IEEE Workshop on Motion and Video Computing,Orlando, FL, December 2002, pp. 105-111.

# Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing

Sangho Park and J.K. Aggarwal
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712, USA
{sh.park | aggarwaljk}@mail.utexas.edu

## Abstract

*This paper presents a system to segment and track multiple body parts of interacting humans in the presence of mutual occlusion and shadow. The color image sequence is processed at three levels: pixel level, blob level, and object level. A Gaussian mixture model is used at the pixel level to train and classify individual pixel colors. Markov Random Field (MRF) framework is used at the blob level to merge the pixels into coherent blobs and to register inter-blob relations. A coarse model of the human body is applied at the object level as empirical domain knowledge to resolve ambiguity due to occlusion and to recover from intermittent tracking failures. A two-fold tracking scheme is used which consists of blob to blob matching in consecutive frames and blob to body part association within a frame. The tracking scheme resembles a multi-target, multi-assignment framework. The result is a tracking system that simultaneously segments and tracks multiple body parts of interacting people. Example sequences illustrate the success of the proposed paradigm.*

***Index Terms*— Segmentation, Tracking, Human Body Parts, Occlusion.**

## I. INTRODUCTION

Video surveillance of human activity requires reliable tracking of moving humans. Tracking non-rigid objects such as moving humans presents several difficulties for computer analysis. Problems include segmentation of human body into meaningful body parts, handling the occlusion, and tracking the body parts along the sequence. Many approaches have been proposed for tracking a human body (See [1,2] for reviews). As of date, research focus has been on tracking a single person in isolation [3,4], or on tracking only a subset of the body parts such as head, torso, hands, etc.[5,6], while research on segmentation or tracking of multiple people has focused on the analysis of the silhouettes [7,8], the contours [9,10], or color [11].

In this paper we present a new framework to segment multiple humans into semantically meaningful body parts and to track them under the conditions of occlusion and shadow. The proposed system processes the input image sequence at three levels: pixel level, blob level, and semantic object level. A Gaussian mixture model is used to classify individual pixels into several color classes. Markov Random Field (MRF) framework is used at the blob level to merge the pixels into coherent blobs of arbitrary shape according to similarity features of the pixels. The blobs are then grouped to form the meaningful body parts by a simple body model. A coarse human-body model is applied as empirical domain knowledge at the object level to assign the blobs to appropriate body parts. Tracking of the body parts is performed in a multi-target, multi-assignment framework [12]. Using the body model helps to resolve the ambiguity due to occlusion and to recover from intermittent tracking failure. The rest of the paper is organized as follows; section II describes the procedure at the pixel level, section III describes the blob formation, section IV presents a method to track multiple blobs, while section V describes the segmentation and tracking of semantic human body parts. Results and conclusion follow in section VI and VII, respectively.

## II. PIXEL CLASSIFICATION

### A. *Color Representation*

Most color cameras provide a RGB (red, green, blue) signal. RGB color space is, however, not effective for human visual perception of color and brightness. Here, the RGB color space is transformed to the HSV (hue, saturation, value) color space to make the intensity or brightness explicit and independent of the chromaticity.

### B. *Background Subtraction*

Background subtraction is performed in each frame to segment the foreground image region. The color distribution of each pixel v($x$,$y$) at image coordinate ($x$,$y$) is modeled as a Gaussian. Using $k_b$ training frames ($k_b = 20$), the mean $\hat{m}(x,y)$ and standard deviation $\hat{s}(x,y)$ of each color channel is calculated at every pixel location ($x$,$y$).

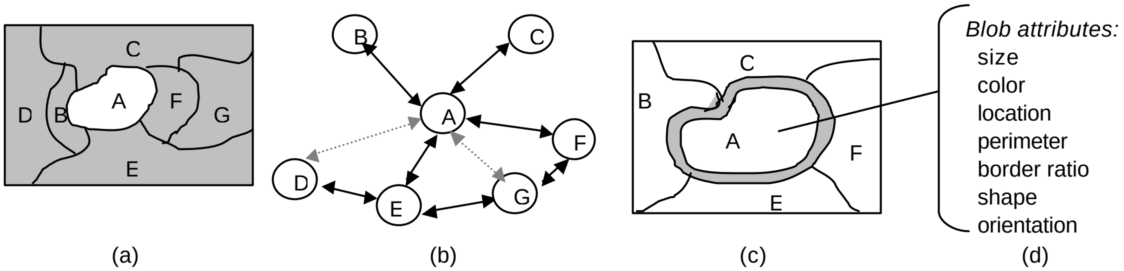

Fig. 1. Given an image region A,: (a) Adjacent regions to A in gray; (b) MRF model to represent some of the blob relations around A. Solid arrow show binary relations, and dotted arrow shows tertiary relations.(c) Border area of blob A in gray; Each portion of the border area connected to adjacent blob defines the *border ratio* between A and the adjacent blob. (d) Blob attributes to describe blob features.

Foreground segregation is performed for every pixel v(*x*,*y*), by using a simple background model, as follows: at each image pixel (*x*,*y*) of a given input frame, the change in pixel intensity is evaluated by computing the Mahalanobis distance *d* from the Gaussian background model.

$$d(x,y) = \frac{\left| v(x,y) - \hat{m}(x,y) \right|}{\hat{s}(x,y)} \tag{1}$$

Foreground image *F*(*x*,*y*) is defined by the maximum of the three distance measures, $\delta_H$, $\delta_S$, and $\delta_V$ for H, S, V channels;

$$F(x,y) = \max[\, d_H(x,y),\ d_S(x,y), d_V(x,y)] \tag{2}$$

*F* is then thresholded to make a binary image. At this stage, morphological operations are performed as a post processing step to remove small regions of noise pixels.

### C. Mixture of Gaussian Modeling for Color Distribution

In HSV space, the color values of a pixel at location (*x*,*y*) are represented by a random variable v = [H, S, V]$^T$ with the vector dimension *d* = 3.

Color distribution of a foreground pixel v is modeled as a mixture of $C_0$ Gaussians weighted by prior probability P($\omega_r$), given by;

$$p(v) = \sum_{r=1}^{C_0} p(v \mid w_r) P(w_r) \tag{3}$$

where the *r-th* conditional probability is assumed as a Gaussian as follows;

$$p(v \mid w_r) = (2p)^{-\frac{d}{2}} \left| \Sigma_r \right|^{-\frac{1}{2}} \exp\left[ -\frac{(v - m_r)^t \Sigma_r^{-1} (v - m_r)}{2} \right], \quad r = 1,\dots,C_0 \tag{4}$$

Each Gaussian component $G_r = \{P(w_r), m_r, \Sigma_r\}$ represents a prior probability *P*($\omega_r$) of the *r-th* color class $\omega_r$, a mean vector $\mu_r$ of the pixel color component, and a covariance matrix $\Sigma_r$ of the color components.

### D. Training the Gaussian Parameters

In order to obtain the Gaussian parameters $G_r = \{P(w_r), m_r, \Sigma_r\}$, expectation-maximization (EM) algorithm is used with the first η frames of the sequence as the training data (η = 5). Initialization (E-step) of the Gaussian parameters is done as follows; all the prior probabilities are assumed as equal, i.e., $P(w_r) = 1/C_0$, the mean is randomly chosen from a uniform distribution within a possible pixel value range, and the covariance matrix is assumed to be an identity matrix. Training (M-step) is performed by iteratively updating the above mentioned parameters. See [13] for details.

The iteration stops if either the change in the value of the means is less than 1 percent with respect to the previous iteration or a user-specified maximum iteration number is exceeded. We start with 10 Gaussian components ($C_0$ = 10) and merge similar Gaussians after the training by the method in [14], resulting in C Gaussians. The parameters of the established C Gaussians are then used to classify pixels into one of the C classes in subsequent frames.

### E. Classification of Individual Pixels

The color classification of the individual pixels is achieved by a maximum *a posteriori* (MAP) classifier. Once the Gaussian mixture model *G* for pixel color is obtained, we compute the MAP probability that each pixel in the subsequent frames belongs to each Gaussian component. The class that produces the largest probability value for a pixel v is chosen as the pixel-color class label $\omega_L$ for that pixel.

$$w_L = \arg\max_r (\log P(w_r \mid v)) \quad 1 \le r \le C \tag{5}$$

## III. BLOB FORMATION

### A. Initial Blob Formation

Foreground pixels with the same color are labeled as being in the same class, even though they are not connected. This causes a problem in labeling blobs. Therefore, connected component analysis is used to relabel the disjoint blobs, if any, with distinct labels, resulting in over-segmented small regions. The number of disjoint blobs generated by the relabeling process depends on the input image, and may vary from frame to frame. The fluctuation of blob numbers causes difficulty in using the usual MRF approach. In order to maintain consistency, we need to merge the over-segmented regions into meaningful and coherent blobs. This requires us to perform a high-level image analysis, taking into account the relationship between the segmented regions.

### B. Markov Random Field Model for Blob Relations

We need image features such as contours and regions, which are more descriptive. Such features are not only described by the properties of the features themselves but are also related to one another by relationships between them.

Markov Random Field (MRF) model has been used for labeling features and for establishing probabilistic distribution of the interacting features [15]. The relational structure, R, in MRF model is specified by node set S, neighborhood system N, and degree of relationship D.

$$R = (S, N, D) \tag{6}$$

Figure 1 shows our application, where *S* corresponds to the set of blobs, *N* the adjacency list for the blobs, and *D* the degree of the relationships which includes unary, binary, and tertiary features. We use tertiary blob-features (D= 3) as the highest level of abstraction to describe the characteristics of the *j-th* blob, $A_j$, as follows;

1. Unary features: determined by a single blob
   - Blob label = $L(A_j) \in Z^+$={natural numbers}
   - Blob size $\alpha(A_j) = |A_j|$, # of pixels in the blob
   - Color =$[m_H, m_S, m_V]^T$, the mean intensities of H,S,V color components of the blob;
   - Centroid = $[\bar{I}, \bar{J}]^T$, the median position of the blob
   - Border pixel set: $\Psi(A_j)$ = {8-connected outermost pixels corresponding to the contour of $A_j$ }
2. Binary features: determined by two adjacent blobs
   - Adjacency list = $\Gamma(A_j)$ = $\{k \in Z^+ \mid A_k$ is adjacent to $A_j, k \neq j\}$
   - Border-ratio of $A_j$ with respect to $A_k = \beta_j(A_k)$ = (# pixels in $\Psi(A_j)$ connected to $A_k$) / (# total pixels in $\Psi(A_j)$ )
3. Tertiary features: determined by three blobs
   - Tertiary relation between $A_j$ and $A_i$:

$$t(A_j, A_i) = \begin{cases} 1 & if\ A_j \in \Gamma(\Gamma(A_i)),\ j \neq i \\ 0 & otherwise \end{cases}$$

### C. Blob Merging Procedure

Merging the over-segmented blobs is a region growing procedure [16] controlled by the local consistency imposed by the MRF formulation. Two blobs are merged by the following criteria; blobs $A_i$ and $A_j$ are merged only if the following criteria are satisfied;

- *Adjacency criterion*: two blobs should be adjacent.
- *Border-ratio criterion*: two blobs should share a large border
  $(b_i(A_j) > T_b) \vee (b_j(A_i) > T_b)$; $T_\beta$ is a threshold
- *Color similarity criterion*: two blobs should be similar in color, where the similarity is defined by the Mahalanobis distance $\delta_\Phi$ of color feature $\Phi$ between the blobs $A_i$ and $A_j$ as follows;

$$d_\Phi = (\Phi_i - \Phi_j)^T (\Sigma_\Phi)^{-1} (\Phi_i - \Phi_j) \tag{7}$$

$$\Phi = [m_H, m_S, m_V]^T \tag{8}$$

where $\Sigma_\Phi$ is the covariance matrix of color channels for all the blobs in the image. If $\delta_\Phi$ is less than a threshold $T_\Phi$, blobs $A_i$ and $A_j$ are similar in color.

We adopt some additional heuristics as follows;

- *Small blob criterion*: a small blob surrounded by a single large blob is merged to it;
- *Skin blob criterion*: Skin blob does not follow the above small blob criteria.

The threshold values were obtained from training data.

### D. Skin Blob Detection

Skin information is very useful in recognizing body parts. We used a simple threshold model for the skin color detection using the chromaticity channels H and S in the HSV color space;

$$((T_{H1} \le m_H < T_{H2}) \wedge (T_{S1} \le m_S < T_{S2})) \rightarrow A_j\ \textit{is a skin blob}.$$

The proper values of the thresholds $T_{H1}$, $T_{H2}$, $T_{S1}$, and $T_{S2}$ are obtained from training data, and are used to segment the skin regions in the new frames. The skin blobs follow the tertiary relation criteria;

- *Tertiary relation criterion*:

$$((A_j\ \textit{is skin}) \wedge (\Gamma(A_j) = A_k) \wedge (\Gamma(A_k) = A_i)) \rightarrow \textit{set}\quad \Gamma(A_j) = A_i$$

The tertiary relation criterion is useful to handle the color smear around skin blobs caused by fast motion of the blobs.

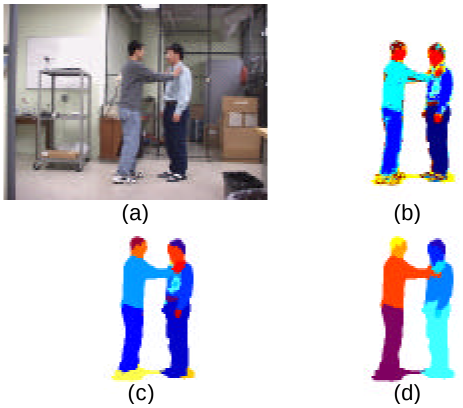

Fig. 2 (a) input image. (b) Initial pixel-level classification shows artifacts due to the shadowing in the second person's chest area. (c) Small blobs are merged to large adjacent blobs. (d) Object level segmentation shows the correct segments of the foreground figures.

## IV. TRACKING MULTIPLE BLOBS

### A. Multi-target, Multi-assignment Tracking Framework

The task of blob-level tracking is to associate a blob $A_i(t)$ at frame t with one of the already tracked blobs $A_j(t\text{-}1)$ at frame t-

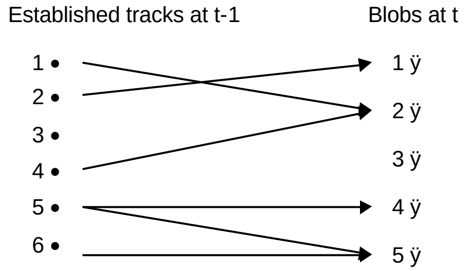

Fig 3. Candidate assignments in many-to-many mapping. Multiple tracks 1 & 4 are assigned to a single blob 2, while a single track 5 is assigned to multiple blobs 4 & 5. Track 3 is missing at t. Blob 3 is missing; i.e., not associated to any established track.

1.

The blob association between $A_i$(t) and $A_j$(t-1) is performed by comparing the similarity between their *unary feature* vectors $m_i$ and $m_j$;

$$m_i = \left[a, m_H, m_S, m_V, \bar{I}, \bar{J}\right]^T \quad for\ A_i \tag{9}$$

$$m_j = \left[a, m_H, m_S, m_V, \bar{I}, \bar{J}\right]^T \quad for\ A_j \tag{10}$$

Given the covariance matrices $\Pi_i$ and $\Pi_j$ of these features for all the blobs in the image, Mahalanobis distance $\Delta_{ij}$ defines the dissimilarity of the two blobs as follows;

$$\Delta_{ij} = \left(m_i - m_j\right)^T \left(\Pi_i + \Pi_j\right)^{-1} \left(m_i - m_j\right) \tag{11}$$

In the actual implementation, the covariance matrices $\Pi_i$ and $\Pi_j$ are assumed to be diagonal, simplifying the computation of $\Delta_{ij}$.

If the blob association is purely based on the above unary features, it may cause problems due to the following reasons.

1. Different number of blobs may be involved in each frame.
2. A single blob at time *t-1* can be split into multiple blobs at time *t* due to shadowing or occlusion, etc.
3. Multiple blobs at time *t-1* can be merged into a single blob at time *t* due to overlap or occlusion, etc.
4. Some blobs at time *t-1* can disappear at time *t*.
5. New blobs may appear at time *t*.

The above phenomena make the blob tracking complicated; we not only have to allow many-to-many mapping, but also have to avoid the situations where scattered blobs in time *t-1* are associated to a blob at time *t*, or vice versa. Figure 2 shows an example of the procedure, and figure 3 describes an example of a possible matching diagram.

In order to associate multiple blobs simultaneously, we adopt a variant of the multi-target tracking framework in [12], which is described below;

Let us call the blobs already tracked up to frame (*t-1*) as *tracks* $T(t-1)$, and the new blobs formed at frame *t* as *blobs* $B(t)$. Let the *i-th* track at frame (*t-1)* be *track* $T_i(t-1) \in T(t-1)$, and the *j-th* blob at frame *t* be *blob* $B_j(t) \in B(t)$.

### 1) *One-to-one Association*

The one-to-one association is formulated as a weighted bipartite matching problem [17]. Given a bipartite graph $G_b = (U, V, E)$, with the node cardinality, $|U| + |V|$, and the edge cardinality, $|E|$, we want to find a set of edges $\Xi \subseteq E$ of maximum cardinality such that no edge in the set shares a vertex with any other edge in the set. This set is a maximum association.

The node sets $U$ and $V$ correspond to $T(t-1)$ and $B(t)$, respectively, and the *nonzero* edge $\varepsilon_{ij} \in \Xi$ connects $U_i$ and $V_j$ for all $i \in [1, |T(t-1)|]$ and $j \in [1, |B(t)|]$.

$$e_{ij} = \begin{cases} 1 & Blob\ B_j(t)\ is\ assigned\ to\ track\ T_i(t-1) \\ 0 & otherwise \end{cases} \tag{12}$$

Each nonzero edge $\varepsilon_{ij}$ is weighted with $w_{ij} \in W$ that registers the dissimilarity measure $\Delta_{ij}$ in eq. (11).

$$W = [w_{ij}] = [\Delta_{ij}] \tag{13}$$

Finding the maximum association $\Xi^*$ is equivalent to minimizing the overall dissimilarity;

$$\Xi^* = \arg\min_{\Xi} \left( \sum_{i=1}^{T_i(t-1)} \sum_{j=1}^{B_j(t)} e_{ij} \Delta_{ij} \right) \tag{14}$$

Various algorithms are available for a weighted bipartite matching problem to find the maximum association $\Xi^*$ [17]. We used the following sequential algorithm;

1. Sort blobs according to area size.
2. Compute similarity (i.e., equation 11) between all the track and blob pairs by breadth-first-search.
3. Pick a unique best match for each blob and track with priority given to larger blobs.

By the one-to-one association, $T(t-1)$ is divided into the set of successfully matched tracks $T^1(t-1)$ and the set of unmatched tracks $T^0(t-1)$, and $B(t)$ into the set of successfully matched blobs $B^1(t)$ and the set of unmatched blobs $B^0(t)$.

Our experimental observation shows that large blobs are reliably tracked in the one-to-one association stage, and they are robust against partial occlusion due to articulation or foreground holes due to background subtraction. We call these large reliable tracks as '*reference tracks*', and the large reliable blobs as '*reference blobs*'. They usually correspond to large homogeneous areas in body parts including torso, arms, or legs, etc., and make the system robust against partial occlusion, shadows and noise.

### 2) *Iterated Multi-association*

The above one-to-one association leaves some unmatched tracks and unmatched blobs due to occlusion, shadow, and noise.

We perform the following additional associations after obtaining the '*reference tracks*' $T^1(t-1)$ and '*reference blobs*' $B^1(t)$;

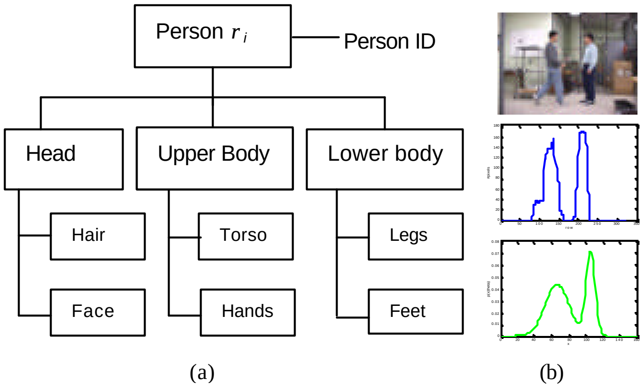

Fig. 4. Human body model. (a) hierarchical data structure. Each slot corresponding to individual body part contains tracked blob list. (b) 1D Gaussian approximation of human bodies: input image, histogram projection, and Gaussian fitting.

1. A one-to-one associations between $T^1(t\text{-}1)$ and $B^0(t)$
2. A one-to-one associations between $T^0(t\text{-}1)$ and $B^1(t)$

The Markovianity property [15] in the MRF constraints regarding the neighborhood system (eq. 6) restricts the search space in the above mentioned additional associations to the adjacent blobs/tracks of currently compared blobs/tracks. That is, for the associations between $T^1(t\text{-}1)$ and $B^0(t)$, an unmatched blob $B_j(t) \in B^0(t)$ is compared only with the subset of $T^1{}_i(t\text{-}1)$ mapped to the adjacency list $\Gamma(B_j(t)) \subseteq B^1(t)$. Similarly, for the associations between $T^0(t\text{-}1)$ and $B^1(t)$, an unmatched track $T^0{}_i(t\text{-}1) \in T^0(t\text{-}1)$ is compared only with the subset of $B^1{}_j(t)$ mapped from the adjacency list $\Gamma(T_i(t\text{-}1)) \subseteq T^1(t)$. These processes continue until no change occurs.

The above steps may still leave some unmatched residuals in $T^0(t\text{-}1)$ and $B^0(t)$ due to the blobs occluded for more than one frame, etc. The unmatched residuals are further processed by inference in the next higher level, i.e., the object level.

## V. SEGMENTING AND TRACKING BODY PARTS

### A. *Human Body Model*

A set of adjacent blobs constitutes a human body part in the initialization step in the first frame. In order to assign the blobs to a certain body part, we use a hierarchical human body model, *r*, in fig. 4(a). A human body is represented by three body parts: head, upper body and lower body.

The three body parts are initialized in terms of region of interest (ROI) in the image as follows; the head is defined by the vertical range from the top to 0.16 times the height of the person's foreground silhouette. The upper body is defined by the vertical range from 0.16 times the height to 0.45 times the height of the silhouette. The lower body is defined by the rest of the vertical range of the silhouette. Basically the division of the three body parts is similar to our previous approach in [9], but the parameters are adjusted and updated automatically in each frame.

Each of the body parts is recursively sub-divided into skin and nonskin parts; for example, head is divided into the face part and the hair part. Figure 4(a) shows the hierarchical data structure of the person model, in which each slot corresponding to individual body part contains a list of tracked blobs that correspond to that part. Some slots can be empty. In this framework, segmenting and tracking body parts amounts to properly updating the associated blob list of each slot across the image sequence. Person identity (Person ID) is maintained by monitoring the overall blob lists in a human body model.

### B. *Initial Assignment of Blobs to Body Parts*

The initial assignment of blobs to body parts is performed when people are isolated before occlusion. As shown in fig. 4(b), each person's foreground silhouette is vertically projected, and modeled by a 1D Gaussian. A mixture of the 1D Gaussians is used to model multiple people. Frame-to-frame update of these Gausian parameters amounts to tracking the whole body translation of each person, which is a coarse-level tracking. This coarse-level tracking module is used to initially assign blobs to corresponding body parts in the first frame together with the specification of the body model. The 1D Gaussian tracker is maintained across the sequence to compute the head part's motion in terms of the velocity and the acceleration, and to provide a re-initialization mechanism in the case of tracking failure. The body-part membership of any reappeared or newly detected blobs are inferred by this re-initialization mechanism.

An MAP classifier is used to assign each blob to a body part, by comparing the centroid of each blob to the human body model parameters. The initially assigned blobs are then tracked along the sequence by the tracking mechanism described in section *IV*.

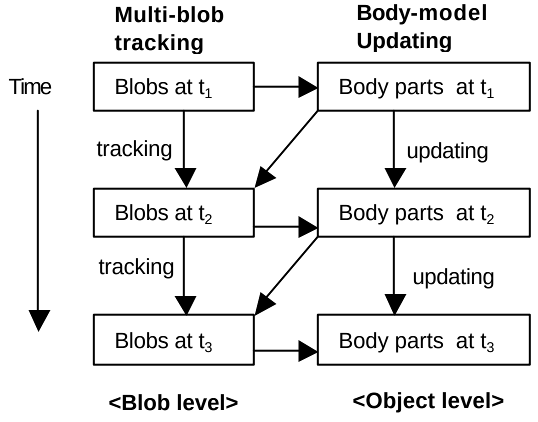

Fig. 5. Inter-weaved procedure of blob-level and object-level tracking. Blob tracking at time $t_2$ is guided by blob tracking results at time $t_1$ and body part configuration at time $t_1$. Body part tracking at time $t_2$ is guided by the blob tracking result at time $t_2$ and previous body configuration at time $t_1$.

### C. *Tracking Body Parts*

Figure 5 shows the inter-weaved mechanism between the blob-level and the object-level tracking. Blob tracking at time $t_2$ is guided by blob tracking results at time $t_1$ and body part configuration at time $t_1$. Body part tracking at time $t_2$ is guided by the blob tracking result at time $t_2$ and previous body configuration at time $t_1$. This mechanism is an integration of bottom-up (i.e., blob to object) and top-down (i.e., object to blob) processes. The updated blob lists are registered in the appropriate slots of human body model in figure 4.

## VI. RESULTS

We have tested our system for various human interaction situations including 'approaching', 'departing', 'hand-shaking', 'pointing', 'pushing', 'hugging', etc. The images used in this work were 320×240 pixels in size, obtained at a rate of 15 frames/sec.

Fig. 6 shows example sequences of 'punching', 'hand-shaking', 'pushing' and 'hugging' behavior obtained from different people. Note that the degree of occlusion and shadow increases from 'punching' through 'hugging' in fig. 6. The first column of each sequence shows the subsampled frames of the color input video, and the second column represents the corresponding frames of the output from the system. The interacting persons' individual body parts as well as person ID (fig. 4) are correctly segmented and tracked across the sequences.

Partial occlusion and shadows caused by some body parts are properly handled in the object level (Compare fig. 2(c)-(d) and fig. 6(c)). Most of the body parts are correctly segmented and reliably tracked over time, except that some group of small skin blobs close to each other are sometimes confused. Confusion also occurs when a blob is completely occluded by another blob. However these confusions are recovered automatically once the blobs move apart. An attempt to overcome this drawback could be a possible extension to the current paper.

## VII. CONCLUSION

We have presented a new framework for segmenting and tracking multiple human body parts. The contributions of this paper are as follows; i) an integration of Markov Random Field approach and multitarget-multiassignment framework is proposed to track multiple objects, ii) a new framework for simultaneous segmentation and tracking of entire human body parts is presented, and iii) a proper treatment of mutual occlusion and shadowing is achieved at the object level.

*Acknowledgement: We would like to thank S. Mohan and Dr. Q. Iqbal for their valuable discussions and comments.*

## REFERENCES

[1] J.K. Aggarwal and Q. Cai, "Human motion analysis: a review", Computer Vision and Image Understanding, 73(3), 1999, pp. 295-304.

[2] D.M. Gavrila, "The visual analysis of human movement: a survey", Computer Vision and Image Understanding, 73(1), 1999, pp. 82-98.

[3] R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts", in Proc. Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, 2000, pp. 721-727.

[4] A.Bakowski and G.A. Jones, "Video surveillance tracking using color region adjacency graphs", in Proc. Image Processing and Its Applications, 1999, pp. 794-798.

[5] A. M. Elgammal and L.S. Davis, "Probabilistic framework for segmenting people under occlusion", in Proc. Int'l Conference on Computer Vision, vol. 2, Vancouver, Canada, 2001, pp. 145-152.

[6] J. Sherrah and S. Gong, "Resolving visual uncertainty and occlusion through probabilistic reasoning, in British Machine Vision Conference, Bristol, UK, 2000, pp. 252-261.

[7] K. Sato and J.K. Aggarwal, "Recognizing two-person interactions in outdoor image sequences", in Proc. IEEE Workshop on Multi-Object Tracking, Vancouver, CA, 2001, pp. 87-94.

[8] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: who, when, where, what: a realtime system for detecting and tracking people", Third Int'l Conference on Automatic Face and Gesture, Nara, 1998, pp. 222-227.

[9] S. Park and J.K. Aggarwal, "Recognition of human interaction using multiple features in grayscale images", in Proc. Int'l Conference on Pattern Recognition, vol. 1, Barcelona, Spain, 2000, pp. 51-54.

[10] N.T. Siebel and S.J. Maybank, "Real-time tracking of pedestrians and vehicles", in Proc. IEEE Workshop on PETS, Kauai, Hawaii, 2001.

[11] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling", IEEE Workshop on PETS, Kauai, Hawaii, 2001.

[12] Y. Bar-Shalom and W.D. Blair, "Multitarget-multisensor tracking: applications and advances", vol. 3, Norwood, MA, 2000, pp. 199-231.

[13] R.O. Duda, P. Hart, and E. Stork, "Pattern Classisifcation", 2nd ed. New York, 2000, pp. 517-583.

[14] S. Khan and M. Shah, "Tracking people in presence of occlusion", in Proc. Asian Conference on Computer Vision, Taipei, Taiwan, 2000

[15] S.Z. Li, "Markov random field modeling in image analysys", Springer-Verlag, Tokyo, 2001, pp. 81-118.

[16] L. Salgado N. Garcia, J. Menedez, and E. Rendon, "Efficient image segmentation for region-based motion estimation and compensation", IEEE Trans. Circuits and Systems for Video Technology, 10(7), 2000, pp. 1029-1039.

[17] J. C. Setubal, "Sequential and Parallel Experimental Results with Bipartite Matching Algorithms". Technical Report IC-96-09, Institute of Computing, State University of Campinas, Brazil, 1996.

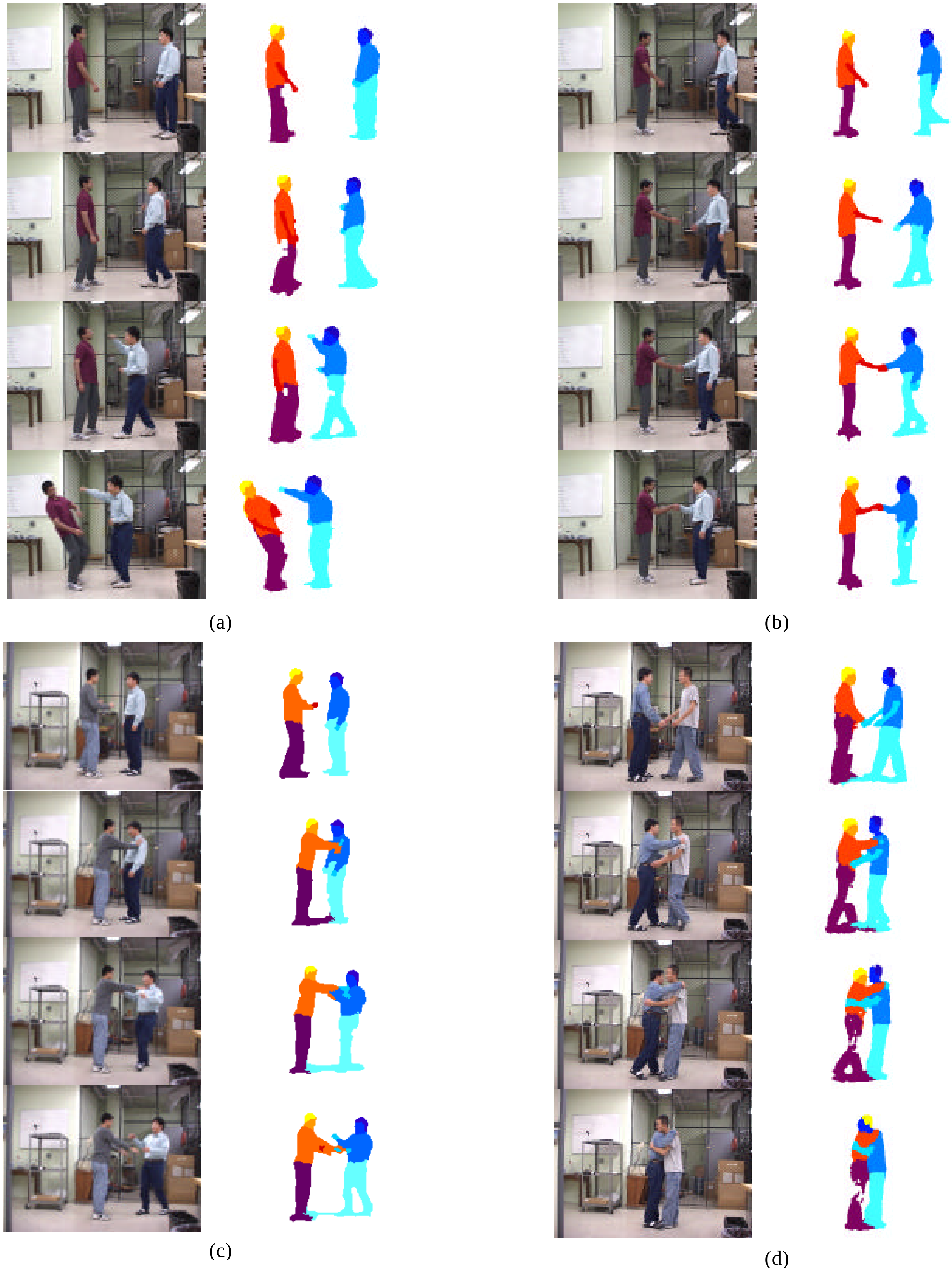

Fig. 6. Example input sequences (1st column) and the corresponding sequences of tracked body parts (2nd column). (a) ‘punching’, (b) ‘hand-shaking’, (c) ‘pushing’, (d) ‘hugging’. Degree of occlusion increases from (a) to (d). The consistency of grayscale (which is color in the original version) of the individual blobs in the 2nd columns in each sequence indicates that each person is correctly tracked and that the interacting persons’ individual body parts are correctly segmented and tracked despite of occlusion, shadow, and foreground pixel loss.