Tracking of Human Hands and Faces through Probabilistic Fusion of Multiple Visual Cues

Haris Baltzakis[1], Antonis Argyros[1,2], Manolis Lourakis[1], and Panos Trahanias[1,2]

[1] Foundation for Research and Technology – Hellas,
Heraklion, Crete, Greece
[2] Department of Computer Science, University of Crete,
Heraklion, Crete, Greece
{xmpalt,argyros,lourakis,trahania}@ics.forth.gr

Abstract. This paper presents a new approach for real time detection and tracking of human hands and faces in image sequences. The proposed method builds upon our previous research on color-based tracking and extends it towards building a system capable of distinguishing between human hands, faces and other skin-colored regions in the image background. To achieve these goals, the proposed approach allows the utilization of additional information cues including motion information given by means of a background subtraction algorithm, and top-down information regarding the formed image segments such as their spatial location, velocity and shape. All information cues are combined under a probabilistic framework which furnishes the proposed approach with the ability to cope with uncertainty due to noise. The proposed approach runs in real time on a standard, personal computer. The presented experimental results, confirm the effectiveness of the proposed methodology and its advantages over previous approaches.

1 Introduction

Real time segmentation of human hands and faces in image sequences is a fundamental step in many vision systems, including systems designed for tasks such as human-computer and human-robot interaction, video-enabled communications, visual surveillance, etc. A variety of approaches have been employed to achieve this task. Several of them rely on the detection of skin-colored areas [1,2,3,4]. The idea behind this family of approaches is to build appropriate color models of human skin and then classify image pixels based on how well they fit to these color models. On top of that, various segmentation techniques are used to cluster skin-colored pixels together into solid regions that correspond to human hands and/or human faces.

A second family of approaches tries to differentiate between static and moving regions based on background subtraction techniques [5,6,7,8]. These techniques involve the calculation of a background model and the comparison (subtraction) of each new frame against this model. Human parts and/or other non-static objects are extracted by thresholding the result.

Both families of approaches have advantages and disadvantages: color based approaches generally work better on constant lighting and predictable background

A. Gasteratos, M. Vincze, and J.K. Tsotsos (Eds.): ICVS 2008, LNCS 5008, pp. 33–42, 2008.

conditions but are unable to distinguish between skin-colored human body parts and similarly-colored objects in the background. On the other hand, background subtraction techniques are not suitable for applications requiring a moving camera. Moreover, depending on the application at hand, additional processing may be required to further-process foreground regions and/or utilize additional information cues in order to distinguish humans from other moving objects and/or distinguish human hands or faces [9,10,11].

In this paper we build upon our previous research on color-based, multiple objects tracking [3]. Based on this, we propose a broader probabilistic framework that allows the utilization of additional information cues (image background model, expected spatial location, velocity and shape of the detected and tracked segments) to efficiently detect and track human body parts, classify them into hands and faces, and avoid problems caused by the existence of skin-colored objects in the image background. Thus, contrary to other color based approaches, the proposed approach operates effectively in cluttered backgrounds that contain skin-colored objects. All information cues are combined by means of a graphical model (bayes network) which provides a clean mathematical formalism that makes it possible to explicitly model the probabilistic relationships among the involved quantities.

The proposed approach runs in real time on a conventional personal computer. Experimental results presented in this paper, confirm the effectiveness of the proposed methodology and its advantages over our previous approach.

The rest of the paper is organized as follows. For completeness purposes, Section 2 provides an overview of the existing approach to color based tracking, on which the new, proposed approach is based. In Section 3 the proposed approach and its individual components are presented in detail. Experimental results are presented in Section 4. Finally, in Section 5, conclusions and hints for future work are given.

2 The Skin-Color Based Tracker

The proposed approach, builds on our previous research on color-based tracking of multiple skin-color regions [3]. According to this approach, tracking is facilitated by three processing layers.

2.1 Processing Layer 1: Assign Probabilities to Pixels

Within the first layer, the input image is processed in order to estimate and assign likelihood values to pixels for depicting skin colored regions like human hands and faces. Assuming that S is a binary random variable that indicates whether a specific pixel depicts skin and C is the random variable that indicates the color of this pixel and c is a perceived color, the probability $P(S{=}1|C{=}c)$ can be estimated according to the Bayes law as:

$$P(S{=}1|C{=}c) = \frac{P(C{=}c|S{=}1)P(S{=}1)}{P(C{=}c)} \tag{1}$$

C is assumed to be two dimensional and discrete, i.e. encoding the Y and U components of the YUV color space within $[0..255]^2$. In the above equation, $P(S{=}1)$ and $P(C{=}c)$

are the prior probabilities of a pixel depicting skin and of a pixel having a specific color c, respectively. $P(C{=}c|S{=}1)$ is the conditional probability of a pixel having color c, given that it depicts skin. All these three probabilities are computed off-line during a separate training phase. A procedure for on-line adaptation of these three priors is also described in [3].

2.2 Processing Layer 2: From Pixels to Blobs

This layer applies hysteresis thresholding on the probabilities determined at layer 1. These probabiliies are initially thresholded by a "strong" threshold T_{max} to select all pixels with $P(S{=}1|C{=}c) > T_{max}$. This yields high-confidence skin pixels that constitute the seeds of potential skin-colored blobs. A second thresholding step, this time with a "weak" threshold T_{min}, along with prior knowledge with respect to object connectivity to form the final blobs. During this step, image points with probability $P(S{=}1|C{=}c) > T_{min}$ where $T_{min} < T_{max}$, that are immediate neighbours of skin-colored image points are recursively added to each blob.

A connected components labeling algorithm is then used to assign different labels to pixels that belong to different blobs. Size filtering on the derived connected components is also performed to eliminate small, isolated blobs that are attributed to noise and do not correspond to meaningful skin-colored regions.

Finally, a feature vector for each blob is computed. This feature vector contains statistical properties regarding the spatial distribution of pixels within the blob and will be used within the next processing layer for data association.

2.3 Processing Layer 3: From Blobs to Object Hypotheses

Within the third and final processing layer, blobs are assigned to object hypotheses which are tracked over time. Tracking over time is realized through a scheme which can handle multiple objects that may move in complex trajectories, occlude each other in the field of view of a possibly moving camera and whose number may vary over time. To achieve this goal, appropriate hypothesis management techniques ensure: (a) the generation of new hypotheses in cases of unmatched evidence (unmatched blobs), (b) the propagation and tracking of existing hypotheses in the presence of multiple, potential occluding objects and (c) the elimination of invalid hypotheses (i.e. when tracked objects disappear from the scene of view).

3 Proposed Approach

As with our previous approach, the proposed approach is organized into three layers as well. These layers are depicted in Fig. 1 and are in direct correspondence with the ones described in the previous section. A major difference is that the first layer is completely replaced by a new layer that implements the pixel model that will be described in the following section. This new pixel model allows the computation of probabilities for human hands and faces instead of skin. That is, distinguishing between hands and faces is facilitated directly from the first layer of the architecture proposed in this paper.

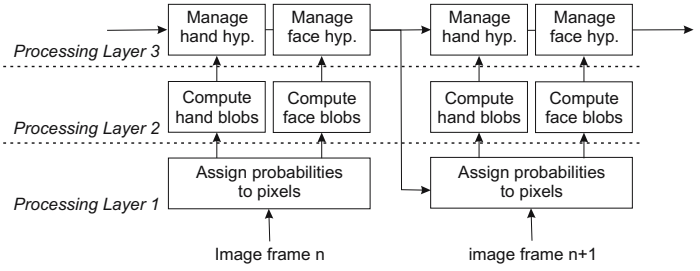

Fig. 1. Block diagram of the proposed approach. Processing is organized into three layers.

The second and the third layers consist of the same components as the ones used in [3]. However, they are split into two different parallel processes, the first being responsible for managing hand blobs and hypotheses and the second for managing face blobs and hypotheses. Another notable difference between our previous approach and the current one is signified by the top-down dependence of the 1st processing level of frame $n + 1$ from the 3rd processing layer of frame n (see Fig. 1). This dependence is essentially responsible for bringing high-level information regarding object hypotheses down to the pixel level.

In the following section, we will emphasize to the new approach for realizing the fist processing layer of Fig. 1 which constitutes the most important contribution of this paper. Details regarding the implementation of the other two layers can be found in [3].

3.1 Notation

Let $\mathcal{U}$ be the set of all pixels of an image. Let $\mathcal{M}$ be the subset of $\mathcal{U}$ corresponding to foreground pixels (i.e a human body) and $\mathcal{S}$ be the subset of $\mathcal{U}$ containing pixels that are skin colored. Accordingly, let $\mathcal{H}$ and $\mathcal{F}$ stand for the sets of pixels that depict human hands and faces, respectively. The relations between the above mentioned sets are illustrated in the Venn diagram in Fig. 2(a). Notice the convention that both $\mathcal{F}$ and

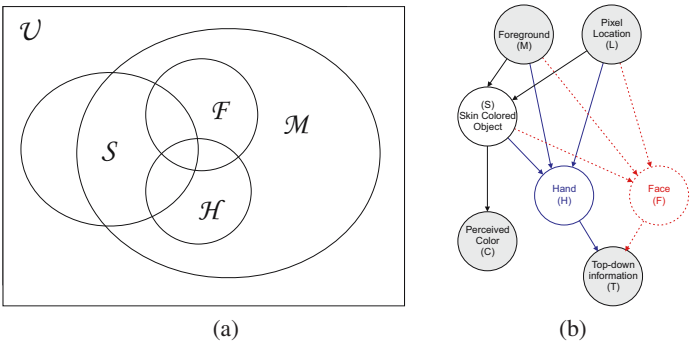

Fig. 2. The proposed approach. (a) Venn diagram representing the relationship between the pixel sets $\mathcal{U}$, M, S, H and F. (b) The proposed Bayes net.

$\mathcal{H}$ are assumed to be subsets of $\mathcal{M}$ (i.e all pixels depicting hands and faces belong to the foreground). It is also important that $\mathcal{F}$ and $\mathcal{H}$ are not mutually exclusive, i.e. a foreground pixel might belong to a hand, to a face or to both (i.e in the case of occlusions). Last but not least, the model does not assume that all hands and face pixels are skin-colored.

3.2 The Pixel Model

Let M, S, F and H be binary random variables (i.e taking values in $\{0, 1\}$), indicating whether a pixel belongs to set $\mathcal{M}$, $\mathcal{S}$, $\mathcal{F}$ and $\mathcal{H}$, respectively. Let L be the 2D location vector containing the pixel image coordinates and let T be a variable that encodes a set of features regarding the currently tracked hypotheses. Note that the source of this piece of information lies in the third processing layer, i.e. the layer that is responsible for the management and the tracking of face and hand hypotheses over time, hence this piece of information traverses the model in a 'top-down" direction.

The goal is to compute whether a pixel belongs to a hand and/or a face, given (a) the color c of a single pixel, (b) the information m on whether this pixel belongs to the background (i.e. $M{=}m$) and, (c) the values l, t of L and T, respectively. More specifically, we need to estimate the conditional probabilities $P_h = P(H{=}1|C{=}c, T{=}t, L{=}l, M{=}m)$ and $P_f = P(F{=}1|C{=}c, T{=}t, L{=}l, M{=}m)$.[1]

To perform this estimation, we use the Bayesian network shown in Fig. 2(b). The nodes of the graph are random variables that represent degrees of belief on particular aspects of the problem. The edges are parameterized by conditional probability distributions that represent dependencies between the involved variables. A notable property of the network of Fig. 2(b) is that if $P(H)$ and $P(F)$ are assumed to be independent, then the network can be decomposed into two different networks one for the computation of hand probabilities and the other for face probabilities. The first one corresponds to what remains from Fig. 2(b) when the red dotted components are excluded and the second one is the one with the blue components excluded.

Regarding the hands, we know that:

$$P(H{=}1|c, t, l, m) = \frac{P(H{=}1, c, t, l, m)}{P(c, t, l, m)} \tag{2}$$

By marginalizing the numerator over both possible values for S and the denominator over all four possible combinations for S and H (the values of S and H are expressed by the summation indices s and h, respectively), Eq. (2) can be expanded as:

$$P_h = \frac{\sum\limits_{s\in\{0,1\}} P(H{=}1, s, c, t, l, m)}{\sum\limits_{s\in\{0,1\}} \sum\limits_{h\in\{0,1\}} P(h, s, c, t, l, m)} \tag{3}$$

[1] Note that capital letters are used to indicate variables and small letters to indicate specific values for these variables. For brevity, we will also use the notation $P(x)$ to refer to probability $P(X{=}x)$ where X any of the above defined variables and x a specific value of this variable.

By applying the chain rule of probability and by taking advantage of variable independence as the ones implied by the graph of Fig. 2(b), we easily obtain:

$$P(h,s,c,t,l,m) = P(m)P(l)P(t|h)P(c|s)P(s|l,m)P(h|l,s,m) \tag{4}$$

Finally, by substituting to Eq. (2), we obtain:

$$P_h = \frac{P(t|H{=}1) \sum_{s\in\{0,1\}} P(c|s)P(s|l,m)P(H{=}1|l,s,m)}{\sum_{h\in\{0,1\}} P(t|h) \sum_{s\in\{0,1\}} P(c|s)P(s|l,m)P(h|l,s,m)} \tag{5}$$

Similarly, for the case of faces, we obtain:

$$P_f = \frac{P(t|F{=}1) \sum_{s\in\{0,1\}} P(c|s)P(s|l,m)P(F{=}1|l,s,m)}{\sum_{f\in\{0,1\}} P(t|f) \sum_{s\in\{0,1\}} P(c|s)P(s|l,m)P(f|l,s,m)} \tag{6}$$

Details regarding the estimation of the individual probabilities that appear in Eq. (5) and in Eq. (6) are provided in the following paragraphs.

Foreground segmentation. It can be easily verified that when $M = 0$ (i.e. a pixel belongs to the background), the numerators of both Eq. (5) and Eq. (6) become zero as well. This is because, as already mentioned, hands and faces have been assumed to always belong to the foreground. This convention simplifies computations because Equations (5) and (6) should only be evaluated for foreground pixels.

In order to compute M, we apply the background subtraction technique proposed by Stauffer and Grimson [6,7] that employs an adaptive Gaussian mixture model on the background color of each image pixel. The number of Gaussians, their parameters and their weights in the mixture are computed online.

The color model. $P(c|s)$ is the probability of a pixel being perceived with color c given the information on whether it belongs to skin or not. This probability is the same as the one in Eq. (1) and can be obtained off-line through a separate training phase with the procedure described in [3]. The result is encoded in the form of two two-dimensional look-up tables; one table for skin-colored objects ($s = 1$) and one table for all other objects ($s = 0$). The rows and the columns of both look-up tables correspond to the Y and U dimensions of the YUV color space.

Top-down information. In this work, for each tracked hypothesis, a feature vector t is generated which is propagated in a "top-down" direction in order to further assist the assignment of hand and face probabilities to pixels at processing layer 1. The feature vector t consists of two different features:

1. The average vertical speed v of the hypothesis, computed as the vertical component of the speed of the centroid of the ellipse modeling the hypothesis. The rationale behind the selection of this feature is that hands are expected to exhibit considerable average vertical speed v compared to other skin colored regions such as heads.

2. The ratio r of the perimeter of the hypothesis contour over the circumference of a hypothetical circle having the same area as the area of the hypothesis. The rationale behind the selection of this feature is that hands are expected to exhibit high r compared to faces. That is, $r = \frac{1}{2}\rho/\sqrt{\pi\alpha}$, where ρ and α are the length of the hypothesis contour and the area, respectively.

Given v and r, $P(t|h)$ and $P(t|f)$ are approximated as:

$$P(t|h) \approx P(v|h)P(r|h) \tag{7}$$

$$P(t|f) \approx P(v|f)P(r|f) \tag{8}$$

In our implementation, all $P(v|h)$, $P(r|h)$, $P(v|f)$ and $P(r|f)$ are given by means of one-dimensional look-up tables that are computed off-line, during training. If there is more than one hypothesis overlapping with the specific pixel under consideration, the hypothesis that yields maximal results is chosen separately for $P(t|h)$ and $P(t|f)$. Moreover, if there is no overlapping hypothesis at all, all of the conditional probabilities of Equation (7) are substituted by the maximum values of their corresponding look-up tables.

The spatial distribution model. A spatial distribution model for skin, hands and faces is needed in order to evaluate $P(s|l,m)$, $P(h|l,s,m)$ and $P(f|l,s,m)$. All these three probabilities express prior probabilities that can be obtained during training and stored explicitly for each each location L (i.e for each image pixel). In order to estimate these probabilities, a set of eight different quantities are computed off-line during a training phase. These quantities are depicted in Table 1 and indicate the number of foreground pixels found in the training sequence for every possible combination of s and h or f. As discussed in Section 3.2, only computations for foreground pixels are necessary. Hence, all training data correspond to $M = 1$. We can easily express $P(s|l, M{=}1)$,

Table 1. Quantities estimated during training for the spatial distribution model

h=0				h=1			
$f=0$		$f=1$		$f=0$		$f=1$	
$s=0$	$s=1$	$s=0$	$s=1$	$s=0$	$s=1$	$s=0$	$s=1$
s_{000}	s_{001}	s_{010}	s_{011}	s_{100}	s_{101}	s_{110}	s_{111}

$P(h|l,s,M{=}1)$ and $P(f|l,s,M{=}1)$ in terms of the eight quantities of Table 1.

$$P(s|l,M{=}1) = \frac{P(s,M{=}1,l)}{P(M{=}1,l)} = \frac{s_{00s}+s_{01s}+s_{10s}+s_{11s}}{s_{000}+s_{001}+s_{010}+s_{011}+s_{100}+s_{101}+s_{110}+s_{111}} \tag{9}$$

Similarly:

$$P(h|l,s,M{=}1|) = \frac{P(h,s,M{=}1,l)}{P(s,M{=}1,l)} = \frac{s_{h0s}+s_{h1s}}{s_{00s}+s_{01s}+s_{10s}+s_{11s}} \tag{10}$$

$$P(f|l,s,M{=}1|) = \frac{P(h,s,M{=}1,l)}{P(s,M{=}1,l)} = \frac{s_{0fs}+s_{1fs}}{s_{00s}+s_{01s}+s_{10s}+s_{11s}} \tag{11}$$

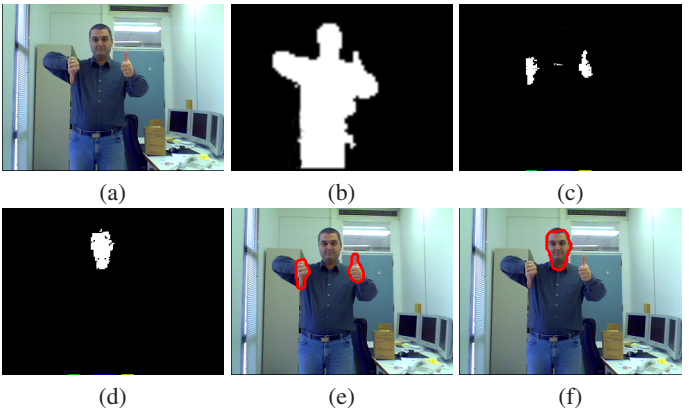

Fig. 3. The proposed approach in operation. (a) original frame, (b) background subtraction result, (c),(d) pixel probabilities for hands and faces, (e),(f) contours of hand and face hypotheses.

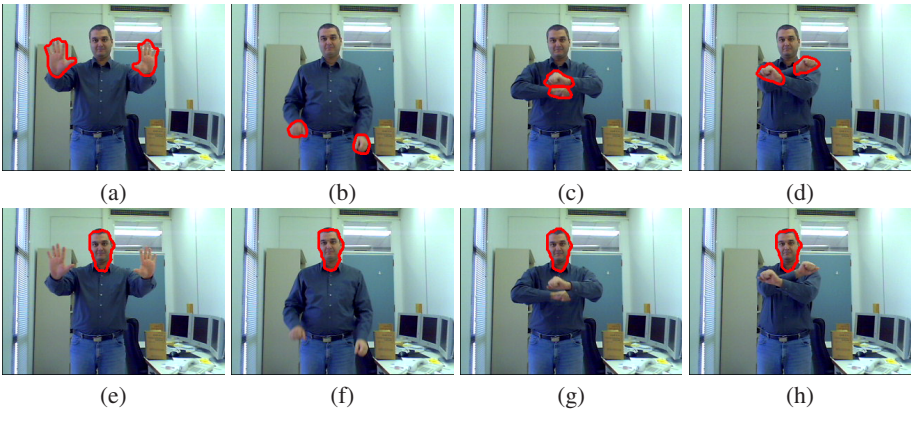

Fig. 4. Four frames for a sequence depicting a man performing gestures in an office environment (a,b,c,d) hand hypotheses, (e,f,g,h) face hypotheses

4 Results and Discussion

The proposed approach has been assessed using several video sequences containing people performing various gestures in indoor environments. Several videos of example runs are available on the web[2].In this section we will present results obtained from a sequence depicting a man performing a variety of hand gestures. The resolution of the sequence is 320×240 and it was obtained with a standard, low-end web camera at 30 frames per second. Figure 4 depicts various intermediate results obtained at different stages of the proposed approach. A frame of the test sequence is shown in Fig. 4(a).

[2] http://http://www.ics.forth.gr/ xmpalt/research/bayesfuse/index.html

Figure 4(b) depicts the result of the background subtraction algorithm, i.e $P(M)$. In order to achieve real-time performance, the background subtraction algorithm operates at down-sampled images of dimensions 160×120.

Figures 4(c) and 4(d) depict P_h and P_f i.e. the results of the first processing layer of the proposed approach. The contours of the blobs that correspond to hand and face hypotheses are shown in Figures 4(e) and 4(f), respectively. As can verified, the algorithm manages to correctly identify both hands and the face of the depicted man. Notice also that, in contrast to what would happen with the original tracking algorithm of [3], the skin-colored books are not detected.

Figure 4 shows additional four frames out of the same sequence. Figures on the left column depict the resulting hand hypotheses, while figures on the right column depict face hypotheses. In all cases, the proposed approach has been successful in correctly identifying the face and the hands of the person. The presented results were obtained at a standard 3GHz personal computer which was able to process images of size 320×240 at 30Hz.

5 Summary

We have presented an approach for visual detection and tracking of human faces and hands. The proposed approach builds on our previous research on color-based, skin-color tracking and extends it towards building a system capable of distinguishing between human hands, faces and other skin-colored regions in the background. This is achieved by the exploitation of additional information cues including motion information as well as spatial location, velocity and shape of detected and tracked objects. All information cues are combined under a probabilistic framework which furnishes the proposed approach with the ability to cope with uncertainty due to noise. Experimental results presented in this paper, confirm the effectiveness of the proposed approach. The resulting system combines the tracking robustness of the approach presented in [3] together with the new capability of discriminating hands from faces. Additionally, the proposed system is not influenced by the presence of skin-colored background objects. These features, together with the maintained real-time performance characteristics, constitute a very attractive framework for building more complex and ambitious vision systems.

Acknowledgements. This work has been partially supported by EU-IST NoE MUSCLE (FP6-507752), the Greek national project XENIOS and the EU-IST project INDIGO (FP6-045388).

References

1. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. International Journal of Computer Vision 46(1), 81–96 (2002)
2. Sigal, L., Sclaroff, S., Athitsos, V.: Skin color-based video segmentation under time-varying illumination. IEEE Trans. Pattern Analysis and Machine Intelligence 26(7), 862–877 (2004)
3. Argyros, A.A., Lourakis, M.I.A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: Proc. European Conference on Computer Vision, Prague, Chech Republic, May 2004, pp. 368–379 (2004)

4. Nickel, K., Seemann, E., Stiefelhagen, R.: 3d-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario. In: Proc. IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 2004, pp. 565–570 (2004)
5. Haritaoglu, I., Harwood, D., Davis, L.: W4s: A real time system for detecting and tracking people in 2.5 d. In: Proc. European Conference on Computer Vision, Freiburg, Germany, June 1999, pp. 877–892 (1999)
6. Grimson, W.E.L., Stauffer, C.: Adaptive background mixture models for real time tracking. In: Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Ft. Collins, USA, June 1999, pp. 246–252 (1999)
7. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Ft. Collins, USA, June 1999, pp. 2246–2252 (1999)
8. Elgammal, A., Harrwood, D., Non-parametric, L.D.: model for background subtraction. In: Proc. European Conference on Computer Vision, Dublin, Ireland (June 2000)
9. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
10. Yoon, S.M., Kim, H.: Real-time multiple people detection using skin color, motion and appearance information. In: Proc. IEE International Workshop on Robot and Human Interactive Communication (ROMAN), Kurashiki, Okayama Japan, September 2004, pp. 331–334 (2004)
11. Zhou, J., Hoang, J.: Real time robust human detection and tracking system. In: Proc. IEEE Computer Vision and Pattern Recognition (CVPR 2005) III, pp. 149–149 (2005)