A robust appearance model for tracking human motions

N. Thome

LIRIS-Sas Foxstream*
Université Lumière Lyon 2
5 av Pierre Mendes France
69576 Bron Cedex, FRANCE

S. Miguet

LIRIS
Université Lumière Lyon 2
5 av Pierre Mendes France
69576 Bron Cedex, FRANCE

Abstract

Abstract— We propose an original method for tracking people based on the construction of a 2-D human appearance model. The general framework, which is a region-based tracking approach, is applicable to any type of object. We show how to specialize the method for taking advantage of the structural properties of the human body. We segment its visible parts, construct and update the appearance model. This latter one provides a discriminative feature capturing both color and shape properties of the different limbs, making it possible to recognize people after they have temporarily disappeared. The method does not make use of skin color detection, which allows us to perform tracking under any viewpoint. The only assumption for the recognition is the approximate viewpoint correspondence during the matching process between the different models. Several results in complex situations prove the efficiency of the algorithm, which runs in near real time. Finally, the model provides an important clue for further human motion analysis process.

1. Introduction

Human motion analysis is currently one of the most active research fields in computer vision. It attempts to detect, track and identify people, and more generally, to interpret human behaviours, from image sequences involving humans. It has attracted great interests from computer vision researchers due to its promising applications in many areas such as visual surveillance, perceptual user interface, content-based image storage and retrieval, video conferencing, athletic performance analysis, virtual reality, etc. In the general framework to achieve a semantic description of a scene, tracking represents the intermediate level between the low-level corresponding to segmentation and object classification and the high-level composed of action recognition. This is a crucial step because it attempts to temporally link features chosen to analyse and interpret human behaviour. Thus, robust algorithms have to be developed in order to accurately perform the highest level process and prevent from misinterpretations.

The remainder of the paper is organized as follows. Section 2 proposes a state of art of the tracking methods. Section 3 describes in more details our approach by developing how objects segmentation and tracking is performed. Section 4 presents several results illustrating the efficiency of the proposed approach and showing its superiority with respect to previous works. Finally we conclude in section 5 by pointing out some limitations of the algorithm and proposes some directions for future works.

2. State of the art

There are numerous ways of tracking video objects, which have been widely studied in the twenty past years. An exhaustive review of the methods is outside of the scope of the paper and the reader can refer to [2, 14]. Concerning the methodology, a classification containing five groups can be establish. Model-based approaches try to impose high-level semantic constraints by exploiting the a priori knowledge about the object being tracked. Motion-based approaches depend on a robust method for grouping visual motion consistently over time, by studying either regions or contours. Appearance-based methods track connected region that roughly correspond to the 2-D shape of video objects based on their dynamic models. The fourth tracking group method is feature-based. Its goal is to track a state vector possibly containing heterogeneous data. Finally the most advanced approaches try to combine advantages of the four previous groups leading to hybrid methods. We propose now to give more details on some recent works being the most related to our.

0-7803-9385-6/05/$20.00 ©2005 IEEE

Previous approaches

Wren et al. [15] use small blob features statistics (position and color) to track people. They represent human shape as a combination of blobs representing various body parts, such as head, torso, hands and legs. Tracking all the small blobs allows them tracking the whole body of a single human. However, their work is only adapted to an indoor environment and is only intended to track single human, which prevents to be used within a large set of situations we would like to manage.

Haritaoglu et al. [4] develop a system named W4 for a real time visual surveillance system operating on monocular greyscale or on infrared video sequences. W4 makes no use of color cues, but employs a combination of shape analysis and tracking to locate people and their body parts. Moreover, the method explicitly uses an appearance model to match objects after an occlusion. However, the model leads to wrong updates in typical cases presented in figure 1. The first column illustrates the fact that the model is very sensitive to segmentation errors or partial occlusions of the limb. As we can see, the two people enter the scene at Frame 100 and the models are initialized. At frame 200, their appearance models are wrongly updated. For the red framed person, it is due to segmentation detection errors. The legs were indeed not detected at the beginning and appear later : as the model is updated with respect to the median coordinate, the new detected parts are updated at the bad place. For the green framed person, the legs were not detected at the beginning because hidden by the desk. It leads to the same difficulties. Finally, the second video illustrates problems to manage large changes in depth due to perspective effects. The model is still badly updated for the same reason (for instance new head's image is at one time updated with part of the torso model).

Zhao et al. [16] suggest a much finer system for tracking. They use an ellipsoid model for the gross human shape and track its parameters with a Kalman filter. They use an appearance model too with use of color clue. The mask of the model is an ellipse instead of a rectangle but this model still suffers from the former drawbacks.

McKenna et al. [9] propose an adaptive background removal algorithm that combines gradient information and color features to deal with shadows in motion segmentation. They differentiate three levels of tracking: regions, single human and human groups. They manage to obtain good results of tracking multiple persons even in the case of occlusions by introducing an appearance model based on a combination of color histogram and Gaussian mixtures. However, this model does not capture spatial properties: a person wearing a yellow tee-shirt and blue pants is interpreted same manner as a person wearing a blue tee-shirt and yellow pants.

Girondel et al. [3] develop a method for tracking multiple persons by Kalman filtering and face pursuit. They use a region-based strategy similar to [4] and [16]. Face detection reduces the applicability of the method to viewpoints where skin color segmentation may be performed which is too restrictive for our purpose. Using a Kalman filter helps them to overcome the occlusion problem. In that case, only partial Kalman filtering may be applied because several measurements are likely to miss. This approach making use of a Kalman filter only in a predictive mode is interesting but is restricted to situations where the occlusion is relatively short, especially if the motion model is simple.

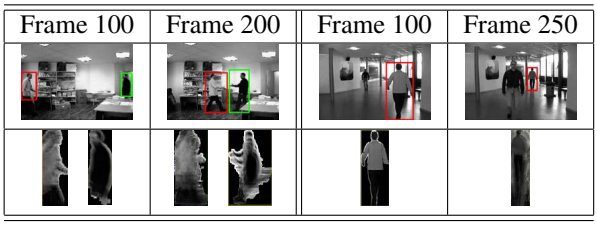


Figure 1: Existing approach limitations

3. Approach overview

3.1 Main contribution

Our approach is an hybrid one. For real-time purpose, we choose to use a region-based method for the basic tracking tasks. To deal with more complex situations, we use an appearance model to describe each tracked person. Our main contribution in order to perform a robust tracking relies on building an invariant appearance model which does not suffer from the drawbacks illustrated in figure 1. We propose to segment all body parts, and then to scale each of them to build a constant-size-model. This enables three major improvements. First, the model update is not sensitive to perspective effects. Second, the movement of the different limbs does not introduce problems when updating the model, because each limb is segmented, scaled and rotated to be aligned to a fixed structure. Finally, the potential errors during the segmentation process do not disturb the model any more as only the detected parts are updated.

3.2 People segmentation

Isolating people in the video is performed in three steps. First, a robust motion segmentation is applied by modelling the background by a mixture of gaussians, allowing us to take into account statistical properties for each pixel. Then a simple shadow removal algorithm based on color considerations is applied to eliminate shadows from the analysis loop. Finally the labelled pixels are grouped into regions by a connected components algorithm, providing the input data

to the tracking module. These three steps are now explained in more details.

3.2.1 Motion detection

We obtain a binary map describing regions where motion occurs by computing a difference between the current frame and a reference one representing the static part of the scene. This approach implies building and updating the background image. Following that direction, many researchers have tried to estimate the Probability Density Function (PDF) of the gray-level (or color) value for each pixel [13, 10, 1]. We choose to model each pixel by a mixture of two gaussians. We found that this was the best compromise between time consuming and having an accurate segmentation. Indeed the method is very fast and makes it possible to compute an adaptative threshold for each pixel depending on the statistical properties computed for the pixel (standard deviation for each color component), which is a great improvement in comparison to a global image threshold. Moreover it provides a dynamic background model able to incorporate or eliminate objects. Thus, the PDF for each pixel is modelled as follows :

$$P_r(X_t) = \frac{1}{N} \sum_{i=1}^{2} w_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \qquad (1)$$

where :

- $w_{i,t}$ represents the weight of each distribution and represents the portion of the data accounted for by this gaussian.
- $\eta(X_t, \mu_{i,t}, \Sigma_{i,t})$ is a normal distribution with mean $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$. For a computational purpose, we choose to model each distribution with a diagonal covariance matrix which assumes that the red, blue and green pixel values are independent. But contrary to [13, 10] we do not impose that they have identical variances.

The background is then modelled by a unique normal distribution and is always the one having highest weight $w_{i,t}$. For complex situations it would not be difficult to model it with several gaussians, but we found it sufficient for our purpose. The training is performed on-line or off-line, leading to computing mean values and variances for each pixel and each color component. During the analysis, each value is tested against the background distribution. If it matches it, the parameters of the corresponding distribution (mean, standard deviation and weight) are updated like in [13]. If not, a new gaussian for non-background (moving objects) is initialized. After that, each new value X_i is tested against the two gaussians, finding the nearest one by minimizing $\left(\frac{X_i - \mu_i}{\sigma_i}\right), i \in \{1, 2\}$.

3.2.2 Shadow removal

Detecting shadows at the same time as people is a well-known difficulty after motion detection. This problem has been largely addressed because it considerably disturbs people silhouettes and then is likely to perturb the analysis process. For us the main drawback in detecting shadows would be to wrongly update the appearance model (see 3.3.2). Shadow detection is performed at the pixel level, and relies on the simple but true consideration that points in shadows must have difference with the background in terms of luminance, but not in terms of chrominance. Salvador [12] proposes to use the following space, which has the property to be invariant in luminance :

$$\begin{aligned} C_1 &= \arctan\left(\frac{R}{\max(G, B)}\right) \\ C_2 &= \arctan\left(\frac{G}{\max(R, B)}\right) \\ C_3 &= \arctan\left(\frac{B}{\max(R, G)}\right) \end{aligned} \qquad (2)$$

3.2.3 Post treatment

As previously mentioned, the result of the detection consists of a binary map where moving pixels have been isolated. No consideration of spatial coherence has been developed at this stage. Morphological operations are applied in that sense. Finally, connected components analysis is used to merge pixels into regions. We only keep the significant ones by thresholding their area.

3.3 People tracking

The tracking module uses as input each region detected at time $t + 1$ and each object identified at previous time t. Its aim is to establish links to determine objects evolutions. To do this we combine two approaches. The first one, inspired by [15, 4] and used again in [3], consists in a two-passes forward/backward tracking. Section 3.3.1 presents in more details how we adapt it to our purpose. This approach which is very fast and works correctly in simple situations but is insufficient for performing tracking as soon as occlusions occur. Such cases, which are easily detected by the previous module, are tackled by using the appearance model. Its building, updating and use to perform segmentation is detailed in section 3.3.2.

3.3.1 Dynamical regions linking

The region-based part of the tracking process aims at giving a description of each object evolution between two consecutive frames. We use a simple first-order motion model to predict objects location in subsequent frame. Then the projection of the p objects O_i detected at time t are compared

with the n regions R_j detected at time $t+1$. For each O_i and R_j, the matching is performed by computing the area overlap $A_{ij} = O_i \cap R_j$. The process then works in a two-passes forward/backward phase :

- For each R_j we determine its predecessors number np_j and sort them with respect to their A_{ij} value (the most probable predecessor is supposed to be the one with the greatest overlap surface). We obtain a list of predecessors $P_k,\ k \in [1, np_j]$.
- For each projected O_i we determine its successors number ns_i and sort them with respect to their A_{ij} value (the most probable successor is supposed to be the one with the greatest overlap surface). We obtain a list of successors $S_l,\ l \in [1, ns_i]$.

3.3.2 Appearance model building

We first describe the geometrical 2-D human body model used to depict gross people shape. Each limb constitutes an individual bounding box inside which independent spatio-temporal templates will be updated. Then we explain how silhouette is related to the model, ie how to detect the different limbs in the image. Finally, we present the model update, and the matching measure chosen to perform recognition.

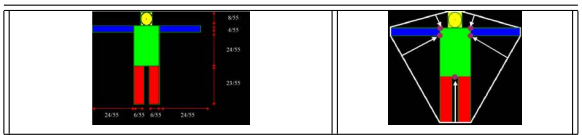
2-D Human body model We use a very simple 2-D model to represent human shape properties. This is a particular case of Cardboard Model [7] which represents the relative position and size of the body parts. It is only composed of six parts : head, torso, two legs and two arms, each one represented as a rectangle. Left column in Figure 2 illustrates our model. The ratio between the Bounding Boxes have been fixed by consulting anthropometrical data. These measures are related to front views. As we explain in more details, this hypothesis does not constitute a significant restriction. The body part detection (see next paragraph) is robust enough to detect limbs in side view as well as in a front or back view.

Body Part detection Body parts detection is performed in two steps. First, joints candidates are extracted from the silhouette's contour. Then our model is used to apply constraints to isolate each limb.

Joints detection results from a study of the contour shape. Joints are supposed to be points with a *high concavity degree*. To find them, we use a two-passes approach consisting in first computing points on the contour belonging to the convex hull. For two convex contour points A and B in the hull that are not direct neighbors, we look for the point C between A and B satisfying :

$$\left\|\overrightarrow{CH}\right\| = \max_{P_i \in [A,B]} \left\|\overrightarrow{P_i H_i}\right\| \qquad (3)$$

where H and H_i are C and P_i's orthogonal projections on $[A, B]$, respectively. C is simply the further point from segment $[A, B]$, in the sense of the Euclidian distance. Right column in Figure 2 shows what is indented for a perfect segmentation. Purple circles represent attempted joints.



(a) 2-D-human body model | (b) Joints detection

Figure 2: Body Parts labelling

After Concave/Convex points silhouette analysis we use our body model to constraint each limb search. Between two successive concave points we compare intended ratio quantities for each limb to computed ratio quantities.

The body model constraints are related to a front view. But we manage to set the parameters in order to detect body parts for a side view as well. As far as we tested, our limb detection module is in fact applicable to every view point. The size of the different body parts for a side view will be warped to those corresponding to a front view, but this will neither alter the quality of the templates update nor than the possible matching process (see next sections).

Textural and Shape template generation We use two temporal templates for each detected body part. Each one is a fixed rectangle, the relative sizes between limbs are related to the model shown in figure 2. The templates have the same meaning as those previously introduced in [4] an [16] for the whole body. The weight template W captures shape information. It represents the foreground probability, i.e the probability for a given pixel of the Bounding Box to be part of a limb. The textural templates T captures color information and gives a discriminative description of the objects appearance. Updating templates proceeds as follow. First, the detected body part bounding box is transformed to be aligned to the template size, which means rotation and scaling. Let us define $I^t(x, y)$ as being the transformed current image at time $t$and $P^t(x, y)$ as :

$$P^t(x,y) = \begin{cases} 1 & \text{if (x,y) is inside current body part} \\ 0 & \text{else} \end{cases}$$

Then the update equations are :

$$\begin{aligned} W^t(x,y) &= W^{t-1}(x,y) + P^t(x,y) \\ T^t(x,y) &= \frac{T^{t-1}(x,y)W^{t-1}(x,y) + I^t(x,y)}{W^{t-1}(x,y) + 1} \end{aligned} \qquad (4)$$

Note that $T^t(x, y)$ is a three-dimensional vector containing color components treated independently.

Matching process The templates allow to discriminate people by their color or shape properties.

We measure a distance between the new object N and an existing one Eo_i by the following way :

$$dist(N, Eo_i) = \frac{\sum_{j=1}^{6} \omega_j \cdot D_j(N, Eo_i)}{\sum_{j=1}^{6} \omega_j} \quad (5)$$

D_j is the distance between N and Eo_i j^{th} body parts $P_{j,N}$ and $P_{j,i}$. ω_j sets the intended influence of each independent limb in the global distance. $[W_{o_i}(x,y), T_{o_i}(x,y)]$ and $[W_n(x,y), T_n(x,y)]$ being $P_{j,i}$ and $P_{j,N}$ templates respectively, D_j is defined by :

$$D_j(N, Eo_i) = \frac{\sum_{(x,y)\in P_j} \| W_{o_i}(x,y) \cdot T_{o_i}(x,y) - W_n(x,y) \cdot T_n(x,y) \|}{\sum_{(x,y)\in P_j} [W_{o_i}(x,y) + W_n(x,y)]} \quad (6)$$

Tracking during occlusion Tracking during occlusion is challenging. We choose to use a particle filter similar to [6]. The weight of each possible particle is computed with a simple color correlation measure. However, this operation does not affect the performance too much because constraints on the particles position can be added due to the partial knowledge of objects position during the merge.

4. Results

This section presents the results obtained with the proposed method.

Figure 3 illustrates how the model is being updated. Each time a limb is detected it is rotated and scaled to update the model shown in the third row. In those examples all the body parts are detected but this is obviously not always the case. This is actually a goal of this module to be robust enough not detecting limbs when the segmentation yields poor results because it prevents from wrongly updating the model. For instance, between frame 130 and 160 the legs are often not detected because the man is squatting himself, resulting on a lonely blob. As we can see, the model update results are very good. The textural template will provide an interesting feature for recognition.

Figure 4 shows an example in an outdoor environment. The textural template is still used to recognize people but some other details are pointed out. The first row shows the results after the two isolated people have entered the place. We can notice that the appearance model, presented in the last column, is very convincing although the silhouette is very small and the view point is not a front one. Frame 300 is the merge time, with creation of a new object and links updating. The third row illustrates the use of the particle

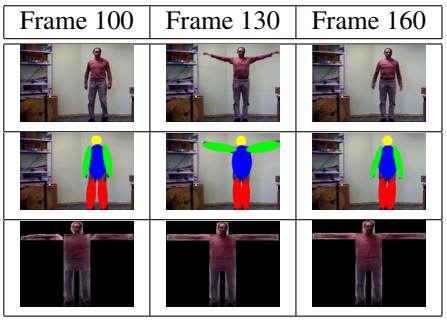


Figure 3: Body Part detection and model updating

filter for tracking during occlusion. At the three successive times, thin rectangles represent the particles for each person, and the thick rectangle is the estimated position returned by the filter. As we can see, even with the simple motion model and the naïve weighting measurement, the tracking is surprisingly satisfying. The probabilistic framework of the filter enables indeed flexibility. Moreover, Condensation algorithm [6] permits multi-modal distribution tracking and is then robust to occlusion. Finally, Frame 360 shows how the template matching step again recognizes the persons.

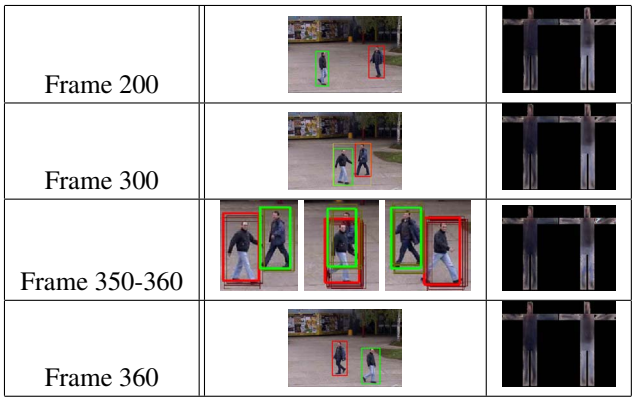


Figure 4: People recognition through every view point and tracking during occlusion

Figure 5 illustrates how our framework for model updating can manage difficult situations such as person's partial occlusion, and important changes in size due to perspective effects. At frame 50, the two persons enter the scene, one being far from the camera with respect to the viewing direction, the other one being very close to it. Let us define P_1 being the red-framed, and P_2 the green-framed one. Between frame 50 and 110, P_2 is partially occluded. At the beginning only its torso, head and legs are visible. Progressively its legs appear until Frame 110 where he is entirely inside the field of view. The third column shows how its ap-

pearance model is being robustly built at Frame 100, meaning that body parts have been truly identified and only visible parts have been updated. Between Frame 110 and 180 , P_1 and P_2 walk in reverse directions, resulting in large change in size of their silhouette due to perspective effects. The models shown in the third column still appear to be good. Indeed, each time a limb is detected, it is rescaled to match its model size resulting to a coherent update, which would not have been the case with an approach similar to [4] or [16].

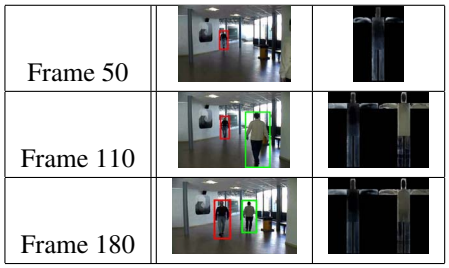


Figure 5: Scale invariance of the model

5. Conclusion and future works

We propose an efficient way to build a 2-D appearance model for human, allowing to track people in complex situations such as occlusions. The matching measure compares two models and then allows us to recognize people. The model building is robust as only detected body parts contribute to update, and is moreover invariant to rotation and scaling. There are many directions we can think about to improve the method. The bottleneck seems to be body parts detection. First, it is only dedicated to upright standing pose. Hartaoglu suggests in [5] a method for detecting body parts in different configurations, after classifying the pose between a set of predefined ones. It could be interesting to extend it without needing knowledge about pose. An other important limitation in our limb detection is the absence of high level verifications after body parts segmentation. We do not verify that the different found body parts constitute a partition of the whole body, nor that there is no aberration on the structural result. Ronfard et al. in [11] attempt to link image features to body part detection by explicitly searching a consistent way to assemble them. It could be a good idea to incorporate such constraints in our framework. Finally, for the tracking module, a great improvement would be to separately track body parts, still making sure that the limbs satisfy whole body constraints. It should lead to significant tracking improvements by being able to accurately predict occlusions as well as segmentation errors.

References

[1] A. Elgammal, D. Harwood, and L. Davis, "Non-Parametric Model for Background Subtraction", *Proc. IEEE Frame Rate Workshop*, 1999.

[2] D. Gavrila, [a]The Visual Analysis of Human Movement: A Survey", *Computer Vision Image Understanding*, vol. 73, no. 1, 1999.

[3] Girondel (V.), Caplier (A.) et Bonnaud (L.), "Real time tracking of multiple persons by kalman filtering and face pursuit for multimedia applications", *In IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 201-205. - Lake Tahoe, Nevada, USA, mars 2004

[4] Ismail Haritaoglu, David Harwood, and Larry S. Davis, "W4: Real-Time Surveillance of People and Their Activities", *IEEE Trans. on Patt. Anal. and Machine Intell.*, 2000.

[5] I.Haritaoglu, D.Harwood, and L.Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes", *Fourteenth International Conference on Pattern Recognition*, Brisbane, August 1998.

[6] M. Isard and A. Blake, "Condensation-Conditional Density Propagation for Visual Tracking" *Intl J. Computer Vision*, vol. 29, no. 1, pp. 5-28, 1998.

[7] S. Ju, M. Black, Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion", *International Conference on Face and Gesture Analysis*, 1996

[8] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. '"Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information". *In Proc. 15th International Conference on Pattern Recognition*, Barcelona, Spain, Vol. 4, pp. 627–630, September 2000.

[9] S. J. McKenna, S. Jabri, Z. Duricand, A. Rosenfeld, and H. Wechsler. "Tracking groups of people". *Computer Vision Image Understanding*, 80:4256, 2000.

[10] V. Morellas, I. Pavlidis and P. Tsiamyrtzis. "Deter: detection of events for threat evaluation and recognition", *Machine Vision and Applications*, 15, pp. 29-45, June 1, 2003.

[11] Remi Ronfard, Cordelia Schmid, Bill Triggs. "Learning to parse pictures of people". *European Conference on Computer Vision*, LNCS 2553 volume 4 pages 700-714 , June 2002

[12] E. Salvador, P. Green and T. Ebrahimi. "Shadow identification and classification using invariant color models" *Proc. of IEEE, Proceedings IEEE ICASSP '01*, Vol. 3, pp. 1545-1548, May 2001

[13] C. Stauffer, W.E.L. Grimson, "Learning patterns of activity using real-ime tracking", *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, no. 8, pp. 747-757, 2000.

[14] Liang Wang, Weiming Hu, Tieniu Tan, "Recent developments in human motion analysis", *Pattern Recognition 36* (2003) p 585-601

[15] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. "Pfinder: Real-time tracking of the human body". *IEEE Trans. on Patt. Anal. and Machine Intell.*, volume 19(7), pages 780785, July 1997.

[16] T. Zhao, R. Nevatia, F. Lv, "Segmentation and Tracking of Multiple Humans in Complex Situations". *IEEE Trans. on Patt. Anal. and Machine Intell.*, september 2004