the extra transmission and reception in the last 3 time slots of node $v \in \mathcal{V}$ in round $t$ as $\tilde{x}_t(v)$ and $\tilde{y}_t(v)$. Then, we consider Scheme 1. Observe that each interior node, as well as each of the boundary nodes during the first 3 time slots, only transmit once. Hence, we use the notation $x_t(v_1, v_2)$ for the transmitted symbol of node $v = (v_1, v_2)$ in round $t$ during the first 3 time slots. Then by (3) we have

$$
x_t(v_1, v_2) = x_{t-1}(v_1-1, v_2) + x_{t-1}(v_1, v_2 + 1) + x_{t-1}(v_1 + 1, v_2 - 1)$$
$$-x_{t-2}(v_1 - 1, v_2 + 1) - x_{t-2}(v_1 + 1, v_2) - x_{t-2}(v_1, v_2 - 1) + x_{t-3}(v_1, v_2).$$

(10)

Now we establish the following lemma.

**Lemma 1** Let $(v_1, v_2) \in \mathcal{V}$.

$$x_t(v_1, v_2) = m_{t-v_1}^1(v_2) + m_{t-v_1+v_2}^2(v_1) + m_{t-v_2}^3(K - v_1 - v_2).$$

(11)

The proof of this lemma is similar to the proof for Lemma 2 in [8], since the coding scheme in (10) is similar to the one used in [8], which considers symbols in $\mathbb{F}_2$ instead of $\mathbb{F}_q$. Thus, we omit the proof of this lemma here to save space.

Now we prove that in each round, a source symbol is decoded at each destination, which validate the scheme. Since the network and our coding schemes are symmetric, w.l.o.g. we consider only the sessions $m_1^1(i)$ from left to right.

**Lemma 2** For the session $m_1^1(v_2)$ and its destination $v = (K - v_2, v_2) \in \mathcal{V}_i$, the symbol $m_{t-v_1}^1(v_2)$ can be decoded at the end of round $t - 1$ by

$$y_{t-2}^{i\oplus 2}(v) - y_{t-2}^{i\oplus 1}(v) + x_{t-3}^i(v) + \tilde{y}_{t-1}(v),$$

(12)

for Scheme 1.

**Proof**: W.l.o.g. we assume $v \in \mathcal{V}_0$. By the definition of the categories, for the four neighbors of node $v$, we have nodes $(v_1 - 1, v_2 + 1), (v_1, v_2 - 1) \in \mathcal{V}_1$ and nodes $(v_1 - 1, v_2), (v_1 + 1, v_2 - 1) \in \mathcal{V}_2$. Thus we have $y_{t-2}^{i}(v_1, v_2) = x_{t-1}(v_1 - 1, v_2) + x_{t-1}(v_1 + 1, v_2 - 1), y_{t-2}(v_1, v_2) = x_{t-2}(v_1 - 1, v_2 + 1) + x_{t-2}(v_1, v_2 - 1)$ and $\tilde{y}_{t-1}(v_1, v_2) = x_{t-1}(v_1 + 1, v_2 - 1).$
Thus, by (4), (5) and Lemma 1 we have (12) equal to
\[
x_{t-1}(v_1 - 1, v_2) + x_{t-1}(v_1 + 1, v_2 - 1) + x_{t-2}(v_1 - 1, v_2 + 1) \\
+x_{t-2}(v_1, v_2 - 1) + x_{t-3}(v_1, v_2) + x_{t-1}(v_1 + 1, v_2 - 1) \\
= m_{t-v_1}^1(v_2) + m_{t-v_2}^2(v_1 - 1) + m_{t-v_2-1}^3(v_1) + m_{t-1}^2(v_1 + 1) - m_{t-2}^2(v_1 - 1) \\
- m_{t-v_1-2}^1(v_2 - 1) - m_{t-3}^2(v_1) - m_{t-v_2-1}^3(v_1) + m_{t-1}^2(v_1 + 1) \\
= m_{t-v_1}^1(v_2).
\] (13)

The proof for the validity of Scheme 2 is similar to Scheme 1 since the two schemes are dual. For Scheme 2, observe that for \( v \in V_t, x_{t+1}^i(v) = -x_{t+2}^i(v) = y_t^i(v) + x_{t+2}^i(v) = -y_{t}^i(v) - x_{t-1}^{i+2}(v) \). We then define \( x_t(v) = x_{t+1}^i(v) = -x_{t+2}^i(v) \), and the symbol \( m_{t-v_1}^1(v_2) \) can be decoded by \( y_t^i(v) + x_{t-2}^{i+1}(v_2) + y_{t-1}(v) \) follows the same steps as the (13)-(15). The validity of Scheme 2 is thus proved.

## 4 Energy Benefit

In this section, we compare our schemes to some existing schemes, in particular, the traditional routing based scheme, and the network coding based scheme proposed in [8]. Here, we consider the energy consumption for the schemes, which is defined as the average energy required by all nodes for each destination to retrieve one source symbol. Here, we ignore the energy consumption in an initial startup phase and consider only the steady-state behavior. The throughput of the network, e.g., the rate of the sessions, is not of our concern in this paper.

Firstly, we consider the scheme based on traditional routing strategy. With traditional routing, clearly, the optimal scheme is that all sessions go along their shortest paths. Since each interior node in the network is on the shortest paths of 3 sessions heading to 3 different directions, 3 transmissions and receptions are needed to relay one symbol for each session. Meanwhile, for the nodes on the borders, which are the sources and the destinations, they only need to transmit and receive the symbols for their corresponding sessions. Since the network and sessions are determined by \( K \) and the pair of transmit and receive energy \( e_t \) and \( e_r \), the energy consumption is thus a function of \( K, e_t \) and \( e_r \). Hence, we have the energy consumption for traditional routing \( E^{TR}(K, e_t, e_r) \) as

\[
E^{TR}(K, e_t, e_r) = 3(K - 1)(e_t + e_r) + (K - 1)(K - 2)(3e_t + 3e_r)/2.
\]

(16)

In [8], a network coding scheme is proposed, in which the interior nodes broadcast the linear sums of the symbols heading different directions, instead of transmit them separately. In each round, which is defined similarly to the round in our schemes, each interior node needs to transmit only once but receive 6 times, and each boundary node needs to transmit twice and receive 4 times. We can thus calculate the energy consumption of this scheme, denoted as \( E^{NC}(K, e_t, e_r) \), and

\[
E^{NC}(K, e_t, e_r) = 3(K - 1)(2e_t + 4e_r) + (K - 1)(K - 2)(e_t + 6e_r)/2.
\]

(17)

Here, we define the energy benefit of a certain scheme as the ratio of \( E^{TR}(K, e_t, e_r) \) to the energy consumption of that scheme when \( K \) tends to infinity. Thus, we directly have the energy benefit of the network coding based scheme proposed in [8] by (16) and (17)

\[
B^{NC}(e_t, e_r) = \lim_{K \to \infty} \frac{E^{TR}(K, e_t, e_r)}{E^{NC}(K, e_t, e_r)} = \frac{3e_t + 3e_r}{e_t + 6e_r}.
\]

(18)
which has also been presented in [9]. From (18), it is clear that this scheme reduces the energy consumption of the network by a factor of 3 when $e_r$ is negligible comparing to $e_t$. However, for the case that $e_r$ is comparable to $e_t$, the performance boost is very limited or completely gone.

Now we consider the schemes in Section 3. According to the schemes, we have the energy consumption of Scheme 1 and 2, denoted as $E_{CF1}(K, e_t, e_r)$ and $E_{CF2}(K, e_t, e_r)$, respectively, and

$$E_{CF1}(K, e_t, e_r) = 3(K - 1)(2e_t + 3e_r) + (K - 1)(K - 2)(e_t + 2e_r)/2,$$

$$E_{CF2}(K, e_t, e_r) = 3(K - 1)(3e_t + 2e_r) + (K - 1)(K - 2)(2e_t + e_r)/2.$$

By the definition of energy benefit and (16), we have the energy benefit of the schemes

$$B_{CF1}(e_t, e_r) = \lim_{K \to \infty} \frac{E_{TR}(K, e_t, e_r)}{E_{CF1}(K, e_t, e_r)} = \frac{3e_t + 3e_r}{e_t + 2e_r},$$

$$B_{CF2}(e_t, e_r) = \lim_{K \to \infty} \frac{E_{TR}(K, e_t, e_r)}{E_{CF2}(K, e_t, e_r)} = \frac{3e_t + 3e_r}{2e_t + e_r}.$$

In Figure 3, we compare the energy benefit of the schemes for different $e_r/e_t$. It shows that we significantly decrease the energy consumption of the network and achieve a higher energy benefit. Compared to the traditional routing based strategy, compute-and-forward will save the energy of the network by a factor between 1.5 and 3, (both schemes have a energy benefit of 2 when $e_r = e_t$), depending on the ratio between transmit energy and receive energy. In other words, applying compute-and-forward in this network is always beneficial for energy saving. This is essentially different from the energy benefit of plain network coding of [9]. Since when $e_r$ is large, the plain network coding based scheme even consumes more energy than traditional routing approach. Furthermore, Scheme 1 also outperforms the plain network coding based scheme at all configurations when $e_r > 0$.

5 Conclusion

In this paper, we have proposed two compute-and-forward based schemes that achieve energy benefit between 1.5 to 3, depending on the transmit and receive energy of the
nodes in the network, which indicates that the energy consumption in the network can be saved by at least a factor of 1.5 compared to traditional routing. Moreover, Scheme 1 also outperforms the plain network coding based scheme for any $e_r > 0$. These results show the superiority of compute-and-forward based schemes used for an energy saving purpose in networks where the receive energy is not negligible, since they are capable to reduce the number of receptions. The energy benefit of compute-and-forward based schemes in other networks is a very interesting problem for further study.

Acknowledgment

This work was supported by ERC Starting Grant 259530-ComCom and by the Netherlands Organisation for Scientific Research (NWO), Grant 612.001.107.

References


Comparative analysis of the optimization of error correcting codes and MIMO SVD schemes for increasing the energy efficiency of communications

Fernando Rosas∗, Richard Demo Souza† and Christian Oberli‡
∗ Department Elektrotechniek, Katholieke Universiteit Leuven, Belgium.
† CPGEI, Federal University of Technology Paraná, Brazil.
‡ Departamento de Ingeniería Eléctrica, Pontificia Universidad Católica de Chile.
fernando.rosas@esat.kuleuven.be

Abstract

Developing techniques for increasing the energy efficiency of wireless communications is interesting for many technological, economical and ecological reasons. In this work we present a model for the energy consumption required for wireless data transfer. We show how the code rate of error correcting codes and multiple antenna schemes can be optimized from an energy-consumption point of view. We show interesting similarities between the optimal strategies, which relate energy efficiency and the multiplexing-diversity trade-off.

1 Introduction

The development of techniques for reducing the energy consumption of wireless communications is a central requirement for technologies like wireless sensor networks (WSN) to prosper into large-scale autonomous networks. The main tasks that the nodes of these networks perform are sensing the environment, processing the data and wireless communicating it. The latter task dominates the overall energy budget and, therefore, optimizing it has a direct impact on a network’s lifetime [1]. Moreover, replacing batteries regularly is impractical in large networks or may even be impossible.

In this work we show how error control coding (ECC) and multiple-input multiple-output singular value decomposition (MIMO SVD) modulation can be optimized in order to increase the energy efficiency. ECC is a popular technique used for reducing the required transmit power by increasing the link reliability [2]. Nevertheless, while stronger codes provide better error correcting performance, usually they require more complex decoders with higher power consumption than simpler codes (c.f. [3]). Meanwhile, the MIMO SVD technique is an efficient method for sending data through a MIMO communications link [4]. The core concept considers the diagonalization of the channel matrix, thus establishing non interfering channels or “eigenchannels” [5]. Using all the eigenchannels maximizes the data rate, but sacrifices symbol error rate (SER). Conversely, using only some of them can yield a better SER but at the cost of decreasing the data-rate (c.f. [5,6]). Interestingly, the study of these two dissimilar scenarios leads to a similar solution: schemes that focus on diversity (low error rates) are optimal over long link distances, while schemes that focus on multiplexing (high data rates) are optimal for short-range links, suggesting an interesting and novel relationship between energy-efficiency with the diversity-multiplexing trade-off (c.f. [4]).

2 Energy consumption model

Our goal is to determine the total energy that is necessary for transferring one bit of data successfully in a point-to-point packet-switched MIMO communication. Such
a bit is called a “goodbit”. Following [7], every frame transmitted in the forward direction is matched by a feedback frame in the reverse direction that acknowledges correct reception or requests a retransmission. It is also assumed that all frames in both directions are detected and that all feedback frames are decoded without error.

In the sequel, Section 2.1 presents the analysis of the energy consumption of the transmitter of forward frames, which also decodes feedback frames. Section 2.2 then synthesizes the total energy consumption model. The analysis has been made for the MIMO transceiver architecture shown in Figure 1, which is frequently used among academic and commercial products (see e.g. [8]).

![Common architecture of a MIMO transceiver.](Image)

Figure 1: Common architecture of a MIMO transceiver.

### 2.1 Components of Energy Consumption of the Forward Transceiver

The energy consumed per goodbit by the transmitter is given by

\[
E_T = N_t E_{st} + E_{enc} + \left( N_t P_{el,tx} + \sum_{k=1}^{N_t} P_{PA}^{(j)} \right) T_b + N_t P_{el,rx} \frac{T_{fb}}{L} \tau .
\]

Above, \(E_{st}\) is the energy needed to wake up each transceiver branch of the transmitter from a low power consumption (sleep) mode, divided by the number of payload bits that are going to be transmitted before the transceiver goes again into low power consumption mode; \(E_{enc}\) is the energy consumed by the baseband electronic processing required for encoding the forward frames; \(P_{PA}^{(j)}\) is the power consumed by the power amplifiers (PAs) of the \(j\)-th transmission branch and \(P_{el,tx}\) (respectively \(P_{el,rx}\)) is the total power per transceiver branch consumed by the remaining baseband and radio-frequency electronic components that perform the forward transmission (respectively the feedback frame reception); \(T_b\) is the average air time per payload bit on a forward frame, \(T_{fb}\) is the air time of the feedback frame and \(L\) is the number of payload bits per frame. It should be noticed that the startup and electronic consumption grows linearly with the number of antennas because the consumption of components that are common to all branches in the chosen architecture is negligible [9]. Finally, \(\tau\) is the number of transmission trials until a frame is decoded without errors in the receiver. This quantity is a random variable whose distribution depends on PHY parameters as the SNR, channel statistics and modulation type, and on link layer parameters such as code rate, frame size and retransmission scheme (see Section 2.2.2).

Let us define \(r = k/n\) as the code rate, where \(n\) is the number of bits per codeword and \(n - k\) is the number of added redundancy bits. Then, each physical-layer forward frame carries \(H\) bits of header and a payload composed by \(rL\) bits of data and \((1-r)L\)
additional bits for coding. The total duration of a forward frame is shared by \( T_L \) seconds for transmitting the \( L \) bits of payload (with a suitable modulation), \( T_H \) seconds for the transmission of the header (with a binary modulation) and \( T_O \) seconds for the transmission of overhead signals for acquisition and tracking (channel estimation, synchronization, etc.). The average air time per data bit in a forward frame is

\[
T_b = \frac{T_L + T_H + T_O}{rL}.
\]

Let us define \( R_s \) as the physical layer symbol-rate of each transmitter antenna branch after the MIMO encoding and \( \omega \) as the number of used eigenchannels (i.e. \( 1 \leq \omega \leq N_t \)). If \( M_1, \ldots, M_\omega \) are the orders of the \( M \)-ary constellations used on the selected eigenchannels, then \( b \) is defined as the average \( b = \omega^{-1} \sum_{k=1}^{\omega} \log_2 M_k \). Hence \( R_b = \omega b R_s \) is the total bit rate of the MIMO system. By considering that header bits are sent using a binary modulation, and noting that \( L/T \) is the feedback time per payload bit, one can express \( T_b \) as

\[
T_b = \frac{1}{r R_s} \left( \frac{1}{\omega b} + \frac{H}{\omega L} + \frac{N_t O_a + O_b}{L} \right),
\]

where \( O_a \) is the acquisition overhead per branch and \( O_b \) is the remaining overhead, which is approximately independent of the antenna array size. Both \( O_a \) and \( O_b \) are measured in bits. Analogously, one finds that

\[
T_{fb} = \frac{F}{r \omega b R_s L}
\]

is the feedback time per payload bit, where \( F \) is the number of bits per feedback frame.

Finally, the costs of encoding forward frames, which is shared among the \( rL \) data bits, are due to the computations required for implementing the ECC and MIMO SVD modulation. This can be calculated as

\[
E_{\text{enc}} = E_{\text{enc,ECC}}(r) + E_{\text{enc,SVD}}(\omega) = \frac{V_{dd} I_0}{r L J_{\text{APU}}} \sum_{j=1}^{J_{\text{APU}}} c_j \left[ n_{j,\text{enc,ECC}}(r) + n_{j,\text{enc,SVD}}(\omega) \right].
\]

Above, \( V_{dd} \) is the arithmetic processing unit (APU) operating voltage and \( I_0 \) is the average current during the execution time of the arithmetic operations, \( J_{\text{APU}} \) is the APU’s clock frequency, \( c_j \) is the number of clock cycles required by the \( j \)-th operation which is performed \( n_{j,\text{enc,ECC}} \) (resp. \( n_{j,\text{enc,SVD}} \)) times during the ECC algorithm (resp. MIMO SVD algorithm) and \( J_{\text{APU}} \) is the number of different operations the APU performs.

### 2.2 Total Energy per Successfully Transferred Bit

In analogy with (1), the total energy used by the receiver for demodulating \( \tau \) forward transmissions and for transmitting the corresponding \( \tau \) feedback frames is

\[
E_R = N_r E_{\text{st}} + \left[ E_{\text{dec}} + N_r P_{\text{el,rx}} T_b + \left( N_r P_{\text{el,tx}} + \sum_{k=1}^{N_r} P^{(j)}_{\text{PA}} \right) \frac{T_{fb}}{L} \right] \tau,
\]

where \( E_{\text{dec}} \) accounts for the energy consumption of decoding the forward frame, which is defined using (5) with obvious modifications. Note that this term is multiplied by \( \tau \) because a new decoding algorithm has to be performed for each transmission trial.
Let us define the total energy per goodbit as $\mathcal{E}_b = \mathcal{E}_T + \mathcal{E}_R$. Because of $\tau$, $\mathcal{E}_b$ is a random variable that depends on the realizations of the channel and the thermal noise. In order to simplify further calculations let us assume $N_t = N_t = N$ (the extension of the presented results to the non-symmetrical case is straightforward). Then, the mean energy consumption per bit is given by

$$
\bar{\mathcal{E}}_b = \mathbb{E}\{\mathcal{E}_T + \mathcal{E}_R\} = NS + \mathcal{E}_{\text{enc}} + \left[\mathcal{E}_{\text{dec}} + \left(NP_{\text{el}} + \sum_{k=1}^{N} P_{\text{PA}}^{(j)}\right)T\right] \bar{\tau}, \tag{7}
$$

where $S = 2\mathcal{E}_{\text{st}}$ it the total startup energy per transceiver branch, $P_{\text{el}} = P_{\text{el,tx}} + P_{\text{el,rx}}$ is the total power consumed by electronic components per branch and $T = T_b + T_{fb}/L$ is the total time per bit. In the remaining of this section we will seek for explicit expressions for the dependence of $P_{\text{PA}}^{(j)}$ and $\bar{\tau}$ on the SNR.

### 2.2.1 PA’s total power consumption as function of the SNR

Let us define $P_{\text{A}}^{(j)}$ as the power radiated by the antenna of the $j$-th branch, which is supplied by a corresponding PA (Figure 1). This quantity can be modeled as $P_{\text{A}}^{(j)} = (\eta/\xi)P_{\text{PA}}^{(j)}$, where $\eta$ is the drain efficiency of the PA and $\xi$ is the peak-to-average ratio of the transmitted signal [10]. Using the result from Appendix A of [5] we have that

$$
\sum_{j=1}^{N} P_{\text{PA}}^{(j)} = \frac{\xi}{\eta} \sum_{k=1}^{\omega} \bar{P}_{\text{tx}}^{(k)}, sy \tag{8}
$$

where $\bar{P}_{\text{tx}}^{(k)}$ is the transmission power that has been allocated to the $k$-th eigenchannel (this result simply states that the irradiated energy is the same if is counted over antennas or eigenchannels). The irradiated power attenuates over the air with path loss and arrives at the receiver with a mean power given by $P_{\text{rx}}^{(k)} = \frac{\bar{P}_{\text{tx}}^{(k)}}{A_0 d^\alpha}$, where $A_0$ is a parameter that depends on the transmitter and receiver antenna gains and the transmission wavelength, $d$ is the distance between transmitter and receiver and $\alpha$ is the path loss exponent.

Let us define $\bar{\gamma} = \omega^{-1} \left(\sum_{k=1}^{\omega} \bar{P}_{\text{tx}}^{(k)}/\sigma_n^2\right)$ as the average SNR among the used eigenchannels at the decision stage, where $\sigma_n^2$ is the noise power. In general, $\sigma_n^2 = N_0 W N_t M_L$, where $N_0$ is the power spectral density of the baseband-equivalent additive white Gaussian noise (AWGN), $W$ is the transmission bandwidth, $N_t$ is the noise figure and $M_L$ is a link margin term which represents any other additive noise or interference [11]. Finally, the total power consumption of the PAs can be written as

$$
\sum_{j=1}^{N} P_{\text{PA}}^{(j)} = \frac{\xi A_0 d^\alpha}{\eta} \sum_{k=1}^{\omega} \bar{P}_{\text{rx}}^{(k)} = \frac{\xi A_0 \sigma_n^2}{\eta} d^\alpha \omega \bar{\gamma} = Ad^\alpha \omega \bar{\gamma}. \tag{9}
$$

### 2.2.2 $\bar{\tau}$ as function of the SNR

A key contributor to the energy consumption is the need for re-transmissions due to forward frames that get decoded with errors. The number of trials, $\tau$, until a frame is correctly decoded is a random variable, whose mean value has been shown to be [7]

$$
\bar{\tau} = 1 + \sum_{j=1}^{\infty} \mathbb{E}\left\{\prod_{i=1}^{j} P_i(i)\right\}, \tag{10}
$$
where $\mathbb{E}\{\cdot\}$ denotes the expectation operator and $P_t(i)$ is the probability of decoding the frame with error during the $i$-th transmission trial. In general, the $P_t(i)$ are random variables that depend on the frame size, modulation type and received SNR during the $i$-th trial. It is to be noted that (10) is a general formula which is valid for correlated or uncorrelated channel fading statistics [7].

Let us assume now that the probabilities of frame error of each transmission trial, \( \{P_t(i)\}_{i=1}^{\infty} \) are a set of i.i.d. random variables. It can be shown that $\tau$ is a geometric random variable with parameter $P_t := \mathbb{E}\{P_t(i)\}$, i.e. \( P\{\tau = r\} = (1 - P_t) P_t^{r-1} \). Its mean value can be found by direct calculation, or using (10) as follows:

\[
\bar{\tau} = 1 + \sum_{j=1}^{\infty} (P_t)^j = \frac{1}{P_t}, \tag{11}
\]

where $P_t^* = 1 - P_t$ is the probability of decoding the frame correctly. In the following we derive expressions for $P_t^*$ for coded single antenna and uncoded MIMO SVD transmissions. Expressions for coded MIMO transmissions are ongoing work.

- Uncoded MIMO SVD systems: The value of the mean frame error rate depends on how the data symbols are fed into the SVD engine. In particular, it has been shown that a pseudo-random feeding of the encoder outperforms an ordered feeding [5]. Hence, we will consider the case in which the bits of each frame are assigned to the $\omega$ used eigenchannels following a different order for each transmission trial in a pseudo-random fashion. For simplicity we will assume that all eigenchannels are used with $M$-ary modulations of equal size and that each used eigenchannel uses an equal amount of irradiation power (the case of different modulation sizes and irradiation power is addressed in [6]). Under those assumptions, it can be shown that [5]

\[
\bar{\tau}_{\text{SVD}} = \left[ 1 - \frac{1}{\omega} \sum_{k=1}^{\omega} P_{\text{bin}}(\lambda_k \bar{\gamma}) \right]^{-H} \left[ 1 - \frac{1}{\omega} \sum_{k=1}^{\omega} P_{s}(\lambda_k \bar{\gamma}) \right]^{-L/n_c}, \tag{12}
\]

where $P_{\text{bin}}$ and $P_s$ are the BPSK and $M$-ary modulation mean symbol error rate and $L = L/\log_2(M)$ is the number of $M$-ary symbols per frame.

- Single antenna systems with linear block codes: For $n < L$ lets define $n_c = L/n$ ($n_c \in \mathbb{N}$) as the number of codewords per payload. Then, to decode a frame correctly one needs $H$ correct header symbols and $n_c$ codewords with at least $(n - t) = \lambda$ correct symbols, where $t$ is the maximum number of bits that the FEC block code is able to correct per codeword. Hence, by taking into account the various permutations, $P_t^*$ can be written as

\[
\bar{\tau}_{\text{ECC}} = \left[ 1 - P_{\text{bin}}^{*}(\bar{\gamma}) \right]^{-H} \sum_{j=0}^{t} \binom{n}{j} \left[ 1 - P_{b}(\bar{\gamma}) \right]^{n - j} \bar{P}_{b}(\bar{\gamma})^j \right]^{-n_c}, \tag{13}
\]

where $P_{b}$ is the mean bit error rate of the $M$-ary modulation.

3 Minimizing the energy per bit

In this section we analyze how the energy consumption of wireless communications depends on the SNR (Section 3.1), the code rate of ECC and the number of used eigenchannels of MIMO SVD (Section 3.2).
3.1 Optimization of the SNR

The mean total energy consumption per goodbit \( \langle E_b \rangle \) can be rewritten using \( N \omega \) and \( \gamma \) so that the dependence on the SNR \( \gamma \), number of used eigenchannels \( \omega \) and antenna array size \( N \) becomes explicit. Concretely:

\[
\langle E_b \rangle = NS + E_{\text{enc}}(\omega, r) + \frac{E_{\text{dec}}(\omega, r) + [NP_{\text{el}} + Ad^\alpha \omega \gamma] T(\omega, r)}{1 - P_1(\gamma, \omega, r)}.
\]  

Just by considering the structure of (14) it can be seen that the energy consumption is large at extreme values of the total SNR. In effect, if the SNR is low then the frame error rate tends to one, and hence the energy consumption is high because of the large number of retransmissions needed for a successful frame reception. On the contrary, at a high SNR the right-hand term of (14) is also large because the radiated power (which is proportional to \( \gamma \)) is excessive. We thus infer that an optimal SNR that minimizes the energy consumption must exist in between. In fact, for typical fading channel models (e.g. Rayleigh or Nakagami-\( m \) models) (14) will be a convex function of the SNR, having a unique minimum. Let us define the optimal SNR at which the system attains a minimal average energy consumption as

\[
\gamma^*_{\omega, r} = \arg \min_{\gamma \in [0, \infty)} \langle E_b \rangle (\gamma, \omega, r) .
\]  

Therefore, \( \gamma^*_{\omega, r} \) represents an optimal trade-off between radiation power and consumption because of retransmissions.

Let us consider now a family of BCH codes with fixed codeword length \( n \) and varying information content \( k \), where each code can be parameterized by its code rate \( r = k/n \). Let us denote the set of all the BCH codes for a given value of \( n \) as \( \mathcal{R}_n \) (the complete list of BCH codes for various values of \( n \) can be found in [2]). Then, for a given codeword length \( n \) and antenna array size \( N \), the optimal code rate and multiplexing gain is given by

\[
(\omega^*, r^*) = \arg \min_{\omega \in \{1, \ldots, N\}} \arg \min_{r \in \mathcal{R}_n} \langle E_b \rangle (\gamma^*_{\omega, r}, \omega, r) .
\]  

Note that in the above definition the consumption is evaluated when the system is using its own optimal mean SNR \( \gamma^*_{\omega, r} \), as defined in (15).

3.2 Compared results of ECC and MIMO SVD

In the following we compare how \( \omega^* \) and \( r^* \) change as function of the link distance for two cases: single antenna systems using BCH codes, for which \( \omega = 1 \) and \( r \) is variable, and uncoded MIMO SVD transmissions, for which \( r = 1 \) and \( \omega \in \{1, 2, \ldots, N\} \). Figures 2a and 2b show evaluations of (14) as function of SNR using parameters of typical low-power devices (c.f. [3,5]).

For the case of single antenna systems using BCH codes it has been show [3] that \( r^* \) increases as the transmission distance decreases (see Figure 2e). Although the rate of change depends on the channel fading statistics, the qualitative scenario is always the same: for long transmission distances it is optimal to use a significant amount of coding (small \( r \)), while for short distances it is best to use almost no coding at all (large \( r \)). Both results agree with intuition. In effect, for long-range communications the power consumed by the power amplifier (equal to \( Ad^\alpha \gamma^* \) Watts) dominates over the consumption of other components. Hence, it is useful to use coding in order to reduce the SNR requirements for attaining a given SER. On the contrary, for short
Figure 2: (a) Energy consumption of a 4 × 4 MIMO SVD system for various possibilities of used eigenchannels and a link distance of d = 45 meters, (b) energy consumption of transmissions using BCH coding and various code rates with d = 50 meters, (c) minimal energy consumption of a 4 × 4 MIMO SVD system that uses uncoded BPSK with equal power allocation, (d) minimal energy consumption of transmissions which use BCH codes with optimal code rate and (e) optimal code rate of BCH codes as function of link distance.

link distances it is found that $\bar{E}_b \approx S + \mathcal{E}_{\text{enc}} + (\mathcal{E}_{\text{dec}} + P_{\text{el}}T)\bar{\tau}$, and therefore it becomes attractive to use less redundancy in order to reduce the total time per bit and the baseband consumption of encoding and decoding.

When optimizing uncoded MIMO SVD transmissions one finds results that resembles the case of ECC. In effect, beamforming ($\omega = 1$) is the energy-optimal transmission scheme for long range communications, because it invests the radiated electromagnetic energy entirely on the best eigenchannel [5]. This reduces the mean SER and thereby limits the average number of retransmissions. For medium link distances (below 350 meters in Figure 2c) the power consumed by electronic components begins to dominate over the radiated power and therefore also over the consumption of the power amplifiers. Therefore, it is attractive to use more eigenchannels simultaneously, because it increases the baud rate and thereby reduces the transmission time per bit. At very short link distances ($d \leq 10$ meters in Figure 2c), this notion leads to using full SVD ($\omega = N$) as the most energy-efficient scheme.

These results make us think in terms of the well known multiplexing-diversity trade-off [12]. Although this trade-off was formulated for MIMO systems, it can be stated in more general terms as communications resources can be invested to achieve many bad-quality bits (high throughput with high error rates) or few low-quality bits (low throughput with low error rates). For the case of ECC, high code rates can be related with diversity and low ones with multiplexing; in the same fashion for MIMO SVD systems, a large number of eigenchannels provides multiplexing capabilities while beamforming achieves diversity. This interpretation unifies the results presented in this work, and suggest the following general principle: energy-efficient long range transmissions are achieved using transmission schemes that focus on diversity, while energy-efficient short-range communications are achieved using schemes that focus on multiplexing.
4 Conclusions

We studied the effects of ECC and MIMO SVD on the energy efficiency of wireless communications. We considered the energy cost of electronic components, of the baseband processing, and the effects of retransmissions. For achieving energy efficiency, results suggest that transmission schemes which focus on diversity are best for long range communications and schemes which focus on multiplexing are better for short distances. In the case of ECC this means that low code rates are optimal for long transmission distances while the optimal code rate increases as the transmission distance shortens. Analogously, the optimal number of used eigenchannels in a MIMO SVD link is small for long distances, while it is larger for short distances. Hence, the link distance is a critical parameter for designing the energy-optimal combination of multiplexing and diversity features of a communication system.

References


1 Introduction

We consider key establishment in a network where any pair \((\xi, \eta)\) of nodes wishing to communicate need to agree on a common key. We focus on non-interactive schemes, in which \(\xi\) and \(\eta\) need not exchange information before they can start employing the generated key.

Blundo et al. [1] describe such a scheme based on symmetric polynomials: a trusted third party (TTP) generates a secret bi-variate symmetric root keying polynomial \(R(x, y) \in \mathbb{Z}_p[x, y]\)

\[
R(x, y) = \sum_{j=0}^{\alpha} \sum_{k=0}^{\alpha} R_{j,k} x^j y^k \text{ with } R_{j,k} = R_{k,j}.
\]

For \(0 \leq k \leq \alpha\), we write

\[
R_k(x) = \sum_{j=0}^{\alpha} R_{j,k} x^j.
\]

The TTP gives to any node with identifier \(\xi \in \mathbb{Z}_p\) the univariate key generating polynomial \(G_\xi(y) \in \mathbb{Z}_p[y]\) defined as

\[
G_\xi(y) = \sum_{k=0}^{\alpha} R_k(\xi) y^k.
\]

If node \(\xi\) wishes to communicate with node \(\eta\), it uses the key \(K_{\xi, \eta} := G_\xi(\eta)\). As \(R\) is symmetric,

\[
K_{\xi, \eta} = G_\xi(\eta) = R(\xi, \eta) = R(\eta, \xi) = G_\eta(\xi) = K_{\eta, \xi}.
\]

Blundo’s scheme operationally is very attractive, but unfortunately it is prone to collusion attacks. Indeed, if \(c \geq \alpha + 1\) colluding nodes \(\eta_1, \ldots, \eta_c\) compute their keys with one particular node \(\eta\), then they know \(G_\xi(\eta_i)\) for \(1 \leq i \leq c\) and hence, with Lagrange interpolation, they can retrieve \(G_\xi\). Even more seriously, if \(c\) nodes \(\eta_1, \ldots, \eta_c\) can share their key generating polynomials, then for \(0 \leq k \leq \alpha\), they know \(R_k(\eta_i)\) for \(1 \leq i \leq c\) and so, using Lagrange interpolation, they can retrieve \(R_k\) whenever \(c \geq \alpha + 1\). That is, \(\alpha + 1\) colluding nodes can retrieve the root polynomial \(R\) and hence break the complete system.

Zhang et al. proposed a variation on Blundo’s scheme in which the TTP gives to nodes \(\xi\) the polynomial \(G_\xi(y) = \sum_{k=0}^{\alpha} R_k(\xi) y^k + b_\xi g(y) + (1 - b_\xi) h(y)\), where \(b_\xi\) is a randomly chosen bit value, \(g\) and \(h\) are two random univariate polynomials of degree \(\alpha\), and the identifiers \(\xi\) are chosen from \(\{y \in \mathbb{Z}_p \mid 0 \leq g(y), h(y) \leq r\}\), where \(r\) is a “noise bound”. Their scheme has been broken by Albrecht et al. [3] using lattice-based collusion attacks.

In Section 2, we describe the HIMMO scheme, that aims to keep the implementation advantages of Blundo’s scheme while being much less vulnerable to collusion attacks. In Section 3, we discuss security results.
2 Description of the HIMMO scheme

In this section, we describe the HIMMO scheme. We use the following notation: for each integer \( x \) and positive integer \( M \) we define \( (x)_M \) as

\[
(x)_M \text{ is the integer } y \in \{0, 1, \ldots, M - 1\} \text{ such that } x \equiv y \mod M.
\]  (1)

Like the Blundo scheme, the HIMMO scheme requires a trusted third party (TTP). In the deployment of HIMMO, we can distinguish between several phases.

In the system set-up phase, the TTP obtains from a system designer positive integers \( B, b, m \) and \( \alpha \) where \( m \geq 2 \). The number \( B \) is the bit length of the identifiers that will be used in the system, while \( b \) denotes the bit length of the keys that will be generated. The TTP randomly generates the public modulus \( N \), an odd number of length exactly \((\alpha + 1)B + b\) bits (so \(2^{(\alpha+1)B+b-1} < N < 2^{(\alpha+1)B+b}\)). It also randomly generates \( m \) distinct secret integers \( \beta_1, \ldots, \beta_m \) with \( 0 \leq \beta_i < 2^B \) and at least one \( \beta_i \) is odd. For \( 1 \leq i \leq m \), the secret modulus \( q_i \) is defined as \( q_i = N - 2^b \beta_i \). Finally, the TTP generates the secret root keying material, that consists of the coefficients of \( m \) bi-variate symmetric polynomials of degree at most \( \alpha \) in each variable. For \( 1 \leq i \leq m \), the \( i \)-th root keying polynomial \( R^{(i)}(x, y) \) is written as

\[
R^{(i)}(x, y) = \sum_{j=0}^{\alpha} \sum_{k=0}^{\alpha} R^{(i)}_{j,k} x^j y^k \text{ with } 0 \leq R^{(i)}_{j,k} = R^{(i)}_{k,j} \leq q_i - 1.
\]

In the keying material distribution phase, to each node \( \xi \) in the system, with \( 0 \leq \xi < 2^B \), the TTP provides the coefficients of the key generating polynomial \( G^{(\xi)} \):

\[
G^{(\xi)}(y) = \sum_{k=0}^{\alpha} G_{\xi,k} y^k \text{ where } G_{\xi,k} = \langle \sum_{i=1}^{m} (R^{(i)}_{k,j}(\xi))_{q_i} \rangle_N.
\]

Here, for \( 1 \leq i \leq m \) and \( 0 \leq k \leq \alpha \), the polynomial \( R^{(i)}_{k}(x) \) is defined as

\[
R^{(i)}_{k}(x) = \sum_{j=0}^{\alpha} R^{(i)}_{j,k} x^j.
\]

In the key generation phase, a node \( \xi \) wishing to communicate with node \( \eta \) with \( 0 \leq \eta < 2^B \), computes

\[
K^{(\xi,\eta)} = \langle (G^{(\xi)}(\eta))_{2^b} \rangle.
\]

With examples, it can be shown that \( K^{(\xi,\eta)} \) and \( K^{(\eta,\xi)} \) are not always equal. However, it can be shown, improving on [4], that for \( 1 \leq \xi, \eta < 2^B \),

\[
K^{(\xi,\eta)} \in \{(K^{(\xi,\eta)} + jN)_{2^b} \mid 0 \leq |j| \leq 2m\}.
\]

In order that \( \xi \) and \( \eta \) will employ the same key, the initiator of the communication (say node \( \xi \)) sends to the other node, simultaneously with an encrypted message, information on \( K^{(\xi,\eta)} \) that enables node \( \eta \) to select \( K^{(\xi,\eta)} \) from the candidate set \( C = \{(K^{(\xi,\eta)} + jN)_{2^b} \mid 0 \leq |j| \leq 2m\} \). For example, node \( \xi \) sends a hash-value \( H \) of \( K^{(\xi,\eta)} \). Node \( \eta \) computes the hash value for all candidates in \( C \) and selects as common key the element of \( C \) with hash value \( H \).
3 Security aspects

We consider two attacks that can be performed by \( c \) colluding nodes \( \xi_1, \ldots, \xi_c \).

3.1 Attacking the root keying material: the MMO problem

In the first attack, the colluding nodes \( \xi_1, \ldots, \xi_c \) try to obtain the root keying material, that is, the polynomials \( P^{(1)}(x), \ldots, P^{(m)}(x) \) and the modules \( q_1, \ldots, q_m \) from the key generating polynomials \( G_{\xi_1}, \ldots, G_{\xi_c} \). We call this the \textit{MMO problem with unknown moduli}. We have not been able to solve this problem.

It can be shown [4] that in case the moduli \( q_1, \ldots, q_m \) are known, the set of all root keying polynomials that generate \( G_{\xi_1}, \ldots, G_{\xi_c} \) correspond to all vectors in an \( m(c + \alpha + 1) \) dimensional lattice in \( \mathbb{R}^{(m+1)c+m(\alpha+1)} \) at \( \ell_\infty \) distance less than \( \frac{1}{2} \) from a well-defined target vector. Finding such a lattice point becomes infeasible for large \( m \). * In [5], we give a more detailed study of the MMO problem with known moduli, that, for sake of simplicity, is restricted to the special case \( m = 2 \), in which by an improved construction the lattice size equals \( c + 2\alpha \).

3.2 Attacking the key generation process of a node: the HI problem

In the second attack, the colluding nodes \( \xi_1, \ldots, \xi_c \) aim to emulate the key generating process of a node \( \eta \not\in \{\xi_1, \ldots, \xi_c\} \) based on \( K_{\xi_i, \eta} = \langle \langle G_{\xi_i}(\eta) \rangle \rangle_{N_{\beta_i}} \) for \( i = 1, \ldots, c \). \footnote{Note that increasing \( m \) does not increase complexity of the key generation process}

This gives rise to the \textit{HI problem}:

let \( G \in \mathbb{Z}_N[x] \) have degree at most \( \alpha \). Given \( c \) distinct identifiers \( \xi_1, \ldots, \xi_c \) and \( \langle \langle G(\xi_j) \rangle \rangle_{N_{\beta_i}} \) for \( 1 \leq i \leq c \), find \( \langle \langle G(\xi) \rangle \rangle_{N_{\beta_i}} \) for a given identifier \( \xi \not\in \{\xi_1, \ldots, \xi_c\} \).

The HI problem is tackled by estimating the coefficients of \( G_\eta \in \mathbb{Z}_N[y] \). In [6], the HI problem is recast as a problem of interpolation and approximation of a polynomial from noisy values. It can be shown that the set of all \( G \in \mathbb{Z}_N[y] \) for which \( \langle \langle G(\xi) \rangle \rangle_{N_{\beta_i}} = \langle \langle G(\xi_j) \rangle \rangle_{N_{\beta_i}} \) for all \( i = 1, 2, \ldots, c \), corresponds to the set of all vectors in an \( (c + \alpha + 1) \)-dimensional lattice at \( \ell_\infty \) distance less than \( \frac{1}{2} \) from a well-defined target vector. In [6], it is studied how large \( c \) must be in order that any solution of the lattice problem predicts the correct key for \( \xi \not\in \{\xi_1, \ldots, \xi_c\} \). With HIMMO, in contrast to the conventional noisy interpolation problem, the bit length of the identifiers is much smaller than the bit length of the coefficients of \( G \). This leads to the study of polynomials over \( \mathbb{Z}_N \) with large coefficients that attain small values for small arguments. Interestingly, it appears that for small \( b \) even for very large \( c \) many polynomials correctly generate the keys for all colluding nodes \( \xi_i \), but have no predictive power for many small arguments.

References


\footnote{We believe that with proper choices for the \( \beta_i \)’s, and for large enough identifiers, the colluding nodes cannot do better by using all bits of \( \langle \langle G_{\xi_1} \rangle \rangle_N \) instead of the \( b \) last bits.}


Abstract

This extended abstract provides an overview of the results presented in [8, 9] regarding the large-coalition asymptotics of collusion-resistant fingerprinting and the asymptotics of group testing for large numbers of defectives.

1 Fingerprinting

To protect copyrighted content against unauthorized redistribution, distributors commonly embed fingerprints in the content, uniquely linking copies to individual users. If the distributor then finds an illegal copy of the content online, he can determine which user was responsible. To combat this solution, a group of \( c \) colluders might cooperate and perform a collusion attack. By comparing their versions of the content, they will detect differences in their copies, which must be part of the fingerprint. They can then try to create a mixed pirate copy, where the resulting fingerprint matches the fingerprints of different colluders in different segments of the content, making it hard for the distributor to find the responsible users. The goal of collusion-resistant fingerprinting is to assign fingerprints of length \( \ell \) to \( n \) users in such a way that, even if \( c \) pirates collude, the pirate copy can still be traced back to the responsible users with high probability.

In 1998, Boneh and Shaw [3] were the first to show that one can construct such a scheme with a code length polynomial in \( c \) and logarithmic in \( n \). In particular, their construction had a code length of the order \( \ell \propto c^4 \ln n \), and they showed that any scheme requires a code of length at least \( \ell \propto c \ln n \). In 2003, Tardos [12] proved a stronger lower bound of the order \( \ell \propto c^2 \ln n \), and described an improved scheme with \( \ell = 100c^2 \ln n \), showing that up to leading constants, his construction is optimal.

Later work on fingerprinting focused on finding the optimal leading constant and finding constructions with shorter code lengths. Huang and Moulin [6] derived explicit expressions for the channel capacities of the related max-min games, and proved that the optimal asymptotic code length is \( \ell \sim 2c^2 \ln n \). By providing a construction matching this lower bound, Oosterwijk et al. [10] later showed that this bound can be achieved using a simple decoder.

The fingerprinting game can naturally be generalized to adaptive settings, where the colluders broadcast their pirate copy in real-time. With the construction of [7] one can efficiently convert arbitrary non-adaptive schemes to adaptive schemes, which may be able to compete with the celebrated scheme of Fiat and Tassa [5].
2 Group testing

A different area of research that has received considerable attention over the last few decades is group testing. Suppose a large population of size $n$ contains $c \ll n$ infected people. To identify these people, it is possible to perform blood tests: testing a subset of the population will lead to a positive test result if this subset contains at least one infected person, and a negative result if all tested people are clean. Since the time to run a single test may be long, the subsets to test need to be chosen in advance, after which all tests are performed simultaneously. Then, when the test results come back, the subset of infected people needs to be identified. The goal of group testing is to identify the infected people using as few group tests $\ell$ as possible.

Already in the 1980s it was known that the optimal code length of such schemes scales as $\ell \sim c \log_2 n$ [11]. Later work focused on slight variations of the classical model such as noisy group testing, where a positive result may not always correspond to the presence of a defective item [2, 4]. For these variants, exact asymptotics on the capacities and constructions achieving these capacities were yet unknown.

3 Contributions

Building upon previous work of Huang and Moulin, in [8] we derived exact asymptotics (for large $c$) for the capacities of various fingerprinting and group testing models. In almost all cases this led to a lower bound on the code length of the order $\ell \propto c \ln n$. For several models it further turned out that there is a strict gap between the capacities of simple and joint decoders. In particular, for the traditional group testing model this gap is a factor $\ln 2$; an optimal joint decoder asymptotically requires a factor $\ln 2 \approx 0.69$ fewer tests than any simple decoder.

With these results in mind, [9] discusses explicit simple and joint decoders for various fingerprinting and group testing models with provable code lengths matching the capacities derived in [8]. The considered decoders are based on log-likelihood ratios, which are well-known from hypothesis testing literature to be optimally discriminative for deciding between two hypotheses $H_0$ and $H_1$ (e.g., distinguishing between guilty and innocent users, or infected and clean persons). Combining these results with [1], this also led to the discovery of a new fingerprinting decoder for arbitrary attacks. This decoder seems to have all the desired properties one can hope for, so maybe, just maybe, this finally ends the search for the optimal decoder in fingerprinting.

References


dos traitor tracing schemes,” 

[8] T. Laarhoven, “Asymptotics of fingerprinting and group testing: tight bounds from channel capacities,” 

[9] T. Laarhoven, “Asymptotics of fingerprinting and group testing: capacity-
achieving log-likelihood decoders,” 


Towards the automation of forensic facial individualisation: Comparing forensic to non forensic eyebrow features

Chris Zeinstra Raymond Veldhuis Luuk Spreeuwers
University of Twente
Services, Cybersecurity and Safety Group, Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede
{c.g.zeinstra, r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl

Abstract
The Facial Identification Scientific Working Group (FISWG) publishes recommendations regarding one-to-one facial comparisons. At this moment a draft version of a facial image comparison feature list for morphological analysis has been published. This feature list is based on casework experience by forensic facial examiners. This paper investigates whether the performance of the FISWG eyebrow feature set can be considered as being "state-of-the-art". We compare the recognition performance of one particular state-of-the-art non forensic eyebrow feature set to a semi-automated version of the forensic FISWG eyebrow feature set. The recognition performance is measured in terms of the forensic relevant log-likelihood-ratio cost metric $C_{llr}$. It is shown the FISWG feature set can be considered as being "state-of-the-art" and there actually exists a collection of feature sets that have similar performance.

1 Introduction

When comparing a facial image from a crime scene with a police photograph, forensic facial examiners pay attention to morphologic-anthropologic features, following a prescribed one to one facial comparison protocol. For example, at the Netherlands Forensic Institute (NFI) a list of facial feature comparisons is independently scored by three examiners. A consensus model is used to arrive at a verbal description of the likelihood that the crime scene image and the police photograph have the same origin. A judge combines this description with other evidence to arrive at a verdict.

This approach has some acknowledged issues such as latent examiner bias and inter examiner differences. Automating this process might mitigate the impact of these issues. Also, the comparison protocol is not standardized between law enforcement agencies. The Facial Identification Scientific Working Group (FISWG) publishes recommendations regarding one-to-one facial comparisons. A draft version of a facial image comparison feature list for morphological analysis [2] has been published by this organisation. Although the FISWG list can be regarded as a mnemonic tool for the forensic facial expert, it is also possible to interpret it as a definition of facial features. This paves the way for (semi-)automation of the facial comparison process.

The FISWG feature list is based on case work experience by forensic facial examiners. We evaluate in this paper the recognition performance of the FISWG eyebrow modality in a semi-automatic setting. To our knowledge, this is the first work to evaluate a FISWG feature description. The choice for the eyebrow modality is additionally motivated by the recent attention from the biometric community for soft biometric modalities in general and the eyebrow in particular. This makes a comparison with
a non forensic feature set possible. Also, whether a more optimal feature set can be found by combining non-forensic with forensic features will be investigated.

2 Related work

Some studies have shown that the eyebrow is a compact and rich container of information, both for humans [10] and for automatic recognition [11]. Early work of [13] based on a Hidden Markov Model reports recognition rates of 92.6% on a set of 54 high quality images. [6] automatically segments eyebrows and uses a Euclidian distance measure to compare contours of eyebrows. On a set of 200 high quality images a recognition rate of 88.1% is reported. The work of [11] is the first to use a substantial dataset (FRGCv2 Experiment 4 protocol) [3]. LBP is applied on spatial and frequency transformed images of the eyebrow strip. In general around a 10-20% TPR is reported at 1% FAR, depending on parameter settings and frequency representations. At first glance this might not seem impressive, but compared with the full face, the eyebrow region has a drop of \( \frac{5}{6} \) in size, but only a \( \frac{1}{6} \) drop in rank-1 identification. [8] selects shape-based eyebrow features for biometric recognition and gender classification. On a subset of the FRGCv2 dataset a rank-1 recognition rate of approximately 75% on the eyebrow is achieved. [15] combines dimensionality reduction techniques with a Radon transform and reports a recognition rate of approximately 87% on the high quality BJUT dataset [1]. [12] uses cross correlation for eyebrow detection and transforms the region of interest into the frequency domain. Recognition rates vary between 96.4% and 98.6% on the BJUT dataset, depending on parameter settings and distance measures.

Although most of the reported performances are impressive, they were obtained on good quality images in which individual hair can be recognised. This is not representative of the forensic situation where the quality (visibility, pose, illumination, expression, resolution) of the trace material is in general less than the reference material. In these limiting circumstances, the Dong Woodard feature set [8] can be considered as ”state-of-the-art”. Moreover, it contains features that could, in principle, be determined by a facial examiner.

3 Methods

3.1 Dong Woodard feature set

The Dong Woodard feature set [8] consists of three feature clusters: global (GL), local (LO) and critical (CR). The global cluster contains three general shape measures: rectangularity, eccentricity and isoperimetric quotient. A bounding box is divided into four equal horizontally (resp. vertically) adjacent subregions. The local feature consists of the relative percentage of eyebrow area in these 8 boxes. The critical features are the coordinates of the left, right, top and centroid point of the shape, expressed in a local coordinate system relative to the eyecorners. The local and critical features are shown in Figure (1).

3.2 FISWG feature set

In essence the FISWG eyebrow feature set [2] consists of four feature clusters: shape description (SH), relative bounding box size (BB), five specific relative distances (AE) and description of hair distribution throughout the eyebrow (HD). The shape description and hair distribution are formulated in a qualitative manner, implying the need
Figure 1: The local (left) and critical (right) features of the Dong Woodard feature set for a quantitative interpretation of these features. We experiment with different implementations of these features.

3.2.1 Shape

Initial experiments indicate that the 2D Fourier Shape Descriptor yields the most promising recognition results. This descriptor interprets the \( n \) points of the shape as a periodic signal in \( \mathbb{C} \). Suppose \( c_0, \cdots, c_{n-1} \) are its Fourier coefficients, then the \( k \) dimensional Fourier Descriptor is given by \( (|c_2|, \cdots, |c_{k+1}|) \). This shape descriptor is invariant under translation, rotation, and scaling [7]. Based on additional experiments we choose equidistant sampling of \( n = 512 \) points on the original shape and the subsequent Fourier Descriptor representation on \( k = 15 \) coefficients.

3.2.2 Bounding box and A-E measures

The second and third feature cluster have an anthropometric nature. The bounding box size (BB) is measured relative to the eye size, in our implementation the horizontal distance between the inner and outer eye corner is used. Furthermore, five special measures (A-E) are shown in Figure (2). In our implementation, these five measures are measured relative to the size of the eye.

Figure 2: A-E features, taken from [2]
3.2.3 Hair distribution

The eyebrow is segmented into 4 equiangular sectors, emanating from the midpoint between the inner and outer eye corner. For each sector the relative number of hair pixels within the eyebrow is determined. A pixel is considered to be hair if the probability being a skin color falls below a threshold. This probability is determined empirically in the same image on a skin patch above the eyebrow. A hue saturation bin of size $64 \times 128$ with a threshold of 0.01 is chosen.

3.3 Likelihoodratio paradigm

The task of the forensic examiner is to estimate likelihood ratios. Trace material from a crime scene (e.g. CCTV still image) and reference material (e.g. frontal image of suspect) form the basis for two hypotheses: the prosecutor hypothesis $\mathcal{H}_p$ (“trace and reference come from the same source”) and the defense hypothesis $\mathcal{H}_d$ (“trace and reference do not come from the same source”). Given the evidence $E$, the forensic examiner estimates the likelihoodratio $L(E) = \frac{p(E|\mathcal{H}_p)}{p(E|\mathcal{H}_d)}$. Based on prior odds $\frac{p(\mathcal{H}_p)}{p(\mathcal{H}_d)}$ and the likelihoodvalue $L(E)$, the judge uses the posterior odds $\frac{p(\mathcal{H}_p|E)}{p(\mathcal{H}_d|E)}$ to arrive at a verdict.

3.4 Likelihoodratio calculation in a (semi-)automatic setting

To determine $L(E)$ in a (semi-)automatic setting, a scorefunction $s(\cdot, \cdot)$ is applied on a training set containing pairs of featurevectors whose labels are known. This yields the score value probability distributions $p(s|\mathcal{H}_p)$ and $p(s|\mathcal{H}_d)$. These distributions are also referred to as ”imposter” and ”genuine”, respectively. Given the distributions and a score value $s^*$ from the case at hand, $L(s^*) = \frac{p(s^*|\mathcal{H}_p)}{p(s^*|\mathcal{H}_d)}$ is interpreted as $L(E)$. We adopt the approach from [14] where the score function $s(\cdot, \cdot)$ is directly modeled as a loglikelihoodratio:

$$s(x_1, x_2) = -\frac{1}{2} (x_1 - x_2)^T \Lambda^{-1} (x_1 - x_2) + \frac{1}{2} x_1^T x_1 - \frac{1}{2} \log(|\Lambda|),$$

where $x_1, x_2 \in \mathbb{R}^k$. It is assumed that the featurevectors have zero mean and unit variance and individuals share a diagonal within variance $\Lambda \in \mathbb{R}^{k \times k}$. Given a score value $s^*$ from the case at hand, we now may interpret this as an estimate for $\log(L(E))$.

3.5 Training, testing, and PAV calibration phase

The score function only acts on whitened data and requires a value for $\Lambda$. The training phase takes care of this. We sketch the procedure given in [14]. Given a training set $X = [X_1 \cdots X_n] \in \mathbb{R}^{m \times n}$ we substract the mean $\mu_X$ from all featurevectors in the training. Next we select two dimensionality reduction parameters $p$ and $l$, $m \geq p > l \geq 1$. The transformation $M \in \mathbb{R}^{l \times m}$ is a composition of a PCA projection from $m$ to $p$ dimensions, whitening, individual mean subtraction, and an LDA projection from $p$ to $l$ dimensions. The within variation $\Lambda$ is estimated from the transformed data $Y = M(X - \mu_X)$.

During the testing phase $\mu_X$, $M$, and $\Lambda$ are known. The query $X_q$ and target $X_{tar}$ datasets are transformed into $Y_q = M(X_q - \mu_X)$ and $Y_{tar} = M(X_{tar} - \mu_X)$, after which the loglikelihoodratio scorefunction is applied. Since we use small datasets, it can be beneficial to calculate the optimal classifier belonging to the convex hull of the ROC by.
means of the Pool of Adjacent Violaters (PAV) algorithm [9]. Moreover, the PAV algorithm also converts scores into loglikelihoodratios [16], a process known as calibration. The output of the testing phase is a calibrated genuine score set $G$ and a calibrated imposter score set $I$.

### 3.6 The $C_{llr}$ performance measure

$C_{llr}$ is a measure that captures both the discriminative power of a classifier and how well the scores are calibrated [5]. Since we use calibrated scores, it will solely measure the discriminative power. It is defined as

$$C_{llr} = \frac{1}{2} \left( \frac{1}{|G|} \sum_{g \in G} \log_2(1 + e^{-s_g}) + \frac{1}{|I|} \sum_{i \in I} \log_2(1 + e^{s_i}) \right)$$

where $G$ and $I$ are the genuine and imposter score sets.

### 4 Experimental setup and results

#### 4.1 Dataset and preprocessing

We select three datasets for our experiments. The first set, denoted by Sel1, consists of 500 images from 125 distinct persons taken from a selection of the FRGCv2 dataset. Each person is represented by two good quality and two lesser quality images. The second set, denoted by Sel2, consists of 400 good quality images from 100 distinct persons again taken from another selection of the FRGCv2 dataset. The final set is a subset of the high quality PUT [4] dataset, approximately 2200 images from 100 distinct persons. In every dataset the right and left eyebrow are manually segmented after which the Dong Woodard and FISWG features are automatically determined.

#### 4.2 Experiments

We conduct two experiments. The purpose of Experiment 1 is twofold. First, we measure the recognition performance of the separate feature clusters of FISWG. Next, we search for a small collection of feature cluster sets that have a promising recognition performance. By varying all possible dimensionality reduction parameters $p$ and $l$ a set of 37472 classifiers is obtained. Experiment 1 uses a 5 fold cross validation scheme and is repeated six times (3 databases, left/right eyebrows).

Experiment 2 builds upon the first experiment. The purpose of Experiment 2 is to assess the performance of the Dong Woodard, FISWG and a small collection of promising feature cluster sets. We train in total 3093 classifiers using these feature combinations on the Sel2 dataset and test the recognition performance on the Sel1 and PUT datasets. This experiment is repeated twice (left/right eyebrows).

#### 4.3 Results Experiment 1

In this experiment the performance of the separate feature clusters of FISWG is measured. Also, we search for promising feature cluster sets. The best classifiers on a given feature set are shown in Figure (3). For comparison purposes the classifiers using the Dong Woodard and FISWG feature sets are also provided. In general, the results on right and left eyebrows are consistent within a dataset. On the Sel1 and Sel2
datasets the recognition performance of the underlying feature clusters is in decreasing order \( \text{AE} \rightarrow \text{SH} \rightarrow \text{BB} \rightarrow \text{HD} \), on the PUT dataset \( \text{SH} \rightarrow \text{AE} \rightarrow \text{HD} \rightarrow \text{BB} \). Two differences are noteworthy. The \( \text{AE} \rightarrow \text{SH} \) difference might be explained by the difference of detailed variation in the original eyebrow shapes. The improved performance of the hair feature on the PUT dataset is explained by a higher quality in terms of resolution and illumination, yielding a clearer distinction between hair and skin pixels.

On the Sel1 dataset, the optimal classifier operates on the feature set \( \{ \text{AE}, \text{SH}, \text{CR} \} \). On the other two datasets the feature set on which the optimal classifier operates differs between the right and left eyebrow. On the Sel2 dataset the best classifier on the right eyebrow is the same as on Sel1. The set \( \{ \text{HD}, \text{AE}, \text{SH}, \text{CR} \} \) is optimal for the left eyebrow of Sel2 and for the right eyebrow of PUT. Finally, the set \( \{ \text{HD}, \text{AE}, \text{SH}, \text{LO} \} \) is optimal for the left eyebrow of PUT. This indicates there does not exist a unique optimal feature set but rather a small collection of optimal feature sets.

When comparing the Dong Woodard and FISWG feature set performances in Figure 3, only on the PUT dataset there seems to be a consistent difference in favor of the FISWG feature set. As mentioned earlier, the FISWG feature set uses texture information, so it is expected to perform better than the Dong Woodard feature set on good quality eyebrow images.

![Figure 3: DET curves for Experiment 1. The columns from left to right are FISWG clusters right, Dong Woodard/FISWG/Optimal features right, FISWG clusters left, Dong Woodard/FISWG/Optimal features left; the rows from top to bottom are Sel1, Sel2, and PUT](image-url)
4.4 Results Experiment 2

In this experiment a limited set of classifiers are trained on the Sel1 dataset and tested on the Sel2 and PUT datasets. In Figure (4) the best classifiers on the Dong Woodard, FISWG and optimal feature cluster set are shown. The performance of the Dong Woodard and FISWG feature sets are comparable. Also, the performance of the optimal feature cluster set is not significantly better than these feature sets.

![Figure 4: DET curves for Experiment 2. From left to right: Sel1/Right, Sel1/Left, PUT/Right, PUT/Left.](image)

5 Conclusions and future work

In our study we implemented the FISWG eyebrow feature and investigated its performance. The components of FISWG ordered in increasing performance are \{AE, SH\} and \{BB, HD\}, the order within the sets being dependent on the dataset used. Our study shows that the performance of the FISWG feature set is comparable to the Dong Woodard feature set, in terms of the $C^{llr}$ performance measure. This shows that the FISWG feature set can be considered as being "state-of-the-art". Also, the performance of optimal feature clusters sets do not differ significantly from the FISWG feature set, emphasising the existence of a small collection of good feature cluster sets.

For future work, we intend to measure the performance of forensic facial examiner and compare their performance with our semi-automatic system.

References


Visual information encoding for face recognition: sparse coding vs vector quantization

Dimche Kostadinov  Sviatoslav Voloshynovskiy  Maurits Diephuis
University of Geneva
Computer Science Department, Stochastic Information Processing Group
7 Route de Drize, Geneva, Switzerland
dimche.kostadinov@unige.ch  svolos@unige.ch  maurits.diephuis@unige.ch

Abstract
In this paper, we investigate the problem of visual information encoding for face recognition. We consider two models of information encoding based on sparse coding and vector quantization and compare their performance and computational complexity. The optimal solution is considered from the point of view of the best achievable classification accuracy by minimizing the probability of error under a given class of distortions. The results from the computer simulations confirm that our approach achieves similar performance with state-of-the-art sparse coding based image classification methods but with the considerably lower complexity.

1 Introduction
Visual information classification is of great practical interest in many multimedia and security applications. Traditionally, human face recognition is considered to be a reference application for testing different recognition frameworks. The main reasons for the interest in automatic human face recognition systems are the wide range of real world practical applications such as identification, verification, posture/gesture recognition, social network linking and multi-modal interaction.

In the past, Nearest Neighbour (NN) [2] and Nearest Feature Subspace (NFS) [7] have been used for classification. NN classifies the query image by only using its Nearest Neighbour. It utilizes the local structure of the training data and is therefore easily affected by noise. NFS approximates the query image by using all the images belonging to an identical class, using the linear structure of the data. Class prediction is achieved by selecting that class of images that minimizes the reconstruction error. NFS might fail in the case that classes are highly correlated to each other. Certain aspects of these problems can be overcome by Sparse Representation based Classification (SRC) [9]. However, on the other hand Qinfei at al. [8] argue that the lack of sparsity in the data means that the compressive sensing approach cannot be guaranteed to recover the exact signal and therefore that sparse approximations may not deliver the desired robustness and performance. It has also been shown [1] that in some cases, the locality of the dictionary codewords is more essential than the sparsity. An extension of SRC, denoted Weighted Sparse Representation based Classification (WSRC) integrates the locality structure of the data into a sparse representation in a unified formulation.

In the most favourable case, when the training and observation models are known, one can design an optimal encoding/representation and a classifier that minimizes the classification error. However, in many applications the training and observation models are unknown or highly non-stationary and one only has a few training samples. In such a set-up, the recognition system basically learns the classifier in a “blind” way using only the available distorted training samples and expects that the observation model will exhibit similar behavior to the training model.
Most of the recent classification frameworks mainly rely on the discriminative nature of the sparse representation to perform classification. Accuracy notwithstanding, it remains an open question whether or not this family of sparse methods attains the best trade-off for memory and computational complexity.

Therefore, considering the case in which the training and observation models are unknown, we focus on the problem of visual information encoding under prior ambiguity. In our formulation, the considered problem is closely related to both machine learning and coding. It should be pointed out that an alternative way of visual information encoding is based on the bag-of-features (BoF) approach. We will proceed with the generalized consideration of the BoF approach for multiple levels of multi-resolution image representation. Since the core of this representation is based on vector quantization, we will refer to this approach as vector classification based recognition.

Practically, we consider and compare a type of face recognition system based principally on sparse coding and a type based on vector quantization. Both approaches are evaluated in terms of their classification accuracy for a certain range of distortions and in their computational and memory requirements.

This paper is organized as follows. Section 2 gives the basic problem formulation. In Section 3, we describe the sparse representation based recognition model, whereas the vector quantization method is introduced in Section 4. The results of the computer simulations for both methods are analysed in Section 5. Finally, Section 6 concludes the paper.

**Notation:** We use capital bold letters to denote real valued matrices, $\mathbf{W} \in \mathbb{R}^{M \times N}$, small bold letters to denote real valued vectors: $\mathbf{x} \in \mathbb{R}^{M}$. We use sub and upper indexed vectors to denoted a single realization out of many from a given distribution e.g. $\mathbf{x}_{i}(m) \in \mathbb{R}^{M}$, where $m$ denotes the sample from some distribution. We denote an element of a vector as $x$. The estimate of $\mathbf{x}$ is denoted as $\hat{\mathbf{x}}$. All vectors have finite length, explicitly defined where appropriate.

## 2 Problem Formulation

The face recognition system consists of two stages: **enrollment** and **identification**.

At the enrollment stage, the photos from each subject are acquired and organized in the form of a codebook. We will assume that the recognition system should recognize $K$ subjects. The photos of each subject $i$, $1 \leq i \leq K$, are acquired under different imaging conditions such as lighting, expression, pose, etc., which will represent the variability of face features and serve as intra-class statistics. We will also assume that the frontal face images are aligned to the same scale, rotation and translation using common computer vision features.

Thus each subject $i$ is defined by $\mathbf{x}_{i}(m) \in \mathbb{R}^{N}$ vectors representing a concatenation of aligned image columns with $1 \leq m \leq M$, where $M$ represents the number of training images per subject that we assume to be the same for all subjects. The samples from all subjects are arranged into a codebook represented by a matrix:

$$
\mathbf{W} = [\mathbf{x}_{1}(1), ..., \mathbf{x}_{1}(M), ..., \mathbf{x}_{i}(1), ..., \mathbf{x}_{i}(M), ..., \mathbf{x}_{K}(1), ..., \mathbf{x}_{K}(M)] \in \mathbb{R}^{N \times (K \times M)}.
$$

At the recognition stage, a probe or query $\mathbf{y} \in \mathbb{R}^{N}$ is presented to the system. The system should identify the subject $i$ as accurate as possible based on $\mathbf{y}$ and $\mathbf{W}$. It is also assumed that $\mathbf{y}$ always corresponds to one of the subjects represented in the database. If it is not a case, a rejection option is integrated into the final decision.
3 Sparse Representation Based Recognition

In this section, face recognition is considered as a classification problem where the classifier should produce a decision in favour of some class \( i \) whose codebook codewords produce the most accurate approximation of the probe \( y \). One important class of approximations is represented by a sparse linear approximation \([9]\), where the probe \( y \) is approximated by \( \hat{y} \) in the form of:

\[
\hat{y} = W\alpha,
\]

where \( \alpha \in \mathbb{R}^{M \times K} \) is a sparse coding vector. The coding vector \( \alpha \) weights the codebook codewords gathered for all classes to favour the contribution of codewords corresponding to the correct class \( \hat{i} \). The model of approximations can be represented as:

\[
y = \hat{y} + r,
\]

where \( r \in \mathbb{R}^N \) is the residual approximation error vector.

For each class \( i \), let \( \delta_i : \mathbb{R}^{K \times M} \to \mathbb{R}^{K \times M} \) be a function that selects the coefficients associated with the \( i \)th class. For \( \alpha \in \mathbb{R}^{K \times N} \), \( \delta_i(\alpha) \) is a new vector whose only non-zero entries are the entries in \( \alpha \) that are associated with class \( i \).

Then the probe \( y \) is classified based on the approximation \( \hat{i} \) that minimizes the \( L_p \)-norm of the residual error vector between \( y \) and \( \hat{y} \):

\[
\hat{i} = \arg \min_{1 \leq i \leq K} \| W\delta_i(\alpha) - y \|_p.
\]

Equation (4) corresponds to the minimum \( L_p \) distance classification, where for \( p = 2 \) one has the Euclidean distance and for \( p = 1 \) one has the Manhattan distance. A natural extension to (4) that might be considered is a bounded distance decoding (BDD) rule:

\[
\hat{i} = \{ i \in \{1, \ldots, M \} : \| W\delta_i(\alpha) - y \|_p \leq \eta N \},
\]

where \( \eta \geq 0 \). The BDD rule is useful when the classifier should reject probes that are unrelated to the database. In the general case, the BDD will produce a list of candidates that satisfy the above condition. To have only one unique \( \hat{i} \) on the list, the parameter \( \eta \) should be chosen accordingly. Geometrically in the \( L_p \) space, it means that the \( L_p \) spheres with radius \( \eta \) around each approximate centroid for each class should not overlap, thus producing a unique classification.

The generalized solution of the approximation problem (2) under the constraint of sparsity of vector \( \alpha \) as a constrained optimization problem was considered in our previous work [5]:

\[
\hat{\alpha} = \arg \min_{\alpha} (\phi(W\alpha - y) + \lambda \psi(G\alpha)),
\]

where \( \phi(.) \) is the penalty function corresponding to the prior distribution of the residual vector, \( \psi(.) \) is a regulator corresponding to the prior distribution of the approximation coefficients \( \alpha \) and \( \lambda \) corresponds to the Lagrangian multiplier. The matrix \( G \) is the regularization matrix, where a simple selection of the regularizer \( G \) corresponds to the identity matrix \( G = I \). A diagonal form of \( G \) might also be used to enforce linear locality constraints [6].

Note that this problem formulation, depending on the penalty function and the regularization term, considers the cases of hard and soft encoding (also global and local as in [5]).
4 Multilevel Vector Quantization based Recognition

In this section, we consider an alternative model of classification based on multilevel vector quantization (MVQ). The proposed approach has a certain similarity with BoF methods and convolutional deep learning neural networks (CNN). The image is partitioned on overlapping or non-overlapping blocks. The main idea behind the proposed method is to learn a codebook of centroids $C^\ell_j = \{c^\ell_{1,j}, \ldots, c^\ell_{Kc,j}\}$ for each block $j$, $1 \leq j \leq B\ell$, where $B\ell$ is the number of blocks at each level of decomposition $\ell$, $1 \leq \ell \leq L$ and $Kc$ stands for the number of centroids chosen to be the same for all blocks and all levels. The different levels correspond to the different block sizes used for the image partitioning. The decomposition of images on local blocks of different sizes is explained by: (a) the necessity to cope with the non-stationary nature of distortions that are approximated by stationary ones using local decompositions and (b) the multilevel decomposition should take the relationship between local coefficients into account, similar to CNN methods. The overall goal of the proposed method is to achieve a competitive classification accuracy together with an acceptable memory storage and complexity.

The MVQ based classification consists of three main steps: (a) codebook generation, (b) block encoding using the basis vectors of the generated codebook and (c) classification.

4.1 Codebook generation

Given the training data $x_i(m)$ for all subjects $1 \leq i \leq K$ with $1 \leq m \leq M$ training samples per subject, each image is partitioned on $B\ell$ blocks of size $2 \times 2$ and $3 \times 3$ corresponding to the levels of decomposition $\ell \in \{1, 2\}$, respectively. Therefore, each block of a training image $x_i(m)$ is denoted as $x^\ell_{i,j}(m)$. The trained codebook for each block $j$ at the level $\ell$ consists of a set of $Kc$ centroids $C^\ell_j = \{c^\ell_{1,j}, \ldots, c^\ell_{Kc,j}\}$, learned with the $k$-means algorithm.

4.2 Encoding

Given a set of training samples $x_i(m)$ for all subjects $1 \leq i \leq K$ with $1 \leq m \leq M$, represented with a multilevel block decomposition, each block is assigned to the nearest centroids using a $k$-NN or $\epsilon$-NN strategy (bounded distance decoding) $^*$, represented by the list:

$$\mathcal{L}(x^\ell_{i,j}(m)) = \{w \in \{1, \ldots, Kc\} : d(x^\ell_{i,j}(m), c^\ell_{w,j}) \leq \epsilon L\ell\},$$

where $1 \leq w \leq Kc$, $L\ell$ is the block total size at level $\ell$ and $\epsilon \geq 0$.

The encoding results in the generation of an encoding vector: $D^\ell_{x_i,j,w}(m) = (d^\ell_{x_{i,j,1}}(m), \ldots, d^\ell_{x_{i,j,Kc}}(m)) \in \{0, 1\}^{L \times Kc}$, with $d^\ell_{x_{i,j,w}}(m) = 1$ for $w \in \mathcal{L}(x^\ell_{i,j}(m))$.

The final stage of encoding includes the pooling of results from all training samples to the final index that is accomplished based on max-pooling (MAXP) or alternatively average-pooling (AVGP):

$$\text{MAXP: } d^\ell_{x_{i,j,w}} = \max_{1 \leq m \leq M} d^\ell_{x_{i,j,w}}(m), \quad \text{AVGP: } d^\ell_{x_{i,j,w}} = \sum_{m=1}^{M} d^\ell_{x_{i,j,w}}(m).$$

$^*$Due to the space limitation, we will proceed with the $\epsilon$-NN only.
The main idea behind this particular form of max pooling is to capture all centroids for a given block and level of decomposition that might represent a subject under various observation distortions. In fact, if the observation model were stationary and known, the representative centroids could be computed analytically.

The final stage of encoding includes the generation of an inverted file look up table where each block $1 \leq j \leq B_\ell$ has a set of centroids $c_{w,j} \in \{1, \cdots, K_c\}$ at the level $\ell \in \{1, 2\}$ containing the indices of corresponding subjects $i \in \{1, \cdots, K\}$. This look up table is very sparse and efficient for memory storage.

### 4.3 Classification

The classification consists of two stages: (a) block decoding and (b) final fusion.

#### 4.3.1 Block decoding

The goal of block decoding is to find the $\epsilon$-NN centroids corresponding to each block of observation image $y$:

$$\mathcal{L}(y_j^\ell) = \{w \in \{1, \cdots, K_c\} : d(y_j^\ell, c_{w,j}^\ell) \leq \epsilon L_\ell\},$$

where $1 \leq w \leq K_c$, $L_\ell$ is the block total size at level $\ell$ and $\epsilon \geq 0$.

The block decoding results in the generation of activation vector:

$$D_{y_j,w}^\ell = (d_{y_j,1}^\ell, \cdots, d_{y_j,K_c}^\ell).$$

It is important to note that the activation vector might be considered as hard decoding, when its elements are assigned 0 or 1, if the above condition is satisfied, or soft decoding, when its element correspond to the reliability or likelihood of observing some centroid $c_{w,j}^\ell$ given the block $y_j^\ell$, for $w \in \mathcal{L}(x_{i,j}^\ell(m))$. In the case of soft decoding, the reliable centroids will obtain weights closer to 1 and non-reliable closer to 0. It is also remarkable that in the case of reliable decoding, the list $\mathcal{L}(y_j^\ell)$ will be sparse indicating that the reliable centroid(s) is(are) found. Otherwise, all elements of this list will have identical weights. Therefore, one can use the notion of sparsity to estimate the reliability of the produced estimate. We refer the interested readers to [4] for more details.

#### 4.3.2 Final fusions

The final decision can be produced at each level of decomposition $\ell$ that would correspond to more conservative recognition architectures or it can be obtained as a result of fusion from several levels.

Therefore, each block of the observation image $y$ at each level $\ell$ produces the lists of image indices that are the most likely candidates for the corresponding blocks. Thus, the final decision is just a result of the most likely index $i \in \{1, \cdots, K\}$ selection that obtains the majority of votes. It should also be pointed out that each decision can be produced as a result of the largest similarity between the observation vector $d_{x_{i,j},w}^\ell$ and $d_{y_j,w}^\ell$ that is estimated as:

$$\hat{i}_\ell = \max_{i \in \{1, \cdots, K\}} t_i^\ell, \text{ where: } t_i^\ell = \sum_{j=1}^{B_\ell} \sum_{w=1}^{K_c} d_{x_{i,j},w}^\ell d_{y_j,w}^\ell.$$ 

85
The decision at the global level is produced as:

$$\hat{i} = \max_{i \in \{1, \ldots, K\}} t_i, \text{ where: } t_i = \sum_{\ell=1}^{L} \sum_{j=1}^{B_\ell} \sum_{w=1}^{K_c} d_{x_i,j,w} d_{y_j,w}. \quad (12)$$

## 5 Computer Simulations

In this section we present the results of the computer simulation. We compare the results of the accuracy using several sparse based representation classification models [5] versus MVQ with a list decoding model. In addition we indicate their computational complexity by measuring the average execution time for a single recognition.

The computer simulation is carried out on the publicly available Extended Yale B database for face recognition [3]. We use all images from this database cropped and normalized to 192x168 pixels. In our set up, the images from the dataset are rescaled to 10x12 pixels using nearest neighbour interpolation. In all of the computer simulations we use raw, basic, elementary image pixel values (block of image pixel values) as features. To be unbiased in our validation of the results we use 5-fold cross validation, where for a single validation for each subject, half of the images are selected at random for training and the remainder for testing.

All of the MVQ models for block $j$ at the level $\ell$ use trained codebooks that consists of a set of $K_c$ centroids $C^\ell_j = \{c^\ell_{1,j}, \ldots, c^\ell_{K_c,j}\}$, learned with the k-means algorithm. The number of centroids at any level $\ell$ for any block $j$ is 512.

Figure 1 shows the resulting accuracy of the MVQ method using one layer independent, overlapping 2×2, 3×3 and two layer joint, overlapping 2×2 and 3×3 blocks with hard and soft decoding, employing bounded distance decoding with different $\epsilon$ values. The parameter $\epsilon$ in equations (7) and (9) is chosen adaptively for each block based on the sparsity level $\epsilon$ as defined in [4].

Table 1 summarizes the best achievable accuracy of classification at different levels for the hard and soft decoding.

<table>
<thead>
<tr>
<th></th>
<th>2 × 2</th>
<th>3 × 3</th>
<th>Fused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard encoding</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Soft encoding</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 1: MVQ recognition accuracy.

Figure 2a gives a comparison of the accuracy of all methods deployed while Figure 2b shows the computation time needed for classifying a single image query.

In conclusion, the accuracy of MVQ based recognition using 2 × 2 and 3 × 3 overlapping blocks is 0.97 which is on par with the best sparse coding based recognition method denoted as "∗", that is, the sparse approximation method that uses the $L_1$ norm as a penalty function and the $L_1$ norm as regulizer applied on overlapping blocks from [5].

It is also noteworthy that the MVQ method is between a factor 200 faster in recognition than competing sparse methods.
(a) MVQ using blocks of size \(2 \times 2\)

(b) MVQ using blocks of size \(3 \times 3\)

(c) MVQ using blocks of size \(2 \times 2\) and \(3 \times 3\)

Figure 1: Accuracy of MVQ recognition with one layer independent, overlapping blocks of size \(2 \times 2, 3 \times 3\) and two layer joint, overlapping blocks of size \(2 \times 2\) and \(3 \times 3\) with hard and soft decoding (red and blue line respectively), employing bounded distance decoding with different \(\varepsilon\) values, computed for each block \(j\).

(a) Comparison of the accuracy using different models

(b) The CPU running time using different models for a PC with an Intel Xeon CPU E5-16200 @ 3.6GH and 32GB RAM memory

Figure 2: Comparison considering the best accuracy and computation time for the MVQ method (using overlapping blocks with sizes \(2 \times 2\) and \(3 \times 3\)) and the methods *, WSRC, SRC and LLC

6 Future work

Future research will consider explore geometrically invariant coding strategies, and further link the MVQ framework with conventional convolutional neural networks. Furthermore, we will strive to optimize the encoding and list decoding strategies by incorporating reliability statistics.

7 Conclusions

In this paper we considered the face recognition problem from a both machine learning and information coding perspective, adopting an alternative way of visual information encoding. Our model of classification is based on multilevel vector quantization (MVQ), conceptually similar to BoF and CNN. The results from the computer simulations confirm that the MVQ based recognition model achieves an accuracy that is comparable to state-of-the-art sparse coding based image classification methods[5]. In addition the complexity in terms of processing time and memory of the MVQ model is significantly lower compared to other state-of-the-art methods based on sparse coding.
Acknowledgement

The work presented in this paper was supported by a grant from Switzerland through the Swiss Contribution to the enlarged European Union (PSPB-125/2010).

References


Automatic Eye Detection Error as a Predictor of Face Recognition Performance

Abhishek Dutta       Raymond Veldhuis       Luuk Spreeuwers
University of Twente, Netherlands.
{a.dutta,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

Abstract
Various facial image quality parameters like pose, illumination, noise, resolution, etc. are known to be a predictor of face recognition performance. However, there still remain many other properties of facial images that are not captured by the existing quality parameters. In this paper, we propose a novel image quality parameter called the Automatic Eye Detection Error (AEDE) which measures the difference between manually located and automatically detected eye coordinates. Our experiment results carried out using FaceVACS recognition system and the MultiPIE dataset show that AEDE is indeed a predictor of face recognition performance.

1 Introduction
The quality of facial images is known to affect the performance of a face recognition system. A large and growing body of literature has investigated the impact of various image quality parameters on the performance of existing face recognition systems [1]. The most commonly used image quality parameters are: facial pose, illumination direction, noise, blur, facial expression, image resolution. However, some aspects of the recognition performance that cannot be explained by the existing image quality measures remain. This shows that still more quality parameters are needed to fully explain the variation in recognition performance.

In this paper, we propose a novel image quality parameter called the Automatic Eye Detection Error (AEDE). Automatic eye detectors are trained to return the location of two eye coordinates in a facial image. To assess the accuracy of automatic eye detectors, we use the manually annotated eye coordinates as the ground truth eye locations. The proposed AEDE measures the error in automatically detected eye coordinates. The main insight underpinning this novel image quality parameter is as follows: Automatic eye detection becomes more difficult for poor quality facial images and hence the eye detection error should be an indicator of image quality and face recognition performance. In other words, we use the knowledge of the accuracy of one classifier (i.e. automatic eye detector) as the predictor of the accuracy of another classifier (i.e. the face recognition system) when both operate on the same pair of facial images. The proposed AEDE quality measure can be seen as providing a summary of many, but not all, properties of a facial image.

This paper is organized as follows: In Section 2, we review some previous work in this area. We explain the proposed AEDE quality measure in Section 3. We describe experiments to study the relationship between AEDE and face recognition performance in Section 4.
2 Related Work

The face recognition research community has been investigating the impact of automatic eye detection error on facial image registration which in turn influences face recognition performance [5, 12, 6, 7, 9, 14, 8]. While some researchers have focused on improving the accuracy of automatic eye detectors [13], others have explored multiple ways to make face recognition systems inherently robust to facial image registration errors [10, 11].

To the best of our knowledge, no previous work has proposed the Automatic Eye Detection Error (AEDE) as a predictor of face recognition performance. However, [12] make a concluding remark that points in this direction. The authors mention that “a face recognition system suffers a lot when the testing images have the lower face lighting quality, relatively smaller facial size in the image,...”. They further note that “the automatic eye-finder suffers from those kinds of images too”. This paper is probably the first to observe that some facial image quality parameters (like illumination, resolution, etc.) impact the performance of both face recognition systems and automatic eye detectors.

3 Methodology

Manually annotated eye coordinates are used as the ground truth for the eye locations in a facial image. Based on this knowledge of true location of the two eyes, we can assess the accuracy of an automatic eye detector. The error in automatic eye detection gives an indication of how difficult it is to automatically detect eyes in that facial image. Some of the image quality variations that make the automatic eye detection difficult also contribute towards the uncertainty in decision about identity made by a face recognition system operating on that facial image. For example: a poorly illuminated facial image not only makes eye detection difficult but it also makes face recognition harder.

Let \( p_{\{l,r\}} \) denote the manually located left and right eye coordinates (i.e. the ground truth). An automatic eye detector is trained to locate the position of the two eye coordinates \( p_{\{l,r\}} \) in a facial image. The error in automatically detected eye coordinates can be quantified using the Automatic Eye Detection Error (AEDE) [4] as follows:

\[
J = \frac{\max \{||p_{\{l,r\}} - p_l||, ||p_{\{l,r\}} - p_r|||}{||p_{\{l,r\}}||}
\]

Let \( J_{\{p,g\}} \) denote the AEDE in a probe and gallery image pair respectively. For this probe and gallery image pair, let \( s^k \) denote the similarity score computed by face recognition system \( k \). We divide \( J \) into \( L \) monotonically increasing intervals (based on quantiles, standard deviation of observed \( J_{\{p,g\}} \), etc.): \( J^l \) where \( l \in \{1, \cdots, L\} \). We partition the set of all similarity scores \( S \) into \( L \times L \) categories of genuine \( G \) and impostor \( I \) scores defined as follows:

\[
G_{(l_1,l_2)} = \{ S(i) : J_p(i) \in J^{l_1} \land J_g(i) \in J^{l_2} \land S(i) \text{ denotes genuine comparison} \},
\]
\[
I_{(l_1,l_2)} = \{ S(i) : J_p(i) \in J^{l_1} \land J_g(i) \in J^{l_2} \land S(i) \text{ denotes impostor comparison} \},
\]

where, \( l_1, l_2 \in \{1, \cdots, L\} \), \( J_{\{p,g\}}(i) \) denotes the normalized eye detection error (or, AEDE) in probe and gallery image respectively corresponding to \( i^{th} \) similarity score \( S(i) \). The performance of a verification experiment is depicted using a Receiver Operating Characteristics (ROC) curve. The ROC curve corresponding to a particular eye detection error interval \( (l_1, l_2) \) is jointly quantified by False Accept Rate (FAR) and False Reject Rate.
(FRR) defined as follows:

\[
\text{FAR}_{(l_1,l_2)}(t) = \frac{n(\{I_{l_1,l_2} : I_{l_1,l_2} > t\})}{n(I_{l_1,l_2})},
\]

\[
\text{FRR}_{(l_1,l_2)}(t) = \frac{n(\{G_{l_1,l_2} : G_{l_1,l_2} < t\})}{n(G_{l_1,l_2})},
\]

where, \( t \) denotes the decision threshold similarity score and \( n(A) \) denotes the cardinality of set \( A \).

Our hypothesis is that the eye detection error \( J \) defined in (1) is correlated with face verification performance defined by (4). Therefore, we expect ROC curves corresponding to different eye detection error intervals to be distinctly different from each other. Furthermore, we also expect recognition performance to degrade monotonically with increase in eye detection error.

The proposed AEDE quality measure should be used with caution because all the factors that make eye detection difficult are not necessarily always involved in making face recognition harder. For example, a facial photograph captured under studio conditions but with the subject’s eyes closed is a difficult image for automatic eye detector while a face recognition system can still make accurate decisions as most important facial features are still clearly visible. Therefore, in addition to the automatic eye detection error, we need more quality parameters in order to reliably predict face recognition performance.

4 Experiments

Fig. 1: MultiPIE camera and flash positions used in this paper.

In this section, we describe experiments that allow us to study the relationship between Automatic Eye Detection Error (AEDE) and the corresponding face recognition performance.

We use the facial images present in the neutral expression subset of the MultiPIE data set [3]. We include all the 337 subjects present in all the four sessions (first recording only). In our experiments, the image quality (i.e. pose and illumination) variations are only present in the probe (or, query) set. The gallery (or, enrollment) set remains fixed and contains only high quality frontal mugshots of the 337 subjects. The probe set contains images of the same 337 subjects captured by the 5 camera and under 5 flash positions (including no-flash condition) as depicted in Fig. 1. Since our gallery set remains constant, we only quantify the normalized eye detection error for facial images in the probe set \( J_p \). Of the total 27630 unique images in the probe set, we discard 69 images for which the automatic eye detector of FaceVACS fails to locate the two eyes.

We have designed our experiment such that there is minimal impact of session variation and image alignment on the face recognition performance. We select the high quality gallery image from the same session as the session of the probe image. Furthermore, we disable the automatically detected eye coordinates based image alignment of FaceVACS.
by supplying manually annotated eye coordinates for both probe and gallery images. This ensures that there is consistency in facial image alignment even for non-frontal view images.

We manually annotate the eye locations $p_{(l,r)}^{(t)}$ in all the facial images present in our data set. Using the eye detector present in the FaceVACS SDK [2], we automatically locate position of the two eyes $p_{(t)}^{d}$ in all facial images. Given the manually annotated and automatically detected eye locations, we quantify the eye detection error $J$ using (1). In Fig. 2, we show the distribution of normalized eye detection error $J_p$ for images in the probe set categorized according to MultiPIE camera and flash identifier. The horizontal and vertical axes of Fig. 2 represent variations in camera and flash respectively. The inset images show a sample probe image with the given pose and illumination.

Now, using FaceVACS [2] recognition system, We now obtain the verification performance corresponding to each unique pair of probe and gallery images. For each verification instance, we have $(J_p, s_{pg}^k)$ where $J_p$ denotes the normalized eye detection error in the probe image and $s_{pg}^k$ is the similarity score (i.e. verification score) computed by $k^{th}$ face recognition system. Since we use only one face recognition system in our experiments, we drop the superscript $k$. Recall that our gallery set remains fixed to high quality images and therefore, we only consider the eye detection error of probe images. This not only simplifies the analysis and presentation of results but also simulates the conditions of a real world verification experiment. We partition the set of all similarity scores $S = \{s_{pg}\}$ into four categories based on the corresponding normalized eye detection error of the probe image $J_p$. If $q_1, q_2, q_3$ denote the 25%, 50%, 75% quantiles of $J_p$, then the four categories correspond to the following interval: $J_1 = [0, q_1), J_2 = [q_1, q_2), J_3 = [q_2, q_3), J_4 = [q_3, 1)$. In Fig. 3, we show the ROC corresponding to the four intervals of $J_p$ as shown in Table 1. The solid lines in Fig. 3 correspond to recognition performance when facial image registration is based on manually annotated eye coordinates. Section 5 describes , it will be clear that we need this result (i.e. the dotted lines)

While discussing our experiment results in Section 5, we need to rule out one possible explanation for the observed results. Therefore, in Fig. 3, we also plot the recognition performance when facial images are registered using automatically detected eye coordinates.

<table>
<thead>
<tr>
<th>Table 1: Interval of $J_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interval</td>
</tr>
<tr>
<td>$J_1$</td>
</tr>
<tr>
<td>$J_2$</td>
</tr>
<tr>
<td>$J_3$</td>
</tr>
<tr>
<td>$J_4$</td>
</tr>
</tbody>
</table>

## 5 Discussion

In this paper, we set out to find if the proposed Automatic Eye Detection Error (AEDE) is a predictor of face recognition performance. Image quality parameters are very strong indicators of face recognition performance. Therefore, we first investigate if AEDE responds to controlled pose and illumination variation in facial images.

We first visually inspect the distribution of AEDE to see if it responds to the quality variations present in our data set. In Fig. 2, we show the distribution of AEDE for images in the probe set categorized according to MultiPIE camera and flash identifier. First, for the frontal camera (05,1), let us compare the distributions corresponding to frontal flash (07) and no-flash. For frontal flash, the distribution of $J_p$ is nearly symmetric and centered around $J_p = 0.5$. For no-flash, the distribution becomes right skewed (i.e. right heavy tail)
indicating that many samples have high eye detection error. For other illumination variations also, we observe small increase in right skewness. This shows that the normalized eye detection error responds to illumination variations. Furthermore, higher values of AEDE corresponds to degrading illumination condition. Now let us compare the distributions for different pose variations under no-flash illumination condition. For frontal pose, the distribution of $J_p$ is already right skewed and it becomes more heavy on the right tail as we move away from the frontal pose. This indicates that AEDE increases as the pose moves away from frontal view. Therefore, we conclude that the proposed AEDE measure responds to, at least, pose and illumination quality variations in facial images.

In Fig. 2, we show the ROC corresponding to the four intervals of the normalized eye detection error in probe image $J_p$. First, we discuss the four ROCs (i.e. solid lines) corresponding to facial images registered using manually annotated eye coordinates. We observe that the four intervals of $J_p$ correspond to four distinct ROC curves. However, contrary to our expectations, the four monotonically increasing intervals of $J_p$ do not correspond to monotonically degrading ROC curves. For example, $J_1$ corresponds to the interval with lowest eye detection error but it does not correspond to the best ROC. In fact, the interval $J_2$ and $J_3$ correspond to best recognition performance. As expected, the largest eye detection error i.e. $J_4$ correspond to the worst recognition performance. These findings are unexpected and suggests that the normalized eye detection error has a non-linear relationship with face recognition performance. Our results further support the argument that a single metric is not sufficient to capture all image quality variations that may affect face recognition performance.
Fig. 3: Recognition performance variation for each monotonically increasing interval of normalized eye detection error $J$.

One could argue that the observed non-linear relationship is due to bias in the manually annotated eye coordinates and FaceVACS would behave differently if allowed to automatically register facial images. To check the validity of this argument, in Fig. 2, we plot the four ROCs (i.e. dotted lines) corresponding to facial images automatically registered by FaceVACS using its own detected eye coordinates. These ROCs also show the same trend and therefore this argument does not explain the non-linear relationship between eye detection error and recognition performance. Further work is required to determine the causes of this non-linearity.

6 Conclusion

In this paper, we have proposed Automatic Eye Detection Error (AEDE) as a predictor of face recognition performance. Our results show that AEDE has a non-linear relationship with face recognition performance and further work is required to fully understand the reasons for this non-linearity.

One of the major limitations of AEDE is that it requires manually annotated eye coordinates in order to quantify the quality of a facial image. For real time biometric applications, the manually annotated eye coordinates are usually not available. However, for forensic face recognition applications, a forensic investigator can manually annotate a small number of facial images relevant to the casework. Availability of such manual eye annotations can greatly help in quantifying the uncertainty in decision about identity using the proposed Automatic Eye Detection Error (AEDE) image quality measure.

The proposed eye detection error cannot capture all types of quality variations that may affect face recognition performance. For example, in a photograph containing facial image with closed eye, the eye detection error will be very high. This does not necessarily translate into a difficult verification problem. Similarly, facial expressions like smile can
greatly affect face recognition performance but may not necessarily impact the performance of an automatic eye detector. Therefore, we need more quality parameters to fully quantify the variability in recognition performance.

Acknowledgements

- This work was supported by the BBfor2 project which is funded by the EC as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) under Grant Agreement number 238803.

- We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendors maximum effort full capability result.

References


Automatic generic Region-Of-Interest selection for video surveillance applications

S. Javanbakht, X. Bao, S. Zinger, P. H. N. de With
Video Coding and Architectures Research group,
Electrical Engineering Dept., Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, the Netherlands
{S.Javanbakht, X.Bao, S.Zinger, P.H.N.de.With}@tue.nl

Abstract

The value of using cameras in surveillance is further augmented when the surveillance system can take decisions autonomously by means of video analysis. For dynamic scene analysis, automatic detection of informative regions such as Regions-Of-Interest (ROI) is a challenging task for surveillance applications due to the large variations of the scene material. Our hypothesis is that if such an ROI is detected, a further advanced video analysis can be applied later exclusively to that ROI. In this paper, we employ a DCT for an ROI detection, since it provides a compact representation of the signal energy and the computation can be implemented at low cost [1]. We verify the usefulness of our hypothesis by cascading our ROI detection techniques with a typical object detector [3] as used in surveillance cases, to evaluate the attractively of the concept. We validate this approach on two different datasets and also compare our algorithm with a number of simple, fast ROI detection techniques. The experimental results show that our proposed approach outperforms the other methods in recall, precision, as well as in computational time.

1 Introduction

Video surveillance has become a key component in ensuring security at airports, banks, casinos etc. and is increasingly used elsewhere in public places. The objective of visual surveillance is to alert police or security officers, where the system applies multiple cameras to look at different locations at the same time. The value of using cameras in surveillance is increasing when the surveillance system can take decisions autonomously by means of video analysis. Studies in psychology and cognition have found that, when looking at an image, our visual system would first quickly focus on one or several “interesting” regions of the image prior to further exploring the image contents. These regions are often called salient regions or Regions-Of-Interest (ROI) [4]. If such an ROI is detected, a further advanced video analysis can be applied later exclusively to that ROI. Our hypothesis is that this provides a significant gain in efficiency in terms of the number of pixels needed to be explored, as compared to analyzing the complete scene. By providing an ROI to a security officer, without the further in-depth analysis, the time for the first-stage visual analysis of the scene is reduced significantly, so that the officer can immediately respond to alarming situations. It can also reduce both the required bandwidth for sending the information and the memory needed for storage.

However, the methods currently used for ROI detection are quite complex and time-consuming. These methods involve training in order to learn the appearances of a specific ROI, before the system can search for it. This learning is time-consuming, ROI detection is also called salient region detection in the literature [4]. ROI detection techniques can be divided into three broad families of approaches: bottom-up, top-down or a combination of both. Bottom-up cues are mainly based on pixel and feature
characteristics of a visual scene, whereas top-down cues are determined by cognitive phenomena, like knowledge, expectations, reward, and current goals [4]. Bottom-up approaches are usually fast in execution, while top-down techniques are slower and task-driven [5]. In the literature, three features are often used for bottom-up ROI detection: intensity (or intensity contrast, or luminance contrast), color, orientation and motion. Intensity is usually implemented as the average of three color channels. Orientation is obtained by a convolution with oriented Gabor filters, or by the application of oriented masks. Motion is based on eye/head movement which can be derived from recordings in freely behaving Rhesus monkey [6]. In this paper, our approach is based on bottom-up techniques using frequency features. For the complete system, we aim at simple and fast methods suitable for real-time operation. The real-time aspect also explains why we have adapted the DCT, as this transform can be efficiently implemented.

To measure the potential efficiency gain, we cascade our ROI detection techniques with a typical object detector [3]. Object detection is challenging, because the appearances and size are highly variable even for the same type of object. Appearance-based techniques have been widely exploited for object detection. Haar-like and Histogram of Oriented Gradients (HOG) features are frequently applied together with classifiers such as AdaBoost or SVM, to train appearance-based detectors. For these appearance-based methods, exhaustive search is performed in an image with e.g. a sliding window with various scaling factors, which is rather computationally expensive. However, if we have prior knowledge of ROIs in the image and apply the detector in those regions, the complexity can be largely reduced to enable real-time applications. In turn, the ROI detection should not miss important objects, when it is applied as a initializing stage of processing.

The paper is organized as follows. Section 2 describes different approaches for ROI detection. Section 3 presents our results. Conclusions and discussion are provided in Section 4.

2 Region-Of-Interest detection approach

For detecting an ROI, we compare a number of simple, fast ROI detection techniques based on various features, such as sum of edge pixels using a Sobel edge detector, the number of connected components based on the intensity values of neighboring pixels, the number of straight lines found by the Hough transform and the entropy of the frequency-domain features of a DCT.

2.1 Sum of edge pixels

In this approach, the sum of edge pixels is used to detect an ROI. First, a Sobel edge detector is applied to an input frame. Then we combine resulting gradients in both horizontal and vertical directions to calculate the gradient magnitude of each pixel. Subsequently, we impose a threshold to the gradient magnitude of each pixel, in order to eliminate weak edges. After removing weak and small edges, we divide each resulting frame into a number of blocks. For deciding whether each block is an ROI or not, the algorithm compares the number of edge pixels in each block with a threshold, and if this number is above the threshold, it is considered as an interesting region.

2.2 Connected components

In this approach, we compute the number of connected components to find ROIs in an image. First, we apply a Sobel edge detector to the input frame. Then we compute the connected components of the resulting frame to label edge pixels, based on the intensity values of neighboring pixels. In this work, an 8-connected neighborhood is employed.
The result of the connected component algorithm is a labeled map, where each label contains a group of connected edge pixels. We explore the number of edge pixels in each label. When the number of edge pixels is lower than a predetermined threshold, then edge pixels from this label are discarded. This criterion avoids the occurrence of small edges that cannot be removed in the previous approach (see Section 2.1). For deciding whether each block is ROI or not after removing the small edges, the resulting map is divided into a number of blocks. Then, for each block the algorithm checks whether it contains any labels or not. The block is considered as an ROI if it contains any labels. This approach is more reliable compared to using only an edge detector.

2.3 Hough lines

In this approach we calculate the number of straight lines found by the Hough transform. The Hough transform is known as an algorithm which can extract lines effectively [7]. Straight lines can be parameterized in the polar Hough parameter space by their length, $\rho$, and orientation, $\theta$, of the normal vector to the line from the image origin [8]. This length parameter $\rho$ is specified by

$$\rho = x\cos(\theta) + y\cos(\theta),$$

where $(x, y)$ denotes a pair from a set of image coordinates, which are lying on a straight line. The transform is quantizing the Hough parameter space into accumulator cells. As the algorithm runs, each pixel $(x, y)$ is transformed into a discretized $(\rho, \theta)$ curve and the accumulator cells which lie along this curve are incremented. Resulting peaks in the accumulator array represent strong evidence that a corresponding straight line exists in the image.

To detect straight lines in an image, we first apply the Sobel edge detector to the image. Then the best scoring results after applying a Hough transform provide the straight lines in the binary image containing the edges. The number of straight lines is limited by a number of local maxima values which is selected to choose the best scoring results. Therefore, to avoid false rejections, first we divide a frame into a number of blocks. Later, we apply the Hough transform to each block. For deciding whether each block is an ROI or not, the number of straight lines associated with each block is compared to a predetermined threshold. If this number exceeds the threshold, the block is considered as an ROI.

2.4 Discrete Cosine Transform (DCT)

Here, the ROI detection algorithm is based on analyzing the information content of the image in the pseudo-frequency domain by computing the entropy of the involved DCT coefficients. The DCT provides a compact representation of the signal energy. To this end, we apply 2D DCT to each pixel block. The 2D DCT transforms the block of the input frame to a coefficient matrix, where each coefficient represents the degree of which a certain cosine function is present in the block of the input frame. If each block has a high peak only in low frequency and no other significant peaks, it is considered as not an informative block. Since the DCT value on location $(u, v) = (0, 0)$ in the coefficient matrix is rather large and represents the average or DC value, we remove this high value to evaluate the rest of the coefficients. To avoid small DCT coefficients at higher frequencies due to minor intensity changes and reduce influence of noise, we again define a threshold. To this end, we compute the entropy of the DCT coefficients for each block [7] to measure the information content. The entropy is defined as [9]:

$$ H = -\sum_{i=1}^{N} p_i \log_2 p_i, $$

(2)
where \( N \) is the size of the image and \( p_i \) contains the probability of the intensity value \( i \) averaged over all pixel locations. The number of bins in the histogram is specified by the image type and its definition. In our case, the input frame is converted to a gray-scale image and we employ 256 bins which corresponds to the number of gray levels.

For defining each block is considered an ROI or not, we compare the magnitude of the DCT coefficients of the current block with the the entropy of this block. If at least one DCT coefficient magnitude of each block is above the entropy of this block, then this block is considered an ROI.

### 3 Experimental results

#### 3.1 Initial test with still images

For evaluating the above four different ROI detection techniques, we have applied them to 30 images from the Caltech Pedestrian Detection Benchmark [2], and created ground truth for these images by manually annotating them. We have defined empirical thresholds for each technique after extensive experimentation to optimize their performance. We have used 10-frame intervals for temporal filtering of the frame-based ROI detections. This prevents flickering in the output due to small temporal changes and noise.

Figure 2 shows the results of applying the proposed ROI detection techniques to a sample frame. In this experiment, the block size for all four techniques is constant and equal to 32 × 32 pixels.

Table 1 presents the average precision and recall rates of the ROI detection approaches on 30 images of Caltech Pedestrian Detection Benchmark [2]. It shows that the results of the ROI detection obtained from the DCT-based approach are similar to the results of the Hough-based technique. Table 2 shows that the edge-based approach is the fastest and the Hough-based technique is the slowest algorithm. By analyzing Table 1 and Table 2, we conclude that the DCT-based approach is the most attractive compared to the other approaches. Therefore, for further analysis of the DCT-based approach, we have also evaluated it on our highway video sequence. This sequence consists of 40 frames with a resolution of 1280 × 960 pixels and a frame rate of 25 fps. We have obtained a precision and recall of 39% and 75%, respectively. The average precision and recall rates of the DCT-based approach for 70 images of both Caltech Pedestrian Detection Benchmark [2] and our dataset are 50% and 77%, respectively. For our application, we consider the recall rate to be more important than precision because we aim at extracting as many ROI regions as possible from the image. For surveillance application, it is important to present to the security officer all ROI regions which are present in a given frame so that no important information is missed.

#### 3.2 Video surveillance use case

To measure the potential efficiency gain when applying ROI detection in video surveillance applications, we cascade our DCT-based ROI detection technique with an object detector [10], as depicted in Figure 1. Our object detector is trained using Haar-like features combined with Adaboost algorithms, while the cascaded detection framework is evaluated for a car detection task. To train the car detector, we have selected 100 images from the “cars 2001(Rear), Caltech” dataset [11] as training set. To test our detection framework, we have chosen 50 frames from our highway video sequence as the test set.

We consider the size of the car in the test set to be in the range of 20 × 20 pixels.
Table 1: Precision and recall rates for the ROI detection algorithms on Caltech Pedestrian Detection Benchmark [2].

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge-based</td>
<td>62</td>
<td>50</td>
</tr>
<tr>
<td>Connected component-based</td>
<td>60</td>
<td>57</td>
</tr>
<tr>
<td>Hough-based</td>
<td>61</td>
<td>80</td>
</tr>
<tr>
<td>DCT-based</td>
<td>61</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 2: Average computational time of the different methods per frame, calculated over 5,454 frames of Caltech Pedestrian Detection Benchmark [2].

<table>
<thead>
<tr>
<th>Time (milliseconds)</th>
<th>Edge-based</th>
<th>Connected component-based</th>
<th>DCT-based</th>
<th>Hough-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63</td>
<td>105</td>
<td>150</td>
<td>225</td>
</tr>
</tbody>
</table>

and 120 × 120 pixels. We use a sliding window with the size of 120 × 120 pixels to scan the image. The scanning step is 10 pixels in both vertical and horizontal directions. If the number of the ROI pixels in the image patch within the sliding window is below 50%, we expect no car inside the image patch, so that we do not apply the car detector. Otherwise, we apply the car detector for the image patch. We have compared the detection results of the new cascaded framework with the results obtained using only the Haar-like object detector.

On the average, for 50 frames of our test set and without using ROI, the car detector is applied to 9,744 image patches. When applying the detector only to ROI regions, only 943 image patches are checked by the detector. The proposed cascaded detection framework using DCT-based technique leads to applying the car detector, which is a computationally expensive algorithm, to only 9.7% of the total amount of image pixels. The average detection rates of car detection with and without using ROI are 80% and 90%, respectively. Figure 3(b) shows the result of the DCT-based approach with 8 × 8 block size on our highway video sequence. Figures 3(c) and (d) present the detected cars in that frame without and with the DCT-based ROI detector in cascade, respectively.

Figure 1: Framework of the cascaded object detection approach.

4 Conclusions and discussion

In this paper, we have presented our ongoing research on scene analysis, namely on employing generic ROIs for outdoor video surveillance as a preceding stage for further precise processing such as an object detector. Our hypothesis is that this will provide a significant gain in efficiency in terms of the number of pixels needed to be explored, compared to the conventional analysis of the complete scene. By already providing a generic informative ROI to a security officer without further in-depth analysis, the first-stage visual analysis of the scene is optimized significantly, so that the officer can
Figure 2: (a) sample frame of Caltech Pedestrian Detection Benchmark [2] video sequence, (b) created ground truth for this frame (black blocks are not ROI), (c) edge-based method, (d) connected component-based method, (e) Hough-based method and (f) DCT-based method.
immediately respond to alarming situations. It can also reduce both bandwidth needed for sending the information and memory needed for storage.

To this end, we have considered a number of simple, fast ROI detection techniques based on various features, such as the entropy of DCT coefficients, sum of the edge pixels, number of connected components and the number of straight lines found by the Hough transform. This means that the criterion for each technique is matched to the nature of that technique. We define empirical thresholds for each of technique after extensive experimentation to optimize the performance. We have used 10-frame intervals for temporal filtering of the frame-based ROI detections in order to prevent the flickering in the output due to small temporal changes and noise.

Our experiments show that the detection results of the ROI detection obtained with the DCT-based approach are similar to the results of the Hough-based technique. Both DCT- and Hough-based approaches provide better results compared to the other ROI detection techniques. Additionally, the DCT-based approach is well-suited for real-time image analysis because it allows parallel processing as the DCT can be computed for each image block independently and thus in parallel. Accordingly, we adopt the DCT-based ROI detection approach for validating its efficiency in a video surveillance use case. The major contribution of this study is developing an accurate ROI detection approach, while maintaining low computational complexity, so that it is suitable for real-time implementation in embedded video surveillance. Furthermore, high adaptivity to new scene types is an additional advantage of the proposed approach.

We have validated the efficiency of our hypothesis by cascading the proposed DCT-based ROI detection technique with an object detector to measure the potential efficiency gain. The experimental results show that such an ROI provides a significant gain in efficiency with respect to the explored number of pixels, as compared to analyzing the complete scene. Even though combining the usage of ROI extraction and the Haar-based object detector slightly decreases the precision, the efficiency is considerably increased, as only 9.7% of the image pixels is analyzed by the object detector. Concerning the future work, we are planning to research a new criteria for threshold selection for the DCT-based ROI detector. Currently the threshold is adaptively chosen.
for each block in which the DCT is performed. In order to better exploit the spatial context of the scene, we will add the information from the neighboring blocks to the algorithm of threshold selection. We expect that this will increase the precision of the ROI detection results.

Acknowledgement

We are grateful to Patrick Langenhuizen for his enthusiasm and hard work while developing the ROI detectors during his internship research.

References


In-plane User Positioning Indoors

Natasa Jovanovic  Tanir Ozcelebi  Johan Lukkien
Eindhoven University of Technology
P.O. Box 513 Eindhoven, The Netherlands
n.jovanovic@tue.nl  t.ozcelebi@tue.nl  j.j.lukkien@tue.nl

Abstract
Indoor positioning is a service required by many smart environment applications for various purposes, such as activity classification, indoor navigation and context awareness. In this paper, we present a novel approach to the user positioning problem based on in-plane detection enabled by a set of infrared light emitters and sensors placed horizontally along the walls. The simulation results show that the proposed system is able to determine locations of multiple users inside the room with high precision and accuracy.

1 Introduction
As an enabling service, user positioning is required by many applications ranging from navigation, surveillance and traffic control outdoors, to robot guidance, user tracking and activity recognition indoors. Outdoor positioning, tracking and navigation today are enabled almost exclusively by Global Positioning System (GPS), with every smartphone and car navigation system containing a GPS receiver. However, using GPS for indoor positioning is neither reliable nor accurate as the technology requires line-of-sight with multiple satellites.

To this date, there is no generally accepted indoor positioning system and a multitude of indoor positioning technologies, sensory devices and algorithms that are suitable for different applications have been developed [1]. Most of these utilize cameras [2, 3], passive infrared (PIR) sensors [4, 5] or radio frequency identification (RFID) tags [6, 7]. Each of the sensing techniques and the positioning methods implied by them have unique advantages and limitations. For instance, vision-based positioning systems are criticized mostly because they disturb privacy of people. PIR sensors are widely used because they are cheap, however, there are known issues with the coverage and hidden objects [8], as well as the issues of reporting false detections that are difficult to filter out [9]. The RFID-based positioning systems provide identification of separate users, however, they rely on users carrying a location device (tag) and are therefore considered obtrusive.

User positioning solutions in the literature treat users as point objects and aim to position them with respect to a two dimensional coordinate system. In reality, people occupy an area which may be symmetric, e.g., when they stand straight, or asymmetric, e.g., when they lean. Instead, we formally define the problem of indoor user positioning as follows. Given a regular three-dimensional indoor environment \( \Gamma \) with walls that are perpendicular to the floor and the ceiling, determine the area \( D \) that is the intersection of all users positioned inside \( \Gamma \) and a horizontal two-dimensional virtual detection plane \( A \) that cuts \( \Gamma \) at a height of \( h \). Suitable values of \( h \) can be different for different applications and indoor settings as explained in Section 2.

In this paper, we explore the possibilities of using in-plane object detection [10] in an indoor environment. The concept of in-plane object detection was developed for positioning and tracking of multiple objects on a two-dimensional rectangular surface. For this, infrared light emitters (LEDs) and sensors are placed at fixed and known positions along the circumference as shown in Figure 1. The LEDs and sensors are
positioned in an alternating fashion and the distance between any neighboring LED-sensor couple is constant. During a so-called detection cycle, the LEDs flash (infrared) light in turns and all sensors report whether or not they sense it each time a LED flashes. The sensor data collected in a detection cycle is saved in the form of a binary matrix where the $i$-th row corresponds to the $i$-th LED and the $j$-th column corresponds to the $j$-th sensor. If there is an object blocking the light between a LED and a sensor, the corresponding entry in the matrix is set to 0, otherwise the entry is set to 1. This binary matrix denoted as $B$ is called the blocking matrix and it is used as an input to the object positioning algorithm. In-plane object detection and positioning became known through implementation in interactive multi-touch screens such as Entertable [11] and Zero-touch [12], where small objects such as fingers and game pawns are repeatedly positioned, i.e., tracked in real time.

Differently from multi-touch screens, it is not possible to place LEDs and sensors at the entire circumference of an indoor environment, due to doors, windows, cabinets and other objects placed at the circumference. Furthermore, the performance is hampered by other objects detected such as furniture. Therefore, the in-plane detection based object positioning algorithm introduced in [10] cannot be directly applied to user positioning for indoors. In this paper, we introduce a user positioning algorithm that is also based on in-plane detection. We evaluate its performance by measuring precision and accuracy metrics in four different simulations of indoor environments. Note that repeated use of the proposed user positioning algorithm in combination with a user identification service can be utilized for a more sophisticated algorithm for tracking identified users in real-time. However, this is beyond the scope of this paper and is left as future work. The proposed algorithm simply aims to determine the area occupied by anonymous users at any instance of time.

2 Algorithm for user positioning

The proposed algorithm is based on the following realistic assumptions. The detection plane $A$ is at a height of $h$ from the floor and contains the infrared LEDs and sensors at the circumference. This height can be conveniently chosen during the hardware installation phase, such that the plane intersects the least possible number of furniture pieces, however, it always intersects the users regardless of their standing and sitting positions. For simplicity, we assume that LEDs and sensors are points on the perimeter.
of the detection plane. For practical reasons, no LEDs and sensors can be placed along
the doors, windows and room furniture.

Figure 2: The bird’s eye view of the hardware setup on the walls and the interior of an
office room: the grey shapes represent the pieces of room furniture that do not intersect
the detection plane (because they are either below or above the detection plane), while
the black shapes represent four users and room furniture that intersect the detection
plane.

The users block the light emitted by the LEDs, therefore, the area occupied by
the users in the detection plane cannot be intersected by the so-called detection lines
connecting individual LEDs and sensors; see Figure 1. The detection lines that are not
blocked (easily determined from the blocking matrix $B$) divide the detection plane into
a large number of convex polygons. The algorithm determines these polygons using a
recursive routine of cutting a polygon into two smaller polygons by one of the detection
lines intersecting that polygon [10]. After cutting the detection plane with all detection
lines, each polygon that is larger than threshold area $\sigma$ is reported as a position in the
output of the positioning algorithm; see Algorithm 1. We choose the value of $\sigma$ to
be equal to the minimum intersection size between a person and the detection plane.
Note that the algorithm determines all polygons larger than $\sigma$, therefore, if there are
other large enough objects besides users in the detection plane, these will be reported
as well. A polygon is represented by an ordered sequence of its vertices, which are
given by their two-dimensional coordinates in $A$. The polygon representation of the
user position represents an improvement over reporting a position in the $(x, y)$-form,
also because it allows the detection of the change in the orientation when the center of
the mass does not move, e.g. when turning around in place.

Additionally, to discard the polygons that are larger than $\sigma$ but may not circum-
scribe any users, e.g. a very long but thin polygon, it can be checked whether the
approximate shape of a user’s cross section with $A$ can be inscribed in each polygon.
This would, however, introduce high computational complexity to improve on identi-
ifying extremely rare occurrences of such polygons.

We consider a user as positioned if the positioning algorithm outputs a polygon
circumscribing the user in the detection plane and this instance is marked as a true
positive. Since the LED-sensor lines blocked by the users are not involved in the cutting
of the detection plane, each user by definition must be entirely contained inside a convex
polygon formed by the detection lines. This means that the described method cannot
miss any user, i.e., not report a user’s position, as long as the threshold size is set to be smaller than the smallest cross section (e.g., waist size) of each user. In other words, the algorithm does not give false negatives. However, a known intrinsic shortcoming of the in-plane detection is that it can result in false positives [13]. A polygon reported by the positioning algorithm is said to be a false positive if it does not circumscribe any users. The false positives occur in situations where all lines intersecting a large enough area are blocked by users (outside the area) and by other objects in the detection plane; see Figure 3.

Algorithm 1 UserPositioning($\mathcal{B}$)

Let $\mathcal{D}$ denote the set of positions in $\mathcal{A}$;
Determine the set $\mathcal{L}$ of all detection lines in $\mathcal{A}$ from the blocking matrix $\mathcal{B}$;

procedure DETECT($\mathcal{A}, \mathcal{L}, \sigma$)

Determine the set $\mathcal{L}_{\text{cut}}$ of lines in $\mathcal{L}$ that intersect $\mathcal{A}$;
if $\text{size}(\mathcal{L}_{\text{cut}}) = 0$ then
if $\text{size}(\mathcal{P}) > \sigma$ then
\[ \mathcal{D} \leftarrow \mathcal{A} \]
end if
return
end if
Choose a line $l$ from $\mathcal{L}_{\text{cut}}$;
Remove $l$ from $\mathcal{L}_{\text{cut}}$;
Cut polygon $\mathcal{A}$ with $l$ and denote the resulting polygons as $\mathcal{D}_1$ and $\mathcal{D}_2$;
if $\text{size}(\mathcal{D}_1) > \sigma$ then
DETECT($\mathcal{D}_1, \mathcal{L}_{\text{cut}}, \sigma$)
end if
if $\text{size}(\mathcal{D}_2) > \sigma$ then
DETECT($\mathcal{D}_2, \mathcal{L}_{\text{cut}}, \sigma$)
end if
end procedure

The performance of the algorithm is measured using precision and accuracy metrics. We define the precision as the ratio between the number $\text{pos}_{\text{true}}$ of true positives and the sum of the numbers $\text{pos}_{\text{true}}$ and $\text{pos}_{\text{false}}$ of the true positives and the false positives, respectively, i.e.,

$$\text{precision} = \frac{\text{pos}_{\text{true}}}{\text{pos}_{\text{true}} + \text{pos}_{\text{false}}} \quad (1)$$

In addition, accuracy is a metric that describes how well the reported positions correspond to the actual area that users occupy in the detection plane. More precisely, if the areas that $N$ users occupy are $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_N$ and the areas of each of the $K$ positions reported in $\mathcal{D}$ are $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$, the accuracy of the user positioning is given by

$$\text{accuracy} = \frac{\mathcal{U}_1 + \mathcal{U}_2 + \cdots + \mathcal{U}_n}{\mathcal{S}_1 + \mathcal{S}_2 + \cdots + \mathcal{S}_k} \quad (2)$$

While higher precision implies a lower rate of false positives, higher accuracy implies a tighter fit between the area determined by the positioning algorithm and the actual area of intersection with users. The described positioning algorithm relies on the presence of a large number of detection lines connecting the LEDs and sensors. Naturally, the performance would improve with respect to precision and accuracy with
increasing the number of detection lines, i.e., by increasing the number of LEDs and sensors that define them. However, regardless of the density of LEDs and sensors, it can be that multiple users are reported as a single polygon and some of the polygons that are reported may not circumscribe users (false positives). In other words, the number of users $N$ can be different from the number of positions reported $K$. Note that detecting multiple users in a single polygon instead of separate polygons affects accuracy, but it does not affect precision. This is because the area occupied by the users is still reported, although jointly within a single polygon.

Figure 3: Four users and four pieces of furniture result in five positions reported; four of these are true positives and one is a false positive.

3 Simulation results

In order to investigate the applicability of the in-plane detection method for user positioning and evaluate the proposed user positioning algorithm, we developed a simulation platform. The simulation platform incorporates three major components that define the simulation environment and that can be changed independently. These components are:

- hardware configuration, defining all parameters related to the LEDs and sensors, such as their count and relative positions;
- room layout, defining all objects that can be present in the room such as the furniture;
- user model, defining all parameters related to the users, i.e. the number of users, user cross section size, shape, position and orientation.

We measured the precision and the accuracy of the positioning algorithm as defined in Equations (1) and (2) in four environments, the result of combining two hardware configurations and two room layouts for an office space of length 500 cm and width 370 cm. The detailed description of these four environments is presented below.

**Environment 1.** This environment is created to explore the precision of user positioning under overly optimistic conditions. The LEDs and sensors are uniformly
placed along the entire circumference on the detection plane in an alternating fashion, including the walls, all windows and doors. There are 142 LEDs and 142 sensors in total, and the distance between a LED and a neighboring sensor is 6 cm. The room contains no furniture that intersects the detection plane, in other words, the detection plane is considered to be empty when there are no users in the room.

**Environment 2.** This environment has the same hardware configuration as Environment 1. However, to create a realistic model of an office, we place two desks, two chairs and two monitors in the room layout. We assume that the height of the detection plane is such that it does not intersect the desks. Hence, the desks cannot block the infrared light emitted by LEDs, however, their presence restricts the movement of users inside the room. The intersection of a chair and the detection plane, as well as the intersection of a monitor and the detection plane are modeled as a rectangle of size 35 cm by 6 cm and a rectangle of size 50 cm by 6 cm, respectively.

**Environment 3.** The hardware configuration of this environment assumes there are no LEDs and sensors placed along the two windows of width 100 cm and one door of the same width. The windows are assumed to be on the wall opposite to the wall containing the door. There are 118 LEDs and 118 sensors in total, and the distance between a LED and its neighboring sensor is 6 cm. The room is assumed to be empty, as it is in Environment 1.

**Environment 4.** This environment represents a realistic model of a small office, with the hardware configuration as in Environment 3 and the room layout as in Environment 2.

For simplicity, a user’s cross section with the detection plane is modeled to have the shape of the convex hull of two disks of radii 10 cm that are tangent to each other from the outside. For each separate case of 1, 2, 3 or 4 users present in one of the four defined environments we measured the precision and the accuracy of the in-plane positioning over 1000 tests, where each test corresponds to a new random set of positions that users occupy. The results are presented in Figure 4 and Table 1.

![Figure 4: The average precision (left) and accuracy (right) of in-plane detection determined for 1000 different user locations.](image)

Table 1: The average precision and accuracy of in-plane detection determined for 1000 different user locations.

<table>
<thead>
<tr>
<th>#Users</th>
<th>Environment 1</th>
<th>Environment 2</th>
<th>Environment 3</th>
<th>Environment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>93.1</td>
<td>99.1</td>
<td>86.9</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>92.5</td>
<td>98.8</td>
<td>84.9</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>91.9</td>
<td>98.6</td>
<td>84.3</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>91.1</td>
<td>97.9</td>
<td>82.3</td>
</tr>
</tbody>
</table>
With the minimum precision of 83.7% and the minimum accuracy of 62.2%, the in-plane positioning method shows very good performance in multiple users positioning in a relatively small office space. Figure 5 shows positioning instances and the corresponding (individual) accuracies.

4 Conclusion and future work

We proposed an algorithm that utilizes in-plane object detection for user positioning in indoor environments. This anonymous and unobtrusive technology is enabled by a set of infrared LEDs and sensors placed in an alternating fashion on the walls of the given indoor space. The proposed algorithm reports the positions in the form of convex polygons circumscribing users within a horizontal detection plane. Using this algorithm, we simulated user positioning in four environments and measured the precision and the accuracy of the positioning method for up to four users.

The precision of 100% and the accuracy of more than 83% in Environments 1 and 3 clearly indicate the direction that should be taken to improve the positioning precision in reality. The relatively large gaps in the LED-sensor frame practically do not affect the positioning method. In contrast, large obstructing objects in detection plane, as the ones in Environments 2 and 4, can cause false positives, usually those same objects being reported as user positions. Therefore, in order to minimize the risk of false positives, the height of the detection plane should be chosen such that it has minimum intersection with objects in the room. In addition, in future work, false positives may be identified by comparing positions reported in consecutive detection cycles and removing those that appear and disappear in the middle of the detection plane causing a kind of blinking polygons effect. Alternatively, multiple detection planes on different heights can be deployed to eliminate the false positives by combining results and to ensure the highest precision and accuracy of user positioning.

References


Abstract

We address the content-identification problem by modeling it as a multi-class classification problem. The goal is to pave the way and establish a general framework to incorporate the powerful algorithms of the machine learning literature in learning from data into this problem. Through this end, a particular successful approach, linked with the coding theory known as ECOC is considered and studied. We argue that the conventional codings used in this approach are suboptimal by analyzing the problem from an information-theoretic viewpoint. We then advise the use of our recently proposed method for this problem. The ECOC approach converts the multi-class problem to several binary problems. We consider these equivalent binary classification tasks in more details and use the Gaussian Mixture Models instead of SVM’s. This latter brings significant reduction in complexity by having an assumption on the distributions.

1 Introduction

Content identification is an active field of research where different methods of robust hashing a.k.a. content fingerprinting, hypothesis testing and computer vision are involved. Conventional approaches towards content identification assume very strict distributions for the data, e.g. they assume that the items in the database and queries corresponding to them both follow a Gaussian distribution. In practice, however, due to different acquisition conditions like the inevitable rotations and posture differences in each acquisition, these assumptions turn out to be oversimplifying. Thus, to cope with these situations, alternative approaches should be utilized.

The power behind machine learning algorithms in learning from data and their generalization capabilities guaranteed by the rigorous statistical learning theory justifies the use of these techniques in content identification applications. Through this end, the problem of content identification was proposed to be considered as a multi-class classification where different acquisitions of each item in the database are regarded as training instances and the query items are the test data in the problem of classification.

In spite of extensive research in the field, multi-class classification is still considered as an open issue in the domain of Machine Learning. One way to address this problem is the so called Error Correcting Output Codes (ECOC) approach that converts a multi-class problem into a set of binary classification problems [1].

In the ECOC approach, the learning problem is considered as a hypothetical communication channel and the imperfections or ambiguity in the learning process are considered as an equivalent channel noise. Therefore, it was argued that, in order to combat noise, one should encode the data before sending them to the channel and thus, using the channel codes. In some recent studies, the authors attempt at straightforward usage of advanced error correction codes like LDPC to achieve the accurate classification [2].
The attractiveness of the ECOC approach comes from the competitive number of binary classifiers needed in comparison to the one-vs-all and one-vs-one strategies. In the limiting case, the number \( l \) of binary classifiers needed in the ECOC case or equivalently the length of the channel code is of \( \mathcal{O}(\log_2(M)) \), where \( M \) is the number of classes. In this sense, the classifiers can be considered as a binary tree that represents the lowest complexity for \( M \)-class classification. Using communication setup terminology, it means that the decoding of random codewords of length \( N \) in the codebook \( M \sim 2^{NC} \) can be accomplished in \( \mathcal{O}(NC) \) checks, versus the typical brute force decoding which is of order \( \mathcal{O}(2^{NC}) \), where \( C \) denotes the channel capacity.

In a previous work [3], we investigated the learning problem based on the ECOC used as a method of content identification. In this information-theoretic viewpoint which is linked with the coding theory, a very important issue is the choice of the coding matrix. There we proposed a novel coding matrix to be used as an alternative to the existing algorithms based on random codes and error correction codewords.

In this work, we further consider the coding approach towards multi-class classification. We analyze the information-theoretic setup for the learning problem in section 2 and argue that due to several factors, including the statistics of channel noise, conventional channel codes cannot be used efficiently and do not have their expected error correcting capability. We then review a proposed coding matrix design which was indicated to have optimal performance. The choice of the underlying binary classifiers used is investigated in section 3. In order to reduce the complexity of binary classifications, instead of the Support Vector Machines (SVM’s) with nonlinear kernels, we classify the training data by fitting a Gaussian Mixture Model (GMM) to each class. We finally conclude the paper in section 4.

2 Multi-class Classification: A Coding Based Approach

In this section we consider the problem of multi-class classification and the coding approach to solve it. The general problem formulation is presented in section 2.1 and the coding approach is presented in section 2.2. We analyze the assumptions of this model in section 2.3.

2.1 Problem Formulation

A set of training instances are given which contain features \( \mathbf{x}(i) \in \mathbb{R}^N \) and their labels \( g(\mathbf{x}(i)) \)'s belonging to the \( M \) classes. These labels are assumed to be generated through an unknown hypothetical mapping function \( g : \mathbb{R}^N \mapsto \{1, 2, ..., M\} \) that we want to learn and approximate based on the given labeled training examples and using pre-selected approximation functions. The labels, unlike the common cases in machine learning, belong to a set of \( M \geq 2 \) members rather than only two classes. Because most of the existing classification algorithms are naturally designed for \( M = 2 \) classes, to generalize them for multi-class cases usually requires to consider the problem as several binary classification problems.

2.2 ECOC Approach Towards Multi-class Classification

The main idea behind the ECOC approach to solve the multi-class problem is to consider it as a communication system where the identity of the correct output class for a given unlabeled example is being transmitted over a hypothetical channel which, due to the imperfections of the training data, the errors in the learning process and non-ideal choice of features is considered to be noisy [1]. Therefore, it is reasonable to
encode the classes using error correcting codes and transmit each of the bits through the channel, i.e., to run the learning algorithm, so that we would be able to cope with the errors in each individual binary classifier. Figure 1 illustrates this idea.

![Diagram](image)

Figure 1: Communication system and its classification equivalent: \( E \) and \( D \) are the Encoding and Decoding stages in communication, respectively. \( T \) is the training procedure, \( C \) is the coding matrix, \( b_y \) is the derived binary code for the test example.

Concretely, we assign randomly to each of the \( M \) classes a row of a coding matrix \( C_{(M \times l)} \). Then we run a binary learning algorithm on all the training samples for every column of \( C \) so that we will have \( l \) mappings from \( \mathbb{R}^N \), the original data space to the one dimensional binary space \( \{0, 1\} \), or equivalently, one mapping rule from \( \mathbb{R}^N \) to the \( l \)-dimensional binary space \( \{0, 1\}^l \). It is impotent to note that the \( l \) learning procedures are essentially binary and symmetric, i.e., the number of instances in each binary class is approximately the same.

Given a new unlabeled example \( y \), we map it from \( \mathbb{R}^N \) to the \( l \)-dimensional binary space through the same mapping rule learned as above. We then compare this binary representation \( b_y \) with the items in the database, or equivalently the rows of \( C \) by minimum Hamming distance rule or any other relevant decoding method.

### 2.3 Experimental Assessment of the Channel Code Model

Although we use the ECOC approach to address the multi-class classification problem, in this section, however, we question this model by arguing that the mismatch in the real and assumed statistics of the channel noise can not guarantee the expected error correcting capability of channel codes.

In particular, we experimentally demonstrate that the samples of learning noise are highly correlated. This is not in accordance with the assumptions behind the binary symmetric channel model used in the design of the common channel codes like the BCH, LDPC or Turbo codes.

In this part we measure the correlation properties of the channel noise of the multi-class learning problem through a simple synthetic simulation. We use a synthesized dataset of \( M \) classes of Gaussian data.

The centers of classes \( x_c(i) \in \mathbb{R}^N \), with \( 1 \leq i \leq M \), are generated as realizations \( X_c \sim \mathcal{N}(0, \sigma_{\text{inter}}^2 I_N) \). The \( j^{th} \) instance of \( i^{th} \) class is generated as \( x_j(i) = x_c(i) + Z_{\text{intra}} \) with \( Z_{\text{intra}} \sim \mathcal{N}(0, \sigma_{\text{intra}}^2 I_N) \). The test data are generated as Additive White Gaussian Noise with covariance matrix \( \sigma_Z^2 I_N \) added to the centres of each class where \( \sigma_Z^2 \) is controlled by \( SNR \).

As was explained in 2.2, the rows of the binary coding matrix denoted as \( b_X(k) \) are fed to the learning channel to produce the \( b_Y \)'s. Thus, we consider the \( k^{th} \) sample of the learning channel noise as

\[
b_Z(k) = b_X(k) \oplus b_Y(k),
\]
with $1 \leq k \leq l$. 

For each of the $M$ classes, we generated 100 test data for different values of SNR. The estimator of the autocorrelation of each noise string is considered as:

$$
\hat{R}_Z(j) = \left\{ \begin{array}{ll}
\frac{1}{l} \sum_{i=1}^{l-j} b_Z(i+j) b_Z(i) & j \geq 0 \\
\hat{R}_Z(-j) & j < 0
\end{array} \right.
$$

with $-l \leq j \leq l$. We then average the $\hat{R}_Z(j)$’s for all the noise realizations.

Figure 2 illustrates the resulting autocorrelation sequences of channel noises. We had $M = 100$ classes and used randomly generated binary values as the elements of the coding matrix $C$. There were 10 instances in each class and 100 testing data per class. The results are compared with the channel noise for the fingerprinting approach in content identification where it is considered that the equivalent Binary Symmetric Channel of identification produces i.i.d. noise. In order to make a meaningful comparison, we evaluated the autocorrelations for each method for results that correspond to the same accuracy for the learning method and probability of correct identification for the identification method ($SNR = -2.14db$ for learning and $SNR = 2.94db$ for fingerprinting). For the fingerprinting method, we considered the center of each class corresponding to the representative of each item.

According to Figure 2, while in this experimental setup, the channel noise for fingerprinting remains i.i.d., the equivalent channel noise for learning problem shows high amounts of correlation in the autocorrelation sequences and thus, far from the i.i.d. assumption.

Figure 2: Autocorrelation of noise, learning channel vs. identification channel

### 2.4 A Novel Approach Towards Coding Matrix Design

When used as the method of choice for binary classification, Support Vector Machine (SVM) with properly tuned kernels draws a hyperplane in a space of higher dimension than the original data space of $\mathbb{R}^N$ where the transformed data are linearly separable, such that this hyperplane has the maximum margin from the boundary training instances of each class (support vectors). As shown in Figure 3, the hyperplane drawn in the higher dimensional space, when considered in the original domain is reflected as complex decision boundaries approximating the Voronoi cells of the support vectors.

Assuming the use of kerneled SVM as the base binary classifier, the design of a good coding matrix should imply efficient learning of the Voronoi cells of the support vectors.
vectors which means learning at least once these cells and at the same time avoiding redundant learning of them. The first requirement links with having good performance via optimal class separability and the latter implies maintaining the number of columns (classifiers) as small as possible, i.e. \( l = \lceil \log_2 M \rceil \) or equivalently, having maximal rate of communication. This optimal design, in fact happens when the rows of \( C \) are chosen as the modulo-2 equivalent of the the class numbers. Therefore, as an example, in a 20-class problem, the codewords will be of length \( l = \lceil \log_2 M \rceil = 5 \), the 1st class should be encoded as the binary string '00000', and the 6th class should be encoded as '00101'.

In fact learning the Voronoi regions of the support vectors is also equivalent to the maximum likelihood rule under Gaussian assumption for the classes. While ML decoding rule requires the knowledge of Voronoi regions for all classes, in our approach, these regions are distributed among the \( l \) binary classifiers and are learned implicitly. The fusion of the results of all binary classifiers, equivalently produces the entire coverage for all classes defined by their Voronoi regions. Therefore, the results of these fused binary decodings should coincide with the optimal ML decoder.

For more detailed explanations on this approach, as well as experimental studies suggesting its efficiency, the reader is refered to [3] and [4].

It is also very important to mention that given a test example to be classified, unlike any other method, the complexity of decoding in this approach comprises only the SVM functional evaluations and does not incur any Hamming distance computation or decoding computations as in LDPC or other coding methods. The reason is due to the fact that a produced codeword is directly referring to a storage point in memory corresponding to the most likely class. However, the SVM functional evaluations is involved and could be very high, if the nonlinear kernels are too complex. In the next section we consider this fact in more details.

### 3 Choice of Binary Classifiers

After converting the multi-class problem to \( l = \lceil \log_2 M \rceil \) binary classification problems, the important issue is the choice of these binary classifiers which is influenced by the specifications of the application. The computational budget of both training and
testing (or equivalently encoding and decoding) stages, the memory storage issues and the amount of prior information we might have about the distribution of the data are among the factors that should be considered in this respect.

The richness behind Vapnik-Chervonenkis learning theory and generalization guarantees it provides, and also the efficiency of the practical implementations of the algorithm makes the SVM the method of choice in many binary classification problems. In [3] and [4], SVM with Gaussian kernels was used and shown to provide good performance.

However, the complexity of both training and testing phases could be intractable in many practical applications. The complexity of training for kernelled binary SVM is between quadratic and cubic of the whole number of instances and the complexity of testing is quadratic to the number of classes because in the current setup, all classes become support vectors. This complexity is due to the fact that the number of free parameters of the algorithm is big, which as was shown in [4], implies that the algorithm might overtrain the data. Moreover, the standard implementation of SVM does not have any assumption on the data distribution, while in identification applications, one might have some prior knowledge of the distribution. Therefore, due to these factors, one might think of using alternative approaches. We consider next the use of Gaussian Mixture Models for the binary classification.

3.1 Gaussian Mixture Model for Classification

After converting the multi-class problem to binary problems using the proposed approach, half of the classes are randomly labeled as ones and the other half as zeros. Therefore, equivalently, in each binary classification there are two classes with very complicated distributions each. Since every distribution can be approximated using a mixture of Gaussians, we fit a GMM to each class from the training set $\mathcal{X}$ and the binary labels $C_j(i)$’s with $1 \leq j \leq l$ and $1 \leq i \leq M$ where $C_j(i)$ is the $i^{th}$ element of the $j^{th}$ column of the coding matrix:

$$p_0(x|\mathcal{X}) = \sum_{i: C_j(i) = 0} \pi_i \phi(x; \mu_i, \Sigma_i),$$

$$p_1(x|\mathcal{X}) = \sum_{i: C_j(i) = 1} \pi_i \phi(x; \mu_i, \Sigma_i),$$

$\phi(x; \mu, \Sigma)$ is a multivariate Gaussian kernel characterised by its mean and covariance, respectively and the $\pi_i$’s are the mixing coefficients.

The simplest situation is the case where each of the original multi-labeled classes is modeled with one Gaussian kernel, so that we will have a total of $M/2$ kernels for each of $p_0$ and $p_1$. The unknown parameters are derived using the Expectation-Maximization algorithm. Here we could incorporate any knowledge about the mean and covariance of classes into the algorithm resulting in its very fast convergence and thus small complexity of training.

Given a test data $y$, we classify it by evaluating the value of each of the two PDFs and finding the one which is maximal. The complexity of testing an example in this case will be linear in $M$ which is much more reasonable than the kerneled SVM.

3.2 Experimental Comparison

Having reduced the multi-class problem to several binary problems using the proposed coding, we compare in this part the performance of the binary SVM classifier with the fitted GMM. The experimental setup is the same as in section 2.3, but here instead
of multi-labeled classes, we assign randomly to each of the generated Gaussian clouds zeros and ones to represent one of the equivalent binary problems.

We once train the data with a kerneled SVM, then we fit one GMM for each of the zero-labeled and one-labeled set of data with the standard Expectation-Maximization algorithm. As initial values of the algorithm, we use the empirical mean of the data points for each cloud. For dimension $N = 10$, two different numbers of instances per cluster, $\sigma_{\text{inter}}^2 = 10$ and $\sigma_{\text{intra}}^2 = 1$ and varied number of total clusters, the training/fitting time for SVM/GMM are sketched in Figure 4(a).

![Figure 4](image)

(a) training/fitting time

(b) testing

Figure 4: Time complexity of algorithms versus the number of clouds

For the test data, again we add Gaussian clouds with covariance $\sigma^2 I_N$ to the class centers. We define the signal to noise ratio as $SNR = 10 \log_{10} \frac{\sigma_{\text{inter}}^2}{\sigma_{\text{intra}}^2}$. For 100 test data per cluster generated with $SNR = 5db$ and varied number of clouds, we compare the testing time for each of the methods in Figure 4(b). The performance of the methods in the same setup are measured by their accuracy as plotted in Figure 5. Due to the specifications of the current setup, the ground truth here is considered to be Euclidean distance comparison with the a priori known cloud centers.

![Figure 5](image)

Figure 5: Accuracy versus cloud numbers
As is clear from the figures, both training and testing time complexities of the GMM are significantly lower than SVM which is polynomial in the number of classes. It is also important to point out that the good performance of GMM is due to the current setup where every original multi-label class is Gaussian. However, for more complicated scenarios, one could model each of the original classes as a mixture of Gaussians instead of one cloud.

4 Summary

The communication system setup for the ECOC approach in multi-class classification was examined to show why the conventional codes are working below expectation. A proposed coding approach was considered which due to the efficiency of the space partitioning and the homogeneity of the equivalent binary problems has an optimal performance. Apart from the coding, the choice of the base binary classifiers was considered in terms of complexity and the assumption on distributions. We showed that the GMM could be considered as an alternative for SVM, especially in identification scenarios where one is concerned with complexity and may have some prior knowledge on the data.

Acknowledgments

The work presented in this paper was supported by a grant from Switzerland through the Swiss Contribution to the enlarged European Union (PSPB-125/2010).

References


Information-theoretic measures for angiogenesis imaging by contrast-enhanced ultrasound

Nabil Bouhouch¹,²  Libertario Demi¹
Maarten P.J. Kuenen¹,²  Hessel Wijkstra²,¹
Tjalling J. Tjalkens¹  Massimo Mischi¹
¹Eindhoven university of technology
Electrical engineering department, signal processing systems group, biomedical diagnostic lab.
Den Dolch 2, 5612 AZ Eindhoven, the Netherlands
²Academic Medical Center, University Hospital of Amsterdam
Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands
N.Bouhouch@student.tue.nl {Initial(s). last name}@tue.nl

Abstract

The fundamental role of angiogenesis in cancer growth has stimulated extensive research aimed at non-invasive cancer detection by blood perfusion imaging. However, the results are still limited due to the complex relationship between angiogenesis and blood perfusion. Recently, contrast ultrasound dispersion imaging (CUDI) was proposed as an alternative for angiogenesis imaging. Determined by the geometrical properties of the microvascular architecture, dispersion may in fact be more suitable for detecting angiogenic processes than perfusion. Two different approaches have been proposed so far to perform CUDI by analysis of ultrasound-contrast-agent (UCA) time intensity curves (TICs). The first approach estimates dispersion by fitting each TIC by a convective-dispersion model in the time domain. The second approach is based on an indirect estimation of dispersion, achieved by means of spatiotemporal analysis of the measured TICs by linear similarity measures such as correlation or spectral coherence. Motivated by the promising results of the second method, this paper presents a novel dispersion estimation method based on spatiotemporal analysis of the mutual information among TICs measured at neighboring pixels. A preliminary validation was performed for prostate cancer imaging by comparison with histology results at a pixel level. 26 datasets obtained from 15 patients referred for radical prostatectomy were analyzed. This validation resulted in a receiver operating characteristic curve area equal to 0.924.

1 Introduction

This study focuses on one of the most common forms of cancer worldwide, namely prostate cancer (PCa). In the United States PCa accounts for 27% and 10% of all cancer diagnoses and deaths in men, respectively [1]. To date, PCa is still diagnosed using systematic biopsies, which consists of taking a dozen of geometrically distributed specimens from the prostate using a core needle. Systematic biopsy is invasive and has a low sensitivity; moreover, it also carries a risk of infection and sepsis [2]. Due to the inability of the current diagnosis to accurately localize and stage PCa, this malignancy is often treated by radical prostatectomy, which exposes the patients to severe side effects such as incontinence and impotence [3]. PCa diagnosis and treatment can be improved by an imaging technique able to detect and localize cancer at an early stage.
However, the reliability of imaging techniques relies on the reliability of prognostic markers.

The key role of angiogenesis in cancer growth has motivated extensive research aiming at non-invasive cancer detection by blood perfusion imaging methods [1], e.g. using dynamic contrast-enhanced ultrasound (DCE-US). However, the results are limited due to the complex relationship between angiogenesis and blood perfusion.

Recently, contrast ultrasound dispersion imaging (CUDI) has been proposed as an alternative method for imaging angiogenesis. Determined by the geometrical properties of the microvascular architecture [4], dispersion may in fact be more suitable for detecting angiogenic processes than perfusion.

Two different approaches have been proposed so far to perform CUDI. The first approach estimates dispersion by fitting each TIC with a convective-dispersion model in time domain [5]. The second approach is based on an indirect estimation of dispersion, achieved by means of spatiotemporal analysis of the measured TICs by linear similarity measures such as spectral coherence and correlation [6, 7].

Motivated by the promising performances of the spatiotemporal analysis approach, this paper presents an extension of this approach by inclusion of nonlinear similarity among TICs obtained at neighboring pixels.

Conditional entropy and mutual information are information theoretic-based tools that are able to assess functional relationships in general (linear and non-linear). In particular, mutual information quantifies the degree of dependence between two random variables [8, 9]. Moreover, as opposed to linear similarity measures, mutual information also allows the exploration of nonlinear connectivity between TICs. Performing mutual information analysis at each pixel permits building parametric dispersion maps based on linear and nonlinear similarities.

A preliminary validation was performed for PCa imaging on 26 datasets from 15 patients referred for radical prostatectomy at the Academic Medical Center (AMC), University Hospital of Amsterdam. The generated parametric maps were validated by comparison with the histology results, considered as the ground truth. The results were then compared with those obtained by different TICs parameters described in the literature.

2 Methodology

After an intravenous injection of a 2.4-ml SonoVue (Bracco, Milan, Italy) bolus, its passage through one plane in the prostate is imaged using an iU22 ultrasound scanner (Philips Healthcare, Bothell, WA) equipped with either a C8-4v or C10-3v probe. Power modulation imaging was adopted at a frequency of 3.5 MHz and a mechanical index of 0.06. Local dispersion is then estimated by means of mutual information analysis applied to TICs as obtained from neighboring pixels.

A kernel is designed that determines the pixels that are considered to estimate the local mutual information at a particular location. A 3x3 mm kernel size is adopted in order to achieve an optimal balance between diagnostic resolution and reliability. In fact, angiogenesis is required for tumors to grow beyond 2-3 mm in diameter. Therefore, to detect early angiogenic processes, a resolution of at least 3 mm should be achieved. A square shape was chosen to estimate mutual information independently of the blood perfusion direction.

The discrete mutual information, here indicated by the symbol $I$, as defined by Shannon [8], can be written as shown in Eqs. 1-3.

The random variable $\Theta(t)$ represents the TIC, expressed in gray-level video density, obtained at the kernels central pixel. The TICs of all other pixels, except the central one, are defined as the random vector $\Psi^{N-1}(t)$. $N$ is the total number of pixels in the kernel. $\Gamma$ is the set of pixel values in gray level ($\Gamma = \{0,1,...,255\}$).
\[
I(\Psi^{N-1}, \Theta) = \sum_{\psi \in \Gamma} \sum_{\theta \in \Gamma} P_{\Psi^{N-1}}(\psi, \theta) \log_2 \left( \frac{P_{\Psi^{N-1}, \Theta}(\psi, \theta)}{P_{\Psi^{N-1}}(\psi)P_{\Theta}(\theta)} \right)
\] (1)

\[
P_{\Psi^{N-1}}(\psi) = \sum_{\theta \in \Gamma} P_{\Psi^{N-1}, \Theta}(\psi, \theta)
\] (2)

\[
P_{\Theta}(\theta) = \sum_{\psi \in \Gamma} P_{\Psi^{N-1}, \Theta}(\psi, \theta)
\] (3)

From these equations, it can be noticed that the mutual information, \( I(\Psi^{N-1}, \Theta) \), is a function of the joint distribution of the two random variables \( \Psi^{N-1} \) and \( \Theta \). This is given by the joint probability mass function (PMF), \( P_{\Psi^{N-1}, \Theta}(\psi, \theta) \), which can be expressed as

\[
P_{\Psi^{N-1}, \Theta}(\psi, \theta) = P_{\Theta}(\theta) \prod_{i=1}^{N-1} P_{\Psi_i|\Theta}(\psi_i|\theta),
\] (4)

where \( P_{\Psi_i|\Theta}(\psi_i|\theta) \) is the conditional PMF of the TIC obtained at the kernels \( i^{th} \) pixel given \( \Theta \), the TIC obtained at the kernels center. In Eq. 4, the TICs in \( \Psi^{N-1} \) are assumed to be mutually independent and only dependent on \( \Theta \). This assumption, which may not be realistic, is made to focus the analysis on the local dependency between the central pixel and its surroundings. Moreover, this assumption reduces the complexity for computing the joint PMF of \( \Psi^{N-1} \) and \( \Theta \).

Conditional entropy \( H(\Psi^{N-1}|\Theta) \), as defined in Eq. 5, was also computed to compare its performance in dispersion estimation with the mutual information performance.

\[
H(\Psi^{N-1}|\Theta) = -\sum_{\psi \in \Gamma} \sum_{\theta \in \Gamma} P_{\Psi^{N-1}, \Theta}(\psi, \theta) \log_2 \left( P_{\Psi^{N-1}, \Theta}(\psi, \theta) \right)
\] (5)

Before starting the analysis, the ultrasound data was filtered temporally and spatially to improve the SNR of measured TICs. Temporally, a low-pass filter with cut-off frequency at 0.5 Hz was adopted [5]. Spatially, a Gaussian low-pass filter with standard deviation of 0.5 mm is used to increase the SNR while maintaining sufficient resolution for angiogenesis detection. Given that the original pixel size is 3 times smaller than the spatial resolution, low-pass filtering enables spatial downsampling of each frame by factor 3, which significantly reduces the computation time.

The mutual information analysis is then performed. First, the kernel center is placed at the pixel where the mutual information must be computed. Then, the kernel TICs are mapped to the probabilistic domain using Eq. 4. Mutual information and conditional entropy are then calculated and their values stored at the central pixels coordinates. By applying this procedure for all pixels, a parametric map is generated for each imaging plane.

This method was validated by comparing mutual information maps to histology results in 26 datasets recorded from 15 patients that were referred to radical prostatectomy at the Academic Medical Center, University of Amsterdam, the Netherlands. After cutting the prostate in 4-mm thick slices, a pathologist marked the presence of PCa [6]. Figure 1 shows an example of dispersion maps obtained by mutual information and conditional entropy analysis, respectively, and their corresponding histology slice. Two regions of interest (ROIs) larger than 0.5 cm², representing healthy and cancerous tissue, were determined on the basis of the histology results and applied on the ultrasound mutual information maps, as shown in Figure 2, to evaluate the performance of the proposed method for PCa localization.
3 Results

Pixel classification using mutual information resulted in a sensitivity and specificity equal to 81% and 87%, respectively. The receiver operating characteristic (ROC) curve area was 0.924. The conditional entropy method resulted in an ROC curve area of 0.79, a sensitivity of 77%, and a specificity of 70%. Previous CUDI methods based on temporal correlation and spectral coherence resulted in a sensitivity of 79% and 76%, in a specificity of 80% and 79%, and in an ROC curve area equal to 0.87 and 0.86, respectively. Figure 1 shows an example of the dispersion maps obtained by the mutual information analysis and its corresponding histology slice.

![Image](image1.png)

Figure 1: Parametric dispersion maps based on (a) mutual information and (b) conditional entropy. (c) Corresponding histology slice with prostate cancer marked with red.

4 Conclusion

The estimation of ultrasound-contrast-agent dispersion by mutual information analysis among neighboring time-intensity curves has shown promising preliminary results for prostate cancer localization, with a ROC curve area of 0.924. Parametric maps as obtained by mutual information show improved performances as compared to conditional entropy alone. Further validation using a larger dataset is required to confirm the clinical value of this method.

![Image](image2.png)

Figure 2: (a) Ultrasound image with superimposed prostate contour (yellow), cancerous (red) and healthy (green) regions as utilized for the analysis on the performances of the method. (b) Corresponding histology slice.
References


Maximum-likelihood feature model selection∗

Amir Jalalirad Tjalling Tjalkens
Eindhoven University of Technology
Dept. EE, Signal Processing Systems Group
P.O.Box 513, 5600 MB Eindhoven, The Netherlands
a.jalalirad@tue.nl t.j.tjalkens@tue.nl

Abstract

In this paper we extend our previous work on efficient model estimation for feature based models to include the possibility to eliminate those features that do not contribute (enough) to the classification decision. This work can be seen as an extension to the naive Bayes model with greedy feature selection. We limit ourselves to the assumption that there exist some natural ordering in the features so that dependent features are consecutive. This assumption is not needed, an algorithm can, and in our previous work, be designed to handle the more general case. However the amount of work in the more general case is still too much to be practical. Here we show a method that, given a set of $N$ observed objects, containing feature values and a classification, produces the penalized maximum-likelihood model with labeling of irrelevant and redundant features in $O(Nk^3)$ time, where $k$ is the number of features.

1 Introduction

The feature based classification problem considers a generative probability distribution $P(C, F^k)$. Here $C$ denotes the categorical class and $F^k$ the vector of categorical features of an object $O$, so we often write $O = (C, F^k)$. A sequence of $N$ objects $O^N$ is drawn independently from this distribution. In this paper we only consider binary classes and binary features. These models are widely used in spam e-mail detection [1].

In a supervised learning setting our goal is to find, implicitly or explicitly, an estimate of $P(C, F^k)$ from a given set of completely specified objects $O^N$, the training set. Usually, the size $k$ of the feature vector is quite large, making it practically impossible to find a decent estimate of $P(C, F^k)$ without additional assumptions on its structure. Two practical approaches are often used and combined: feature selection and the naive Bayes model [2]. In the feature selection approach one tries to find a suitable small subset of features that encompass most or all of the discriminatory power of the features. The naive Bayes model assumes that the features are conditionally (on the class) independent, so $P(C, F^k) = P(C) \prod_{i=1,k} P(F_i|C)$.

Feature selection methods often work in a greedy way, adding to or removing from the set of wanted features in a single best choice manner, cf. [3]. This approach is not guaranteed to give the optimal result, though in practice the performance is quite satisfactory. The naive Bayes assumption of independence is likewise unjustified but appears to work well for many practical situations.

In this paper we intend to explore a more fundamental approach where the model constraints are not as severe as in the naive Bayes model and feature selection is done in an optimal way (not greedy). The approach, penalized maximum likelihood, is in the flavor of the Minimum Description Length (MDL) principle, [4]. As such it considers the feature and model selection problem as a single optimization. In section

∗This work is in part supported by ENIAC Joint Undertaking, grant 270707-2, EnLight.
we introduce some special notation and the feature model. Section 3 introduces the concept of irrelevant and redundant features and section 4 shows how this can be used in the classification problem. The derivation of an estimated probability based on a sequence of objects is the topic of section 5. Section 6 discusses how the various models are compared efficiently using recursive calculations based on a factor graph like approach. In section 7 we summarize the results.

2 Notation and model structure

In this paper we will discuss feature vectors and sequences of feature vectors. We use the following notations.

$\mathbf{F}$: A vector of features, the length of the vector is implicit and assumed to be $k$.

$F_i$: The $i$th element of a vector.

$\mathbf{F}^k$: A vector of $k$ elements,

$$\mathbf{F}^k := (F_1, F_2, \ldots, F_k).$$

$\mathbf{F}_i^j$: A subvector of features,

$$\mathbf{F}_i^j := (F_i, F_{i+1}, \ldots, F_j), \quad 1 \leq i \leq j \leq k.$$

$\mathbf{F}$: A sequence of feature vectors, the sequence length is assumed to be $N$.

$\mathbf{F}_i$: The $i$th feature vector of a sequence of feature vectors.

$\mathbf{F}^N$: A vector of $N$ feature vectors,

$$\mathbf{F}^N := (\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_N).$$

As an extension to the naive Bayes model we assume that $P(C, \mathbf{F}^k)$ can be split into independent subsets of features, e.g. let $s^g$ be a sequence of $g$ increasing indices from 1 to $k$ inclusive, so $s^g = s_1, s_2, \ldots, s_g, \quad 1 = s_1 < s_2 < \cdots < s_g \leq k$. We implicitly extend $s^g$ with $s_{g+1} = k + 1$ and then write $P(C, \mathbf{F}^k) = P(C) \prod_{i=1, g} P(\mathbf{F}_{s_i}^{s_{i+1}-1} | C)$. If the index sequence $s^g$ is understood, we abbreviate $\mathbf{F}_{s_i}^{s_{i+1}-1}$ by $\mathbf{F}^{[i]}$. This non standard notation will be used in more situations so we will refer to it as equation 1

$$\mathbf{F}^{[i]} := \mathbf{F}_{s_i}^{s_{i+1}-1}. \quad (1)$$

We extend this notion to the sequence of feature vectors and can write

$$\mathbf{F}^{[i]} := \left( \mathbf{F}_1^{[i]}, \mathbf{F}_2^{[i]}, \ldots, \mathbf{F}_N^{[i]} \right).$$

for the sequence of subvectors of features.

The implied ordering of the features will be a constraint that we explore in our solution. In a more general setting the subsets may contain arbitrary features. Though this is very similar in description and method the resulting complexity is still prohibiting.
3 Feature selection

As criteria to remove a feature from the set of discriminatory features we consider

Irrelevance: when a feature subset does not depend on the class, or

\[ P(F[i]|C) = P(F[i]), \]

where the group is an actual group of the model,

Redundancy: when one or more features, \( F^n_m \), in a group, \( F[i] \), do not contribute to the discriminatory power, or

\[ P(F^n_m|C, F^{m-1}_{s_i}, F^{s_{i+1}-1}_{n+1}) = P(F^n_m|F^{m-1}_{s_i}, F^{s_{i+1}-1}_{n+1}). \]

In the following we will constrain the redundant subgroup by requiring that it is either the first part of a group or the last part, it cannot occur in the middle of the group and it contains consecutive features. So in equation 3 we require either \( m = s_i \) or \( n = s_{i+1} - 1 \).

4 Classification and the class posterior

An example will clarify the model. Let the generative distribution \( P(C, F^8) \) of a eight features model be given as

\[
\begin{align*}
P(C, F^8) &= P_C(C)P(F^8|C) \\
P(F^8|C) &= P(F_1|C)P(F_2, F_3|C)P(F_4|P(F^8_5|C) \\
&= P(F_1|C)P(F_2, F_3|C)P(F_4)P(F^6_5|C)P(F_7, F_8|F_5, F_6)
\end{align*}
\]

In this example equation 4 explains that we will focus on the conditional feature distribution. Equation 5 shows the dependent groups \( F_1, F^3_2, F_4 \), and \( F^7_5 \). So the index sequence of this model is \( s^4 = (1, 2, 4, 5) \).

We observe from equation 6 that group \( F_4 \) is irrelevant, and the part \( F^8_7 \) from group \( F^8_5 \) is redundant.

In classification we observe the features \( f^8 \) of an object and wish to decide on its class \( c \). Therefor we consider the ratio of probabilities, using equation 6,

\[
\begin{align*}
\frac{\text{Pr}\{C = 0, F^8 = f^8\}}{\text{Pr}\{C = 1, F^8 = f^8\}} &= \frac{\text{Pr}\{C = 0\} \text{Pr}\{F^8 = f^8|C = 0\}}{\text{Pr}\{C = 1\} \text{Pr}\{F^8 = f^8|C = 1\}} = \frac{P_C(0) P(f^8|0)}{P_C(1) P(f^8|1)} \\
&= \frac{P_C(0) P(f_1|0)P(f_2, f_3|0)P(f_4)P(f^6_5|0)P(f_7, f_8, f_5, f_6)}{P_C(1) P(f_1|1)P(f_2, f_3|1)P(f_4)P(f^6_5|1)P(f_7, f_8, f_5, f_6)} \\
&= \frac{P_C(0) P(f_1|0)P(f^3_2|0)P(f^6_5|0)}{P_C(1) P(f_1|1)P(f^3_2|1)P(f^6_5|1)}
\end{align*}
\]

So we see that irrelevant and redundant features do not play a role in the classification decision.
The same holds for the complete posterior of $C$.

$$P(C|E^8) = \frac{P_C(C)P(E^8|C)}{P(E^8)}$$

$$P(E^8|C) = P(F_1)P(E^8_1|E^6_1)P_C(c)P(F_1|c)P(E^6_2|c)P(F_1^6|c)$$

$$P(E^8) = \sum_{c \in \{0,1\}} P_C(c)P(F_1|c)P(E^6_2|c)P(F_4|c)P(E^6_5|c)$$

$$= P(F_4)P(E^8_7|E^6_7) \sum_{c \in \{0,1\}} P_C(c)P(F_1|c)P(E^6_3|c)P(E^6_5|c)$$

$$P(C|E^8) = \frac{P_C(C)P(F_1|C)P(E^6_2|C)P(E^6_5|C)}{\sum_{c \in \{0,1\}} P_C(c)P(F_1|c)P(E^6_2|c)P(E^6_5|c)}$$

Again we see that the irrelevant and redundant features don’t play a role in the final result.

5 Sequence probability estimation

We are given a sequence of objects $O^N = ((C_1, E_1), (C_2, E_2), (C_N, E_N))$. Here $C_i$ denotes the class of the $i$th object and $E_i$ its feature vector. Because the generative distribution is unknown, we define an estimated probability for the sequence of objects as

$$P_e(O^N) := P_e(C^N)P_e(F^N|C^N).$$

(7)

Here we abuse the notation $P_e()$ as it will be clear from the context which precise sequence probability is intended.

In [5] the authors introduce an efficient sequence probability estimation for binary sequences, which we will use for $P_e(C^N)$. In [6] a bound is given for the ratio of the maximum likelihood sequence probability $P_{ML}$ and $P_e(C^N)$.

We will briefly repeat and expand on these results. Let $p_B(\theta)$ be the beta distribution on $\theta$, with $0 \leq \theta \leq 1$. Also let $a$ denote the number of zeros in $C^N$ and $b$ denote the number of ones in $C^N$, so $a + b = N$. We have, cf. [5, 6]

$$p_B(\theta) := \frac{1}{\pi \sqrt{(1-\theta)\theta}},$$

$$P_e(C^N) := \int_0^1 p_B(\theta)(1-\theta)^a\theta^b d\theta$$

$$= \frac{\Gamma(a + \frac{1}{2})\Gamma(b + \frac{1}{2})}{\pi \Gamma(N + 1)},$$

(8)

$$P_{ML}(C^N) := \left( \frac{a}{N} \right)^a \left( \frac{b}{N} \right)^b.$$  

(9)

We will use a result from [6] that we repeat here.

$$\frac{1}{2} \sqrt{N} \leq \frac{P_e(C^N)}{P_{ML}(C^N)} \leq \sqrt{\frac{2}{\pi}} \sqrt{N}.$$  

(10)
Given an index sequence \( s^g \) we write
\[
P_e(\mathbf{F}^N|C^N) = \prod_{i=1}^{g} P_e(\mathbf{F}^i|C^N).
\]

\( P_e(\mathbf{F}^i|C^N) \) is the estimated conditional probability of the sequence of subfeatures \( \mathbf{F}^i \).

Using a binary decomposition technique from [7] we can write \( P_e(\mathbf{F}^i|C^N) \) as a product of binary sequence probabilities.

We will explain this by example. Let \( \mathbf{F} = (F_1, F_2, F_3) \) and consider the decomposition graph in figure 1. From this graph is is clear that the decisions at any internal node of the tree are independent and identically distributed with the corresponding \( \theta \) parameter. The number of free parameters in the decomposition equals the number of these thetas. Without further detail and proof we state the following lemma.

**Lemma 1** Let the probability part \( P(\mathbf{F}^i|C) \) have \( \alpha \) free parameters, then
\[
\left( \frac{1}{2} \sqrt{N} \right)^{\alpha} \leq \frac{P_e(\mathbf{F}^i|C^N)}{P_{ML}(\mathbf{F}^i|C^N)} \leq \left( \sqrt{\frac{2}{\pi N}} \right)^{\alpha}.
\]

We distinguish the following situations for \( P(\mathbf{F}^i|C) \).

**Fully dependent group:** Say that the group contains \( m \) features, then there are \( 2^m - 1 \) free parameters that describe the feature probabilities per class value. There are two classes so the number of free parameters is
\[
\alpha = 2(2^m - 1) = 2^{m+1} - 2.
\]

![Figure 1: The decomposition graph](image-url)
**Irrelevant group:** The class value does not matter so the number of free parameters is

\[ \alpha = 2^m - 1. \]

**Redundant subset:** Say that of the \( m \) features, \( a \) depend on the class and the remaining \( m - a \), conditioned on the \( a \) features do not depend on the class. So the number of free parameters is

\[ \alpha = 2(2^a - 1)(2^{m-a} - 1) = 2^{m+1} - 2^{a+1} - 2^{m-a+1} + 2. \]

So we see that irrelevance and redundancy reduces the number of free parameters of a feature group.

### 6 Model comparison

A data generating distribution \( P(C, \underline{F}^k) \) is described by

1. its index sequence \( s^g \) that defines the partitioning of the feature vectors into groups,

2. per group the decision whether it is an irrelevant group, contains a redundant subgroup, or is a fully dependent group.

If the group contains an redundant subgroup we must specify the subgroup. Every model contains a certain number of free parameters being the sum of the free parameters per group as explained below lemma 1.

If we define as optimality rule the sequence probability and say that the model that assigns the highest probability to the observed sequence \( O^N \), then we are likely to select an overly complex model. We follow the solution to this over-estimation problem along the Bayesian Information Criterion, [8], and the Minimum Description Length principle, [4].

We use as optimality criterion the estimated probability \( P_e(O^N) \) as defined in equation 7. Using lemma 1 it can be shown that this solves the over-estimation problem and we will find the most suitable, but not overly complex, model. Note that \( P_e(C^N) \) does not depend on the model so we will focus on the part \( P_e(\underline{F}^N|C^N) \).

In [9, 10] we introduced an computationally efficient network method to compute a Bayes mixture over all models [9] and to find the penalized maximum-likelihood model [10]. In this paper we extend this approach to include the possibility for irrelevant and redundant features. Consider the network in figure 2, taken from [9]. The node label,

![Network Diagram](image)

Figure 2: The network
e.g. \( N_{23} \), define the features, in this case \( F_2, F_3 \), considered by that node. The arrows coming into the node specify the possible splits of that node into smaller subgroups, and recursively into all smaller subgroups. Every arrow can be identified by an integer \( j = 1, 2, \ldots, m - 1 \), where \( m \) is the number of features considered in the node. This integer specifies the nodes that are connected to the arrows: it specifies the number of first features in one node and the remainder in the second node. e.g. \( N_{234} \) has the arrows \( i = 1 \) and \( i = 2 \), \( i = 1 \) specifies \( N_2 \) and \( N_{34} \), while \( i = 2 \) specifies \( N_{23} \) and \( N_4 \).

First every node contains, i.e. calculates or updates, its local sequence probabilities \( P_e(F[p]) \) and \( P_e(F[p]|C^N) \). Note: with some abuse of notation we use \( p \) here to indicate the considered features. Using the binary decomposition described in section 5 it can be seen that this amounts to a number of basic calculations of \( \mathcal{O}(Nk^3) \).

On every incoming arrow \( j \) the node receives information from the nodes \( l_j \) and \( r_j \), the first and last part respectively. Using this information the node calculates a \( P_{\text{max}}(F[p]) \). The information received per node, say node \( p' \), is \( P_{\text{max}}(F[p']) \) and

\[
\beta(p') := \frac{P_e(F[p]|C^N)}{P_e(F[p]).}
\]

The node \( p \), with incoming arrows \( j = 1, \ldots, m - 1 \) and corresponding nodes \( l_j \) and \( r_j \) calculates \( P_{\text{max}}(F[p]) \) as follows:

\[
P_{\text{max}}(F[p]) = \max \{ P_e(F[p]), \quad F[p] \text{ irrelevant} \}
\]

\[
P_e(F[p]) \times \beta(p), \quad F[p] \text{ conditional group}
\]

\[
P_e(F[p]) \times \beta(l_j), \quad F[r_j] \text{ is a redundant group}
\]

\[
P_e(F[p]) \times \beta(r_j), \quad F[l_j] \text{ is a redundant group}
\]

\[
P_{\text{max}}(F[l_j]) \times P_{\text{max}}(F[r_j]), \quad F[p] \text{ must be split}
\]

Note that the equations 14 to 16 are repeated for all valid arrow indices. The motivation for this expression is as follows:

if expression 12 achieves the maximum, then \( F[p] \) belongs to the maximum likelihood model, but those features are irrelevant, \( P_e(F[p]) \times \beta(p) = P_e(F[p]|C^N) \) so, if expression 13 achieves the maximum, then \( F[p] \) belongs to the maximum likelihood model, and contribute to the classification, \( P_e(F[p]) \times \beta(l_j) = P_e(F[l_j]|C^N)P_e(F[r_j]|F[l_j]) \) so, if expression 14 achieves the maximum, then \( P_e(F[r_j]|F[l_j]) > P_e(F[r_j]|C^N,F[l_j]) \) so \( F[p] \) belongs to the maximum likelihood model, but has a redundant subgroup \( F[r_j] \), if expression 15 achieves the maximum, then likewise \( F[p] \) belongs to the maximum likelihood model, but has a redundant subgroup \( F[l_j] \), if expression 16 achieves the maximum, then \( F[p] \) must be split into the subgroups \( l_j \) and \( r_j \).

The top layer nodes do not have arrows coming in so they only evaluate the equations 12 and 13. Eventually the bottom node contains the sequence probability \( P_{\text{max}}(F^N|C^N) \) of the penalized maximum-likelihood model. By investigating the network starting at the bottom node and following the decisions every node made the maximum-likelihood model can be recovered, as explained in [10].

As was shown in [9, 10] the computational complexity of the network calculations are also of \( \mathcal{O}(k^3) \) per execution. This calculation can be done after each new object is entered into the algorithm, i.e. sequential update or online classification with feedback, or only once after observing \( O^N \), i.e. training the model.
7 Conclusions

Finding the best subdivision of features into conditionally independent groups and finding redundant and irrelevant features is a connected problem that must be solved in a combined way. The number of possible group compositions including the selection of features results in a computationally infeasible brute force solution. We have discussed a method that relies on a uniform means to define the maximum likelihood, the decision on redundancy and irrelevance, and is able to use all similarities in the various models.

The method basically takes a maximum-likelihood model selection approach but solves the overestimation problem by using penalized sequence probabilities. The efficiency of the method comes from a recursive factorization of the feature probabilities.

The final result is an method that finds the penalized maximum likelihood model with the minimum number of features and an amount of computations that can be performed in polynomial time.

References


Abstract

Content fingerprinting and digital watermarking are techniques that are used for content protection and distribution monitoring. Over the past few years, both techniques have been well studied and their shortcomings understood. Recently, a new content fingerprinting scheme called active content fingerprinting was introduced to overcome these shortcomings. Active content fingerprinting aims to modify a content to extract robust fingerprints than the conventional content fingerprinting. Moreover, contrary to digital watermarking, active content fingerprinting does not embed any message independent of contents thus does not face host interference.

The main goal of this paper is to analyze fundamental limits of active content fingerprinting in an information theoretical framework.

1 Introduction

Generally, identification systems [1] are facing numerous requirements related to identification rate, complexity, privacy, security as well as memory storage. To address the trade-off between these conflicting requirements, content fingerprints are used [2], [3]. A content fingerprint is a short, robust and distinctive content description. In conventional content fingerprinting, the fingerprint is extracted directly from an original content and does not require any content modification that preserves the original content quality. In this sense, it can be considered as a passive content fingerprinting (PCFP). The extracted fingerprints resemble random codes, for which no efficient decoding algorithm is known as for structured codes. Moreover, the performance of PCFP in terms of identification rate is not satisfactory due to acquisition device imperfections.

For these reasons, active content fingerprinting (ACFP) was proposed in [4], [5], where the basic idea was introduced and a feasibility study revealed higher performance w.r.t. PCFP and digital watermarking. ACFP by modifying digital contents takes the best from two worlds of content fingerprinting and digital watermarking to overcome some of fundamental restrictions of these techniques in terms of performance and complexity. In the proposed fingerprinting scheme, contents are modified in a way to increase the identification rate and reduce the decoding complexity with respect to conventional content fingerprinting.

The main goal of this sequel is to analyze ACFP in an information theoretical framework to investigate its fundamental limits in identification systems. In this paper, we investigate the identification capacity based on ACFP in which a modified content can be modelled as an output of a discrete memoryless channel with an original content as its input. Moreover, we investigate the optimal encoding scheme under the assumptions that content sequences can be modeled as a Gaussian memoryless source and the observation channel as an additive white Gaussian. And, we introduce an optimal scheme that can achieve the identification capacity based on ACFP.
The rest of this paper is organized as follows. Section 2 presents the identification system based on active content fingerprinting and we will state our main result. Section 3 contains the proof of this result. Concluding remarks follow in Section 4.

**Notations:** Throughout this paper, we adopt the convention that a scalar random variable is denoted by a capital letter $X$, a specific value it may take is denoted by the lower case letter $x$, and its alphabet is designated by the script letter $\mathcal{X}$. As for vectors, a capital letter $X^N$ with a corresponding superscript will denote an $N$-dimensional random vector $X^N = (X_1, \ldots, X_N)$. A lower case letter $x^N$ will represent its particular realization $x^N = (x_1, \ldots, x_N)$. The expectation operator is designated by $E[\cdot]$.

### 2 Model Description

![Model Diagram](image)

Figure 1: Model of an identification system using modified content-sequences.

In an identification system, see Fig. 1, there are $M$ items indexed $w \in \{1, 2, \ldots, M\}$ to be identified. A randomly generated content-sequence (vector) of length $N$ corresponds to each item. This sequence has symbols $x_n, n = 1, 2, \ldots, N$ taking values in the discrete alphabet $\mathcal{X}$, and the probability that content-sequence $X^N = (x_1, x_2, \ldots, x_N)$ occurs for item $w$ is

$$\Pr\{X^N(w) = x^N\} = \prod_{n=1}^{N} Q_s(x_n),$$

hence the components $X_1, X_2, \ldots, X_N$ are independent and identically distributed according to $\{Q_s(x), x \in \mathcal{X}\}$. Note that this probability does not depend on the index $w$.

An encoder $e(\cdot)$ transforms each content-sequence $x^N$ into a modified content-sequence $y^N = (y_1, y_2, \ldots, y_N)$, where $y_n, n = 1, 2, \ldots, N$ taking values in the discrete alphabet $\mathcal{Y}$. The distortion between modified content sequence and content-sequence cannot be too large. The modification distortion $D_{xy}$ is defined as

$$D_{xy} = \frac{1}{N} E \left[ \sum_{n=1}^{N} D_{xy}(X_n, Y_n) \right],$$

where $D_{xy}(X_n, Y_n)$ is the distortion between $X_n$ and $Y_n$. The objective is to minimize $D_{xy}$.
where \( \{D_{xy}(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\} \) is the distortion matrix specifying the distortion per component. We assume that the distortion matrix has only finite non-negative entries. Moreover, we assume that all modified content-sequences are generated prior to the identification procedure. These modified sequences form a codebook that we call the “database” here. This database \( C \) consists of the list of entries, hence

\[
C = \{ y^N(1), y^N(2), \ldots, y^N(M) \}. \tag{3}
\]

In the identification process the probabilities for the items to be presented for identification are all equal, hence

\[
\Pr\{W = w\} = 1/M \text{ for } w \in \{1, 2, \ldots, M\}. \tag{4}
\]

When item \( w \) is presented for identification, its corresponding modified content-sequence \( y^N(w) \) is “selected” from the database \( C \) and presented to the system, hence

\[
y^N = s(w). \tag{5}
\]

The system observes \( y^N \) via a memoryless observation channel \( \{Q_c(z|y), y \in \mathcal{Y}, z \in \mathcal{Z}\} \), with discrete alphabet \( \mathcal{Z} \), and the resulting channel output sequence is \( z^N = (z_1, z_2, \ldots, z_N) \), where \( z_n \in \mathcal{Z} \) for \( n = 1, 2, \ldots, N \). Now

\[
\Pr\{Z^N = z^N|Y^N = y^N\} = \prod_{n=1}^{N} Q_c(z_n|y_n). \tag{6}
\]

After observing \( Z^N \), the decoder decides that \( Z^N \) is related to which modified content-sequence. If this is \( Y^N(w) \), the decoder outputs \( \hat{W} = w \). The reliability of our identification system is measured by the error probability

\[
P_e = \Pr\{\hat{W} \neq W\}. \tag{7}
\]

### 2.1 Statement of Result

An identification rate-distortion pair \((R, \Delta)\) is said to be achievable if for all \( \epsilon > 0 \) there exist for all \( N \) large enough an encoder and a decoder such that

\[
D_{xy} \leq \Delta + \epsilon, \\
\log_2(M) \geq N(R - \epsilon), \text{ and} \\
\Pr\{\hat{W} \neq W\} \leq \epsilon. \tag{8}
\]

We are now ready to state the main result of this submission, the proof follows in section III.

**Theorem 1.** The region of achievable rate-distortion pair \((R, \Delta)\) for the identification system based on modified content-sequence is given by

\[
\left\{ (R, \Delta) : R \leq I(Y; Z), \right. \\
\Delta \geq \sum_{x,y} Q_s(x)P_t(y | x)D_{xy}(x, y), \\
\left. \text{for } P(x, y, z) = Q_s(x)P_t(y | x)Q_c(z | y) \right\}. \tag{9}
\]

The “capacity” of identification based on ACFP, the maximum of possible identification rate, for a given distortion \( \Delta \) is given by

\[
C_{ACFP}(\Delta) = \max_{P_t(y|x): \sum_{x,y} Q_s(x)P_t(y|x)D_{xy}(x, y) \leq \Delta} I(Y; Z). \tag{10}
\]
3 Proof

The proof consists of the achievability part and a converse. We start with the converse.

3.0.1 Converse Part

First, we define the random variable $I$ that takes values in $\{1, 2, \ldots, N\}$ with probability $1/N$. Then the random triple $(X, Y, Z)$ is defined as $(X, Y, Z) \overset{d}{=} (X_i, Y_i, Z_i)$ if $I = i$. Hence, the joint distribution of $(X, Y, Z)$ is given by

$$P(x, y, z) = \frac{1}{N} \sum_{i=1}^{N} \Pr\{X_i = x, Y_i = y, Z_i = z\}$$

$$= \frac{1}{N} \sum_{i=1}^{N} Q_s(x_i)P_t(y_i | x_i)Q_c(z_i | y_i)$$

$$= Q_s(x)P_t(y | x)Q_c(z | y)$$ (11)

Consider the $M$ number of modified contents. Using Fano’s inequality $H(\hat{W} | W) \leq F$ where $F = 1 + \Pr\{\hat{W} \neq W\}$, we have the following series of (in)equalities:

$$\log_2(M) = H(W)$$

$$\leq I(W; Z^N, Y^N(1), \ldots, Y^N(M)) + F$$

$$= I(W; Y^N(1), \ldots, Y^N(M)) + I(W; Z^N | Y^N(1), \ldots, Y^N(M)) + F$$

$$\overset{(a)}{\leq} H(Z^N) - H(Z^N | Y^N(1), \ldots, Y^N(M), W) + F$$

$$= H(Z^N) - H(Z^N | Y^N(W)) + F$$

$$\leq \sum_{i=1}^{N} H(Z_i) - \sum_{i=1}^{N} H(Z_i | Y_i(W)) + F$$

$$= \sum_{i=1}^{N} I(Y_i(W); Z_i) + F$$

$$= NH(Z | I) - NH(Z | Y, I) + F$$

$$\overset{(b)}{\leq} NI(Y; Z) + F,$$ (12)

where (a) and (b) follow from the facts that conditioning reduces entropy and the channel is memoryless.

Now for the distortion part we have

$$\overline{D_{xy}} = \frac{1}{N} E \left[ \sum_{n=1}^{N} D_{x,y}(X_n, Y_n) \right]$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{x_n,y_n} Q_s(x_n)P_t(y_n | x_n)D_{x,y}(x_n, y_n)$$

$$= \sum_{x,y} Q_s(x)P_t(y | x)D_{x,y}(x, y)$$
Now, assume that \((R, \Delta)\) is achievable. Then \(F \leq 1 + \epsilon \log_2(M)\) and \(\Delta \geq D_{xy} - \epsilon\). For all blocklengths \(N\) and small enough \(\epsilon > 0\), we obtain from (12) that

\[
N(R - \epsilon) \leq \log_2(M) \leq \frac{1}{1 - \epsilon} (NI(Y; Z) + 1),
\]

for some \(P(x, y, z) = Q_s(x)P_t(y \mid x)Q_v(z \mid y)\). If we now let \(\epsilon \downarrow 0\) and \(N \rightarrow \infty\), then we obtain the converse of Theorem 1 from (13) and (14).

### 3.0.2 Achievability

We can only give an outline of the achievability proof here. For each content-sequence \(X^N(w)\), a modified content-sequence \(Y^N(w)\) is generated using conditional distribution \(P_t(y \mid x)\). These modified sequences are codewords in a random codebook that guarantees a rate that can be as large as \(I(Y; Z)\). The distortion is as expected (i.e. \(D_{xy}(X,Y)\)) because of the law of large numbers.

### 3.1 Gaussian Source

Let’s assume the content-sequences are distributed i.i.d. according to a Gaussian distribution with variance \(V_X\) and mean zero. Moreover, the observation channel \(Q_v(z \mid y)\) can be modelled as an additive white Gaussian noise (AWGN) with variance \(V_N\).

**Theorem 2.** Considering distortion as the mean-squared error, the capacity of identification based on ACFP is given by

\[
C_{ACFP}(\Delta) = \frac{1}{2} \log_2 \left( 1 + \frac{(\sqrt{V_X} + \sqrt{\Delta})^2}{V_N} \right),
\]

that can be achieved by scaling the content by a factor \(f\), i.e. \(Y^N = fX^N\), such that \((f - 1)^2V_X = \Delta\).

**Proof.** We can upper bound the identification rate as follows

\[
I(Y; Z) = h(Z) - h(Z \mid Y)
\]

\[
= h(Z) - \frac{1}{2} \log_2 2\pi e V_N
\]

\[
\leq \frac{1}{2} \log_2 2\pi e (V_Y + V_N) - \frac{1}{2} \log_2 2\pi e V_N
\]

\[
\leq \frac{1}{2} \log_2 \left( 1 + \frac{(\sqrt{\Delta} + \sqrt{V_X})^2}{V_N} \right),
\]

where \(V_Y = E[Y^2]\) and (a) follows from the fact that

\[
\]

\(V_Y\) attains the maximum if the equality holds in the above equation and \(Y\) is aligned with \(X\), i.e. \(Y^N = fX^N\), such that \((f - 1)^2V_X = \Delta\). Note that by setting \(Y^N = fX^N\) equalities in (16) hold.

Figure 2 shows the capacity of identification systems using PCFP versus ACFP for different values of distortion \(\Delta\).
Figure 2: (a) the capacity of identification based on ACFP versus PCFP for a Gaussian source and AWGN with SNR=10 dB.

4 Conclusions

In this paper, we evaluated the capacity of identification systems based on active content fingerprinting. In active content fingerprinting, the main goal is to modify a digital content to improve the performance in terms of identification rate. We assumed that the modification can be modeled by a memoryless channel. Then, we investigated the optimal encoding scheme under Gaussian assumptions that can achieve the identification capacity based on ACFP.

References


Error Checking and Near Matching in Helper Data Systems for Biometric Authentication

Chara Papatsimpa Joep de Groot Jean-Paul Linnartz
Eindhoven University of Technology
Signal Processing Systems
Den Dolech 2, 5612 AZ Eindhoven, The Netherlands
C.Papatsimpa@tue.nl j.a.d.groot@tue.nl J.P.Linnartz@tue.nl

Abstract
Helper data systems mitigate the risk that biometric templates are stolen from a biometric data base. Yet, current systems face the drawback that strong Error Correction is needed in order to mitigate variations in the measured biometric during verification. Error correction codes are not always attractive, as these may severely reduce the effective entropy extracted from the biometric. We study an alternative, namely to check the (set of) most likely errors, based on a posterior side information, using soft decision during verification.

1 Introduction

Biometrics have become a popular solution for authentication or identification. The storage of biometric features at large scale however, has introduced both privacy and security risks. Examples of possible security risks include identity theft, cross matching and the fact that some biometric characteristics may provide additional information about the background of an individual. Those security and privacy problems cannot be solved with a simple encryption/decryption scheme, since encryption cannot prevent insider attacks, attacks by people who have access to the database and thus possess the decryption keys. We cannot trust, and in general do not want to rely on trusting, the verifier. Once biometric data has been compromised, it is public forever and cannot be used in a security application anymore. Biometric data is inherently part of a user, e.g. one cannot change a users fingerprints, therefore we have to make sure that whatever is stored cannot be traced back to the original features. The situation is quite similar to the problem of password storage, for which the standard solution is to store hashed passwords. An attacker who gets hold of a hashed password cannot deduce the original password from it. However, this method cannot be straightforwardly applied to biometrics. The cryptographic hash functions are highly sensitive to small deviations in input and produce different results for two different input values, regardless of how similar they may be. This property requires that the data presented during enrollment to be exactly equal to the data presented during verification. The problem with biometrics, however, is that it has inherent limitations: the measurements present small deviations (“noise”) between enrollment and verification, making cryptographic template protection a challenging task.

Scientific literature distinguishes between digital and analog variants of Helper data Systems. Analog HDSs, such as Biased Quantization typically use a user-dependent bias in the quantization of biometrics in the analog domain, i.e., before quantization.
A prime example is [2]. In digital HDSs, e.g. [1, 3] during the verification, the quantified biometrics are "shifted" towards a valid code word. This shifting can be a xor operation with a vector of low-hamming weight that shifts the extracted biometric to the nearest code word. This operation occurs fully in the digital domain. After this shift an error correcting code (ECC) can be applied. Practical systems need to combine those two: Biased Quantization as HDS-1 can provide a sufficiently good BER to allow the operation of an error correction code as HDS-2.

The most commonly accepted verification schemes consist of an enrollment and verification phase. In the enrollment phase the prover provides his biometric data \( \mathbf{x} = (x_0, \ldots, x_{M-1}) \). From this data, the system extracts a secret \( \hat{s} = Q(\mathbf{x}) \), where \( Q \) is a known and predefined quantization function. Consequently the system stores this safely in the hashed form \( (h(\hat{s} \| z), w) \), where \( w \) is the helper data, which is generated as \( w = g(\mathbf{x}) \), where \( g \) is a publicly known function and \( z \) is the salt. The salt is a system and/or user specific random string to prevent cross-matching between different databases. In the verification phase, the prover provides his (correlated) biometric data \( \mathbf{y} = (y_0, \ldots, y_{M-1}) \) to prove his identity. All variables, except for the salt \( z \), are length \( M \) vectors extracted by some means of preprocessing to ensure that the components are (nearly) independent, but not necessarily identically distributed. Independence can be obtained for example by applying a principle component analysis (PCA) to the raw data. This allows us to analyze each dimension separately. Yet the verification sample may not be exactly equal to the enrollment sample. Hence the \( r \) vector may not map the biometric exactly on a code word. Some other error can occur, but \( r \) will mostly likely shift the verification biometric to the vicinity of the code word chosen during enrollment. An ECC operation that (hopefully) corrects the error. The outcome of the ECC is hashed and compared to the template.

## 2 Error Checking

Skoric et al. [4] introduced a different approach for error-correction, better suited to combat errors in digitized biometrics. The idea is that when considering an N-dimensional biometric feature vector, it often happens that certain errors appear to be more likely than others. This paper investigates the idea that during verification the verifier simply

---

Figure 1: Fig. 1 depicts an example of a helper data verification scheme, in which the auxiliary data \( w \) generated during enrollment is applied during verification to reproduce the secret \( s \). It contains an analog stage HDS-1 and a Digital stage HDS-2.
checks a set of likely bit patterns, (instead of applying an ECC on the code shifted bit pattern). Information about "where the errors are" is available a posteriori based on soft decision information during verification, i.e., before hard quantizing, one may see that certain samples lie closer to decision boundaries than others. Bits extracted from such dimension are more likely to contain errors than bits extracted from dimensions where the sample lies in the middle of a large quantization intervals. We will investigate this idea for the zero leakage quantization scheme presented in [5] and derive a model for the reproduction probability based on that scheme.

2.1 Zero Leakage Biometric Verification

Helper data \( w \) and secret \( s \) are given by:

\[
\begin{align*}
s &= Q(x) \tag{1} \\
w &= g(x) \tag{2}
\end{align*}
\]

According to the ZL-scheme[5]

\[
g(x) = \frac{F_X(x) - F_X(q_s)}{p_s}. \tag{3}
\]

Quantization function \( Q \) uses quantization boundaries \( q_s \) as lower bounds on contiguous quantization intervals. In the special case of a FE \( p_s = 1/N \), which simplifies the enrollment functions to

\[
\begin{align*}
s &= Q(x) = \lfloor N \cdot F_X(x) \rfloor, \tag{4} \\
w &= g(x) = N \cdot F_X(x) - s. \tag{5}
\end{align*}
\]

in which \( F_X \) is the cumulative distribution function (CDF) of feature \( x \). For HDS's with non-uniform secrets as well as for the FE, this yields a continuous helper data \( w \) that reveals no information about the enrolled secret \( s \). In fact, one can exactly reconstruct \( N \) possible \( x \) values, each in a different quantization interval. This reconstruction is given by

\[
x_s(w) = g_s^{-1}(w) \tag{6}
\]

For ease of expression we will refer to these points as 'sibling points'. Based on the \( N \) sibling points, out of which (only) one is the enrolled secret value, and the verification sample \( y \), the verifier can decide which is the most likely value of the secret \( s \), so

\[
s = \arg\max_s p_{s|y,w} = \arg\max_s f_{Y|X}(y|g_s^{-1}(w)). \tag{7}
\]

As shown in [5] for symmetric and fading noise, this even leads to a simple comparison with by helper-data based thresholds \( \tau_s \), with

\[
\tau_s = \rho \frac{g_s^{-1}(w) + g_{s-1}^{-1}(w)}{2}. \tag{8}
\]
2.2 Gaussian Error

Given a dimension $i$ with quantization intervals $N_i$ and correlation $\rho_i$, a verifier wants to know how likely each $\hat{s}_i$ is by using the helper data $w_i$ and noisy measurement $y_i$. We assume the verification sample $y$ to be related to enrollment sample $x$ by the joint density function

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2 \sqrt{1-\rho^2}} \exp \left( -\frac{(x - \mu)^2 + (y - \mu)^2 - 2\rho(x - \mu)(y - \mu)}{2(1-\rho^2)\sigma^2} \right).$$

(9)

For ease of the analysis we might assume $\mu = 0$ and $\sigma = 1$, which simplifies the above equation to

$$f_{XY}(x, y) = \frac{1}{2\sqrt{1-\rho^2}} \exp \left( -\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)} \right),$$

(10)

from which we can derive the conditional density function

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp \left( -\frac{(y - \rho x)^2}{2(1-\rho^2)} \right).$$

(11)

By using the conservation of probability

$$f_{XY}(x, y)dx\,dy = f_{SWY}(s, w, y)dw\,dy$$

(12)

and the definition of the helper data, $w = g(x) = (F_X(x) - F_X(q_s))/p_s$, hence

$$\frac{dx}{dw} = g'(x) = \frac{f_X(x)}{p_s}$$

(13)

we can define a joint density function for the reproduction candidates and the noisy measurement

$$f_{SWY}(s, w, y) = p_s \frac{f_{XY}(g^{-1}_s(w), y)}{f_X(g^{-1}_s(w))} = p_s f_{Y|X}(y | g^{-1}_s(w)).$$

(14)

From this joint density function we can derive

$$f_{WY}(w, y) = \sum_{s \in S} f_{SWY}(s, w, y)$$

(15)

and subsequently the verifier’s a posterior reproduction probabilities

$$p_{s|yw} = \frac{f_{SWY}(s, w, y)}{f_{WY}(w, y)} = \frac{f_{SWY}(s, w, y)}{\sum_{s \in S} f_{SWY}(s, w, y)}.$$

(16)

2.3 $\Delta L_C(w, y)$ Detector

The initial estimate of our secret during verification is given by the most probable secret on dimension $i$, hence

$$\hat{s}_i = \arg \max_{s \in S} p_{s|yw}. $$

(17)
Since we assumed the \( M \) dimensions to be independent, the a posterior probability of this estimate is

\[
P_{0|\mathbf{w}y} = \prod_{i=1}^{M} p_{\hat{s}_i|w,y_i} = \exp \left( \ln \left( \prod_{i=1}^{M} p_{\hat{s}_i|w,y_i} \right) \right) = \exp \left( \sum_{i=1}^{M} \ln p_{\hat{s}_i|w,y_i} \right)
\]  

(18) (19) (20)

Yet, we are not only interested in the most likely string, as it may possibly contain an error. If we change the \( k \)-th bit decision in the string to a target value \( t_k \neq \hat{s}_k \), this new string has probability

\[
p_{1|\mathbf{w}y} = \exp \left( \sum_{i=1}^{M} \ln p_{\hat{s}_i|w,y_i} - \ln p_{s_k|w,y_k} + \ln p_{t_k|w,y_k} \right)
\]

(21)

Instead of looking at the probabilities we can look at the log likelihood value that this estimate is correct, hence \( L_C(w,y) = \ln p_{0|\mathbf{w}y} \). The penalty, the change in the initial \( L_C(w,y) \) value due to a single bit flip can be expressed as (we omit the \( k \) indices for clarity)

\[
\Delta L_C(w,y) = \ln p_{1|\mathbf{w}y} - \ln p_{0|\mathbf{w}y} = \ln p_{1|\mathbf{w}y} - \ln p_{0|\mathbf{w}y}
\]

(22) (23)

\[
= \ln \frac{f_{SWY}(t,w,y)}{f_{SWY}(\hat{s},w,y)} - \ln \frac{f_{SWY}(\hat{s},w,y)}{f_{SWY}(w,y)}
\]

(24)

\[
= \ln f_{SWY}(t,w,y) - \ln f_{SWY}(\hat{s},w,y)
\]

(25)

\[
= \ln p_t + \ln f_{Y|X}(y|g_{\hat{s}}^{-1}(w)) - \ln p_{\hat{s}} - \ln f_{Y|X}(y|g_{\hat{s}}^{-1}(w))
\]

(26)

\[
= \ln \frac{p_t}{p_{\hat{s}}} \left( \frac{(y - \rho g_{\hat{s}}^{-1}(w))^2}{2(1 - \rho^2)} + \frac{(y - \rho g_t^{-1}(w))^2}{2(1 - \rho^2)} \right)
\]

(27)

\[
= \ln \frac{p_t}{p_{\hat{s}}} \left( \frac{(y - \rho g_{\hat{s}}^{-1}(w))^2 - (y - \rho g_t^{-1}(w))^2}{2(1 - \rho^2)} \right)
\]

(28)

\[
= \ln \frac{p_t}{p_{\hat{s}}} \left( \frac{\rho \left( g_{\hat{s}}^{-1}(w) - g_t^{-1}(w) \right) \left( y - \rho g_{\hat{s}}^{-1}(w) \right)}{1 - \rho^2} \right)
\]

(29)

So, if we assume equiprobable secrets, i.e., \( p_s = 1/N \) for \( s \in \mathcal{S} \), the verification thresholds are given by (8), and

\[
t = \hat{s} - 1 \quad \Rightarrow \quad \Delta L_C(w,y) = -\frac{\rho}{1 - \rho^2} \left( g_{\hat{s}}^{-1}(w) - g_t^{-1}(w) \right) (y - \tau_s(w))
\]

(31)

\[
t = \hat{s} + 1 \quad \Rightarrow \quad \Delta L_C(w,y) = -\frac{\rho}{1 - \rho^2} \left( g_{\hat{s}+1}(w) - g_{\hat{s}}^{-1}(w) \right) (\tau_{s+1}(w) - y)
\]

(32)
We notice that $\Delta L_C(w, y)$ consists of three terms. The first term shows the dependence from $\rho$ or equally the noise variance. The second term is the distance between the sibling points; equally it represents the length of our quantization intervals or the distance between the decision thresholds. Finally the last term is the distance of the verification sample from the decision threshold. Thus, in order to choose the most likely bits to flip, we want to consider verification samples close to a decision threshold but given that the SNR of that dimension is low and the distance between the thresholds is relatively large.

\[
\rho x_{s-1} \quad \tau_s \quad \rho x_s \quad \tau_{s+1} \quad \rho x_{s+1}
\]

\[
\Delta L_C(w, y) \quad \Delta L_C(w, y')
\]

Figure 2: The likelihood penalty is linearly based on the distance of $y$ to its nearest quantization boundary namely $\tau_s$ or $\tau_{s+1}$.

3 Performance Evaluation

To examine the performance of our error checking verifier we conducted experiments on both synthetic and real data sets. We compare a one-bit error correcting system with a system that tolerates the occurrences of a single bit error by smart trial of changing bits that are likely to be in error. The ECC system reduces the entropy of a 63 bit biometric secret to 57 bits. On the other hand, our error checking verifier, uses soft decision to determine which $n$ bits are most likely to contain an error and applies an error checking scheme that attempts a single bit flip in the $n$ most likely dimensions.

3.1 Synthetic data experiments

We carried out experiments on synthetic Gaussian data for two cases, namely 63 i.i.d features each having a SNR of 10, 12 and 15 dB, and also with 63 independent features but with non-equal SNR. Non-i.i.d case, the noise variances from 17 to 5 dB according to measured variances in PCA of realistic biometric.

Figure 3 shows the resulting detection rate (dr) versus the number of trials required in order to achieve the same detection rate as when applying an ECC.
When snr varies from 17 to 5 db are able to achieve a detection rate of 92% (result when using an 1-error ECC) by trying the seven more likely error locations. At this situation the verifier accepts seven different bit strings, thus the user capacity is reduced to 60 bits. An attacker, who is just operating on data base template data would still need an exhaustive search over $2^{63}$ bit combinations, such that the effective secrecy entropy remains 63 bits while an ECC scheme would only deliver 57 bits of secrecy entropy. Trying only the first three most likely error locations would result to a 90% dr, a result quite close to the desired. In the case of having the same snr in each feature, we reach the desired detection rate with a limited number of trials.

### 3.2 Real data experiments

The real data experiments were conducted on a small face database we enrolled, containing variable images of 60 different users.

Fig. 4 shows the resulting detection rate (dr) versus the number of trials required in order to achieve the same detection rate as when applying an 1-ECC. We are able to achieve a detection rate of 80% (result when using an 1-error ECC) with a limited number of trials.
4 Conclusions

Biometric keys often suffer from low secret entropy, thus reducing the complexity of the secret over brute force attacks. Most current biometric systems are based on error correction codes. However, tests in practical systems revealed that the code rate of the ECC is far below any desirable value. Hence the biometric value looses a lot of its entropy, which jeopardizes the strength of any cryptographic system. In a 1-bit ECC over 63 dimensions, the search space of an attacker reduces by 6 bits. This penalty is severe, particularly because the entropy that can be extracted from biometrics is often lower than one would prefer to use for a strong security system. However, an attacker who launches an exhaustive search in a n error checking system still has to search through the entire space span by the original 63 bits. Hence we may argue that error checking does not reduce the key entry, while an error correction system does. An error checking system reduces the computational gap between executing a verification and an exhaustive search attack. It increases the verification effort, but does not change the search space compared to a system without error protection.

Our results have shown that we were able to extract reliable soft decision values and we noticed improvements, both for synthetic and real (facial) biometrics in the case of a single bit flip.

References


Online Self-supervised Learning for Road Detection

Marco L.L. Rompen, Willem P. Sanberg, Gijs Dubbelman, Peter H.N. de With
Eindhoven University of Technology
SPS/VCA, Department of Electrical Engineering
Eindhoven, Netherlands
m.l.l.rompen@student.tue.nl, {w.p.sanberg, g.dubbelman, p.h.n.de.with}@tue.nl

Abstract

We present a computer vision system for intelligent vehicles that distinguishes obstacles from roads by exploring online and self-supervised learning. It uses geometric information, derived from stereo-based obstacle detection, to obtain weak training labels for an SVM classifier. Subsequently, the SVM improves the road detection result by classifying image regions on basis of appearance information. In this work, we experimentally evaluate different image features to model road and obstacle appearances. It is shown that using both geometric information and Hue-Saturation appearance information improves the road detection task.

1 Introduction

Vehicles are relying increasingly on computer vision to ensure a safer way of driving. Examples are Advanced Driver Assistance Systems (ADAS), such as lane departure warning and pedestrian detection. These and related technologies will be improved and extended in the future, leading to autonomously driving vehicles. Such intelligent vehicles have the potential to significantly reduce traffic congestions, accidents, and pollution levels.

In this work, we focus on computer vision methods necessary to distinguish obstacles from drivable surfaces. Our approach is designed to work with asphalt roads, paved roads, country roads, or any other drivable terrain. This is enabled by online and self-supervised learning of the visual appearance of drivable surfaces versus obstacles. In contrast to [2], it does not depend on human-annotated train images, and is fully adaptive to different road surface types, environments, weather and lighting conditions. An obstacle detection system based on a stereo camera [1] provides a basis for reliably separating obstacles from roads by exploring a weak labeling, which is then improved by a Support Vector Machine (SVM) classification using color features. A similar approach was used in [3], where the focus was purely on outdoor terrain in which the differences between drivable surfaces and obstacles are typically more pronounced than in urban (gray world) environments, which is our main focus. In [4] also a similar approach was used, building on the assumptions that the road in front of the vehicle is always drivable, has uniform appearance and differs significantly from obstacles. By combining [3] and [4] and improving it with [1] a system is created applicable to the urban environment.

The complete system is described in Section II and III. The results of our experiments are provided in Section IV and finally Section V brings forward our conclusions and recommendations.
2 SYSTEM OVERVIEW

Our vision-based drivable surface (i.e. road) detection system with its key processing blocks is depicted in Fig. 1. Its general concepts are discussed here. Detailed descriptions of its key processing blocks are provided in Section III.

The input to the system is a continuous binocular video stream. It keeps a buffer of the past 60 frames $I_{n-60}, ..., I_{n-1}$ to online learn an appearance model that is used to classify pixels in the current frame $I_n$ as being part of a drivable surface or part of an obstacle.

The classifier used by the system is a two-class SVM. It requires a set of training examples to learn a decision surface separating the training samples of the two classes {road, obstacles}. This decision surface is then used to classify new samples. In contrast to the typical usage of SVM, where training samples are obtained by manually annotating images, in our system the training samples are automatically obtained from the stereo-based obstacle detection processing block.

This stereo subsystem estimates disparity maps from which the depth of pixels can be obtained. From this geometric information, the local slope of surfaces is calculated. If the slope exceeds a certain threshold, pixels are labeled as obstacles [1]. As stereo-based depth estimation suffers from many artifacts, the obstacle labeling is not perfect. Therefore, we use these (weak) labels, automatically obtained from stereo geometric information, to train an SVM on appearance-based information. The appearance model for pixels is a histogram representing the color or gradient distribution in a local rectangular area around each pixel. In this work different color spaces and histogram creation strategies are experimentally evaluated.

3 ROAD DETECTION WITH SELF-SUPERVISED LEARNING

The online learning and classification pipeline starts with generating training masks, using the obstacle detection and road Region Of Interest (ROI) processing blocks, as depicted in Fig. 1. The next step is feature extraction followed by online SVM training and finally classification. These four key steps in the system are described in detail below.
3.1 Training mask

Training masks assign labels \{road, obstacle, unknown\} to pixels to facilitate creating training examples for the online self-supervised SVM classifier. The training mask creation exists of two processing blocks, obstacle detection and road ROI, which are combined into one mask for every train frame. The obstacle map provided by the obstacle detection processing block is eroded to reduce the number of pixels that are wrongly labeled as obstacles. The road ROI is based on the assumption that the surface directly in front of the car is always drivable, unless the obstacle detector determines otherwise. All parts not specifically designated as obstacle or road are given the unknown label. In Fig. 2 an example frame with its obstacle map and its training mask is depicted.

![Fig. 2. Example of training mask generation. (a) The rectified left image of the stereo input frame. (b) The obstacle map (white is obstacle and black is non-obstacle) generated by the obstacle detection processing block. (c) The training mask containing the weak labels. In this mask black is a weak obstacle label, white is a weak road label, and gray is unknown. (d) Training mask as an overlay on the rectified left image.](image)

3.2 Feature extraction

The feature extraction processing block has two functions. Its first function is to provide unlabeled training samples for the SVM training processing block. The labels for the training samples are obtained from the training masks. Next to that, it provides features for the road classification processing block.

The feature extraction processing block divides frames in (non-overlapping) square blocks of 17-by-17 pixels. For each of these blocks it computes a histogram, i.e. the image feature, as a representation of their appearance. In this work, we consider different histogram content and creation methods.

In order to fill the histograms we need to assign an Index (I) and Magnitude (M) to each pixel on basis of its (multidimensional) value. The Index (I) is used to determine the corresponding bin (B) of the pixel in the histogram, which can be a non-linear mapping. The Magnitude (M) of a pixel is its contribution to its corresponding bin in the histogram. For color-based histograms, the value of a pixel is used as the Index (I) and the Magnitude (M) is defined as 1. For histograms of gradients (HOG), the orientation of the gradient is used as the Index (I) and the Magnitude (M) of a pixel is the magnitude of its gradient.

A histogram can be created by mapping each feature space dimension to a histogram dimension, thereby forming a multidimensional histogram. An alternative is to concatenate all feature space dimensions into one histogram dimension, thereby forming a 1-D histogram. The potential benefit of a multidimensional histogram is that it can capture correlations between feature space dimensions. In a 1-D histogram all individual feature spaces are modeled as being independent of each other. With this we gain computational
efficiency, but we potentially lose valuable statistical information. In this work we experiment with both approaches.

Another aspect of histogram creation that we evaluate is the mapping from pixel indexes (I) to histogram bins (B). For this we compare a simple linear mapping with an adaptive data dependent method. The adaptive method can be seen as an offline dimension reduction strategy [6]. Its aim is to only have meaningful histogram bins. This is performed by taking random pixels out of the dataset and determining the bin boundaries such that every bin has an equal (non-zero) number of pixels in it.

Finally, the histograms are normalized. For 1-D histograms that are made up of multiple feature space dimensions, the separate histograms of each feature dimension are first normalized separately, after which the complete histogram is normalized. This ensures that training of the SVM classifier starts with a situation in which each independent feature space is equally weighted. An overview of all histogram based feature spaces that we use for our experiments are provided in Table I. They are a combination of gradient histograms and histograms over the color spaces: RGB, HSV, YIQ (NTSC) and subspaces of these.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of feature space dimensions</th>
<th>Bins per feature space dimension</th>
<th>Histogram dimensionality</th>
<th>Total number of bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS100</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>HS144</td>
<td>2</td>
<td>50</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>HSV96</td>
<td>3</td>
<td>64</td>
<td>1</td>
<td>128</td>
</tr>
<tr>
<td>HS-HOG96</td>
<td>3</td>
<td>32</td>
<td>1</td>
<td>96</td>
</tr>
<tr>
<td>HSV216</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>216</td>
</tr>
<tr>
<td>YIQ216</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>216</td>
</tr>
<tr>
<td>RGB216</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>216</td>
</tr>
<tr>
<td>RGB96</td>
<td>3</td>
<td>32</td>
<td>1</td>
<td>96</td>
</tr>
</tbody>
</table>

3.3 Training

The goal of the training processing block is to train a SVM, online and self-supervised, that can be used to classify pixels as being part of a road or an obstacle. A training sample consists of a feature and a label. Its feature is a histogram computed over a 17-by-17 pixel block and its label is the value \{road, obstacle, unknown\} of the center pixel of the block in the automatically obtained training mask. Blocks for which their label is unknown are ignored during training.

The goal of SVM is to fit a hyper plane in such a way that two classes \{road, obstacle\} are maximally separated by the hyper plane [7]. In this work we experiment with a linear SVM and a Radial Basis Function (RBF-) SVM. The linear-SVM fits a hyper plane directly in the feature space. Its benefit is that it is more computational efficient than RBF-SVM but it can only model linear decision boundaries. Such linear decision boundaries are not appropriate for every classification task; some tasks require non-linear decision boundaries. An RBF-SVM implicitly maps the feature space to a higher dimensional feature space such that a non-linear decision boundary in the original feature space can be represented by a linear decision boundary in the higher dimensional feature space. This is known as the kernel trick. Although kernel methods like RBF are relatively efficient, they still take significantly more
computations than a linear-SVM. For more information on SVM and the kernel trick we recommend [7].

3.4 Classification

On basis of an SVM trained on past image data, i.e. $I_{n-60} \ldots I_{n-1}$, the task is to classify pixels in the current frame $I_n$ as being part of a road or an obstacle. This process starts with feature extraction on $I_n$ providing a feature for each non-overlapping 17-by-17 pixel block in $I_n$. These features are then classified by the SVM resulting in a classification of each pixel block in $I_n$. In this work, we assume that the road is always below the horizon; therefore we only classify pixel blocks that are below the horizon. Every pixel block above the horizon is determined to be part of an object.

After classification, we apply post processing to incorporate some level of regularization. This starts with creation of a binary mask from the classification result in which 1 represents a road and 0 an obstacle pixel block. Regularization is then performed by applying median filters and enforces local smoothing of the classification result at the pixel block level.

4 Evaluation

The primary goal of our experiments is to evaluate different image features (color-based and gradient-based histograms) for the road detection task. We also experiment with using a linear-SVM and a RBF-SVM. Our system is currently implemented in MATLAB utilizing 4 cores and a GPU through MATLAB’s parallel computing toolbox.

Fig. 3. The automatically obtained train mask sets (a) in which red is a weak obstacle label, green a weak road label and blue is unknown. The classification result (b) and the ground truth (c).

Our experimental evaluation is based on 20 datasets with 123 manually annotated ground truth frames in total. The datasets are recorded at a resolution of 640-by-480 pixels at a frame rate of 30Hz. For a single experiment, we take a ground truth image $G_n$ and train the SVM on 60 preceding frames $I_{n-60} \ldots I_{n-1}$ and classify the current frame $I_n$. This classification is then compared pixel-wise against $G_n$, resulting in an error rate, i.e. the percentage of wrongly classified pixels in each frame. This per-frame error rate is averaged over all frames in a particular dataset, resulting in a per-dataset average error rate. An example of this classification process is shown in Fig. 3.

We use five different performance criteria to evaluate our system:

1. **Average error rate:** The average over the per-dataset average error rates. This assures that each dataset is equally weighted in the metric regardless of its number of ground truth...
frames. Frames for which the SVM did not converge, and which were not classified, are given an error of 100%.

2. **Average maximum error rate**: The average over the per-dataset maximum error rates. Frames for which the SVM did not converge are ignored.

3. **Average maximum false negative rate**: Similar as average maximum error rate but now only considering false negatives (road classified as obstacle).

4. **Average maximum false positive rate**: Similar as average maximum error rate but now only considering false positives (obstacle classified as road).

5. **Total number of unclassified frames**: Frames for which the SVM did not converge. This is an important measure for the robustness of the system.

### 3.3 SVM configuration

In the first experiment we compare a linear-SVM with a RBF-SVM on six randomly selected datasets. The two important parameters of the SVM are the box constraint (C) and the kernel width (σ) of the RBF. For these parameters, sweeps were performed and we only report the performance that was obtained with the optimal parameter values (C=90 and σ=30). The results of the RBF-SVM over several feature spaces are shown in Table II and that of the linear-SVM in Table III. It can clearly been seen that the RBF-SVM significantly outperforms the linear-SVM on multiple performance criteria.

### Table II: Results RBF-SVM (Sorted on average error rate)

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Average error rate (%)</th>
<th>Average maximum error rate (%)</th>
<th>Average maximum false negative rate (%)</th>
<th>Average maximum false positive rate (%)</th>
<th>Total number of unclassified frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS144</td>
<td>5.62</td>
<td>10.93</td>
<td>10.28</td>
<td>3.40</td>
<td>0</td>
</tr>
<tr>
<td>HS128</td>
<td>5.72</td>
<td>10.92</td>
<td>10.77</td>
<td>1.97</td>
<td>0</td>
</tr>
<tr>
<td>HSV96</td>
<td>6.05</td>
<td>13.96</td>
<td>13.95</td>
<td>1.96</td>
<td>0</td>
</tr>
<tr>
<td>HS-HOG96</td>
<td>7.71</td>
<td>12.82</td>
<td>11.69</td>
<td>2.50</td>
<td>1</td>
</tr>
<tr>
<td>HSV216</td>
<td>8.18</td>
<td>10.46</td>
<td>10.01</td>
<td>3.56</td>
<td>1</td>
</tr>
<tr>
<td>YIQ216</td>
<td>9.06</td>
<td>23.40</td>
<td>23.25</td>
<td>4.34</td>
<td>0</td>
</tr>
<tr>
<td>RGB216</td>
<td>10.18</td>
<td>24.11</td>
<td>22.23</td>
<td>6.77</td>
<td>1</td>
</tr>
<tr>
<td>IQ144</td>
<td>10.34</td>
<td>23.78</td>
<td>23.15</td>
<td>2.52</td>
<td>0</td>
</tr>
<tr>
<td>RGB96</td>
<td>21.40</td>
<td>19.68</td>
<td>18.36</td>
<td>3.77</td>
<td>5</td>
</tr>
</tbody>
</table>

### Table III: Results Linear-SVM (Sorted on average error rate)

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Average error rate (%)</th>
<th>Average maximum error rate (%)</th>
<th>Average maximum false negative rate (%)</th>
<th>Average maximum false positive rate (%)</th>
<th>Total number of unclassified frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-HOG96</td>
<td>11.90</td>
<td>11.60</td>
<td>11.60</td>
<td>1.78</td>
<td>3</td>
</tr>
<tr>
<td>HSV96</td>
<td>15.86</td>
<td>10.84</td>
<td>10.07</td>
<td>1.89</td>
<td>5</td>
</tr>
<tr>
<td>HSV216</td>
<td>17.97</td>
<td>19.30</td>
<td>19.06</td>
<td>2.24</td>
<td>5</td>
</tr>
<tr>
<td>HS144</td>
<td>19.51</td>
<td>9.51</td>
<td>9.29</td>
<td>2.91</td>
<td>7</td>
</tr>
<tr>
<td>HS128</td>
<td>19.90</td>
<td>13.46</td>
<td>13.42</td>
<td>1.87</td>
<td>7</td>
</tr>
<tr>
<td>RGB216</td>
<td>20.12</td>
<td>10.85</td>
<td>7.95</td>
<td>5.34</td>
<td>5</td>
</tr>
<tr>
<td>IQ144</td>
<td>39.32</td>
<td>19.10</td>
<td>19.03</td>
<td>1.51</td>
<td>15</td>
</tr>
<tr>
<td>RGB96</td>
<td>43.96</td>
<td>18.67</td>
<td>16.50</td>
<td>3.88</td>
<td>16</td>
</tr>
<tr>
<td>YIQ216</td>
<td>45.83</td>
<td>13.94</td>
<td>13.94</td>
<td>1.24</td>
<td>18</td>
</tr>
</tbody>
</table>

An interesting observation that can be made from Table II and Table III is that the best performing linear-SVM method has twice the average error rate of the best performing RBF-SVM method but the feature spaces are different. For the linear method, feature spaces that have more dimensions (e.g. both color and gradient) perform better. The RBF-SVM method performs better on feature spaces with fewer dimensions. A closer inspection reveals that the performance of the linear-SVM approach is harmed by the high number of frames for which it did not converge. For our system, robustness is as important as accuracy, so the total...
number of unclassified frames is an important performance criterion and makes the linear-
SVM approach unattractive.

Looking more in depth in the results of linear-SVM method, we see that the HOG based
(HS-HOG96) and intensity based features (HSV96 and HSV216) outperform the features
that do not use gradients or intensity. This shows that intensity and gradients are better
linearly separable than e.g. Hue and Saturation for our application domain. However, this
benefit is lost when using non-linear decision boundaries of RBF-SVM.

3.4 Feature sets

In our second experiment we focus on different feature spaces. We only discuss the results
obtained with the RBF-SVM that are provided in Table II. The results show that feature sets
with at least Hue and Saturation give the best results. Even so, only using Hue and
Saturation provides optimal results. It is interesting to note that the HS128 feature, that uses
a 1-D histogram with 128 bins in total, provides practically the same performance as the
HS144 feature, which uses a 2-D histogram with 144 bins. This indicates that modeling
dependencies between the Hue and Saturation subspaces adds little to no useful statistical
information for the classifier. The slightly better average error rate of HS144 compared to
HS128 is probably solely due to using more bins. Qualitative results obtained with using a
RBF-SVM and the HS144 feature space are provided in Fig. 4.

It can also be observed that gradient information used in the HS-HOG96 feature adds no
benefits when compared to the HSV96 feature for the RBF-SVM. Apparently, only using
Hue and Saturation and allowing for a non-linear separation boundary is better than using
gradient/intensity information either with a linear or a non-linear separation boundary. This
can be explained from the fact that gradient information is more sensitive to perspective
distortion than color information.

3.5 Histogram creation

In this experiment we evaluate the potential benefit of linear compared to adaptive histogram
creation strategies over all 123 ground truth frames. We use the Hue Saturation feature space
and evaluate 1-D and 2-D histograms with a RBF-SVM. Here we make sure that each tested
approach has an equal number of 100 histogram bins. The results are provided in Table IV.

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>RESULTS LINEAR VS. ADAPTIVE HISTOGRAMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions</td>
<td>Bin distribution</td>
</tr>
<tr>
<td>1</td>
<td>Linear</td>
</tr>
<tr>
<td>1</td>
<td>Adaptive</td>
</tr>
<tr>
<td>2</td>
<td>Linear</td>
</tr>
<tr>
<td>2</td>
<td>Adaptive</td>
</tr>
</tbody>
</table>

It can be seen that using adaptive bins has no advantage over linear bins. However, in
preliminary experiments where we used very few bins (10-25 bins), we observed that using
adaptive bins is advantageous. However, the absolute performance when using so few bins
is significantly worse than when using 100 bins. This indicates that using adaptive bins is
only useful when a system has strict limitations on computation resources.
5 Conclusions

We developed an online and self-supervised computer vision system that can aid intelligent vehicles in distinguishing obstacles from surfaces on which can be driven. Critical design choices regarding the used classifier and appearance models are evaluated. It is shown that optimal accuracy of 95.6% is achieved, when using a RBF-SVM classifier with 1-D color histograms containing Hue and Saturation. In future work, we will develop a real-time version of our system and extend it with methods that regularize the classification over time.

References

Comparison of Linear Prediction Methods in Terms of Sparsity, Stability and Robustness to Reverberation

Andreas I. Koutrouvelis  Richard Heusdens  Nikolay D. Gaubitch
Delft University of Technology
Faculty of EEMCS

Abstract

The aim of this paper is to provide an experimental evaluation of five linear prediction methods in terms of sparsity, stability and robustness to reverberation. Moreover, we show that all the considered methods can be derived from a general linear prediction optimization problem. It is empirically found that the sparsest and also one of the most stable and robust to reverberation method is linear prediction based on $L_1$ minimization combined with a double pre-emphasis of speech.

1 Introduction

Linear prediction (LP) analysis is used extensively in many applications of speech technology such as coding, recognition, speaker identification, epoch extraction, speaker localization or speech dereverberation [1–4]. LP analysis is based on the source filter model, in which a speech signal $s[n]$, $n = 0, ..., N - 1$, can be generated as the convolution of an excitation signal and a vocal tract filter [1, 5]. The excitation of voiced speech, the excitation of unvoiced speech, and the vocal tract filter are usually modeled as a quasi-periodic impulse train, a white Gaussian noise (WGN), and an all-pole filter, respectively. Based on the source-filter model assumption, a speech signal can be written using an auto-regressive model

$$s[n] = \sum_{k=1}^{p} a_k s[n - k] + A\epsilon[n],$$

where $\epsilon[n]$ is the excitation signal (also called prediction error sequence), $a_k$, $k = 1, ..., p$, are the linear prediction coefficients (LPCs) and $A$ is the filter gain which is commonly assumed to be unity [6–9].

LP analysis methods estimate the LPCs such that $\epsilon[n]$ is minimized. The resulting minimum $\epsilon[n]$, say $\hat{\epsilon}[n]$, is called the residual. The classical LP analysis method minimizes the variance (i.e., the squared $L_2$ norm) of $\epsilon[n]$, and therefore, results in a linear least-squares estimator (LLSE) [10]. The LLSE estimator works well for unvoiced speech, and it is equivalent to the maximum likelihood estimator (MLE) if the excitation is indeed WGN [11]. On the other hand, in voiced speech the excitation consists of quasi-periodic strong peaks and the LLSE suffers from outliers, i.e., it overemphasizes the large errors and puts less emphasis on smaller errors [10], producing a non-spiky residual. The property of spikiness on just a few samples and all the other samples to have small values, is also called sparsity. Thus, in case of voiced speech, a desired property for an LP estimator is to estimate the LPCs such that the residual is sparse. The methods that follow this philosophy are called sparse LP methods.

Sparsity can be measured with various metrics [12]. The $L_0$ quasi-norm is one such metric and is defined as $\|x\|_0 = |\{i, x_i \neq 0\}|$, which is the number of non-zero elements of the vector $x = [x[0], x[1], ..., x[M - 1]]^T$. The Gini measure [12] is another sparsity
measure that satisfies all the properties desirable for a sparsity measure [12] (i.e., Robin Hood, scaling, rising tide, cloning, Bill Gates and babies), and is defined as

$$Gini(x) = 1 - 2 \sum_{m=1}^{M} \frac{x_o(m)}{\|x_o\|_1} \left( \frac{M - m + 1/2}{M} \right),$$

where $x_o$ is the sorted version of $x$ in increasing order. The Gini measure takes values in the interval [0,1]. These two sparsity measures, however, are not convex and are not efficient to optimize with sparse LP analysis methods. Instead, the $L_1$-norm is often used [7, 8, 13–15] which is an approximation of the $L_0$ norm. The estimator that minimizes the $L_1$-norm of the prediction error sequence is called the least absolute deviations (LAD) estimator and is equivalent to the MLE estimator when the prediction error sequence follows a Laplacian distribution [16].

Sparsity as a property to LP estimators in voiced speech is very useful in many applications such as LP coding [1, 7], epoch extraction [2] and speaker localization [3]. In the first application, it is desired to transmit as little information of the speech residual as possible without discriminating the ineligibility and perceptual quality of speech. The aim of the second application is to estimate the glottal closure instants (GCIs) and the pitch periods through the residual. To achieve a good performance, it is necessary for the residual to have strong epochs coinciding on the glottal closure instants and all the other samples to be very small. In the third application, the speech signal is acquired from multiple microphones and the location of the speaker is estimated by using the time difference of arrivals (TDOAs) of the strong epochs of all residuals. Moreover, stability (i.e., the poles lie inside the unit circle) of the LP estimated filters is essential to these applications. Apart from the classical LP autocorrelation method which is guaranteed to be stable, a common strategy to guarantee stability is to reflect the poles that occur outside the unit circle inside the unit circle [17]. This heuristic modification clearly changes the residual and might effect its degree of sparsity [a].

The performance of these three applications is degraded under the existence of reverberation caused by reflected surfaces. As a consequence, dereverberation techniques as a front-end of these applications might increase their performance. In particular, a dereverberation method [4] which was based on the beamforming of the Hilbert envelopes of the speech residuals, acquired from the differently spaced microphones, was used in order to estimate a weighting function. This weighting function was applied on the reverbered residuals, to emphasize the true epochs of these residuals and to smooth all the other values including the secondary peaks caused by reverberation. Therefore, it is interesting for future research to explore if the dereverberation method performs more accurately for sparser residuals. In order to help this future investigation, five LP methods [6, 13–15] are evaluated with and without reverberation in terms of sparsity and stability. To the authors best knowledge, there has not been any previous experimental comparison in terms of sparsity, stability and robustness to reverberation with all these methods.

The rest of the paper is organized as follows. In Section 2, we present the general LP problem and its configuration for each LP method. In Section 3, the experimental evaluation is demonstrated. Concluding remarks are given in Section 4.

## 2 Linear Prediction Methods

LP analysis estimates the LPCs by minimizing the prediction error sequence $\epsilon[n]$. In [7, 13, 14], there was a previous effort to generalize the LP problem. Here we propose a slightly different formula which covers all the non-constrained LP methods that we present in this paper [6–8, 10, 13–15] and many others which are not included in this
The proposed general LP optimization problem is given by
\[
\hat{a}, \hat{\epsilon} = \arg \min_a \|W\epsilon\|^p = \arg \min_a \|W(s_0 - Sa)\|^p,
\] (2)
where \(W\) is a weighting matrix, \(\|\cdot\|^p\) is the \(L_p\) norm and \(\epsilon = s_0 - Sa\) is the prediction error vector. Moreover, \(s_i\) and \(S\) are given by
\[
s_i = [s[N_a - i] \ldots s[N_b - i]]^T,
\]
\[
S = [s_1 s_2 \ldots s_p],
\]
where if \(N_a = 0\) and \(N_b = N + p - 1\), then (2) is set to the autocorrelation mode, and if \(N_a = p\) and \(N_b = N - 1\), to the covariance mode. Furthermore, \(N_a\) and \(N_b\) may change in each speech signal frame if we use a closed-phase LP method [18]. In this paper we consider only the autocorrelation mode for all LP methods and, therefore, omit to mention it in the following descriptions of the methods. This choice is based on the fact that the covariance mode could result in more unstable all-pole filters for LP methods based on the \(L_1\) minimization [13]. Moreover, pre-emphasis is applied to speech before LP analysis by using the filter
\[
D(z) = 1 - \alpha z^{-1},
\] (3)
where \(\alpha = 0.99\) in this paper. We observed that pre-emphasis increases slightly the degree of sparsity of the residual. In one of the five methods the speech signal is doubly pre-emphasized (see method D). We compare the following LP methods.

A. Classical \(L_2\) LP Method (LP2)

The classical LP (LP2) method [10] has the configuration \((W = I, p = 2)\), where \(I\) is the identity matrix. Although, it does not produce optimally sparse residuals in the case of voiced speech, it is used in many speech processing applications due to its proven stability and its efficient implementation using the Levinson-Durbin algorithm [19].

B. Sparse \(L_1\) LP Method (SLP1)

The sparse \(L_1\) LP (SLP1) method, with configuration \((W = I, p = 1)\), is equivalent to the LAD estimator [16]. To the authors knowledge, the application of the LAD estimator in speech was first explored in [8]. In particular, the authors of [8] used the Burg algorithm in order to solve the \(L_1\)-problem. Although the stability of this algorithm is proven, it behaves somewhere in between the \(L_2\) and the \(L_1\) minimization in terms of sparsity [13]. For this reason, in [13] the \(L_1\)-problem is solved with an interior point algorithm which gives sparser residuals and a low (approximately 2%) percentage of unstable frames. In this paper we consider the later approach.

C. Iterative Sparse Weighted \(L_1\) LP Method (ISWLP1)

The iterative sparse weighted \(L_1\) LP (ISWLP1) method [14,20] has the configuration \((W^{(k)}, p = 1)\), where \(W^{(k)}\) is a diagonal matrix and its diagonal elements \(\{w_{ii}\}\) are updated in each iteration \(k\) as
\[
w_{ii}^{(k)} = \frac{1}{|\hat{\epsilon}^{k-1}[i]| + c},
\]
where \(\hat{\epsilon}^{k-1}\) is the residual of the previous iteration and the constant \(c\) is selected to be in the order of the expected nonzero magnitude of \(\epsilon\) [20]. ISWLP1 is solved iteratively.
until $\|\epsilon\|_1$ becomes smaller than a threshold or until a maximum number of iterations is reached. Note that in the first iteration $k = 0$ and $W^{(0)} = I$.

D. Iterative Sparse Second Derivative $L_1$ LP Method (ISSDLP1)

The iterative sparse second-derivative $L_1$ LP (ISSDLP1) method [15] has the configuration $(W^{(k)}, p = 1)$, where $W^{(k)}$ is the same diagonal matrix as in the ISWLP1 method. The only difference with SLP1 for $k = 0$ and with ISWLP1 for $k > 0$ is that instead of pre-emphasizing the speech signal with one filter (3), it passes the speech signal through this filter two times before LP analysis. As in the ISWLP1 method, the same stopping criteria is used. This method also has another very good property: if the estimated residual is passed through the inverse of the double pre-emphasis filter, then we get an estimate of the glottal flow derivative [15] without first estimating the closed phase interval of the glottal flow as in [18].

E. Weighted $L_2$ Sparse LP Method (WSLP2)

The weighted $L_2$ sparse LP (WSLP2) method [6] has the configuration $(W, p = 2)$. $W$ is a diagonal matrix which de-emphasizes the large errors in GCIs. The diagonal of this matrix is the square root of the weighting function given by

$$w[n] = 1 - \sum_{k=1}^{N_{gc}} g[n - n_k],$$

where $n_k$ are the GCI positions in $x[n]$ which are determined via the SEDREAMS algorithm [21]. The function $g[n]$ is a Gaussian function, i.e., $g[n] = \kappa e^{-n^2/\sigma^2}$.

3 Experimental Results

In this section we evaluate the five methods reviewed in Section 2. We use 3424 voiced-speech frames taken from 5 males and 5 females from the APLAWD database [22], with time duration 32 ms each. The sampling frequency is $f_s = 8$ KHz and the LP order is $p = 12$. We evaluate the ISWLP1 and the ISSDLP1 methods for $iter = 3, 5$ and $iter = 1, 3, 5$ respectively, where $iter$ is the maximum allowable number of iterations. The WSLP2 method is evaluated using $\sigma = 5, 10, 15$, and the same $\kappa = 0.9$. All $L_1$ optimization problems are solved using the primal-dual interior point algorithm of the 11-magic toolbox [23]. For consistency, the average Gini values for each method are computed over the maximum possible intersected set of frames that are stable for all methods. The experiments are carried out in a reverberant and a non-reverberant environment in order to test the robustness of these methods in terms of stability and sparsity. The source-image method [24, 25] is used for simulating a reverberant room with dimensions 6x5x4 m. The reverberation time $T_{60}$ varies between 0.2 s and 0.8 s with steps of 0.1 s. Ten different talker-microphone position pairs are used for the evaluation and they are selected randomly in the inner concentric room box of 5x4x3 m.

Tables 1 and 2 show the average Gini values per frame and probabilities of stability (i.e., the ratio (number of stable frames)/(total number of frames=3424)) for each method. A frame is unstable when its corresponding estimated filter has at least one pole outside the unit circle [1]. It is clear from Table 1 that the sparsest method is ISSDLP1. In particular, it is slightly sparser than IWSLP1 for $iter = 3, 5$ and SLP1 for $iter = 1$. It is also more stable than IWSLP1 and SLP1. Except of LP2 which is guaranteed to be stable, the next most stable methods are WSLP2 and ISSDLP. It is
worth noting that reverberation does not decrease stability. In case of reverberation, all sparse LP methods remain significantly sparser than the LP2 method except of the WSLP2 method which is almost as sparse as the LP2 method. This is due to the fact that the performance of the WSLP2 method is connected to the GCI estimation algorithm (i.e., in this case the SEDREAMS algorithm).

Figure 1 depicts a speech frame without reverberation and the corresponding residuals of the five LP methods for two different prediction orders $p = 8$ and $12$. It is clear again that the ISSDLp method is the sparsest one for both orders. Especially for $p = 8$ it is much sparser than all the other methods. This is a very important observation especially for applications in which the computational complexity, which depends on the filter order, should be low. Finally, the increased sparsity in the cases of ISWLP1 and ISSDLp1 methods compared to the the SLp1 and ISSDLp1 with one iteration methods, respectively, is not noticeable.

<table>
<thead>
<tr>
<th>Method</th>
<th>$T_{60}$</th>
<th>0</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP2</td>
<td></td>
<td>0.488</td>
<td>0.433</td>
<td>0.425</td>
<td>0.422</td>
<td>0.421</td>
<td>0.420</td>
<td>0.419</td>
<td>0.419</td>
</tr>
<tr>
<td>SLP1</td>
<td></td>
<td>0.540</td>
<td>0.470</td>
<td>0.461</td>
<td>0.457</td>
<td>0.456</td>
<td>0.456</td>
<td>0.455</td>
<td>0.454</td>
</tr>
<tr>
<td>IWSLP1, iter=3</td>
<td></td>
<td>0.551</td>
<td>0.477</td>
<td>0.468</td>
<td>0.466</td>
<td>0.464</td>
<td>0.464</td>
<td>0.463</td>
<td>0.462</td>
</tr>
<tr>
<td>IWSLP1, iter=5</td>
<td></td>
<td>0.553</td>
<td>0.478</td>
<td>0.470</td>
<td>0.468</td>
<td>0.466</td>
<td>0.466</td>
<td>0.465</td>
<td>0.464</td>
</tr>
<tr>
<td>ISSDLp1, iter=1</td>
<td></td>
<td>0.546</td>
<td>0.472</td>
<td>0.461</td>
<td>0.457</td>
<td>0.457</td>
<td>0.456</td>
<td>0.455</td>
<td>0.455</td>
</tr>
<tr>
<td>ISSDLp1, iter=3</td>
<td></td>
<td>0.555</td>
<td>0.478</td>
<td>0.468</td>
<td>0.465</td>
<td>0.464</td>
<td>0.464</td>
<td>0.463</td>
<td>0.463</td>
</tr>
<tr>
<td>ISSDLp1, iter=5</td>
<td></td>
<td><strong>0.557</strong></td>
<td><strong>0.480</strong></td>
<td><strong>0.470</strong></td>
<td><strong>0.467</strong></td>
<td><strong>0.466</strong></td>
<td><strong>0.466</strong></td>
<td><strong>0.465</strong></td>
<td><strong>0.465</strong></td>
</tr>
<tr>
<td>WSLP2, $\sigma = 5$</td>
<td></td>
<td>0.503</td>
<td>0.437</td>
<td>0.428</td>
<td>0.423</td>
<td>0.422</td>
<td>0.421</td>
<td>0.420</td>
<td>0.420</td>
</tr>
<tr>
<td>WSLP2, $\sigma = 10$</td>
<td></td>
<td>0.504</td>
<td>0.437</td>
<td>0.428</td>
<td>0.423</td>
<td>0.421</td>
<td>0.421</td>
<td>0.420</td>
<td>0.419</td>
</tr>
<tr>
<td>WSLP2, $\sigma = 15$</td>
<td></td>
<td>0.502</td>
<td>0.437</td>
<td>0.428</td>
<td>0.423</td>
<td>0.421</td>
<td>0.421</td>
<td>0.420</td>
<td>0.419</td>
</tr>
</tbody>
</table>

Table 1: Average Gini values for various reverberation times $T_{60}$. Bold numbers indicate the largest Gini value per column.

<table>
<thead>
<tr>
<th>Method</th>
<th>$T_{60}$</th>
<th>0</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLP1</td>
<td></td>
<td>0.987</td>
<td>0.990</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td>IWSLP1, iter=3</td>
<td></td>
<td>0.963</td>
<td>0.983</td>
<td>0.986</td>
<td>0.986</td>
<td>0.987</td>
<td>0.986</td>
<td>0.985</td>
<td>0.987</td>
</tr>
<tr>
<td>IWSLP1, iter=5</td>
<td></td>
<td>0.957</td>
<td>0.980</td>
<td>0.982</td>
<td>0.983</td>
<td>0.983</td>
<td>0.982</td>
<td>0.982</td>
<td>0.983</td>
</tr>
<tr>
<td>ISSDLp1, iter=1</td>
<td></td>
<td><strong>0.998</strong></td>
<td><strong>0.992</strong></td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>ISSDLp1, iter=3</td>
<td></td>
<td>0.992</td>
<td>0.983</td>
<td>0.984</td>
<td>0.983</td>
<td>0.986</td>
<td>0.984</td>
<td>0.985</td>
<td>0.986</td>
</tr>
<tr>
<td>ISSDLp1, iter=5</td>
<td></td>
<td>0.992</td>
<td>0.980</td>
<td>0.981</td>
<td>0.979</td>
<td>0.983</td>
<td>0.980</td>
<td>0.981</td>
<td>0.981</td>
</tr>
<tr>
<td>WSLP2, $\sigma = 5$</td>
<td></td>
<td>0.996</td>
<td><strong>0.992</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.996</strong></td>
<td><strong>0.996</strong></td>
<td><strong>0.996</strong></td>
<td><strong>0.997</strong></td>
<td><strong>0.997</strong></td>
</tr>
<tr>
<td>WSLP2, $\sigma = 10$</td>
<td></td>
<td>0.995</td>
<td>0.983</td>
<td>0.988</td>
<td>0.991</td>
<td>0.991</td>
<td>0.992</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>WSLP2, $\sigma = 15$</td>
<td></td>
<td>0.995</td>
<td>0.984</td>
<td>0.988</td>
<td>0.990</td>
<td>0.991</td>
<td>0.991</td>
<td>0.991</td>
<td>0.992</td>
</tr>
</tbody>
</table>

Table 2: Probabilities of Stability for different reverberation times $T_{60}$. Bold numbers indicate the largest probability of stability per column (not including the LP2 method).
4 Conclusion and Future Work

In this paper we compared the sparsity and stability of five different LP methods which are special cases of a general LP optimization problem. It was empirically shown that the iterative sparse second-derivative LP (ISSDLP) method was the sparsest method and also very stable (less than 0.5% of the frames are unstable) with and without reverberation. We also showed that the ISSDLP method retains its sparsity for low filter orders which means that it is proper for low complexity applications such as LP coding with low power consumption. The WSLP2 method does not maintain its increased sparsity compared to the LP2 method in reverberant environments due to its connection to the GCI estimation algorithm. Generally, none of these methods gave more than 5% unstable filters. Depending on the application, this very low percentage of unstable filters can be handled either by reflecting the poles outside the unit circle into the unit circle, or by simply rejecting these frames (e.g. in speaker localization).

It seems desirable to further investigate the application of the ISSDLP method in applications such as speaker localization, speech dereverberation, and epoch extraction due to the strong peaks of its residual. Moreover, the ISSDLP method should be compared with the sparse LP methods of the work in [7] in the context of speech LP coding.

Figure 1: Residuals of various LP methods.
References


Compressed Sensing Using Adaptive Sparse Measurements

Dejan Vukobratovic  
DEET-FTN  
University of Novi Sad  
Serbia  
dejanv@uns.ac.rs

Aleksandra Pizurica  
TELIN  
Ghent University  
Belgium  
sanja@telin.ugent.be

Abstract

Compressed sensing (CS) using sparse measurement matrices and iterative message-passing reconstruction algorithms have been recently investigated as a low-complexity alternative to traditional CS methods. In this paper, we investigate the adaptive version of well-known Sudocodes scheme, where the sparse measurement matrix is progressively created based on the outcomes of previous measurements. Inspired by resemblance with rateless coding, we provide a detailed analysis of the adaptive Sudocodes approach in combination with the verification-based LMz reconstruction. The results show that the adaptivity is a promising feature for reducing complexity and improving performance of CS methods based on sparse measurement matrices.

1 Introduction

Recovery of strictly sparse signals using sparse measurement matrices and iterative message-passing reconstruction algorithms was first discussed by Sarvotham et al [1]. The reconstruction algorithm applied in their Sudocodes scheme was later identified to be a version of the verification decoding algorithm (LM2) proposed for the decoding of Low-Density Parity Check (LDPC) codes [2]. Inspired by the iterative Belief-Propagation (BP) decoding of LDPC codes, the same authors extended the Sudocodes scheme into a more general framework called Compressed Sensing via Belief Propagation (CSBP) [3], that triggered significant interest in CS methods using sparse measurement matrices by both compressed sensing and sparse-graph coding community.

CSBP is a powerful solution that introduces factor graph modeling and BP reconstruction to recover sparse or approximately sparse signals from noisy measurements. However, CSBP is rather complex both from the analysis and implementation perspective, since the messages exchanged across factor graph represent continuous probability distributions. As a simplification of CSBP amenable to fast implementation and rigorous analysis, verification-based decoding algorithms known as LM1 and LM2 algorithm [2] for noise-free CS recovery of strictly sparse signals have been investigated [4,5]. A related class of reconstruction algorithms are so called Interval-Passing Algorithms (IPA) proposed in [6] and recently analyzed using coding-theoretic tools [7].

The approaches mentioned above apply non-adaptive design of measurement matrices. Recently, benefits of CS methods with adaptive measurement matrices have been demonstrated (see [8–12] and the references therein). In an adaptive scenario, the measurements are constructed sequentially making use of the feedback from the observations, such that the sensing energy can be focused on the suspected non-zero components. Most of the reported adaptive CS approaches focus on improving the performance of the support recovery in the presence of noise (reducing the required signal to noise ratio), either by using adaptive Gaussian random matrices [9], collections of independent structured random matrices [10] or repeated bisection of the signal support [11,12]. Much less work has been done on exploring the benefits of adaptive designs for the reduction of the number of measurements. The results in [8] indicate great potentials in this respect. To the best of our knowledge, adaptive designs for the CS schemes based on message passing, such as Sudocodes or generalizations...
thereof have not been reported yet. In this paper, we introduce and analyze an adaptive version of Sudocodes scheme with LM1 recovery. We provide in-depth analysis of this scheme using tools borrowed from analysis of rateless codes. Our results demonstrate that the adaptivity represents a promising feature for further improvement of CS methods based on sparse measurement matrices.

2 Adaptive CS System Model

We observe a setup where the goal is to recover a strictly sparse signal $x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$ containing $K \ll N$ non-zero components (the remaining $N - K$ being zero) from a set of measurements $y = (y_1, y_2, \ldots, y_M) \in \mathbb{R}^M$ where $K < M \ll N$. The CS system model comprises the adaptive measurement subsystem, communication channel and the reconstruction subsystem.

The adaptive measurement subsystem sequentially samples the signal $x$ using a sequence of measurement vectors $\phi_i$, $1 \leq i \leq M$, in order to produce a stream of measurements $y$, where $\phi_i$ is a sparse vector of length $N$ and $y_i = \phi_i \cdot x^T$. The number of measurements $M$ is not fixed in advance and may be arbitrarily extended. Measurement vectors are adaptively designed, i.e., $\phi_i = f(\{\phi_1, \ldots, \phi_i\})$. The measurement vector $\phi_i$ is defined by the degree $d_i$: the number of non-zero elements where, unless otherwise stated, the positions of $d_i$ non-zero values are selected uniformly at random among the elements of $x$. Overall, the measurement process can be described as $y = \Phi \cdot x^T$, where the measurement matrix $\Phi$ is a sparse $M \times N$ matrix.

The measurements $y_i$ are sequentially communicated to the reconstruction subsystem via a communication channel. We consider an ideal noise-free channel that conveys real numbers without errors/erasures, leaving considerations of imperfect channels and quantization for future work.

For the reconstruction subsystem, we consider iterative verification-based LM1 and LM2 algorithms [2, 4]. Both algorithms exchange messages across a measurement graph, which is a bipartite graph consisting of $N$ coefficient nodes that correspond to signal coefficients $x$ and $M$ measurement nodes that correspond to a sequence of measurements $y$. Edges of the graph connect each measurement node $y_i$ to its neighbor set $N(y_i)$ of coefficients determined by non-zero positions of the measurement vector $\phi_i$. Thus the degree $d_i = |N(y_i)|$ is the number of edges incident to $y_i$. Edges in the graph are in one-to-one correspondence with non-zero entries of the measurement matrix $\Phi$. If the non-zero entries of $\Phi$ are arbitrary reals, then the graph is weighted and the edge weights correspond to real entries of $\Phi$. For simplicity, we assume binary matrix $\Phi$, thus weights are not needed. Below, we briefly describe LM1 and LM2 algorithms noting that we assume they operate progressively with the arrival of each new measurement until the signal is fully recovered (i.e., all coefficients are verified).

**LM1**: The LM1 operates iteratively over the measurement graph as follows: 1) If $y_i = 0$ then $\forall x_j \in N(y_i): x_j = 0$; Verify all $x_j: x_j \in N(y_i)$. 2) If $y_i \neq 0$ and $(|N(y_i)| = 1)$ then $x_j = y_i$ for the node $x_j \in N(y_i)$; Verify $x_j$. 3) Remove verified coefficient nodes and their incident edges from the graph; Subtract out verified values from remaining measurements. 4) Repeat until the LM1 successfully recovers the signal or does not progress in two consecutive iterations.

**LM2**: Besides the above LM1 rules, LM2 adds the additional one: If $(N(y_i) \cap N(y_j) = \{x_k\}) \wedge (y_i = y_j)$ then $x_k = y_i = y_j$ and $\forall x_i \in \{N(y_i) \cup N(y_j) \setminus x_k\}: x_i = 0$; Verify all $x_i \in \{N(y_i) \cup N(y_j) \setminus x_k\}$.

Note that, in the above setup, if $\phi_i$’s are created non-adaptively using fixed degree $d_i = L$, we obtain the (first stage of the) Sudocodes scheme. Furthermore, if $x \in \mathbb{F}_2^N$, $K = N$ (i.e., the signal is non-sparse), $\phi_i$’s are created independently (non-adaptively) using degree’s $d_i$ drawn from a given degree distribution $\Omega(x)$, the channel is $2^q$-ary erasure channel, and we apply LM1 decoder, we obtain standard rateless (LT) coding scenario [13].
3 Asymptotic Analysis

Using sparse-graph coding methodology, the measurement graph may be described using coefficient and measurement node degree distributions

$$\Lambda(x) = \sum_{i=1}^{d_{\max}} \Lambda_i \cdot x^i$$

and

$$\Omega(x) = \sum_{i=1}^{d_{\max}} \Omega_i \cdot x^i,$$

where \( \Lambda_i \) and \( \Omega_i \) are the fraction of coefficient and measurement nodes of degree \( i \), respectively, while \( d_{\max}^{(c)} \) and \( d_{\max}^{(m)} \) are maximum coefficient and measurement node degrees. It is also useful to define so called edge-oriented degree distributions

$$\lambda(x) = \sum_{i=1}^{d_{\max}} \lambda_i \cdot x^{-1} = \mathcal{N}(x)/\mathcal{N}(1)$$

and

$$\omega(x) = \sum_{i=1}^{d_{\max}} \omega_i \cdot x^{-1} = \mathcal{O}(x)/\mathcal{O}(1).$$

For a given measurement graph, one can (asymptotically) analyze probabilities of reconstruction of signal coefficients using well-established tools from coding theory. For LM reconstruction algorithm, the iterative “graph-peeling” process is the same as the rateless decoding over erasure channels, with the only difference being signal sparsity and existence of zero-valued measurements. Below, we provide and-or-tree analysis of non-adaptive Sudocodes CS scheme in combination with LM1 recovery:

Lemma 3.1 Let \( p^{(z)}_l \) and \( p^{(z)}_l \) denote the probabilities that a zero and non-zero signal coefficient, respectively, is not recovered after \( l \) iterations of LM1 recovery. Then

$$p^{(z)}_l = \lambda \left( 1 - \sum_{i=1}^{d_{\max}^{(m)}} \sum_{j=0}^{i-1} \binom{i-1}{j} (\alpha p^{(z)}_{l-1})^j \cdot (\bar{\alpha} p^{(z)}_{l-1})^{i-1-j} \right)$$

and

$$p^{(z)}_l = \lambda \left( 1 - \sum_{i=1}^{d_{\max}^{(m)}} \sum_{j=0}^{i-1} \binom{i-1}{j} (\alpha p^{(z)}_{l-1})^j \cdot \bar{\alpha}^{i-1-j} \right),$$

where \( \alpha = K/N \) is sparsity-factor, we use compact notation \( \bar{x} = 1 - x \), and the recursion is initialized at \( p^{(z)}_0 = p^{(z)}_0 = 1 \). Finally, \( p_l = \alpha p^{(z)}_l + \bar{\alpha} p^{(z)}_l \) is the average probability that a signal coefficient is unrecovered after \( l \) iterations of LM1.

Proof: The proof follows directly from and-or-tree analysis [14] by exhaustively analyzing all cases when zero and non-zero valued coefficient nodes are recovered.

*The above Lemma is exact and improves over an approximate version provided in [15] (Lemma 2), which does not exhaustively cover all recovery scenarios for zero-valued nodes.
Example 3.1 Fig. 1 shows asymptotic recovery probabilities (as $N \to \infty$) obtained from Lemma 1 for Sudocodes scheme that applies $\Omega(x) = x^{20}$ after LM1 recovery of input message of sparsity-factor $\alpha = 0.05$ for zero and non-zero input symbols and the average value.

In CS, similarly as in rateless coding, the measurement subsystem can control only $\Omega(x)$, while $\Lambda(x)$ depends on how signal coefficients are sampled to participate in each measurement. If this selection is uniformly at random, as in the case of Sudocodes (as well as rateless codes), then $\Lambda(x)$ asymptotically tends to the Poisson distribution, and $\Lambda(x) = e^{\mu(x-1)}$, where $\mu = \Omega(1)$ is the average measurement node degree, and $\epsilon = M/N$ is known as the reception overhead in rateless coding, while in CS, $\epsilon$ is the product of the oversampling ratio $\theta = M/K$ and the sparsity factor $\alpha = K/N$.

4 Adaptive Sparse Measurements

In the following, we propose usage of adaptive sparse measurements in Sudocodes scheme and analyze its performance in combination with the LM1 recovery. The adaptive version of Sudocodes relies on two simple modifications to the measurement subsystem:

Modification 1: If the measurement subsystem records a measurement $y_i = 0$, the signal coefficients $x_j \in N(y_i)$ are not considered in following measurements. Consequently, by eliminating known zeros, the size of the problem $N$ decreases and the sparsity-factor $\alpha = K/N$ increases as the measurement process evolves.

Modification 2: As the sparsity-factor $\alpha$ changes, the measurement subsystem selects the optimal degree $d^* = d^*(\alpha)$ of each measurement row-vector $\phi_i$ that maximizes the expected number of zero-valued signal coefficients that will be recovered by the following measurement.

For a given sparsity-factor $\alpha = K/N$, the optimal degree $d^*$ is obtained as follows. A measurement of degree $d$ is a zero-measurement with probability $P_0 = \binom{N-K}{d}/\binom{N}{d}$. The expected number of zero-valued signal coefficients recovered by a measurement of degree $d$ is $n_0 = P_0 \cdot d$. The optimal strategy selects $d = d^*$ that maximizes $n_0$:

$$d^* = \arg \max_d \{n_0\} = \arg \max_d \left\{ \frac{(N-K)}{d} \cdot d \right\}. \tag{2}$$

For a fixed and small $\alpha$, using Stirling approximation\textsuperscript{1} of binomial coefficients that holds asymptotically as $N \to \infty$, and taking the logarithm of the argument of maximization (which does not change the optimal $d$), we obtain:

$$\log(n_0) = d \log\left( \frac{N-K}{d} - \frac{1}{2} \right) - d \log\left( \frac{N}{d} - \frac{1}{2} \right) + \log d \tag{3}$$

$$= d \log\left( \frac{2N(1-\alpha) - 1}{2N - 1} \right) + \log d \tag{4}$$

$$\approx d \log(1 - \alpha) + \log d. \tag{5}$$

Taking partial derivative of the above with respect to $d$ and making it equal to zero, we obtain:

$$d^* \approx \left( \log \frac{1}{1 - \alpha} \right)^{-1}, \tag{6}$$

where the approximation\textsuperscript{2} is asymptotically tight for small $\alpha$ and $N \to \infty$. The optimal degree $d^*$ asymptotically depends only on the sparsity-factor $\alpha$. In addition, $d^*$ calculated using (6) closely matches the one calculated by (2) for finite $N, K$.

\textsuperscript{1}$\log(\binom{n}{d}) \approx k \cdot \log(n/k - 0.5) + k - 0.5 \log(2\pi k)$.

\textsuperscript{2}We note that the above result is derived in [16](Lemma 2) in order to optimize the (non-adaptive) Sudocodes scheme. We rederive it here using different approximation that we use in the sequel of the paper.
To analyze the proposed adaptive Sudocodes scheme, we follow the evolution of the measurement matrix $\Phi$ over the measurement process. Starting from the initial sparsity-factor $\alpha_0 = K/N$, the first measurements are performed using the degree calculated from (6) and rounded to the nearest integer $d_0^* = \lfloor d^*(\alpha_0) \rfloor$. As the measurement system removes known zero-coefficients during the measurement process, the process passes through the sequence of increasing $\alpha$-values $\{\alpha_1, \alpha_2, \ldots\}$ at which the optimal $d^*$-values decrement, resulting in the corresponding set $\{d_1^*, d_2^*, \ldots\}$. The total of $m_i$ measurements are generated using the degree $d_i^*$ during which the sparsity-factor is increased from $\alpha_i$ to $\alpha_{i+1}$. The evolution of sparsity-factor $\alpha = K/N(m)$ with the number of measurements $m, m \geq 1$, can be described recursively:

\[ N(m + 1) = N(m) - n_0^* = N(m) - \left( \frac{N(m) - K}{d^*} \right) \cdot d^*. \tag{7} \]

Replacing (6) into (5), we obtain the approximation of the second term in (7), which, after dividing (7) by $K$ results in recursive evolution of $\alpha(m)$:

\[ \alpha^{-1}(m + 1) = \alpha^{-1}(m) - \frac{1}{K} e^{-(1+\log(\log(1/\alpha(m))))}. \tag{8} \]

Although $m_i$ can be implicitly obtained from (8) by counting the number of recursions in the interval $(\alpha_i, \alpha_{i+1})$, it can be approximated by assuming that $\alpha_i$ remains constant during generation of degree-$d_i^*$ measurements:

\[ m_i = \frac{K(\alpha_{i+1} - \alpha_i)}{\alpha_i^2 e^{-(1+\log(\log(1/\alpha_i)))}}. \tag{9} \]

We approximate the evolution of $\Phi$ during the measurement process as follows. Firstly, observe a sequence of matrices $\Phi_{d_i^*}, i \geq 0$ of dimension $m_i \times N_i$, where $N_i = K/\alpha_i$, that represent groups of measurements of the same degree $d_i^*$. Then, we define the sequence of measurement matrices $\Phi_i, i \geq 0$, where $\Phi_i$ is obtained from the set of $\Phi_{d_i^*}$, $0 \leq i \leq l$ by removing the columns corresponding to all the zero-valued coefficients identified after the first $M_i = \sum_{i=0}^{l} m_i$ measurements. This is shown in Fig. 2, where the columns are sorted so that the rightmost $K$ columns correspond to the non-zero signal coefficient. In addition, starting from the first column, the zero-valued coefficients are sorted in the order they are identified by a sequence of measurements.

The $\alpha_i$ values can be obtained from (6) as $\alpha_i = 1 - e^{-1/d_i^{(th)}}$, where $d_i^{(th)} = d_{i-1}^* + \frac{d_i^*}{2}$.
The degree distributions of $\Phi_t$ sequence are approximated as follows. In the non-adaptive case, $\Omega(x)$ tends to the Poisson distribution with the average value $\mu = \mu M/N = \mu M/N = M_p$, where $p$ is the probability that a coefficient node is selected by a measurement. For the adaptive case, $p$ is a function of $\alpha$:

$$p(\alpha) = \frac{d(\alpha)}{N(\alpha)} = \frac{\alpha}{K} \cdot \frac{1}{\log \frac{1}{1-\alpha}}. \quad (10)$$

However, by Taylor expansion of $\log(1 + x)$ around $x \approx 0$ and neglecting higher-order terms, one obtains $\log(1 + x) \approx x$, thus for large interval around $\alpha > 0$, $p$ behaves as $p \approx 1/K$. Thus we approximate $\Omega(x)$ of the matrix $\Phi_t$ as a Poisson distribution with the mean value $M_t p = M_t / K$. Regarding $\Lambda(x)$ distribution of $\Phi_t$, it is a mixture of $l$ hypergeometric distributions $\mathcal{H}(N_i, \bar{N}_i, d_i^*)$, $1 \leq i \leq l$, weighted by their relative number of measurements $n_i/M_i$.

Finally, we have all ingredients to track the recovery probability of adaptive Sudocodes scheme by replacing parameters of the sequence of $\Phi_i$’s into Lemma 3.1. Note from the equations in this section that the entire analysis depends only on initial and final sparsity-factors $\alpha_0$ and $\alpha_l$ and the number of non-zero valued coefficients $K$.

5 Numerical and Simulation Results

We illustrate the performance of Sudocodes (SC) and adaptive-SC (aSC) scheme using asymptotic analysis following Lemma 3.1 and equations in Sec. 4. For SC scheme and the signal of sparsity factor $\alpha = 0.05$, the optimal degree $d^* = 20$. This is observable from Fig. 3 where we present recovery probability $P_r = 1 - p_i \to \infty$ for different values of $d = \{5, 10, 15, 20, 30\}$ as a function of the reception overhead $\epsilon$. The SC performance curves hold asymptotically as $K, N \to \infty$. For finite $K = 50$ ($N = 1000$), in Fig. 3 we provide an example curve obtained by simulations (we note that simulated curves converge to the asymptotic one for large $K$).

For aSC scheme, we start from $\alpha_0 = 0.05$ and adaptively generate measurements until successful recovery. We present the performance for both asymptotic and simulated case ($K = 50$). Firstly, note that for the simulated case, aSC slightly outperforms SC (we obtain similar performance advantage of aSC over SC for larger values of $N$). This comes in combination with significantly sparser measurement matrix since, during the measurement process, the initial $d^* = 20$ will gradually drop to as low as $d^* = 1$ (if the recovery process is not terminated earlier). Asymptotic analysis predicts well the simulated curve, however, at large $\epsilon$-values, the performance prediction of the error-floor is conservative due to the approximation of (10) using $p = 1/K$ that holds for small $\alpha$.

Fig. 3 also shows SC scheme for $d = 30 > d^*$. In this case, the performance deteriorates for intermediate $\epsilon$-values and reaches $P_r$ close to one for larger $\epsilon$ compared to the $d^*$ case, however, the larger $d$ values exhibit lower error-floors.\footnote{Note that there are no zero-measurements in $\Phi_t$. Thus, in eq. (2) of Lemma 3.1, we exclude the case $j = 0$ corresponding to a zero-measurement.} In rateless coding, the error-floor is removed by precoding \cite{17} (which could be also applied in CS by taking a set of intermediate “pre-measurements”), whereas in Sudocodes CS, it is cleared by postprocessing in the second phase that applies small fraction of dense measurements and stronger recovery algorithms \cite{1} \cite{16} \cite{18}.

\footnote{The error-floor follows from non-zero probability that a signal coefficient does not participate in any of the generated measurements \cite{17}.}
6 Conclusions

We introduced and analyzed adaptive Sudocodes scheme. For LM1 decoder, the adaptation is shown to be simple yet resulting in slightly improved performance and reduced encoding/recovery complexity. For the future work, we will explore adaptation in Sudocodes scheme with LM2 decoder which increases learning opportunities at the measurement subsystem, however, under more complex analysis of LM2 decoding. We will also further investigate connections between rateless coding and CS and how the two concepts could converge in efficient joint source channel coding solutions (as indicated by our system model in this work.)

References


Recognition of Lane Markings in Panoramic Images

C. Li I. Creusen L. Hazelhoff P.H.N. de With
CycloMedia Technology B.V. Eindhoven University of Technology
Dept: Derived Content Dept: Electrical Engineering
Van Voordenpark 1b, Zaltbommel Group: Video Coding and Architecture
Den Dolech 2, Eindhoven
licheng.kit@gmail.com
{icreusen,lhazelhoff}@cyclomedia.com, p.h.n.de.with@tue.nl

Abstract

The detection of road lane markings has many practical applications, such as advanced driver assistance systems, as well as road maintenance. In this paper we propose an algorithm to detect and recognize lane markings from panoramic images. Our system consists of three steps. First after inverse perspective mapping, the potential lane markings are segmented by finding the high-intensity pixels of the image. Second, we extract the inner distance as a feature of each region of interest from the segmented image. Third, we classify features through a Support Vector Machine (SVM). The experiments show that the system is capable of recognizing 97% of the parts of long solid lines and 94% of the dashed lines and can be extended to recognize other road markings such as arrows.

1 Introduction

Lane marking detection is an essential task in vision-based intelligent systems for transportation, such as advanced driver assistance systems and autonomous vehicles. It helps to improve traffic safety, accurate GPS navigation, and road marking maintenance which particularly interests the government. Currently, locating faded road markings is performed manually, which is very labor intensive. Cyclomedia annually records a country-wide dataset of panoramic images, captured every 5 m by cameras on vehicles driving on all public roads in the Netherlands. Each image is captured as a spherical panorama, well calibrated and positioned by GPS with an accuracy of 10 cm. An example panoramic image is shown in Fig. 1.

To avoid the above manual labor, we are aiming at designing a system to automatically detect and locate the positions of faded road markings, by comparing and analyzing panoramic images of subsequent years. Several road marking detection and recognition approaches have been proposed in literature. In [1], Rebut et al. propose a system based on the Hough transform for detecting lines and utilizing a Fourier descriptor. In [2], MSER is applied for segmentation and classify features by a matching algorithm. The objective of this paper is to design a robust detection algorithm for faded road markings that is sufficiently reliable to handle faded markings and painted road signs. In Section 2 of this paper, we describe the details of the processing stages of the system. These stages are Inverse Perspective Mapping, the segmentation, feature extraction of each segment, and segment classification using a Support Vector Machine. In Section 3, the experimental results are discussed for solid and dashed lines.
2 System Description

Let us now briefly discuss the road markings detection system in more detail. Because the panoramic image shown in Fig. 1 is matched to a sphere, the flat to 2D representation shows a perspective transformation which bends the road markings. Therefore, a top view of the image is generated using an Inverse Perspective Mapping (IPM), so that the road markings become again a straight line in the image. Next, since road markings are usually connected regions, an initial segmentation of the road markings can now be made with a line segmentation algorithm. Finally, the segmented regions are classified, based on their shape with a Support Vector Machine (SVM)[3], in the shape classification module. The details of each step are provided below. A block diagram of the entire system is shown in Fig. 2.

Inverse Perspective Mapping (IPM)

As explained above, the original spherical panoramic images show a perspective bending in a flat plane. We transform the panoramic image to a top view by applying an Inverse Perspective Mapping to reduce the distortion, similar to [1][2]. The mapping matrix between spherical panoramic images and the top view image only depends on the camera height of the ground plane and the azimuth angle of the camera with respect to the ground plane. Since our panoramic images are well calibrated, 360° panoramics, and the orientation of the car in the image is recorded by the car positioning system,
we simply remap to a top view according to the car’s position. Fig. 3 shows an example of a reconstructed top view image. The relation between original panoramic image and top view image is specified by the following equations:

\[
\begin{align*}
y &= \left( 2400 - \frac{\arctan(d/h)}{2\pi} \times 4800 \right) \mod 2400, \\
x &= \left( x_{\text{car}} + \frac{\arctan(y_o/x_o)}{2\pi} \times 4800 \right) \mod 4800.
\end{align*}
\]

In the equation, \( x_{\text{car}} \) denotes the position of the car in the panoramic image. \((x, y)\) and \((x_o, y_o)\) denote the pixel position of the original panoramic image and top view image respectively and parameter \( h \) is the height of the camera with respect to the ground plane and \( d \) the distance between the pixel \((x_o, y_o)\) and the center of the car in the top view.

**Line Segmentation Algorithm**

Road markings are designed to be clearly noticed and thus made reflective, so they can be easily observed by drivers. We design our system to work under various lighting conditions. Usually, pixels of road markings are brighter than their neighboring pixels. Based on this assumption, we divide our segmentation algorithm into three steps.

1. High-intensity pixels are obtained by obtaining the average gray-scale level of the surrounding pixels and comparing that with the central pixel. If the residual gray-scale level of that pixel exceeding the average is larger than a threshold, then we consider this pixel as a potential candidate of road markings. For instance, Fig. 4 shows an example of an image in which each small square represents a
We choose a $55 \times 55$ pixel window centered at pixel $P$ and then calculate the average pixel value within the window. If $P$ has a higher intensity than its surroundings, after subtracting the average gray-scale within the window, the residual value of $P$ should also be larger than its surroundings.

2. After subtracting the previous gray-scale average, we adopt Otsu’s method [4] for binary classification to obtain the potential high intensity regions describing the road markings. As a consequence, the result of second step is a binary image.

3. Finally, we remove the segments of interest which are too small to be considered as road marking regions. Fig. 5 shows an example of the areas of interest extracted with our segmentation algorithm.

**Shape classification**

Each of the detected regions shown in Fig. 5 is classified into three categories: dashed line segments, long solid line segments, and non-line segments. Each region is represented by a collection of high-intensity points, forming line pieces of varying lines or noise blobs. Scale normalization should first be applied to obtain scale invariance of the obtained features. The features are then extracted in a small rectangle from the region by centering it, and calculating the distance $d_i$ from the center to the border at certain angles (see Fig. 6(a)). In our experiment, we choose these angles from $0^\circ$ to $360^\circ$ with a step size of $30^\circ$. The region can then be described by the vector $\vec{v} = (d_1, d_2, ... d_{12})$, as shown in Fig. 6.

To distinguish the road markings from other regions, we apply the Support Vector Machine (SVM) algorithm to classify the shapes. In our experiment, a non-linear SVM with radial basis function kernel is used. Large regions, such as solid lines, can be split into smaller blocks which can be processed in the same way as smaller shapes.

3  Experiments

The current ground-truth dataset contains $142$ panoramic images with a spatial resolution of $2,400 \times 4,800$ pixels, containing highway scenery. We have selected 42
panoramic images as a training set. The training set contains 131 dashed lines and 84 solid lines. For instance, in Fig. 7(a), there are 2 solid lines and 5 full dashed lines segments. After segmentation, 84 solid lines can be split into 2,433 blocks. Detected regions are compared to the manually annotated ground truth, and the percentage of detected regions that corresponds to the ground truth is calculated. Using the remaining 100 panoramic images as a test set, there are a total of 197 solid lines and 637 dashed lines segmentation. In the test set, 97.6% of the blocks which belong to solid lines are correctly detected, as well as 94.6% of the dashed lines segments. Due to the shape analysis of our system, there are 4.2% and 8% of the found line segments acting as false positives, which do not belong to a solid road line block or a dashed road line segment. Fig. 7(b) shows an example of the promising results obtained with our system.

Fig. 6: (a): Feature vector of dashed line element shape. (b): Feature vector of arbitrary shape.

Fig. 7: (a): Top down image. (b): Lines detected on the image.
4 Conclusions

In this paper, we have presented an algorithm to detect road markings, such as dashed and solid lines which are visible in panoramic images at street level. The algorithm first performs an Inverse Perspective Mapping and then succeeds with a special segmentation algorithm followed by an SVM. The key to our segmentation algorithm is to assume that road markings are made of high-intensity pixels, which lead to binarized images with Otsu’s method. To this, we have added a block-oriented line-piece segmentation algorithm which maps the distances in the block to the block center into a feature vector for classification. The algorithm results show that 97.6% of the solid line blocks together with 94.6% of the dashed lines are correctly detected. The false positives can be further reduced by combining prior knowledge of road markings positions, which are not yet exploited. The algorithm proposed here can be adapted to recognize additional shapes like arrows and triangles, which are typically used in road marking as well. In future work we will experiment with a larger data set, and we will also extend our detection towards painted road signs.

References


Gbps Throughput Architecture for Turbo Decoder

Antonio Callejón, Meng Li, Yanxiang Huang, Sofie Pollin, Liesbet Van der Perre
Interuniversity Micro-Electronics Center (IMEC) vzw
Kapeldreef-75, Leuven, B-3001, Belgium
{antonio.manuel.callejonmontoya.ext; meng.li; yanxiang.huang; sofie.pollin
liesbet.vanderperre}@imec.be

Abstract

Turbo codes were widely used in 3G mobile communication due to its high error correction capability. As the demand of faster and more reliable information access, high throughput Turbo decoder is mandatory in LTE-Advanced and even in the wireless communication beyond 2020 (5G). Increasing decoding parallelism, efficient window slicing scheme and NII (Next Iteration Initialization) are the ways to increase the decoding throughput. In this paper, intensive study of the state-of-the-art turbo decoder in term of algorithm and architecture has been carried out in order to perform an optimal turbo decoder which may be used in futures applications. SW-MAP (Sliding Window MAP) and NSW-MAP (Non Sliding Window MAP), and the so called XMAP have been investigated in terms of performance, complexity and throughput in order to find the best architecture for high throughput applications. The highest throughput is achieved using XMAP but it is also the architecture with the highest complexity. In the other hand, SW-MAP and NSW-MAP, for the same number of MAP decoder, present a lower complexity, area and power than XMAP. In terms of performance, XMAP presents the worst degradation of each of them while SW-MAP and NSW-MAP have a lower degradation on their performance. Additionally, Radix-4 MAP decoder is one efficient way to increase the throughput compared to Radix-2 MAP decoder. Radix-4 based NSW-MAP architecture plus NII technology turns out the optimal solution for high throughput application for high coding length. The proposed architecture is implemented using 28nm technology, achieving a maximum decoding throughput of 2.56 Gbps, which validates the feasibility of Gbps throughput Turbo implementation in future applications.

1 Introduction

High throughput requirements is the most important feature for emerging wireless communications standard such as 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) and LTE-Advanced, which promises to provide up to 1Gbps data rate. To ensure reliable communications, turbo codes have been adopted in several wireless mobile communications standards due to their outstanding error correction performance, and they are also mandatory in LTE-Advanced. For high throughput requirements, highly parallel architectures schemes are the key. In addition, implementing an optimal turbo decoder with low complexity and low performance degradation is a challenging task. In this paper after study of the state-of-the-art, a turbo decoder implementation for 3GPP LTE and LTE-Advanced is presented. The paper is organized as follow. Section II presents a fundamental background of Turbo Codes; section III describes the principal architectures to avoid high decoding latency and then they are compared in terms of correction performance, throughput and complexity; section IV presents the hardware implementation and in section V the synthesis results are given, finally, section VI contains the conclusions.
2 Turbo Codes

In LTE standard, the scheme of turbo encoder is a Parallel Concatenated Convolutional Code that consist in two 8-state recursive systematic convolutional (RSC) encoders connected by an interleaver, Fig.1(a) (The first RSC encoder produces a systematic symbol\((s_k)\) and parity symbol\((p_{1k})\) to the channel. The information bits are permuted in pseudo-random way by the interleaver and then the second RSC produces only a parity symbol\((p_{2k})\). In this way for each N-bit information block, turbo encoder produces a codeword of \(3N + 12\) bits, where 12 bits are used to converge the trellis in a state known.

![Figure 1: Block Diagram: (a) Turbo Encoder. (b) Turbo Decoder](image)

The decoding scheme, Fig.1(b), is composed of two soft-input soft-output (SISO) decoders which are connected by the interleaver and de-interleaver. The decoding process is the next: the first SISO uses the systematic\((y_k)\) and parity\((c_{1k})\) symbol (soft-input) to produce extrinsic information (soft-output) about this input bits. Then, the second SISO decoder uses the systematic symbol in the interleaving form, parity\((c_{2k})\) and the output produced by the previous decoder, also in the interleaving form, to compute again extrinsic information. Then this output is fed into the first SISO decoder in the de-interleaving form. This iterative process continues until extrinsic information has converged or a maximum number of iterations has been achieved. After this, a hard decision of decoded symbol is made.

Interleaver is the main component to achieve good correction performance but it is also the critical unit which restricts high throughput requirements, because memory collision may occur in the interleaving or de-interleaving stage. This problem is solved in 3GPP LTE standard through the use of Quadratic Permutation Polynomial (QPP), where the i-th interleaving position satisfies the next expression:

\[
\pi(i) = (f_1 \cdot i + f_2 \cdot i^2) \mod N
\]

The parameters \(f_1\) and \(f_2\) depend on the block size and they are defined in the LTE standard. There are some algebraic properties for this interleaver, although the most important one to achieve high throughput is the contention-free property, which avoids the collision memory problem [1].

\[
\frac{[\pi(i + jL)]}{L} \neq \frac{[\pi(i + kL)]}{L}
\]

Where \(L\) is a window size, \(0 \leq i < L, 0 \leq j, k < P, P = N/L\) and \(j \neq k\). This expression is the memory indices that are accessed simultaneously by the \(P\) SISO decoders. In order to avoid collision in the memory, these indices have to be unique. This is satisfied if \(P\) is a factor of the interleaver length \(N\) [1]. The next sections are based on this property.
In order to exchange the soft-output between SISO decoders, the MAP algorithm [2] is applied. The original expressions to compute the soft-output, called log-likelihood ratio (LLR), is too expensive to perform in hardware because of it uses exponentiation and logarithmic calculations. Approximations of the MAP algorithm without degradation in performance correction are presented in [3]. The proposed turbo decoder uses the Max-Log-Map algorithm which is feasible to implement in hardware.

\[
LLR(u_k) = \max_{u,u_k=1} [\alpha_k(S_k) + \gamma_{k+1}(S_k, S_{k+1}) + \beta_{k+1}(S_{k+1})] \\
- \max_{u,u_k=0} [\alpha_k(S_k) + \gamma_{k+1}(S_k, S_{k+1}) + \beta_{k+1}(S_{k+1})]
\]

(3)

\(\alpha_k\) and \(\beta_k\) are the forward and backward state metrics and they are computed recursively, Fig.2(a), using the next expressions:

\[
\alpha_k(S_k) = \max_{S_{k-1}}[\alpha_{k-1}(S_{k-1}) + \gamma_k(S_{k-1}, S_k)]
\]

(4)

\[
\beta_k(S_k) = \max_{S_{k+1}}[\beta_{k+1}(S_{k+1}) + \gamma_k(S_k, S_{k+1})]
\]

(5)

After receiving systematic and parity symbol the decoder can be derived \(\gamma_k\) of each branch transition between trellis stages.

![Trellis Diagram](image)

*Figure 2: (a) Forward, backward recursion on the trellis diagram. (b) Radix-4 trellis diagram*

### 3 High Throughput Architectures

In order to achieve high throughput decoding, an efficient parallelism scheme has to be chosen. A study of the state-of-art in terms of architecture, throughput and complexity has been performed.

#### 3.1 Radix-2 Radix-4 architecture

Basically there are two ways to perform \(\alpha\) and \(\beta\) recursion at the trellis stage: radix-2 and radix-4. The second one doubles the throughput for a given clock rate [4] because it processes two received symbol whereas radix-2 processes one symbol. Equations (3), (4) and (5) have to be changed in order to include two received symbols, Fig.2(b),

\[
\alpha_{k+2}(S_{k+2}) = \max_{S_k} [\alpha_k(S_k) + \gamma_{k+1}(S_k, S_{k+1}) + \gamma_{k+2}(S_{k+1}, S_{k+2})]
\]

(6)

\[
\beta_k(S_k) = \max_{S_{k+2}} [\beta_{k+2}(S_{k+2}) + \gamma_{k+2}(S_{k+1}, S_{k+2}) + \gamma_{k+1}(S_k, S_{k+1})]
\]

(7)
\[ \text{LLR}(u_k) = \max_{u: u_k = 1} \left[ \alpha_k(S_k) + \gamma_{k+1}(S_k, S_{k+1}) + \gamma_{k+2}(S_{k+1}, S_{k+2}) + \beta_{k+2}(S_{k+2}) \right] \\
- \max_{u: u_k = 0} \left[ \alpha_k(S_k) + \gamma_{k+1}(S_k, S_{k+1}) + \gamma_{k+2}(S_{k+1}, S_{k+2}) + \beta_{k+2}(S_{k+2}) \right] \] (8)

Figure 3: (a) Proposed Add-Compare-Select radix-4. (b) Extrinsic unit pipeline

This approach reduces the memory needed to save the metrics, although the complexity to perform a recursion is higher than radix-2. This may result in an increased delay, specially in the LLR because each \( \alpha_{k+1}(S_{k+1}) \) and \( \beta_{k+1}(S_{k+1}) \) nodes needs to be computed from two \( \alpha_k(S_k) \) and \( \beta_{k+2}(S_{k+2}) \) respectively. For this reason, a fully pipeline unit is implemented in order to reduce the critical path, Fig. 3(b).

### 3.2 Avoiding high decoding latency

The main idea to reduce the decoding latency is to split the block \( N \) into sub-blocks \( L \) in such a way that, decoding is distributed on parallel working SISO units. In Sliding Window MAP architecture (SW-MAP), each sub-block is again divided into shorter windows lengths \( W_L \) in order to compute \( \beta \) and LLR metrics before \( \alpha \) metric finishes [5, 6]. The throughput using radix-4 is:

\[ \text{SW}_{\text{throughput}} = \frac{N}{\text{decodingtime}} \approx \frac{N \cdot \text{frequency}}{I \cdot (N/(2P) + W_L)} \] (9)

where \( I \) is the total half iterations.

Avoiding the additional delay introduced by SW-MAP, term \( W_L \) in (9), Non Sliding Window MAP (NSW-MAP) proposes a double recursion simultaneously [5]. This proposal increases the throughput, however, the fact that it uses a double recursion gives rise to doubling the hardware cost. The throughput is:

\[ \text{NSW}_{\text{throughput}} \approx \frac{N \cdot \text{frequency}}{I \cdot (N/(2P))} \] (10)

In contrast to the previous architectures, X-MAP processes a single window at a time. In this way, a completely window is fed into a fully pipeline decoder in each clock cycle [7]. With the aim of increasing the throughput achieved in [7], this work suggests to
use more than one X-MAP unit as May et. al. propose. In this case, the throughput is measured as:

$$\text{XMAP}_{\text{throughput}} \approx \frac{N \cdot \text{frequency}}{I \cdot \left( \frac{N}{PW_x} + \frac{W_x}{2} \right)}$$

where $W_x$ is the window size of X-MAP engine.

Backward and forward recursion have to be initialized from previous and next SISO due to the block length division. In order to avoid the classic training calculation, which increases the latency, the proposal decoder uses Next Iteration Initialization (NII). This method propagates the $\alpha$ and $\beta$ metrics to the next iteration [8], Fig.6(b).

### 3.3 Architecture Comparison

The main step to design an optimal turbo decoder is to analyze the features of the different architectures in order to find the right trade-off. Firstly, a throughput comparison is presented in Fig.4a. The maximum parallelism for X-MAP is limited (up to 32) due to the contention-free property, even so, it is the maximum throughput architecture. After throughput comparison, the next factor to take into consideration is the error correction performance. Fig.5 presents the floating point simulation results of distinct architectures for $N=6144$, at 6 iterations, for different codes rates. The modulation used is BPSK over AWGN channel. As can be seen, the BER performance of X-MAP is worse than NSW-MAP and SW-MAP, which is the closest architecture to the no window scheme. As parallelism increases ($P=96$), correction performance degradation is higher and for high code rate, Fig.5c, the performance is unacceptable.

![Normalized Throughput](image1)

(a) Normalized Throughput

![Normalized Complexity](image2)

(b) Normalized Complexity

*Figure 4: Normalized throughput and complexity*

To give an idea of the complexity, Fig.4b shows the normalized complexity of each architecture in terms of units (recursion units, LLR units and memory to save metrics) needed in each of them. X-MAP has the highest complexity because it is a fully pipeline engine while SW-MAP and NSW-MAP have similar results.

Based on the previous results, the highest throughput is achieved by X-MAP architecture, however it is also the highest complexity and the worse correction performance. For these reasons, X-MAP is discarded as final implementation. On the other hand, SW-MAP has lower throughput, complexity and degradation on the performance than NSW-MAP. As a design trade-off in order to achieve high throughput requirements, 64 NSW-MAP architecture is chosen in the final hardware implementation.
4 Hardware Implementation

The proposed turbo decoder has been implemented in an ASIC design using Verilog HDL code. Fixed-point implementation is according to Tempo IMEC project [8]. The proposed NSW-MAP hardware architecture and its corresponding schedule is shown in Fig.6(a), Fig.6(b) and Fig.6(c). Due to the double recursion, two branch metric units, two LLR units and two LIFO to save $\alpha$ and $\beta$ metrics are needed. It is very important to notice that in each cycle four systematic symbol, four parity symbol and four extrinsic data are read from memory. In addition, four LLR values will be written in memory after $L/2$. In order to satisfied the contention-free property (2), each memory is split in four different banks, Fig. 6(d). This approach allows four reading and four writing without collision in the interleaving iteration. According to radix-4, add-compare-selection (ACS) operations are implemented as shown Fig.3(a). Interconnection network to permute the data between the NSW-MAP decoders and the memory modules is very important in terms of complexity in the final implementation. A crossbar network is proposed in [5], but in the proposed decoder this would result in 64x64 switches with their control signal, which can significantly increase the area. In this work, a barrel-shift network is used because of it needs $P \cdot \log_2(P) = 256$ 2-to-1 multiplexer. Moreover, in [10] is proved $P/2$ signals are sufficient to control all multiplexer.
5 Synthesis results

The implementation is synthesized in a TSMC 28-nm CMOS technology for a 300MHz clock frequency. The decoder works with block size of 6144 using 64 NSW-MAP units. The maximum clock frequency, at least, is 300MHz and the corresponding maximum throughput is 2.56Gbps, at 6 iterations, with a core area of $13\text{mm}^2$. Table 1 presents the synthesis results and a comparison with existing turbo decoder.

6 Conclusion

In this paper, a high throughput turbo decoder for 3GPP/LTE-Advanced has been presented using highly parallel architecture. The proposed turbo decoder uses 64 NSW-MAP radix-4 units in order to achieve high throughput. According to the synthesis results, the design can work, at least, at 300MHz achieving a maximum throughput of 2.56Gbps with a core of $13\text{mm}^2$ in a 28-nm technology.
Block Size & 6144 & 6144 & 6144 & 5144 \\
Algorithm & Max-Log MAP & Log-MAP & Log-MAP & Log-MAP \\
Radix & 4 & 2 & 4 & 4 \\
Sliding Window & NSW-MAP & SW-MAP & SW-MAP & NSW-MAP \\
MAP cores & 64 & 64 & 8 & 32 \\
Max. Iterations & 6 & 6 & 6 & 5.5 \\
Technology & 28nm & 65nm & 90nm & 45nm \\
Clock frequency & 300MHz & 300MHz & 400MHz & 600MHz \\
Max. Throughput & 2.56Gbps & 1.28Gbps & 1Gbps & 1.67Gbps \\
Core area & 14mm² & 8.5mm² & N/A & 2.4mm² \\
Gate count & 9.3M & 9.3M & 9.8M & 1.4M \\

Table 1: Synthesis results and comparison with existing turbo decoders

References


Comparisons of Pre-Coders under Power Normalization in Massive-MIMO Systems

Edward De Brouwer and Jérôme Louveaux
Institute of Communications Technology, Electronics and Applied Mathematics (ICTEAM)
Université catholique de Louvain

Abstract

This paper presents the comparison of various pre-coders for Massive MIMO systems. 3 pre-coders are simulated: MMSE, ZF and Conjugate Beamforming. As power normalization schemes are known to modify the pre-coding operation by reducing the number of degrees of freedom in the systems, the pre-coders are simulated under power normalization requirements. Per-antenna, per-carrier and per-user normalization schemes are considered. ZF and MMSE are shown to perform best under global power normalization. However, when per-carrier and per-antenna power normalization is used with sufficient number of antennas at the base station, conjugate beamforming is found to be optimal due to its low complexity. Finally, work perspectives are given relative to an enhanced conjugate beamforming pre-coder.

1 Introduction

It is now well known that MIMO can offer many advantages over the classical SISO systems. This technology is becoming mature and the perspective of large MIMO systems is presently studied. However, as MIMO was previously used in point-to-point operation, that is with only one transmitter and one receiver, the spatial diversity offered by MIMO systems is now used to provide multi-users communication. Increasing the number of antennas at one side of the transmission to simultaneously serving multiple autonomous terminals with fewer antennas (and thus lower complexity) is the challenge of Massive MIMO.

Massive MIMO systems are multi-users systems using a large antenna array ($N_T$ antennas, possibly up to 100) at the base station to communicate with several users equipped with fewer antennas ($N_R$ users with typically a single antenna each). It has already been theoretically shown that these systems are promising from different points of view. Their data rate performance and capability to work under low power consumptions while keeping a limited complexity at the terminals side is one of the main advantages of Massive MIMO.

From a power consumption point of view, it has been shown in recent papers that the total power consumption at the base station could be scaled with the number of antennas at the base station (BS) [3]. It means that doubling $N_T$ allows to reduce by two the total emitted power at the BS when downlink transmission is considered (and the the emitted power per-user when uplink transmission is considered). Indeed, the SINR at each user is shown to be $\text{SINR} \sim N_T$ for downlink and uplink [3] and [4]. This holds when perfect CSI is available at both sides of the link. What is more, when only imperfect CSI is available, the power consumption of each user can be scaled with $\frac{1}{\sqrt{N_T}}$ [3]. For downlink, we are not aware of papers analyzing the expected power reduction with a growing number of antennas under imperfect CSI.

With sufficient number of antennas at the BS, linear pre-coders were shown to be asymptotically optimal. Indeed, the remaining inter-users interference term is known to
vanish when large antennas arrays are considered [5]. This allows a reduced complexity at both transmission side while providing near-optimal data rates. For these systems that manage interference cancellation, the performance will then be driven by thermal noise present at each antenna and by the imperfect channel state information. Pre-coders choice will thus be important for these parameter optimization.

Energy and spatial efficiency of Zero-Forcing and Conjugate pre-coders have already been analyzed in [4]. Conjugate beamforming was shown to perform best under low SNR conditions while ZF was best suited in the high SNR range.

In practical systems however, pre-coding scheme often cannot be applied in their intrinsic form. Power constraints require some power allocation that differ from the one that would be obtained with pre-coders only. Power constraints can arise from various considerations. Power amplifiers can require normalized emitted power antenna and one could impose fairness in the system by allowing same power to each user. Flat frequency spectrum can also be asked. It will be satisfied using per-carrier power normalization. These power normalizations schemes obviously affect the behaviour of the pre-coding schemes. For instance, ZF is not expected to inverse the channel perfectly anymore.

This paper highlights the near-optimal behaviour of linear pre-coders in Massive MIMO systems under power normalization requirements. With sufficient number of antennas at the BS, conjugate beamforming will be shown to perform as well as the ZF even in the high SNR region. As it is also the pre-coder that shows the lowest complexity, the energy consumption of both transmitter and receiver could be lowered.

This paper is organized as follows: in Section 2, the Massive MIMO system model is described. The 3 pre-coders simulated are presented: Conjugate Beamforming (CB), Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE). Theoretical capacity bounds are given. They are then discussed. The power-normalization schemes are also presented. Their purpose and their computation procedures are explained. Section 3 gives the simulations results and discuss them. The pre-coders are compared under the different power normalization schemes before the conclusion in Section 4.

## 2 System Model

A massive MIMO system is considered with $N_T$ antennas at the BS and $N_R$ single-antenna users. Only the downlink transmission is studied here. A the transmitter (BS), a $(N_c \cdot N_b) \times 1$ vector of symbols $s_k$ is generated for each user $k$. $N_b$ stands for the number of OFDM blocks to be sent by user. $N_c$ is the length of the IFFT in the OFDM operation, or equivalently the number of subcarriers. The precoding is performed for each subcarrier independently. The data for subcarrier $j$ is stored in a $N_R \times N_b$ matrix $S_j$ whose lines are the OFDM vectors $s_k(j,:)$ of the different users for this subcarrier.

The pre-coding matrix $G_j$ for this subcarrier is then applied. $G_j$ is a $N_T \times N_R$ matrix. For each sub-carrier, $S_j$ is pre-multiplied by $G_j$. The result is a $N_T \times N_b$ matrix. After IFFT and cyclic prefix, the data is ready to be sent through the channel.

The channel matrix in the frequency domain is denoted by $H_j$ for subcarrier $j$. It is a $N_R \times N_T$. The channel is assumed to be known at the transmitter in order to compute the pre-coder, so channel estimation has to be performed. In our simulations, the channel statistics are assumed to be Rayleigh fading with unit energy for each channel path. Considering only one OFDM block ($N_b = 1$), the $N_R \times 1$ vector of signals received at the different users ($k = 1, \ldots, N_R$) for subcarrier $j$ is given by

$$y_j = H_j G_j S_j + N_j$$

where $N_j$ is the vector of additive white gaussian noise samples for the different users for subcarrier $j$. All samples are assumed to have variance $N_0$. 
2.1 Pre-Coders

The pre-coders are computed by the way of a pre-coding matrix. This is possible because the pre-coders used here are linear. Each pre-coder will lead to a specific pre-coding matrix $G_j$.

2.1.1 Conjugate Beamforming

We first consider conjugate beamforming. $G_j$ then consists in the complex conjugate of the channel frequency response. Formally for each sub-carrier $j$:

$$G_{CB,j} = H_j^H.$$  

With conjugate beamforming pre-coding, the phase of the received symbols will be aligned at each users leading in a coherent addition and thus maximizing received power. However, some inter-users interference will still persist and degrade the performance. Conjugate beamforming is known to maximize SNR at the expenses of some losses with respect to the SIR (Signal to Interference Ratio). What is more, using conjugate operation, it is readily seen that more power will be allowed to the strong carriers than to the weak ones.

Taking into account training procedure, it has been shown in [4] that a lower bound capacity for conjugate beamforming (with MMSE for channel estimation) is given by

$$C_{\text{sum,conj}} = N_R \log_2 \left( 1 + \frac{N_T}{N_R} \frac{\rho_f \tau_r \rho_r}{\rho_f + 1 + \tau_r \rho_r} \right)$$

where $\rho_f$ and $\rho_r$ denote the SNR in the downlink and the uplink respectively and $\tau_r$ is the number of symbols in the uplink pilot sequence used for channel estimation.

The effective SINR at each terminal is also given by

$$\text{SINR}_{\text{conj}} = \frac{N_T}{N_R} \frac{\rho_f \tau_r \rho_r}{\rho_f + 1 + \tau_r \rho_r}.$$  

2.1.2 Zero-Forcing

The Zero-Forcing pre-coder is now considered. In this case, $G_j$ consists in a pseudo-inversion of the channel matrix. For each sub-carrier $j$ we have

$$G_{ZF,j} = H_j^H (H_j H_j^H)^{-1}.$$  

The matrix inversion ensures that there is no inter-user interference. However, the power of the useful symbols is not maximized and it thus not optimal from a SNR point of view. Performing inversion, more power will be allowed to the weak carriers. What is more, channel matrix inversion requires heavy computation if the dimension grows large.

In terms of theoretical achievable rates, it was shown [4] that ZF allows a capacity bound of (with MMSE for channel estimation):

$$C_{\text{sum,ZF}} = N_R \log_2 \left( 1 + \frac{N_T - N_R}{N_R} \frac{\rho_f \tau_r \rho_r}{\rho_f + 1 + \tau_r \rho_r} \right)$$

The effective SINR of such a transmission is also found to be:

$$\text{SINR}_{ZF} = \frac{N_T - N_R}{N_R} \frac{\rho_f \tau_r \rho_r}{\rho_f + 1 + \tau_r \rho_r}.$$
2.1.3 MMSE

Finally the MMSE pre-coder is known to minimize the sum of the mean square errors of the decoded symbols at the different users. This is obtained by using the following pre-coding matrix $G_j$ for each sub-carrier $j$:

$$G_{MMSE,j} = (H_j^H H_j + \frac{1}{\rho_{f,j}} I_{N_T})^{-1} H_j^H$$  \hspace{1cm} (8)

where $\rho_{f,j}$ the SNR of the downlink transmission on subcarrier $j$ and $I_{N_T}$ is an identity matrix of size $N_T$. MMSE performs a combination between ZF and MMSE. Performing an inversion, it will allow more power on the weak carriers. However, the $\frac{1}{\rho_{f,j}} I_{N_T}$ term will make sure that the strong carriers are still enforced.

2.1.4 Pre-Coders Comparison

In the low SNR region, the system is driven by the noise power. The SNR should thus be maximized and the conjugate beamforming performs best for this task. Conjugate-beamforming will allow more power on the strong sub-carriers in order to ensure as reliable transmission as possible. However, as the SNR grows, the system will tend to be interference driven and the SIR should be maximized. This is best done using MMSE and ZF. What is more, MMSE and ZF will allow more power on the weak carriers so that each of them can transmit in a significant way. In the medium SNR region, MMSE will perform best as it provides a kind of trade off between the power allocation used in ZF and Conjugate. As the SNR grows large, MMSE and ZF tend to behave the same way as the $\frac{1}{\rho_{f,j}} I_{N_T}$ term will vanish.

Finally, from a computational cost point of view, let’s note that MMSE and ZF are significantly more complex than conjugate beamforming. In order to avoid complex transmitter, it may be advantageous to use the conjugate pre-coding scheme.

2.2 Power Normalization Schemes

In practical systems, normalized emitted power may be required. Different power normalization schemes are considered here: Per-antenna, per-user and per-carrier power normalization.

First of all, a global power normalization requirement is used in the system. The total emitted energy from the base station is always set to the number of users to serve. This assumption comes from fairness considerations. Indeed, when 2 users are to be served, it seems fair to be able to consume twice the emitted energy used when only one user is present. This is required in order to perform fair comparisons between systems involving different number of users.

The constellation symbols being normalized to 1 and the IFFT operation conserving the emitted power of the signal, the emitted energy corresponds to energy of the pre-coding matrices $G_j$, summed over the antennas, users and subcarriers. The total constraint becomes

$$\sum_{j=1}^{N_c} \frac{1}{N_R} ||G_j||^2 = 1$$  \hspace{1cm} (9)

2.2.1 Per-User Power Normalization

To obtain fairness in the system, the same amount of energy should be mapped on each user. Per-User normalization is then introduced. Each user will be allowed a total pre-coder energy of 1. Thus $N_R$ pre-coded energies will be computed. For each
of them, there will be a summation over the carriers and the antennas. The resulting pre-coded energy per user should be equal to 1. Practically, the computation of the normalized pre-coding matrix is done as follows:

- For each element along the $N_R$ dimension (i.e. each user), select the corresponding columns of the precoding matrices $G_j$ for all subcarriers.
- Sum the square norms of the columns across the antennas and subcarriers.
- Divide each coefficient by the square root of the computed energy. The total energy should now be normalized to 1.
- Concatenate the normalized columns together to find the normalized pre-coding matrix.

2.2.2 Per-Antenna Power Normalization

Power amplifier are normally designed for the same output power on each antenna. They may not accept significant power fluctuations. For this purpose, it is interesting to be able to simulate per-antenna normalized systems. Similarly as in the per-user case, $N_T$ pre-coded energies will be computed and each of them should lead the value $\frac{1}{N_T}$. Practically

- For each element along the $N_T$ dimension, select the corresponding line of the precoding matrix $G_j$ for all subcarriers.
- Sum the square norms of these lines across the carriers, and average along the $N_R$ (user) dimension.
- Divide each coefficient by the square root of the computed energy. Then divide by $\sqrt{N_T}$. The energy should be normalized to $\frac{1}{N_T}$ for each antenna.
- Concatenate the normalized lines together to find the normalized pre-coding matrix.

2.2.3 Per-Carrier Power Normalization

Imposing the same power allocation on each sub-carrier makes sure the power spectral density is constant in the frequency band. Again this is implemented in a similar way as for the other normalization schemes. $N_c$ pre-coded energies will be computed and should have a value of $1/N_c$. Practically, each precoding matrix $G_j$ should be normalized such that

$$\frac{1}{N_R}||G_j||^2 = \frac{1}{N_c}. \quad (10)$$

2.2.4 Combined Power Normalization

The different normalization configurations presented above can be activated simultaneously. Indeed, in some cases, distinct practical considerations require simultaneous normalization schemes. In this case, a combination of the previously presented computations is performed. Ultimately, the most restrictive constraint will request normalized energy per antenna, per carrier and per user. Then, $N_R \cdot N_T \cdot N_c$ pre-coded energies will be computed corresponding to all the values of the pre-coding matrix. They thus all should have an energy (squared amplitude) of $\frac{1}{N_T N_c}$.

190
As highlighted above, except the global power normalization requirement, all the power allocation schemes modify the properties of the pre-coding matrix. The number of degrees of freedom in the system will be reduced and the performance of the pre-coders will thus be different. In the case of per-antenna power normalization, it will result in performance degradation as shown next. Indeed, as the degrees of freedom of the system are reduced due to the per-antenna normalization, the pre-coder coefficients can no longer necessarily apply more energy to the better antennas for a given subcarrier. This is thus no longer optimal from a SNR point of view. Similarly, some inter-user interference may appear when performing the normalization differently across the antennas.

3 Simulations and Results

3.1 Simulations

A two-users massive MIMO system was simulated under Rayleigh fading channel and various antennas configurations at the BS. At first no power normalization are used, only the global power requirement is activated. The result is shown in Figure 1. As expected, in the low SNR region, the conjugate beamforming performs best. In the medium SNR region, MMSE behaves better than the other and finally the MMSE and ZF tend to become the best pre-coders when the SNR of the transmission grows. Note the important performance difference between MMSE/ZF and conjugate beamforming. It is highlighted in table 1. As $N_T$ grows, the difference tends to shrink but a significant difference remains observable.

![Figure 1: Computing Procedure of Independent Normalization](image)

Power normalizations are then added to the system. Per-antenna and per-carrier power normalizations are considered. The resulting BER curves are presented in Figure 2. The results of figures 1 and 2 are compared in table 1.

First of all, conjugate pre-coder performs best when power constraints are added. Indeed, the per-carrier power normalization makes the conjugate-beamforming behave
like the Zero-Forcing scheme. The per-carrier power normalization will impose a constant power allocation on each carrier. Compared to classic conjugate beamforming, more power will thus be allocated on the weak carriers and less on the strong ones so that they all carry the same amount of energy. As Zero-Forcing performs better than conjugate in the medium-high SNR region, higher performance are obtained when per-carrier power normalization is added. Secondly, the three implemented pre-coders tend to behave the same way when the most restrictive power constraints are added and when $N_T$ grows. However, if only the global power constraints is to be respected, there is a significant gap between conjugate-beamforming and ZF/MMSE performances. For instance, there is a 4dB difference for the 128 × 2 configuration with global power normalization. Pre-coders will behave quite differently as ZF and MMSE will allow more power on weak sub-carriers while conjugate beamforming will do the opposite. When per-carrier normalization is added, these differences tend to vanish. Finally, differences between power normalized (per-carrier and per-antenna) systems and globally power normalized systems are quite obvious. Indeed, 1.5dB is approximately observed between power normalized and globally normalized systems. Systems are thus sensitive to power constraints as it is highlighted in table 1.

![Diagram](Image)

Figure 2: Computing Procedure of Independent Normalization

### 3.2 Conclusions

When no power normalization is used, the SNR range in which conjugate beamforming performs best is very narrow. What is more, the performance difference between conjugate and ZF/MMSE is such that MMSE/ZF will nearly always be preferred. However, when per-antenna and per-user power normalization is added, conjugate beamforming seems to perform better. Indeed, more power will be allocated to weak carriers than in the standard case. When sufficient number of antennas at the base station $N_T$ is taken, conjugate beamforming and ZF tend to behave the same way. Due to its low complexity, conjugate beamforming will thus be preferred. These results let us expect great performances with a enhanced conjugate beamforming. Indeed, the limitation of this pre-coder seems to be it’s un-optimal power allocation at the sub-carrier level. We
<table>
<thead>
<tr>
<th>$N_T$</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conj.</td>
<td>/</td>
<td>7.7dB</td>
<td>2.25dB</td>
<td>-2.5dB</td>
<td>-5.7dB</td>
</tr>
<tr>
<td>ZF</td>
<td>3.9dB</td>
<td>1dB</td>
<td>-2.7dB</td>
<td>-6.2dB</td>
<td>-9.6dB</td>
</tr>
<tr>
<td>MMSE</td>
<td>4.05dB</td>
<td>0.1dB</td>
<td>-2.9dB</td>
<td>-6.3dB</td>
<td>-9.3dB</td>
</tr>
<tr>
<td>Conj. Norm.</td>
<td>/</td>
<td>5.6dB</td>
<td>-0.2dB</td>
<td>-3.9dB</td>
<td>-7.9dB</td>
</tr>
<tr>
<td>ZF. Norm.</td>
<td>6.6dB</td>
<td>2.1dB</td>
<td>-1.9dB</td>
<td>-4.6dB</td>
<td>-8.1dB</td>
</tr>
<tr>
<td>MMSE. Norm.</td>
<td>6.4dB</td>
<td>2.3dB</td>
<td>-1.6dB</td>
<td>-5dB</td>
<td>-8dB</td>
</tr>
</tbody>
</table>

Table 1: SNR needed to obtain a $10^{-4}$ BER for different two-users configurations. Normalized pre-coders include per-antenna and per-carrier power normalization. Rayleigh Channel is used.

could imagine a more efficient allocation and thus minimize the throughput. An non-optimal version of this principle is used when per-carrier normalization is activated. What is more, it would be interesting to modify the constellation for distinct carriers. Strong channel paths should be coded with an higher constellation code in order to maximize the throughput.

References


<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stan Baggen</td>
<td>Philips Research</td>
<td><a href="mailto:stan.baggen@philips.com">stan.baggen@philips.com</a></td>
</tr>
<tr>
<td>Nabil Bouhouch</td>
<td>TU/e</td>
<td><a href="mailto:n.bouhouch@student.tue.nl">n.bouhouch@student.tue.nl</a></td>
</tr>
<tr>
<td>Antonio Callejon</td>
<td>IMEC</td>
<td><a href="mailto:antonio.manuel.callejonmontoza.ext@imec.be">antonio.manuel.callejonmontoza.ext@imec.be</a></td>
</tr>
<tr>
<td>Ivo Creusen</td>
<td>Cyclomedia Technology BV</td>
<td><a href="mailto:icreusen@cyclomedia.com">icreusen@cyclomedia.com</a></td>
</tr>
<tr>
<td>Abhishek Dutta</td>
<td>TU Twente</td>
<td><a href="mailto:a.dutta@utwente.nl">a.dutta@utwente.nl</a></td>
</tr>
<tr>
<td>Farzad Farhadzadeh</td>
<td>TU/e</td>
<td><a href="mailto:f.farhadzadeh@tue.nl">f.farhadzadeh@tue.nl</a></td>
</tr>
<tr>
<td>Sohrab Ferdowsi</td>
<td>University of Geneva</td>
<td><a href="mailto:sohrab.ferdowsi@unige.ch">sohrab.ferdowsi@unige.ch</a></td>
</tr>
<tr>
<td>Wessel Ganzervoort</td>
<td>Benchmark Electronics BV</td>
<td><a href="mailto:wessel.ganzervoort@bench.com">wessel.ganzervoort@bench.com</a></td>
</tr>
<tr>
<td>Jasper Goseling</td>
<td>TU Twente - TU Delft</td>
<td><a href="mailto:j.goseling@utwente.nl">j.goseling@utwente.nl</a></td>
</tr>
<tr>
<td>Richard Heusdens</td>
<td>TU Delft</td>
<td><a href="mailto:r.heusdens@tudelft.com">r.heusdens@tudelft.com</a></td>
</tr>
<tr>
<td>Tanya Ignatenko</td>
<td>TU/e</td>
<td><a href="mailto:t.ignatenko@tue.nl">t.ignatenko@tue.nl</a></td>
</tr>
<tr>
<td>Amir Jalalirad</td>
<td>TU/e</td>
<td><a href="mailto:a.jalalirad@tue.nl">a.jalalirad@tue.nl</a></td>
</tr>
<tr>
<td>Solmaz Javanbakhti</td>
<td>TU/e</td>
<td><a href="mailto:s.javanbakhti@tue.nl">s.javanbakhti@tue.nl</a></td>
</tr>
<tr>
<td>Natasa Jovanovic</td>
<td>TU/e</td>
<td><a href="mailto:n.jovanovic@tue.nl">n.jovanovic@tue.nl</a></td>
</tr>
<tr>
<td>Arie Koppelaar</td>
<td>NXP</td>
<td><a href="mailto:arie.koppelaar@nxp.com">arie.koppelaar@nxp.com</a></td>
</tr>
<tr>
<td>Alex Kot</td>
<td>Nanyang Tech. University</td>
<td><a href="mailto:eackot@ntu.edu.sg">eackot@ntu.edu.sg</a></td>
</tr>
<tr>
<td>Andreas Koutrouvelis</td>
<td>TU Delft</td>
<td><a href="mailto:a.koutrouvelis@student.tudelft.nl">a.koutrouvelis@student.tudelft.nl</a></td>
</tr>
<tr>
<td>Thijis Laarhoven</td>
<td>TU/e</td>
<td><a href="mailto:t.m.m.laarhoven@tue.nl">t.m.m.laarhoven@tue.nl</a></td>
</tr>
<tr>
<td>Jerome Louveaux</td>
<td>UC Louvain</td>
<td><a href="mailto:jerome.louveaux@uclouvain.be">jerome.louveaux@uclouvain.be</a></td>
</tr>
<tr>
<td>Mihaela Mitici</td>
<td>TU Twente</td>
<td><a href="mailto:m.a.mitici@utwente.nl">m.a.mitici@utwente.nl</a></td>
</tr>
<tr>
<td>Giel Op’t Veld</td>
<td>EPFL/LINX</td>
<td><a href="mailto:giel.optveld@epfl.ch">giel.optveld@epfl.ch</a></td>
</tr>
<tr>
<td>Chariklea Papatsimpa</td>
<td>TU/e</td>
<td><a href="mailto:c.papatsimpa@tue.nl">c.papatsimpa@tue.nl</a></td>
</tr>
<tr>
<td>Aleksandra Pizurica</td>
<td>Universiteit Gent</td>
<td><a href="mailto:sanja@telin.ugent.be">sanja@telin.ugent.be</a></td>
</tr>
<tr>
<td>Sofie Pollin</td>
<td>KU Leuven</td>
<td><a href="mailto:sofie.pollin@esat.kuleuven.be">sofie.pollin@esat.kuleuven.be</a></td>
</tr>
<tr>
<td>Bhaskar Rao</td>
<td>Univ. of California, San Diego</td>
<td><a href="mailto:brao@ucsd.edu">brao@ucsd.edu</a></td>
</tr>
<tr>
<td>Zhijie Ren</td>
<td>TU Delft</td>
<td><a href="mailto:z.ren@tudelft.nl">z.ren@tudelft.nl</a></td>
</tr>
<tr>
<td>Jeremie Roland</td>
<td>ULB - QuIC</td>
<td><a href="mailto:jroland@ulb.ac.be">jroland@ulb.ac.be</a></td>
</tr>
<tr>
<td>Marco Rompen</td>
<td>TU/e</td>
<td><a href="mailto:m.l.l.rompen@student.tue.nl">m.l.l.rompen@student.tue.nl</a></td>
</tr>
<tr>
<td>Fernando Rosas</td>
<td>KU Leuven</td>
<td><a href="mailto:fernando.rosas@esat.kuleuven.be">fernando.rosas@esat.kuleuven.be</a></td>
</tr>
<tr>
<td>Francois Rottenberg</td>
<td>UC Louvain</td>
<td><a href="mailto:francois.rottenberg@student.uclouvain.be">francois.rottenberg@student.uclouvain.be</a></td>
</tr>
<tr>
<td>Willem Sanberg</td>
<td>TU/e</td>
<td><a href="mailto:w.p.sanberg@tue.nl">w.p.sanberg@tue.nl</a></td>
</tr>
<tr>
<td>Boris Skoric</td>
<td>TU/e</td>
<td><a href="mailto:b.skoric@tue.nl">b.skoric@tue.nl</a></td>
</tr>
<tr>
<td>Ludo Tolhuizen</td>
<td>Philips Research</td>
<td><a href="mailto:ludo.tolhuizen@philips.com">ludo.tolhuizen@philips.com</a></td>
</tr>
<tr>
<td>Anagnostis Tsiatmas</td>
<td>TU/e</td>
<td><a href="mailto:a.tsiatmas@tue.nl">a.tsiatmas@tue.nl</a></td>
</tr>
<tr>
<td>Raymond Veldhuis</td>
<td>TU Twente</td>
<td><a href="mailto:r.n.j.veldhuis@utwente.nl">r.n.j.veldhuis@utwente.nl</a></td>
</tr>
<tr>
<td>Tom Vermeulen</td>
<td>KU Leuven</td>
<td><a href="mailto:tom.vermeulen@esat.kuleuven.be">tom.vermeulen@esat.kuleuven.be</a></td>
</tr>
<tr>
<td>Svyatoslav Voloshynovskiy</td>
<td>University of Geneva</td>
<td><a href="mailto:svolos@unige.ch">svolos@unige.ch</a></td>
</tr>
<tr>
<td>Jos Weber</td>
<td>TU Delft</td>
<td><a href="mailto:j.h.weber@tudelft.nl">j.h.weber@tudelft.nl</a></td>
</tr>
<tr>
<td>Frans Willems</td>
<td>TU/e</td>
<td><a href="mailto:f.m.j.willems@tue.nl">f.m.j.willems@tue.nl</a></td>
</tr>
<tr>
<td>Peter De With</td>
<td>TU/e</td>
<td><a href="mailto:p.h.n.de.with@tue.nl">p.h.n.de.with@tue.nl</a></td>
</tr>
<tr>
<td>Zhijie Xu</td>
<td>IMEC</td>
<td><a href="mailto:zhijie.xu@imec.be">zhijie.xu@imec.be</a></td>
</tr>
<tr>
<td>Chris Zeinstra</td>
<td>TU Twente</td>
<td><a href="mailto:c.g.zeinstra@utwente.nl">c.g.zeinstra@utwente.nl</a></td>
</tr>
</tbody>
</table>