A real-time video surveillance system with human occlusion handling using nonlinear regression

Citation for published version (APA):

DOI:
10.1109/ICME.2008.4607432

Document status and date:
Published: 01/01/2008

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

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.
A REAL-TIME VIDEO SURVEILLANCE SYSTEM WITH HUMAN OCCLUSION HANDLING USING NONLINEAR REGRESSION

Jungong Han\textsuperscript{a}, Minwei Feng\textsuperscript{a}, and Peter H.N. de Wijs\textsuperscript{a,b}

\textsuperscript{a}University of Technology Eindhoven, 5600MB Eindhoven, The Netherlands
\textsuperscript{b}CycloMedia Technology, 4180 BB Waardenburg, The Netherlands
jg.han@tue.nl

ABSTRACT

This paper presents a real-time single-camera surveillance system, aiming at detecting and partly analyzing a group of people. A set of moving persons is segmented using a combination of the Gaussian Mixture Model (GMM) and the Dynamic Markov Random Fields (DMRF) technique. For a better extraction of the human silhouettes, the energy function of DMRF is extended with texture information. The mean-shift algorithm is utilized to track multiple people over the sequence. To address the human-occlusion problem, we model the horizontal projection histograms of the human silhouettes using a nonlinear regression algorithm. This model enables to automatically locate the people during the occlusions. Experiments show that the proposal has nearly same performance (also with occlusion) as the particle-filter with the benefit of being a factor of 10-20 faster in computing.

1. INTRODUCTION

Video surveillance is becoming popular in ensuring security at airports, train stations, banks and even normal offices. However, two research problems remain to be addressed before we can fully rely upon such video surveillance systems. First, a good balance between complexity and accuracy of the system is required, because only real-time systems with sufficient accuracy can contribute to real applications. Second, the tracking is more difficult when there are small groups of people moving together, or interacting with each other. In these cases, individual people are not visually isolated, but are partially or totally occluded by other people.

We can classify existing surveillance systems, according to their efficiency (real time or not) and their functionality (track single person, multiple persons, handle occlusion). Pfister [1] solves the problem of real-time tracking of people in complex scenes in which there is a single unoccluded person and fixed camera. The system called $W^d$ [2] is a real-time visual surveillance system for detecting and tracking people and their body parts, and monitoring their activities in an outdoor environment. However, this system has a limited capability to handle occlusions. Mean-shift [3] is a real-time non-parametric technique that searches along density gradients to find the peak of probability distributions. This approach is computationally effective, but it is susceptible to converge to a local maximum and the occlusion remains problematic. The particle-filter technique [4] performs a random search guided by a stochastic motion model to obtain an estimate of the posterior distribution describing the object’s configuration. This method is robust in the sense that it allows to handle clusters in the background, and it recovers from temporal loss of tracking. Unfortunately, there are high computational demands in the approach, and this is the bottleneck to apply it in real-time systems. In the layer tracking approaches, such as [5], shape, motion and appearance models of the different layers corresponding to foreground objects and the background are estimated along with the ordering of the layers. Foreground layers behind background layers are labeled as occlusion. The drawback of these methods is the high computational cost and maintaining the layer visibility. Summarizing, none of the approaches solves simultaneously the two primary problems mentioned at the start of this section.

This paper proposes a real-time system to detect, track and analyze a group of people, which contributes in two aspects. First, unlike the existing object-segmentation algorithms that only consider the spatial and temporal information of the object, our method seamlessly incorporates texture information as a pixel feature into the DMRF framework. With our method, the silhouettes of the human can be accurately extracted. Second, we propose an occlusion-handling approach, that models the horizontal projection histograms of the human silhouettes, using a nonlinear regression algorithm. Our model effectively changes the person locating during the occlusion into the finding of the relative maximum points on a smooth curve (function), so that visual persons in the partial or complete occlusion can still be precisely captured. Then, a hypothesized person can be identified by matching its appearance information with a predefined human template. Our algorithm is designed to be robust against the cases where the silhouettes of the human are fragmented.

2. SYSTEM OVERVIEW

Fig. 1 shows the flowchart of the proposed system, which consists of four components.

1. **Object segmentation.** We perform an adaptive background subtraction [6] to produce the initial masks for the moving
persons. These masks are input to a DMRF-based segmentation algorithm [7] that incorporates the spatial coherence for robust foreground extraction.

2. **Oclusion detection.** This component distinguishes the occlusion and the non-occlusion cases, since we use different methods to treat these two cases. In our method, if the distance between two persons is smaller than a predefined threshold in the current frame, we consider that the occlusion may happen in the next frame. Otherwise, we keep the object descriptions separated.

3. **Object tracking.** The mean-shift algorithm [3] is explored to track moving persons in the non-occlusion case, and our nonlinear regression algorithm is used to deal with occlusion.

4. **Trajectory generation.** We base the trajectory generation on our previous work [8], which adopts a Double Exponential Smoothing (DES) operator to model the position of people collected from the tracking component. Using this method, we can obtain more accurate and smooth moving trajectories.

In this paper, the emphasis is on our contributions: texture-controlled object segmentation and human occlusion handling. More details about other algorithms, e.g., adaptive-background subtraction, mean-shift tracking and DES-based trajectory generation, can be found in [6][3][8].

### 3. Texture-Controlled Object Segmentation

Markov Random Fields (MRF) is a popular model used for moving-object segmentation, which takes spatial coherence of the pixels into account instead of considering the pixels independently. The problem in this model is to minimize an energy function, which can be solved by graph-cut algorithms. However, this method cannot be applied to real-time applications due to its high computational requests. Dynamic Markov Random Fields (DMRF) technique [7] is an extension work for MRF, which is suitable for real-time applications. The basic idea of DMRF is to start the energy-minimization procedure of the current frame from the optimization results of the previous frame, since there are only slightly different energy terms between two successive frames in a video.

In this paper, we base our segmentation algorithm on the DMRF [7] technique, but further strengthen it in two ways. First, we improve the initialization step in [7] by using an adaptive background subtraction [6] algorithm, in order to achieve more accurate initial masks for moving persons. This algorithm maintains a Gaussian-mixture probability function for each pixel separately, where the parameters for each Gaussian distribution are updated in a recursive way. This algorithm also associates with a shadow-detection component to remove shadow pixels. Secondly, when designing the energy function for DMRF, we involve the texture information of objects rather than only considering the spatial and temporal information [7]. More specifically, if a pixel is labeled as the boundary pixel by our texture analyzer, the relation energy covered by an edge-energy term, between this pixel and its neighborhoods is assumed to be smaller, because we expect that the boundary pixel is more relies on its node energy, but reduces the influences from its neighborhoods. With our idea, the boundary of the object can be extracted perfectly, thereby facilitating our silhouette-based occlusion handling. Mathematically, the energy function to be minimized is defined as:

\[ U(f|d) = \sum_{i} V_i(f, d) + \sum_{(i,j) \in C_i} V_{ij}(f, d) \quad \text{(1)} \]

where

\[ V_i(f, d) = \begin{cases} \alpha_1 & \text{if } f_i = I_i \text{ and } T_i = 255, \\ \alpha_2 & \text{if } f_i = I_i \text{ and } T_i = 0, \\ 0 & \text{otherwise} \end{cases} \quad \text{(2)} \]

Here, \( U(f|d) \) is the posterior energy function. Parameter \( f \) indicates the label set in object segmentation, only the foreground label \( F \) and the background label \( B \) are considered; \( d \) is the set of observations on each grid. \( C_i \) is the clique of location \( i \). \( V_i(f, d) \) represents the node energy term, and is defined as:

\[ V_i(f, d) = \begin{cases} \beta_1 & \text{if } f_i \neq I_i \text{ and } T_i = 255, \\ \beta_2 & \text{if } f_i \neq I_i \text{ and } T_i = 0, \\ 0 & \text{otherwise} \end{cases} \quad \text{(3)} \]

In our implementation, \( \alpha_1 = 3, \beta_1 = 1, \) and \( \alpha_2 = \beta_2 = 2 \). Fig. 2 shows people extraction results generated by different algorithms. Clearly, our method achieves better results, especially at the boundary of the head and foot area.

### 4. Occlusion Handling Using Nonlinear Regression

Tracking people becomes more challenging when they are occluded by or interacting with other people in the scene. This problem has been addressed recently in literature [3][5][9], where the work of [9] is highly related to our research. In [9], a silhouettes-based occlusion-handling algorithm is proposed, in which both the shape of the silhouette boundary and the projection histograms of the silhouettes are used to locate the people during the occlusion. This method assumes that the complete body silhouette of each person is detected and the size of the people in the picture should be sufficiently large (at least 75 \( \times \) 50 pixels). Unfortunately, such assumptions do not always hold in real applications.

In this paper, we also employ the silhouette-related image features. However, our technique differentiates the method in [9] in two aspects: (1) we do not use natural vertices at the silhouette boundary of the people, since we have found that this image feature is too sensitive to the noise and the shape of the silhouettes; (2) unlike the method [9] that directly treats the projection histograms of the silhouettes, we firstly model the projection histograms by a nonlinear regression algorithm. Afterwards, we search relative maximum points on a smooth curve which is automatically generated by the regression technique. In our model, each relative maximum point corresponds to one visual person in the occlusion. Our method is robust against the case where the silhouette of the person is fragmented, e.g., only a part of the head is segmented.
4.1. Locate the people during the occlusion

By projecting the human silhouette on an axis perpendicular to the assumed torso axis, one should observe a peak in the projection in the vicinity of a head. Therefore, the shape of a 2D binary silhouette is represented by its horizontal projection histogram, which is computed by projecting the binary foreground region on the x-axis. It is noted that the y-coordinate of each bin on our histogram map does not start at the zero point (in [9], zero point is the start point), but at the y-coordinate of its lowest projection point. By doing so, the spatial information of the human body is preserved, e.g., the arm part of the person in Fig. 3. The benefit is that the achieved human geometric shape is more similar to the curve (model) used in our regression algorithm, thereby leading to more accurate results. Now, let us describe our modeling of the projection histograms.

Let \( \{ x_i, y_i \}_{i=1}^{m} \) represent the horizontal projection histogram of a person silhouette, where \( x_i \) denotes the bin index of the histogram, \( x_i \) is the x-coordinate of the \( i^{th} \) bin on the histogram map, and \( y_i \) is its y-coordinate. We intend to find a smooth curve (function) to model the data points \( (x_i, y_i) \). Mathematically, this is a nonlinear least-squares regression problem. This problem can be solved by minimizing \( \chi^2 \), which is the sum of the squared residuals over \( m \) experimental data points \( (x_i, y_i) \) for the model \( F(e, x) \),

\[
\chi^2 = \sum_{i=1}^{m} (y_i - F(e, x_i))^2. \tag{4}
\]

The parameters of the model are in the vector \( e = \{ c_0, c_1, \ldots \} \). The model \( F(e, x) \) used in this algorithm is a Gaussian model, whose term number equals to the number of persons in the occlusion. The number of the person in the occlusion can be easily determined by the occlusion-detection component. For instance, if we know that there are two persons involved in the occlusion, the applied Gaussian model would have two terms, and be written as:

\[
F(e, x) = c_1 \exp\left(-\frac{(x - c_2)^2}{2\sigma_2^2}\right) + c_2 \exp\left(-\frac{(x - c_3)^2}{2\sigma_3^2}\right). \tag{5}
\]

The Levenberg-Marquardt optimization algorithm helps to solve the minimization problem mentioned above, returning the best-fit vector \( e \) and their covariance matrix of the individual components. Once we obtain all the parameters of \( F(e, x) \), the next step is to find the relative maximum (peak) points on this curve, each corresponding visually to the head of one person. Suppose that there is a point \( x_i \), if there exists some \( e > 0 \), such that \( F(e, x_i) \) is \( \text{max} \) when \( |x_j - x_i| < e \), the value of the \( F(e, x) \) is the relative maximum of the function. Our model is fully automatic, e.g., it can output two peak points when two persons are partially occluded, but provide only one peak point when they are completely occluded.

In practice, the initialization of this algorithm is critical because it determines not only the performance but also the convergence speed of the optimization. Our basic idea is that we split the vector of data points \( \{ x_i, y_i \}_{i=1}^{m} \) into two datasets (suppose that we have two persons involved in the occlusion), which are \( \{ x_i, y_i \}_{i=1}^{m/2} \) and \( \{ x_i, y_i \}_{i=m/2}^{m} \). We have found that there is only partial occlusion when we detect the beginning of an occlusion over time. Therefore, each dataset mentioned above can be linked to one partial person in the picture. In other words, each dataset should be associated with one Gaussian term in Eqn. (5). Having these datasets, we can compute the \( \mu \) and \( \sigma \) for each Gaussian, thereby achieving the initial guess for \( e \). Once we get the optimized parameters for \( F(e, x) \) in the current frame, we use it as the initial parameters for the next frame, because there is only a slight difference between two succeeding frames. This strategy can efficiently provide the initial parameters, and it can be easily extended to the situation where more than two persons are involved in the occlusion.

4.2. Recognize the people during the occlusion

Apart from locating the person during the occlusion, it is also required to recognize the person. We base our method on the template-matching approach using color histogram distributions [3]. Assuming that there are \( n \) templates for \( n \) persons, and all the templates are generated and maintained by the mean-shift algorithm before the occlusion occurs. The probability of the feature \( \{ u_i \}_{i=1}^{m} \) in the \( n^{th} \) template is \( \hat{T}_{n}(u_i) \). Here, \( u_i \) represents the color histogram distribution and \( i \) denotes the bin number of the histogram. Furthermore, we use the same method to model the appearance of the blobs obtained by the people-locating module, and represent them by \( \hat{W}_j(u_i) \). The aim of template matching is to find the best match to the template. Mathematically, finding the best match means maximizing the correspondence between the templates and the blobs in the occlusion, so that we compute

\[
C_j = \max_n \rho(\hat{W}_j(u_i), \hat{T}_{n}(u_i)). \tag{6}
\]

The term \( \rho(\hat{W}_j(u_i), \hat{T}_{n}(u_i)) \) is a metric to measure the matching between the templates and the target candidate. For metric \( \rho \), the Bhattacharyya coefficient [3] is used because of its performance.

5. EXPERIMENTAL COMPARISONS

The presented video surveillance system has been tested on six video clips (each one has more than 1,000 frames) captured by a TRV30E video camera from a leading manufacturer. In our dataset, we have designed two occlusion scenarios. The first one describes that two persons walk towards each other and immediately split after occlusion. The second case is that two persons join, and then walk circularly after they meet. Obviously, the second one is more challenging, in which occlusion occupies more than 70% frames. In our test set, the video sequences had 320×240 resolution with 25-Hz frame rate.
Fig. 4. Human tracking during the occlusion. Left: binary map. Middle: our nonlinear regression result (the white dot represents the projection histogram, the (red) curve is a smooth curve generated by our model, and the top (yellow) dots represent the detected peak points). Right: tracking result.

We compared our system with the mean-shift algorithm [3] and also the particle-filter algorithm (200 particles per object) [4]. Table 1 shows the system capability of treating the occlusion. The ground-truth data was manually generated. The evaluation criterion is that at least 70% of the human body is included in the detection window. To sum up, the mean-shift algorithm achieves an averaged 82.7% tracking accuracy during the occlusion events. The particle-filter algorithm achieves an averaged 96.4% accurate rate, and our regression-based algorithm is correct for 93.3% frames when the occlusion occurs. Fig. 4 portrays three examples, where some difficult scenes have been selected. In the first two examples, we can see that the silhouettes of the person are not extracted perfectly. The people-tracking algorithm in [9] may fail at these cases, because it may find many peak points on the projection histogram instead of two. On the contrary, our method successfully handles this case. The last example demonstrates an extremely difficult case where two persons are completely occluded each other. It can be found that two detection windows are converged exactly to the same position.

In addition to comparing the performance of these three algorithms, we have also evaluated their efficiency. Fig. 5 gives the time consumed for each tracking algorithm during the occlusion, and also the execution time of the whole system. In the occlusion situation, the average execution times per frame for the tracking component of mean-shift, particle-filter and our technique are 20.45, 504.92 and 9.07 ms, respectively. Moreover, the execution times per frame for the complete algorithm including object segmentation, object tracking and trajectory generation, are 54.52, 553.78 and 46.77 ms, respectively. The experiments were performed on a P-IV 3GHz PC. The graph in Fig. 5 indicates that our proposal is a near real-time system with small fluctuations between successive frames.

6. CONCLUSIONS

In this paper, a new system for single-camera surveillance has been introduced, thereby concentrating on two primary contributions. First, we take the texture information into account when designing the energy function for DMRF-based object segmentation, so that the silhouettes of the object can be extracted accurately. Second, to address the occlusion problem, we have proposed an approach, where we model the silhouette-related image features into a feature vector which is optimized using a nonlinear regression algorithm. Therefore, the object position can be located through the finding of the peak points of a smooth curve when object occlusion occurs. It has been shown that our algorithm can operate at near real-time speed with more than 90% tracking accuracy during the occlusion.

7. REFERENCES