Optimization of hierarchical 3DRS motion estimators for picture rate conversion

Citation for published version (APA):

DOI:
10.1109/JSTSP.2010.2063014

Document status and date:
Published: 01/01/2010

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 08. Jun. 2019
Optimization of Hierarchical 3DRS Motion Estimators for Picture Rate Conversion

Adrienne Heinrich, Member, IEEE, Chris Bartels,
René J. van der Vleuten, Member, IEEE, Claus Nico Cordes, Member, IEEE,
and Gerard de Haan

Abstract

There is a continuous pressure to lower the implementation complexity and improve the quality of motion-compensated picture rate conversion methods. Since the concept of hierarchy can be advantageously applied to many motion estimation methods, we have extended and improved the current state-of-the-art motion estimation method in this field, 3-Dimensional Recursive Search (3DRS), with this concept. We have explored the extensive parameter space and present an analysis of the importance and influence of the various parameters for the application of picture rate conversion. Since well-performing motion estimation methods for picture rate conversion show a trade-off between prediction accuracy and spatial motion field consistency, determining the optimal trade-off is an important part of the analysis. We found that the proposed motion estimators are superior to multiple existing techniques as well as standard 3DRS with regard to performance at a low computational complexity.

I. INTRODUCTION

Motion estimation (ME) is an essential part of the picture rate conversion methods that are applied to eliminate film judder in high-end televisions [1]. Because of the increasing spatial resolution (from SD to HD and Full HD) and picture rates (from 24 fps to more than 200 fps) of video shown on those televisions, as well as the increasing size and quality of the television displays, there is a continuous pressure to lower the implementation complexity and improve the
quality of the ME algorithm. To this end, spatio-temporal prediction methods such as recursive search, e.g. [2], [3], [4], are typically applied in practice (e.g. [5], [6]). Generally, spatio-temporal predictors have proven to be a powerful tool in the design of motion estimation algorithms [7], [8], [9].

Combinations of 3-Dimensional Recursive Search (3DRS) [10] with concepts borrowed from alternative ME methods have shown to be beneficial in earlier publications, e.g. [11]. In this paper, we extend the 3DRS technique with the concept of hierarchy [12] which has been advantageously applied to many ME methods. This work shows the effect of parameter optimization on the quality of hierarchical motion estimation, and most importantly, that the commonly used manual parameter optimization is unlikely to arrive at an optimum due to the vast size of the optimization space. This work thus clearly indicates that sufficient attention has to be placed on the optimization of a hierarchical motion estimator, as otherwise the result is likely to be suboptimal. Therefore, we explore the extensive parameter space and provide insights with respect to the importance and the influence of the individual parameters. Since well-performing ME methods for picture rate conversion show a trade-off between prediction accuracy and spatial motion field consistency, the optimal trade-off is analyzed.

In Section II, we introduce the concept of hierarchy for ME and describe its integration with 3DRS. The designed motion estimators and their parameters are formally defined in Section III. In Section IV, we explore the parameter space in order to find the optimal range of parameter settings. A further in-depth analysis of several parameters, the performance and complexity analysis of the proposed motion estimators and existing techniques is then given in Section V and Section VI, respectively. Section VII summarizes our conclusions.

II. Hierarchical 3DRS Motion Estimation

Hierarchical ME is introduced in Section II-A, followed by a discussion of its integration into 3DRS in Section II-B.

A. Multi-scale and Multi-grid Hierarchical Motion Estimation

The ME methods discussed in this paper are based on the principle of block-matching [13]. According to this principle, the image is divided into blocks and for each block a reference image is searched for the best-matching block (according to a pre-defined cost function). In
case of motion in the sequence, the best-matching block will be located at a different spatial position in the reference image than in the current image. The vector over which the matching block has been shifted compared to the block in the current image is called the motion vector. The complexity of finding the best-matching block obviously depends directly on the number of different vectors (with corresponding cost function evaluations) that are examined. Hierarchical ME methods can take large search ranges efficiently into account [14] which reduces the risk of being trapped in suboptimal local minima.

In this paper, we investigate a hierarchical ME approach using resolution down-scaling, which we call multi-scale block-matching ME. Using down-scaling, the coarser motion vectors are obtained from block-matching at a lower spatial resolution and can be successively refined at higher resolutions. We will combine the multi-scale approach with a hierarchical ME method known as multi-grid block-matching [13]. In this method, a coarse motion vector is first found using a large block size and this vector is successively refined for the smaller blocks into which the larger blocks are decomposed (using a quad-tree decomposition). By combining the multi-scale and multi-grid approaches, we are flexible in investigating the effects of using different block sizes and scale factors.

B. Hierarchical 3DRS block-matching

3DRS ([2]) selects the output motion vector $\vec{d}$ from a candidate vector set $C$, that is based on prediction vectors from a spatio-temporal neighborhood. This process comprises two steps:

1) For each (block/pixel) location $\vec{x}$ in frame number $n$, construct a candidate set $C$, e.g.

$$C = \left\{ \vec{d}(\vec{x} + k \cdot \vec{u}_x - \vec{u}_y, n), \vec{d}(\vec{x} - \vec{u}_x, n), \vec{d}(\vec{x}, n - 1), \vec{d}(\vec{x} + \vec{u}_x, n - 1), \vec{d}(\vec{x} + l \cdot \vec{u}_x + \vec{u}_y, n - 1), \vec{d}(\vec{x} - \vec{u}_x, n) + \vec{\eta}, \vec{d}(\vec{x} - \vec{u}_y, n) + \vec{\eta} \right\}$$

$\vec{\eta}$

where $\vec{u}_x, \vec{u}_y$ are unit vectors on the block/pixel grid, and $\vec{\eta}$ is a random value. Usually this random value is drawn from a fixed update set [2].
Fig. 1. Configuration of the spatial and temporal candidates for the scanning direction indicated by the gray arrows in the block grid. The light gray block is the current block. The spatial candidates are indicated with the gray circles, the temporal candidates with the white circles (see Eq. (1)).

2) The estimated value for $\vec{d}(\vec{x}, n)$ then is

$$\vec{d}(\vec{x}, n) = \arg \min_{\vec{d}_c \in C} ( E_m(\vec{x}, \vec{d}_c, n) + E_p(\vec{d}_c^r) )$$

(2)

where $E_m$ is a common match term for which we use the Sum of Absolute Differences (SAD), while $E_p$ is a block size dependent penalty term to bias the preference among the different types of candidates $\vec{d}_c$ which is denoted by $\vec{d}_c^r$. Regarding the three candidate types, we distinguish between the spatial, temporal and update predictors which will be elaborated on in the following. We refer to the sum of the match term and the penalty term as the energy function.

Important is that steps 1 & 2 are performed sequentially for each location. Hence, the newly estimated value is assigned to the location before moving to the next location. Therefore this new value becomes part of the candidate set of the next location, directly influencing the estimate for that next location.

The underlying idea of 3DRS is that “objects are larger than blocks” and therefore already estimated neighboring vectors are good predictions for the current value to be estimated. These neighboring values are called spatial candidates ($\vec{d}(\cdot, n)$ in Eq. (1)). Unfortunately, not all neighboring values have already been estimated. However, previous estimates both in time and iteration are also good predictions, assuming “objects have inertia”. These predictions from previous estimates are called temporal candidates ($\vec{d}(\cdot, n - 1)$ in Eq. (1)), but are somewhat less reliable than spatial candidates, because of the motion of the objects and the change in motion. The reliability of the different types of predictors is taken into account by the penalty in Eq (2).
The scanning direction determines the order in which block-based motion estimation is performed. Fig. 1 shows the configuration of spatial and temporal candidates when processing from top-left to bottom-right. This is the scanning order assumed in Eq. (1). Processing solely in this order means that good motion vector estimates can only propagate in one direction. If a good estimate is found at the bottom of the image it can take some time before it is propagated to the top.

To improve the propagation of good estimates it is beneficial to vary the scanning direction. In practice two mechanisms are used for this. In the first option after a scan from top to bottom, the next scan is run from bottom to top. This is alternated continuously. The second option is called meandering meaning that after scanning a line from left to right, the next line is scanned from right to left. If both mechanisms are used, good estimates can propagate in four different directions, which ensures a quick spreading of good estimates all over the image. For an even faster propagation, two scans (both top to bottom and bottom to top) per image can be performed in a meandering manner. The example candidate set from Eq. (1) assumes scanning from top to bottom and from left to right, i.e. along the unit vectors defining the axes of the image. If the scanning direction is changed the unit vectors \( \vec{u}_x \) and \( \vec{u}_y \) in Eq. (1) should be changed in unit vectors defining the current scanning direction, \( \vec{s}_x \) and \( \vec{s}_y \).

Another important aspect concerns the update candidates (\( \vec{d}(\cdot, n) + \vec{\eta} \) in Eq. (1)). Both spatial and temporal candidates contain values that already have been estimated. However, new values need to be introduced as well to find vectors for appearing objects and to accommodate for acceleration. This is achieved by adding small random values to spatial candidates. These random values can be drawn from a random distribution, e.g. a Gaussian distribution (\( \mathcal{N} \)), but typically they are drawn from a fixed update set (\( US \)) [2].

\[
\eta_{x,y} \sim \mathcal{N}(0, \sigma) \quad \text{where} \quad \vec{\eta} = \begin{pmatrix} \eta_x \\ \eta_y \end{pmatrix}, \quad \text{or} \quad \vec{\eta} \in US
\]  

These vectors can be small since objects have inertia and in order to promote smoothness (\( \sigma \leq 2 \)). Hence, the motion of objects will only change gradually. To find a motion vector that differs significantly from previously estimated vectors, takes several consecutive updates. This process is called convergence. Updated vectors are considered the least reliable predictors and are therefore assigned the highest penalty.
Typically the penalty for spatial candidates is fixed to zero. For 16-bit image data, we empirically determined that the penalty $E_{p}$ of 128 per pixel in a block for temporal candidates and 512 per pixel in a block for update candidates produces good results.

The evaluation of the energy function in Eq. (2) is the most expensive part. In the case of 3DRS the energy function only needs to be evaluated for a few candidates, regardless of the range of the motion vectors. The size of the candidate set can be tuned to achieve good quality with a minimum number of candidates. The candidate set from Eq. (1) contains 11 candidates. However because of the smoothness of the motion field, often neighboring candidate locations result in the same prediction. Therefore the number of candidates can be sub-sampled without significant loss in quality (see the 3DRS candidate structure in Fig. 3). In this paper, we add an additional candidate vector from an ‘external’ source, i.e. a hierarchical candidate. The hierarchical candidate vector can be obtained by both multi-grid (same resolution, multiple block-sizes) and multi-scale (multiple resolution levels) approaches. The 3DRS block matching method with the additional hierarchical candidate is performed on each hierarchical level (except on the coarsest level, where the hierarchical candidate is not available and thus the non-hierarchical 3DRS is performed). On each hierarchical level, the temporal candidate vectors are propagated from the previously computed vectors on the highest-resolution image and down-scaled with regard to all the scales used in the current hierarchical ME scan. This shows better performance than when possibly unconverged temporal candidates from the same scale are used.

The introduction of a hierarchical candidate vector increases the number of candidate vector evaluations that are performed, compared to non-hierarchical 3DRS. In order to profit from the hierarchical candidate without a complexity increase, we could e.g. skip the ME on the full-resolution image and use motion vectors that are up-scaled from a lower-resolution estimation or modify the motion vector candidate structures (see Section III-B).

III. HIERARCHICAL MOTION ESTIMATION DEFINITION AND PARAMETERS

First, the hierarchical motion estimators, as well as their parameters, are defined in Section III-A. Next, the chosen parameter values are discussed in Section III-B Candidate Structures, Section III-C Scans, Section III-D Scale Parameter Sets, and Section III-E Block Sizes. Finally, an example motion estimator configuration is given in Section III-F.
Fig. 2. Illustration of multi-grid and multi-scale approach.

TABLE I
PARAMETERS FOR THE HIERARCHICAL MOTION ESTIMATOR DESIGN

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>fine</td>
<td>Lowest/Finest scale on which ME is performed</td>
<td>Scalar</td>
</tr>
<tr>
<td>coarse</td>
<td>Highest/Coarsest scale on which ME is performed</td>
<td>Scalar</td>
</tr>
<tr>
<td>$s_f_w$, $s_f_h$</td>
<td>Scaling factor width and height for resizing scales</td>
<td>Array</td>
</tr>
<tr>
<td>$b_{lk_w}$, $b_{lk_h}$</td>
<td>Block width and height of each scale</td>
<td>Array</td>
</tr>
<tr>
<td>scan</td>
<td>Amount of ME scans performed per scale</td>
<td>Array</td>
</tr>
</tbody>
</table>

A. Definitions

In order to describe the multi-scale or multi-grid approach, a scale pyramid is used, as shown in Fig. 2, where ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates. The parameters involved in the design of hierarchical motion estimators are explained in the following and an overview is given in Table I.

The relevant parameters for the scale structure are the fine scale, the coarse scale and the scaling factors $s_{f_w}$ and $s_{f_h}$. The fine scale and the coarse scale denote the levels of the pyramid (see Fig. 2) where ME is performed, e.g. $fine = 1$, $coarse = 2$. fine is the finer scale (for multi-scale ME) or the one with a finer block grid than coarse in the case that the coarse and fine scale have the same size (multi-grid ME). If the full resolution is included as a scale on which ME is performed, $fine = 0$ is chosen (otherwise $fine = 1$). The scale factors $s_{f_w}$ and $s_{f_h}$ determine the size of the scales. The scaling factors $s_{f_w}$ and $s_{f_h}$ for width and height
are arrays which indicate how much one scale is down-scaled in comparison to the next lower scale in the pyramid. The first component, i.e. $s_{fw}[0]$, denotes the scaling factor between the full resolution image and the following higher level of the pyramid, the second component, i.e. $s_{fw}[1]$, the scaling factor between the first down-scaled image and the next higher scale in the pyramid etc. In the case of a multi-scale motion estimator (right image in Fig. 2), the image dimensions become smaller as we ascend in the pyramid. However, when a multi-grid (left image in Fig. 2) motion estimator is designed, two scales have the same dimension, thus the corresponding scaling factor component equals 1. As the spatial resolution of two vector fields from two different scales may not be equal, this may require scaling of the vector field as well, which is implemented as nearest neighbor scaling.

The block width and block height dimensions $blk_w$ and $blk_h$ are arrays where the elements $blk_w[i], i = 0, \ldots, \text{coarse}$, indicate the block sizes for each scale $i$ in the pyramid. The equivalent block width dimensions on the full resolution image can be computed as in

$$blk_w'[0] = blk_w[\text{coarse}] \cdot \prod_{i=0}^{\text{coarse}-1} s_{fw}[i], \text{for } \text{coarse} > 0.$$ (4)

The equivalent formula is applied to the block height dimensions. The number of ME scans performed on each scale is defined by the parameter $scan$ which is also an array. In this experiment, all the elements were chosen to take identical values, i.e. either all equal to 1 or all equal to 2.

The random update vectors in both, positive and negative, horizontal and vertical direction are chosen with quarter-pixel accuracy. The lengths of the update vectors are discretized to 0.25, 0.5, 1 and 2. The length of the update vectors is not changed throughout the scales and thus not down-scaled proportionally to the scaling factor, which should favor a fast convergence speed. In order to find the zero motion of stationary image parts such as subtitles and logos faster, the zero vector is included as an additional motion vector candidate with a high penalty, set equal to the update penalty.

B. Candidate Structures

Different numbers of candidates and different approaches are applied to determine the motion vector candidates, as shown in Fig. 3. In contrast to the usual 3DRS candidate structure (also shown in Fig. 3), the temporal candidate is closer to the current block for all the hierarchical
Fig. 3. The usual 3DRS candidate structure, as well as nine different subsamplings of the spatio-temporal neighborhood of a block (candidate structures 1,...,9) are shown. C denotes the current block for which candidate motion vectors are determined, S a spatial candidate, U a random update vector added to the spatial candidate, T a temporal candidate, and H the hierarchical candidate resulting from the ME scan on a coarser grid or on a coarser scale. For candidate structures 4 and 8, \( d = 1/60 \).

approaches because, for coarse scales, the temporal candidate may come to lie outside the object in which the current block is located.

Candidate structures 1 and 2 are intended to determine the performance of simple candidate structures that resemble the usual 3DRS structure. Candidate structures 5 and 6, on the other hand, are intended to determine the performance of very complex candidate structures with many candidates.

Candidate structures 3, 4, 7, 8 and 9 are quite complex in their design. The goal of candidate structures 3 and 7 is to speed up the convergence. Candidate structure 3 includes a candidate which least resembles the spatial candidate S, with respect to its angle, whereas candidate structure 7 adds the longest vector which is computed by comparing the sum \( |v_x| + |v_y| \) of the
absolute value of the vector components $v_x, v_y$. Candidate structures 4 and 8 choose different types of candidates (8) or a different location of the candidates (4) depending on the ratio between the block size and the scale dimension $\text{dim(block)}/\text{dim(scale)}$. Candidate structure 9 includes a motion-compensated candidate by projecting the motion vectors found in the previous scan to the new block locations in the current image.

C. Scans

In our experiments, the motion estimation scans are performed in a meandering manner either once or twice for each scale of a designed motion estimator.

D. Scale Parameter Sets

Different scale structures were selected by varying the values of $\text{fine, coarse, sf}_w$ and $\text{sf}_h$. We chose simple structures involving at most two scales on which ME is performed. The benefit of multi-scale and multi-grid ME in comparison with 3DRS ME on a down-scaled version of the input image were investigated. The multi-scale estimators are B and D in Table II (2 scales), where B performs the last ME step on the full resolution and D on a down-scaled version. The multi-grid estimators are C and E in Table II (1 scale), where C performs ME only on a down-scaled image and E on the full resolution. Finally, the 3DRS ME on a down-scaled version of the input is described by scale structure A in Table II.

The investigated parameter settings of the scale structures shown in Table II are given in Table III.
E. Block Sizes

For each row in Table III, the block width and block height are selected from the set of possible block sizes \{2, 4, 8, 16, 32, 64\}. Non-square blocks are included as well, however only when the block width is twice as large as the block height. When the last ME scan is performed on the full resolution image (\textit{fine} = 0), the block width and height for scale 0 are chosen to be either 2, 4 or 8.
An example of a hierarchical motion estimator configuration is given in Table IV. The candidate and scale structures and block sizes have been explained in Section III-B, Section III-C, Section III-D, and Section III-E, respectively.

IV. Quantitative Analysis

In this section, the performance of the different motion estimator (ME) parameter settings will be evaluated for the application of picture rate conversion (see [15], Chapter 4). Therefore, we selected ten Full HD test sequences (see Figure 4) with a duration of 3 frames that address common challenges in ME, such as several layers with different motion, repetitive structures, small objects, subtitles and ticker tapes, de-interlaced images with typical de-interlacers of average quality (e.g. [16]), large motion, and occlusion areas. To ensure a satisfactory performance with less challenging test material, we also included fairly straightforward sequences for ME as well as a repeated still image. We expect a well-performing ME to have a good average performance for all challenges. For individual challenges, we acknowledge that other ME parameter settings may render a better result, however, the objective in the ME design for picture rate conversion remains a good overall performance. Therefore, the average performance over all test sequences is compared. In order to analyze the behavior of the motion estimators with respect to convergence speed and steady state performance, two motion vector initializations are chosen as described in Section IV-A. The objective measures used to evaluate the motion estimator performance are introduced in Section IV-B, followed by the evaluation itself in Section IV-C.

A. Motion Vector Initialization

In order to examine the different motion estimators with respect to the convergence speed of the motion vectors and regarding their performance in the steady state, when the motion field is already converged, two different initializations are chosen. Firstly, for evaluating the convergence speed, a zero vector initialization is used. Such unconverged states occur frequently, not only in scene changes but, more importantly, when a tracked object reappears from behind an occluding area or when accelerations and irregular motions are involved (e.g. up and downwards moving head of walking person). Secondly, an initialization with converged motion vectors is performed.
Fig. 4. Test sequences used for the quantitative evaluation. Note that the bottom row is reused for still images and de-interlacing with a typical de-interlacer, e.g. [16].

To save computation time for the analysis, the motion vectors used for the second initialization are computed with a fixed multi-grid ME of which the parameter settings are given in Table IV.

B. Performance Measures

Two fundamental characteristics are recognized as the basis of ME design: The brightness constancy assumption when the true motion is found and the smoothness constraints to enforce consistent motion fields within a moving object. The trade-off between smoothness terms and brightness constancy in the form of luminance comparisons has already become apparent in the early optical flow advances [17]. The metrics developed for high-performance ME methods for retiming show comparable features and the known trade-off between prediction accuracy and spatial motion field consistency. It is recognized in [3] and [10], that accurate predictions at a highest possible consistency are necessary for a satisfactory viewing experience. Relevant metrics addressing the temporal continuity and spatial consistency of the motion vectors are documented in [3] and [10]. The prediction accuracy and temporal continuity are quantitatively assessed with the ‘M2SE’ [10],

\[
M2SE(n) = \frac{1}{n_h \cdot n_w} \sum_{\bar{x} \in W} (F_0(\bar{x}, n) - F_1(\bar{x}, n))^2, \tag{5}
\]

and the spatial inconsistency measure ‘SI’ is based on [10].
\[\text{SI}(n) = \sum_{x_b \in W_b} \sum_{k=-1}^{1} \left( \frac{\mid \Delta_x(x_b^k, k, l, n) \mid + \mid \Delta_y(x_b^k, k, l, n) \mid}{8 \cdot N_b} \right),\] 

(6)

where \(n_h\) and \(n_w\) are the image height and width in pixels, respectively, \(W\) is the set of all the pixels in the entire image, \(F_o(\vec{x}, n)\) the luminance of the original image at position \(\vec{x}\) and at the temporal position \(n\). \(F_i\) is the motion compensated average of frames \(n - 1\) and \(n + 1\) by applying the vectors estimated for frame \(n\), \(\vec{x}_n^b\) the position of the block \(b\) among the set of all the blocks \(W_b\) in the entire image, \(N_b\) the number of blocks in an image and

\[\Delta_x(x_b^k, k, l, n) = d_x(x_b^k, n) - d_x(x_b^k + \left(\frac{k}{l}\right), n),\] 

(7)

\[\Delta_y(x_b^k, k, l, n) = d_y(x_b^k, n) - d_y(x_b^k + \left(\frac{k}{l}\right), n),\] 

(8)

where \(d_x\) and \(d_y\) are the computed motion vectors. Different block sizes in the SI metric, e.g. 8x8-pixel blocks vs. 1x1-pixel blocks, return different results due to the metric bias towards larger motion vector blocks, thus appropriate block dimensions should be chosen for the set of MEs one would like to compare (8x8-pixel blocks in this paper).

The PSNR depends on the number of bits \(NB\) used for representing the video data and is calculated from the M2SE (\(\text{PSNR}(n) = 10 \cdot \log_{10} \left( \frac{(2^{NB} - 1)^2}{\text{M2SE}(n)} \right) \)). It measures how well the interpolation result corresponds to true motion using temporally extrapolated motion vectors, whereas the SI indicates the spatial smoothness of the computed motion field. The motion field and interpolated images are evaluated after performing ME on the second image of the input sequence since the pixels from a previous image are included in the M2SE computation. Note that, for these measures, ME is performed at the original image position and not at the interpolated position. Thus, a motion vector is assigned to each occurring element in the original image rendering it unlikely to miss small objects which may be the case when ME is performed on the interpolated position.

All the motion vectors are computed without applying any post processing such as block erosion [18] in order to facilitate an easier analysis of the results. This is assumed correct because the SI and PSNR measures are expected to indicate the same tendency and ranking of the MEs with and without applying block erosion.
C. Performance Evaluation

Since both a high PSNR as well as a consistent motion field are characteristics of a good ME, a PSNR - Consistency plot as shown in Fig. 5 is introduced as a means to capture the achieved PSNR performance in relation to the consistency of the motion field. The inverse mean of the PSNR and the mean inconsistency values (SI) are plotted in the following sections by computing the average performance of all parameter setting combinations with regard to the different test sequences. The optimal ME with a high PSNR and a low inconsistency is located in the bottom left corner. A ME which surpasses all the others in one regard (either consistency or PSNR) is called a Pareto-optimal or an ‘optimal’ motion estimator. The set of optimal MEs lies on the ‘optimal trade-off curve’ as described by [19].

The PSNR-Consistency graphs in Fig. 5 and Fig. 6 depict the metric results of 13320 hierarchical MEs (6660 MEs in the steady state and 6660 MEs in the un converged state) which are created based on all possible parameter combinations (i.e. varying candidate structures, scale structures, block sizes and scans) described in Section III. For each ME, the average performance with respect to the different test sequences was computed. A wide spread of the MEs in the un converged state is visible in Fig. 5. The best MEs lie close to the optimal trade-off curve. Therefore, the optimal contour lines of the hierarchical MEs are depicted in Fig. 6 where a compromise between PSNR and consistency performance is attained. It is visible that
an improvement in both motion field consistency and PSNR is achievable for a hierarchical ME with regard to the traditional 3DRS ME.

The contour lines in Fig. 6 indicate often a better consistency performance in the unconverged than in the steady state. Note that the unconverged state denotes merely that the motion vector initialization was chosen to be zero but does not exclude the fact that a converged motion field may result already in the second image. This was the case for the MEs on the contour line. Furthermore, the lower consistency in the steady state might be related to the fact that a converged default motion field is used in the initialization which is not computed with the tested ME but with the one given in Section IV-A.

Since it is not evident which part of the optimal contour line a good visual quality corresponds to, seven optimal MEs were chosen for an initial analysis. Their characteristics are summarized in Table V.

Along the curve of Fig. 6, from low to high PSNR up to Opt. ME 6, the following motion field improvements are observed. The related picture rate conversion benefits are confirmed when playing back the interpolated sequence.

- Improved spatial consistency of the motion field
- Better alignment of motion vectors with the edge of moving objects, resulting in reduced artifacts in occlusion regions
TABLE V

SELECTED MEs ON OPTIMAL CONTOUR LINE. ALL OF THEM INCORPORATE THE ZERO VECTOR AS AN ADDITIONAL CANDIDATE.

(a) Steady state motion field, Opt. ME 5  (b) Steady state motion field, 3DRS

(c) Unconverged motion field, Opt. ME 5  (d) Unconverged motion field, 3DRS

Fig. 7. Motion field visualized with color overlay of 3DRS and Opt. ME 5 in example sequence chosen for subjective assessment.
• Improved temporal consistency of the motion field, resulting in reduced flickering
• Higher convergence speed (for the steep part of the contour line)

When the metric results indicate a high spatial consistency of the motion field, there are large temporal motion field inconsistencies and artifacts due to erroneous consistent motion vectors across the motion edges. Opt. ME 1-3 show indeed spatially consistent vector fields (e.g. large zero vector areas) but their quality is unacceptable due to the produced local judder when played back. Along the contour line, an improvement of the motion vectors regarding the object alignment (clearly better with Opt. ME 4) is visible which causes less occlusion artifacts. The motion field is temporally still quite inconsistent which produces flickering and visibly varying artifacts at motion boundaries. Overall, Opt. ME 6 shows the best visual quality. The hierarchical MEs Opt. ME 5 and Opt. ME 6 obtain a similar PSNR value as 3DRS but a higher consistency measure for both the unconverged and steady states. When comparing 3DRS with Opt. ME 5 (Fig. 7(a) and Fig. 7(b)) in the steady state, a smoother motion field (encoded in color) is clearly visible with Opt. ME 5 in the background and in the legs of the leopard. For the unconverged state, the motion fields obtained by performing ME between the second and the third image are shown in Fig. 7(c) and Fig. 7(d). The consistency increase and faster convergence of Opt. ME 5 in comparison with 3DRS is apparent as well. This corresponds well with the expected added value of the hierarchical candidate with respect to larger search ranges and the rare selection of false local minima. The performance degrades when going beyond Opt. ME 6. The motion field of Opt. ME 7 is perceived as noticeably more inconsistent which leads to disturbing flickering artifacts. The relevance of the spatial inconsistency metric is confirmed since a high PSNR at the cost of consistency is not preferred.

We expect the performance gain achieved with the addition of hierarchical layers to stagnate when more than a couple of layers are applied. In order to get a visual impression of the qualitative contribution when more than 2 scales are used, the visually best performing ME, Opt. ME 6, was extended to a ME with 5 scales. An informal subjective evaluation showed hardly any visible differences in terms of the listed motion field improvements mentioned earlier. This is in correspondence with the quantitative improvements of 0% in PSNR and 7% in SI, indicating that using more than 2 scales has only minor performance benefits.
Fig. 8. Range of optimal hierarchical MEs in steady state (red) and unconverged state (blue)

V. DETAILED PARAMETER ANALYSIS

In order to carry out a more representative analysis and to allow for slight imperfections in the metrics, also the MEs within a particular distance from the optimal contour lines are investigated. For the 16-bit HD data, the considered range was chosen to be $\delta(1/\text{PSNR}) = 0.0005$ and $\delta(\text{SI}) = 0.1$. The MEs within this range are shown in Fig. 8. The contour lines are further divided into two segments as shown with the dashed line. When comparing common settings among the MEs, it is chosen to take into account all the MEs on the right hand side of the dashed line where the PSNR hardly decreases and room for a rather large improvement in terms of consistency is given.

Section V-A discusses the candidate structures, Section V-B the amount of scans, Section V-C investigates the scale parameter sets, and Section V-D the block sizes. The resulting optimal hierarchical MEs are summarized in Section V-E.

A. Candidate Structures

An overview of the performance of the optimal MEs regarding the candidate structures is given in Fig. 9. The contour lines of all the optimal MEs for each candidate structure are given in Fig. 9. The description of the different candidate structures can be found in Fig. 3. For comparison, the performance of 3DRS is illustrated as well. It is clearly visible that the two candidate structures with the least (4) candidates (1S1TC and 2Unlike) perform worst. This indicates the necessary number of prediction vectors for a satisfactory performance. The importance of spatial predictors on a large resolution with small block sizes is apparent in the suboptimal performance of the
candidate structure 5H1TC. The optimal MEs of the other candidate structures in Fig. 9 achieve a more or less similar metric result. Especially for the unconverged state, a significant increase in consistency (around 1) and PSNR (around 1.2 dB) compared with 3DRS is found. In the steady state there is mainly room for a consistency increase.

ME candidate structures that yield more often an optimal ME are preferred as they are assumed less sensitive to varying settings than other candidate structures and thus more robust. The goal is to find optimum settings for both the unconverged and the steady state. For a practical
implementation in real-time applications, however, it can be useful to discriminate between the unconverged and the steady state and choose the best candidate structure for each state. With respect to the range of optimal MEs, the distribution of the candidate structures is given in Fig. 10 (the steady state case is comparable to the unconverged state). The numbers on the x-axis correspond to the candidate structure numbers in Fig. 3 and the y-axis to the ME count. For both states, the distribution in the interesting segment indicates that a good performance can be achieved with the candidate structures 2, 4, 5, 6 and 9. Hence, it may be interesting to use the most straightforward candidate structure, 2, as it contains the least number of candidates and does not require a complex implementation (for candidate structure 9 which only involves one more candidate a higher computational complexity is expected due to the motion compensated candidate). These results suggest that a minimum number of prediction vectors is needed for a satisfactory performance and that most of the necessary information is contained in this minimum candidate set.

B. Scans

Performing two ME scans per scale generally renders a better overall performance than only one scan per scale (occurrence rate of 67% vs. 33% respectively) since good motion vectors found close to object edges can be refined and propagated to other parts within the object. Two scans with an occurrence rate of 81% in the unconverged state are found to be particularly useful for a fast convergence.

Note that the total number of scans performed when computing the resulting motion field of one image is dependent on the number of scans per scale and the number of scales or block
grids used. For e.g. the multi-grid approach with 1 scan, the total number of scans is equivalent to the case with the scale structure A (indicated in Table II) and 2 scans.

C. Scale Parameter Sets

In this section, the scale structures and scaling factors are discussed. Particularly the unconverged state shows a clear discrepancy between the scale structures. The contour lines in Fig. 11 depict the optimal quantitative performances of the MEs for each scale structure in the unconverged state. Particularly the scale structures ‘2 scales, fine 0’ and ‘1 scale, fine 0, multi-grid’ show a significant decrease in consistency and/or PSNR suggesting suboptimal high-frequency content (e.g. noise) in the full resolution image. With the selected test sequences addressing natural content, removing the higher frequencies by downscaling the input image thus does not show any visible drawbacks. In the steady state there is no noticeable difference among the other three scale structures. However, when the motion vector is not yet converged the multi-scale/multi-grid approach (‘1 scale, fine 1, multi-grid’, ‘2 scales, fine 1’) seems beneficial for both PSNR and consistency which confirms the hypothesis of the added value of an hierarchical candidate.

In the analysis of the distribution of the five scale structures (analogously to the candidate structures in Fig. 10) we found that the scale structures C and D are the most represented groups (84 %). Scale structure D occurs around 35 % more often than C, thus using two different scales seems to be of advantage.

When analyzing the distributions of the scaling factors the following is observed. The finest scale on which ME is performed is dominated by the scaling factor 2. The full resolution with scaling factor 1 is rarely chosen. Apparently, the highest frequencies (such as noise) in the image do not contribute to a more accurate ME. In the multi-scale approach, the coarse scale which is added for fast convergence and consistency shows, as expected, higher scaling factors (4 and 8 occur approximately equally often).

D. Block Sizes

We assume that larger blocks and/or coarser scales would improve the convergence speed and large object area smoothness, and smaller blocks on the fine scale would serve as a refinement of the motion field obtained on the coarse scale. For the fine scale in the context of HD sequences,
block dimensions in the neighborhood of 8x8 blocks on the full resolution would be plausible since experience on SD content has shown that 8x8 blocks are a good trade-off between PSNR and SI [20].

The distribution of the block dimensions for multi-scale MEs is illustrated in Fig. 12. The data reveals that the block sizes of MEs using 1 scale are similarly distributed as the ones of the fine level of the multi-scale MEs (in order to avoid repetition and limit the figures, only the graphs corresponding to the multi-scale case are shown). The dominant width and height dimensions in the well performing segment range from [8,32] and [4,16] respectively. The block sizes of the multi-grid motion estimators which are included in the 1-scale case are more concentrated than the ones of the multi-scale MEs (see the large spread of the coarse level block sizes) indicating that more similar block dimensions are selected when the same scale is re-used. On average, the block width and height of the selected MEs on the coarse scale range from [32, 128] and [16,64], respectively. Based on these observations, we conclude that multi-scale MEs make use of the varying frequency content and are more robust when different block sizes are used.

E. Optimal Hierarchical Motion Estimators

Based on the parameter analysis in the previous sections, we propose to employ the multi-scale MEs with candidate structure 2S1TC, scale structure D and 2 scans. An overview of the proposed parameter settings of this ME type is given in Table VI where block settings, performance and
complexity are rendered. The fourth row of Table VI shows the mean performance for the 62 most robust MEs. The range of block width and height settings were analyzed in more detail. Therefore, their distributions were considered as probability distributions of settings for good performing MEs and their expectation value a good approximation of a robust well-performing ME given in the seventh row. The selected block sizes indicate that larger blocks are suited for HD content. When applying one of the two scale factor settings of scale structure D given in Table III, the resulting ME happens to coincide with Opt. ME 6 in Figure 6 which was visually perceived as the most pleasing ME among the seven MEs on the contour line.

VI. RESULTS

We measured the quantitative performance and computational complexity of various MEs. The computational complexity is expressed in the number of block comparisons $n_{\text{BC}}$. For the proposed MEs, this can also be derived from

$$ n_{\text{BC}} = \sum_{i \in W_{\text{scals}}} n_h \cdot n_w \cdot \frac{\text{scan}(i) \cdot n_{\text{cand}}}{\text{blk}_{w,\text{fullRes}}(i) \cdot \text{blk}_{h,\text{fullRes}}(i)}, $$

where $W_{\text{scals}}$ is the set of all the used scales, $\text{scan}(i)$ the number of scans on scale $i$, $n_{\text{cand}}$ the number of motion vector candidates ($n_{\text{cand}} = 7$ for the hierarchical MEs due to the addition of the zero vector candidate), $\text{blk}_{w,\text{fullRes}}$ and $\text{blk}_{h,\text{fullRes}}$ the width and height of the equivalent block sizes for the full resolution scale.

![Fig. 12](image_url) Range of optimal MEs using 2 scales. Distribution of equivalent block sizes for full resolution image.
### TABLE VI

Performance and complexity analysis of proposed MEs, 3DRS and various techniques documented in literature. Block width and block height indicate the equivalent block sizes for the full resolution where the selected settings for the fine and course scales can be a range ([..]) of values.
Table VI gives an overview of the SI and M2SE-PSNR values and the number of block comparisons of the different recursive-search MEs that are proposed, as well as the benchmark results from several methods described in literature. These include full-search (FS) and reduced-search pattern based methods, i.e. three-step-search (TSS) [23], one-at-a-time-search search (OTS) [24], diamond search (DS) [25] and hexagon-based-search (HEXBS) [26], as well as algorithms based on spatio-temporal predictors, i.e. the predictive (zonal) search methods MVFAST [27] and EPZS [9], the recursive search methods 3DRS [2] and HRNM [21], and combined hierarchical-predictive methods, i.e. the MRST-method proposed in [14] and MPMVP from [4]. In the steady state, we simulate the convergence mode for these methods by iterating the corresponding MEs ten times on the first image. Note that the M2SE-PSNR metric favors ‘true’ motion, i.e. MEs with a better vector field consistency can outperform a full-search method. Furthermore, all methods from literature were adapted and tested with smaller block dimensions (e.g. 8x8), however, no improvement in PSNR and SI was observed.

In comparison with standard 3DRS with two scans, the proposed hierarchical MEs achieve a complexity reduction of 38% while outperforming 3DRS on average by 0.7 dB. This holds particularly for the unconverged state with an improvement of more than 1 dB and 7% in consistency. Even the sophisticated HRNM ME [21], with a significantly higher complexity due to 3-picture estimates, is surpassed in the unconverged state (PSNR difference of 0.9 dB). However, in the steady state, HRNM shows a clearly better performance than any of the hierarchical MEs. From these observations we conclude that for the unconverged state, a combination of the hierarchical approach and HRNM may be beneficial for both computational complexity and performance.

The benchmark further shows that the non-predictive (reduced-)search methods FS, TSS, OTS, DS, and HEXBS are generally unsuitable for picture rate conversion. As these methods purely optimize for minimal ‘residue’ in the match criterion, they produce highly inconsistent vector fields (with PSNR values smaller than 24 and/or SI values larger than 4). The predictive search methods generally perform better, as they (implicitly) enforce vector field consistency, with the methods EPZS and MRST achieving the best steady-state PSNR performance (slightly below the proposed MEs). Among these, when taking the computational complexity into account, EPZS is identified as the ME achieving the best compromise between performance and complexity. Yet, its spatial inconsistency is more than 50% larger than the SI values of the proposed MEs,
and this has a large impact on the perceived picture quality. We conclude from these results that
the proposed hierarchical MEs are superior to multiple existing techniques as well as standard
3DRS with regard to combined PSNR/SI performance at a low computational complexity.

VII. CONCLUSION

Hierarchical ME promises fast convergence of motion vectors, a high motion field consistency
and a small M2SE error. In this paper, we introduced the concept of hierarchical ME to
3D-Recursive Search (3DRS), and we performed a design-space exploration of the extensive
parameter space to provide insights into the importance and influence of individual parameters.
In particular, a quantitative analysis was performed by determining the PSNR and Spatial
Inconsistency (SI) of 13,320 hierarchical MEs to show the trade-off between spatial consistency
and match quality.

In general, we found that applying the hierarchical approach to 3DRS does not require complex
candidate structures in order to perform well. In fact, straightforward candidate structures having
relatively few candidates already offer a good overall performance, i.e. one that is close to the
optimal trade-off curve. Furthermore, we identified that multi-scale MEs are amongst the best
performing hierarchical MEs, closely followed by multi-grid MEs on down-scaled images, with
these being hindered by a lower robustness with respect to varying block sizes.

Based on the design-space exploration, a ME configuration is proposed that offers an im-
provement of more than 1 dB over 3DRS in the unconverged state, and of 0.7 dB on average.
At the same time, the computational complexity is reduced by 38%. When benchmarking the
proposed MEs to various other techniques, the results show a superior combination of PSNR/SI
performance while offering a low computational complexity.

We also showed that, in comparison to a sophisticated ME approach using 3-picture estimates
(HRNM), the proposed hierarchical MEs offer better results in terms of both image quality and
complexity in the unconverged state. Therefore, as future work, a combination of the hierarchical
approach and HRNM may be investigated to identify whether the combination offers further
improvements in performance and/or computational complexity.

ACKNOWLEDGMENT

The authors would like to thank Rui Guerreiro for his contribution to this work.
REFERENCES


Adrienne Heinrich was born in 1980 in Zurich, Switzerland. She studied Electrical Engineering and Information Technology at the ETH Zurich and received her M.Sc. degree in 2006 from the ETHZ. In the same year, she started working in the Video and Image Processing group of the Philips Research Laboratories in Eindhoven, The Netherlands. Currently, she is working as a Research Scientist to improve algorithms in the area of motion estimation and picture rate conversion.

Chris Bartels was born in 1981 in Bunde, The Netherlands. He studied Electrical Engineering at Eindhoven University of Technology and received the M.Sc. degree in 2006. In the same year, he started pursuing the Ph.D. degree at the Electronic Systems group of the same institute on a grant from Philips Consumer Electronics. His research work focuses on signal processing algorithms for real-time video enhancement. He has published multiple papers on the topics of on-chip communication infrastructure, motion estimation, occlusion detection and frame rate conversion.
René J. van der Vleuten was born in Valkenswaard, the Netherlands. He received the M.Sc. degree in Electrical Engineering from Eindhoven University of Technology and the Ph.D. degree in Electrical Engineering from Delft University of Technology.

Since 1995, he has been with Philips Research in Eindhoven, working on compression of digital audio, images, and video as well as on various video processing topics, including compression-artifact reduction, wide-gamut color acquisition and conversion, and motion-compensated frame-rate conversion. Currently, he is involved in the standardization of the next generation of wireless high-definition video links in the WirelessHD consortium.

René has a broad experience in leading multinational research projects. He has led projects involving colleagues and students of various national origins who were working at Philips Research laboratories in the Netherlands, France, and the U.S.A. and he is currently leading the European project 3D4YOU. This project, which is developing the key elements of advanced future 3D television systems, involves eight industrial and academic partners from Germany, France, the U.K. and the Netherlands.

René is a member of the IEEE. He has (co-)authored 6 journal papers and around 35 conference papers. He also made 11 contributions to MPEG audio and video standardization. He holds 28 U.S. Patents and his work has been applied in all Super Audio CD players, millions of Philips TVs, and millions of high-end video processing ICs from NXP Semiconductors.

Claus Nico Cordes graduated in the field of video processing algorithms and started his career as a research scientist at the Philips Research Laboratories in Eindhoven. Here, he initially focused on the signal processing aspects of OLED and LCD televisions, including color management, contrast enhancement and motion portrayal.

In 2005, he switched to the area of video scan-rate conversion, where he led a project on motion estimation, motion compensation and related algorithms. His activities included the design of new algorithms, their cost optimization and competitor analysis. In 2008, he continued this work as a senior scientist at NXP Semiconductors. Nico’s work has resulted in 13 published patent applications.
Gerard de Haan received B.Sc., M.Sc. (Cum Laude), and Ph.D. degrees from Delft University of Technology in 1977, 1979 and 1992, respectively. He joined Philips Research in 1979. He has led research projects in the area of video processing, and participated in European projects. Since 1988, he teaches post-academic courses for the Philips Centre for Technical Training at various locations in Europe, Asia and the US. In 2000, he was appointed Research Fellow in the Video Processing and Analysis group of Philips Research Eindhoven, and part-time full Professor at the Eindhoven University of Technology teaching “Video Processing for Multimedia Systems”. He has a particular interest in algorithms for motion estimation, video format conversion, image enhancement/analysis and computer vision. His work in these areas has resulted in 3 books, 2 book chapters, about 145 scientific papers, more than 100 patents and patent applications, and various commercially available ICs. He received 5 Best Paper Awards, the Gilles Holst Award, the IEEE Chester Sall Award, bronze, silver and gold patent medals, while his work on motion received the EISA European Video Innovation Award, and the Wall Street Journal Business Innovation Award. Gerard de Haan serves in the program committees of various international conferences on image/video processing and analysis.