Journal of Computer Science 6 (9): 1002-1007, 2010
ISSN 1549-3636

Hand Gesture Recognition for Human-Computer Interaction

S. Mohamed Mansoor Roomi, R. Jyothi Priya and H. Jayalakshmi
Department of Electronics and Communication, Thiagarajar College of Engineering,
Madurai, Tamil Nadu, India

Abstract: Problem statement: With the development of ubiquitous computing, current user interaction approaches with keyboard, mouse and pen are not sufficient. Due to the limitation of these devices the useable command set is also limited. Direct use of hands can be used as an input device for providing natural interaction. **Approach:** In this study, Gaussian Mixture Model (GMM) was used to extract hand from the video sequence. Extreme points were extracted from the segmented hand using star skeletonization and recognition was performed by distance signature. **Results:** The proposed method was tested on the dataset captured in the closed environment with the assumption that the user should be in the Field Of View (FOV). This study was performed for 5 different datasets in varying lighting conditions. **Conclusion:** This study specifically proposed a real time vision system for hand gesture based computer interaction to control an event like navigation of slides in Power Point Presentation.

Key words: Gaussian mixture model, EM algorithm, vocabulary, star skeletonization

INTRODUCTION

Human gestures have long been an important way of communication, adding emphasis to voice messages or even being a complete message by itself. Such human gestures could be used to improve human machine interface. These may be used to control a wide variety of devices remotely. Vision-based framework can be developed to allow the users to interact with computers through human gestures. This study focuses in understanding such human gesture recognition, typically hand gesture. Hand gesture recognition generally involves various stages like video acquisition, background subtraction, feature extraction and gesture recognition. The rationale in background subtraction is detecting the moving objects from the difference between the current frame and a reference frame, often called the background image or background model. Wren *et al*. (1997) have proposed to model the background independently at each pixel location. The model is based on ideally fitting a Gaussian probability density function (pdf) on the last few pixel's values. Lo and Velastin (2001) proposed to use the median value of the last 'n' frames as the background model. Cucchiara *et al*. (2003) argued that such a median value provides an adequate background model even if the subsequent frames are sub sampled with respect to the original frame rate by a factor of 10. The main disadvantage of a median-based approach is that its computation requires a buffer with the recent pixel values. Stauffer and Grimson (1999) proposed Gaussian Mixture Model (GMM) in which scene background is modeled by classifying the pixels as object or background by computing posterior probabilities. The advantage of using GMM is that it provides multiple background model to cope with multi background objects. Then the features are extracted from the foreground objects. Skin color based features can be extracted from the foreground objects as in (Jones and Rehg, 1999), but it lacks the robustness to varying illumination conditions and it requires an exhaustive training phase. Extreme points of the foreground object, typically hand, can be used to best describe the gesture. Skeletonization is used to extract the extreme points as it provides a mechanism for controlling scale sensitivity. In Sánchez-Nielsen *et al*. (2004), gestures are recognized by Hausdorff distance measure but it is too sensitive to the shape of the hand gesture. The proposed method employs Gaussian Mixture Model to segment the hand region. Star skeletonization is used to extract the extreme points of the hand region. Gestures are recognized based on the distance signature.

MATERIALS AND METHODS

This study proposes a method to automatically recognize the hand gestures which could be used to control any event like power point presentation. The

Corresponding Author: S. Mohamed Mansoor Roomi, Department of Electronics and Communication,
Thiagarajar College of Engineering, Madurai, Tamil Nadu, India

proposed method has three stages viz. Gaussian Mixture Model to detect the hand, Star skeletonization for feature extraction and Distance signature for hand gesture recognition. The overall block diagram of the work is given in Fig. 1.

Gaussian mixture model: Background of an image is modeled using Gaussian Mixture Model. Each pixel x is modeled by a mixture of K Gaussian distribution. Different Gaussian are assumed to represent different colors. The mixtures are weighted based on the time proportions of colors in consequent frames. The probable background colors stay longer or more static in video sequences. The probability of the mixture model p(x) with 'M' number of components (classes) is given in Eq. 1:

$$p(x) = \sum_{m=1}^{M} \alpha_m p(x/m) \qquad (1)$$

where, $\alpha_m \in [0,1]$ (m = 1,2,…..M) is the mixing proportions subject to the condition given by Eq. 2:

$$\sum_{m=1}^{M} \alpha_m = 1 \qquad (2)$$

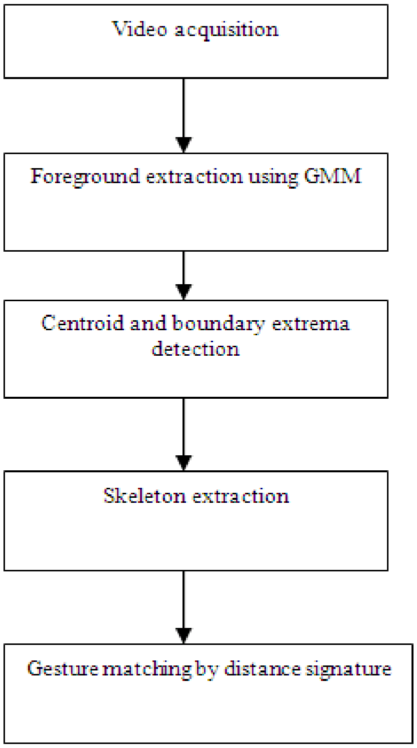

Video acquisition
Foreground extraction using GMM
Centroid and boundary extrema detection
Skeleton extraction
Gesture matching by distance signature

Fig. 1: Overall block diagram

For Gaussian mixtures, each component density p(x/m) is a normal probability distribution as in Eq. 3:

$$p(x/\theta_m) = \frac{1}{(2\Pi)^{n/2}\det(C_m)^{1/2}}\exp\begin{bmatrix} -\frac{1}{2}(x-\mu_m)^T \\ C_m^{-1}(x-\mu_m) \end{bmatrix} \qquad (3)$$

where, 'T' denotes the transpose operation. Here the mean, μ_m and covariance C_m parameters are encapsulated into a parameter vector, as $\theta_m = (\mu_m, C_m)$. The parameters θ_m and α_m are concatenated as $\Theta = (\alpha_1, \alpha_2, \ldots \alpha_m\ \theta_1, \theta_2, \ldots\ \theta m)$. Using Θ, Eq. 1 can be rewritten as Eq. 4:

$$p(x/\Theta) = \sum_{m=1}^{M} \alpha_m p(x/\theta_m) \qquad (4)$$

If the component from which 'x' originated is known, then it is feasible to determine the parameters Θ and vice versa. Since the parameters are unknown it is difficult to estimate. The EM algorithm is incorporated to overcome this difficulty through the concept of missing data.

The EM algorithm: Expectation Maximization (EM) algorithm (Dempster *et al*., 1977) is a widely used class of iterative algorithms for Maximum Likelihood (ML) or Maximum Posteriori (MAP) estimation in problems with missing data. Given a set of samples $X = (x_1, x_2, \ldots, x_k)$, the complete data set Z = (X, Y) consists of the sample set X and a set Y of variables indicating from which component of the mixtures the sample came. The estimation of parameters of the Gaussian mixtures with the EM algorithm is discussed in (Zhang *et al*., 2003).

The EM algorithm consists of an E-step and M step. Suppose that $\Theta^{(t)}$ denotes the estimation of Θ obtained after the t^{th} iteration of the algorithm. Then at the $(t+1)^{th}$ iteration, the E-step computes the expected complete data log-likelihood function given by Eq. 5:

$$Q\left(\Theta/\Theta^{(t)}\right) = \sum_{k=1}^{K}\sum_{m=1}^{M}\{\log\alpha_m p(x_k/\theta_m)\}P\left(m/x_k;\Theta^{(t)}\right) \qquad (5)$$

where, $P(m/x_k;\ \Theta^{(t)})$ is a posterior probability and is computed as in Eq. 6:

$$P\left(m/x_k;\Theta^{(t)}\right) = \frac{\alpha_m^{(t)} p\left(x_k/\theta_m^{(t)}\right)}{\sum_{l=1}^{M}\alpha_m^{(t)} p\left(x_k/\theta_l^{(t)}\right)} \qquad (6)$$

And the M-step finds the $(t+1)^{th}$ estimation $\Theta^{(t+1)}$ of Θ by maximizing through Eq. 7-9:

$$\alpha_m^{(t+1)} = \frac{1}{K}\sum_{k=1}^{K} P\left(m/x_k;\Theta^{(t)}\right) \quad (7)$$

$$\mu_m^{(t+1)} = \frac{\sum_{k=1}^{K} x_k P\left(m/x_k;\Theta^{(t)}\right)}{\sum_{k=1}^{K} P\left(m/x_k;\Theta^{(t)}\right)} \quad (8)$$

$$C_m^{(t+1)} = \frac{\sum_{k=1}^{K} P\left(m/x_k;\Theta^{(t)}\right)\left\{\left(x_k - \mu_m^{(t+1)}\right)\left(x_k - \mu_m^{(t+1)}\right)^T\right\}}{\sum_{k=1}^{K} P\left(m/x_k;\Theta^{(t)}\right)} \quad (9)$$

The parameters are maximized and their optimal values are obtained once the convergence is achieved. The pixel ‘x’ is fitted to the corresponding component by optimal weight, mean and covariance. The extracted foreground object from the Gaussian Mixture Model is applied to star skeletonization for feature extraction.

Star skeletonization: Star skeleton, a simple but robust technique extracts the feature points from the foreground object. The features consist of the several vectors which are the distances from the extremities of human contour to its centroid. The basis of the star skeleton is to connect the extremities of human contour with its centroid. To find the extremities, distance from boundary point to the centroid is calculated through boundary tracking in a clockwise or counter-clockwise order. In distance function, the extremities are located at local maxima. The distance function is smoothed by a low pass filter for noise reduction. Consequently, the final extremities are detected by finding local maxima in smoothed distance function.

Boundary extraction: The first pre-processing step is morphological dilation followed by erosion to clean up anomalies in the targets. This removes any small holes in the object and smoothes any interlacing anomalies. This closing operation is performed on the binarized image, i.e., detected hand (A) is dilated followed by erosion using the structuring element B given in Eq. 10:

$$D_{AB} = (A \oplus B); E_{DB} = (D_{AB} \Theta B) \quad (10)$$

Where:
D_{AB} = The dilation of ‘A’ with ‘B’
E_{DB} = The erosion of D_{AB} by ‘B’
$\oplus$ = The morphological dilation on the object
Θ = The morphological erosion on the object

This effectively makes the algorithm robust for small features of the object. The boundary of the object (O_B) is extracted by the image subtraction of D_{AB} and E_{DB} which is given in Eq. 11:

$$O_B = D_{AB} - E_{DB} \quad (11)$$

Boundary extrema detection: Centroid is an intersection of all the straight lines that divides an object into parts of equal moment about the line. The centroid of the object boundary (x_c, y_c) is given by:

$$(x_c, y_c) = \left(\frac{1}{N_b}\sum_{i=1}^{N_b} x_i, \frac{1}{N_b}\sum_{i=1}^{N_b} y_i\right) \quad (12)$$

Where:
(x_c, y_c) = The average boundary pixel position
‘N_b’ = The number of boundary pixels
(x_i, y_i) = The ith point lies on the boundary

To find the extrema points in the object, the distances between the centroid and boundary points are calculated using Eq. 13:

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (13)$$

The distance of boundary to centroid, ‘d_i’ gives the information of the extremal points in the objects. From the distance plot the extrema points are considered as skeleton points.

Skeleton extraction: The distance between the boundary points and centroid is calculated and plotted. The distance plot of an object contour has noises. Therefore these noises are removed by smoothing in frequency domain. Fourier transform is performed on the measured distance as given in Eq. 14 and is smoothed by a low pass filter:

$$D(u) = \frac{1}{L}\sum_{x=0}^{L-1} d(x) e^{-j2\Pi ux/L} \quad (14)$$

where, ‘L’ is the size of distance vector d(x). The low pass filter in frequency domain is represented as in Eq. 15:

$$H(u) = \begin{cases} 1 & \text{if } \mathrm{Dist}(u) \le c \\ 0 & \text{Otherwise} \end{cases} \quad (15)$$

Where:

$$\mathrm{Dist}(u) = \sqrt{u - \left(L/2\right)^2}$$

‘c’ = Cut-off frequency

Then the smoothening is carried out in frequency domain followed by the inverse Fourier transform as given in Eq. 16 and 17:

$$D_{smooth}(u) = D(u) \bullet H(u) \quad (16)$$

$$d_{smooth}(x) = \frac{1}{L}\sum_{u=0}^{L-1} D_{smooth}(u) e^{j2\Pi ux/L} \quad (17)$$

Where:

$\bullet$ = The multiplication operator

D_{smooth} = The filtered 1-D distance in frequency domain

d_{smooth} = The smoothed distance in spatial domain

Local maxima of d_{smooth} are taken as extrema points and the Star skeleton is constructed by connecting them to the object centroid (x_c,y_c). Local maxima are detected by finding zero-crossings of the difference function mentioned in Eq. 18:

$$\delta(x) = d_{smooth}(x) - d_{smooth}(x-1) \quad (18)$$

RESULTS AND DISCUSSION

The dataset for the proposed study is acquired using a web cam and simulated using Matlab 7.0. The open and close fists are used to represent the navigation to next slide and previous slide respectively. These gestures shown in Fig. 2 and 3 are used as a vocabulary for human computer interaction.

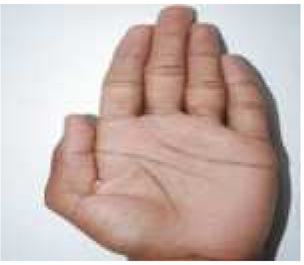


Fig. 2: Gesture to move to next slide

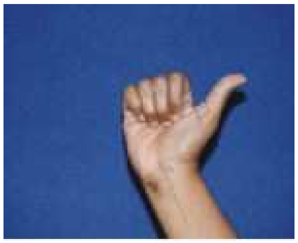


Fig: 3 Gesture to move to previous slide

Gaussian Mixture Model is applied on the input video to extract the foreground. The input frame is shown in Fig. 4a and 5a. This algorithm is trained to segment the object which exhibits drastic movements. Fig. 4b and 5b shows the segmented hand image in the input video which depicts the gesture to move next slide and previous slide respectively. The extracted moving object is given to the star skeletonization algorithm. Morphological operations are applied to extract the contour of the segmented hand region as shown in Fig. 4c and 5c. The plot of distance between centroid and the boundary of the object is shown in Fig. 4d and 5d. The smoothed distance plot is shown in Fig. 4e and 5e. From the smoothed distance plot, the extrema points are extracted.

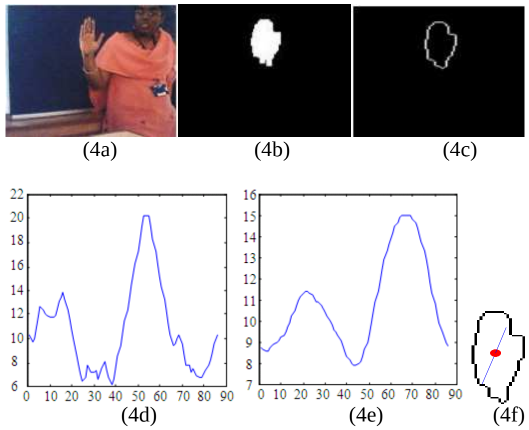


Fig. 4: (a) Input frame for open fist (b) Segmented hand image; (c) Boundary extracted image; (d) distance Plot; (e) Smoothed distance plot (f) Star skeleton for open fist

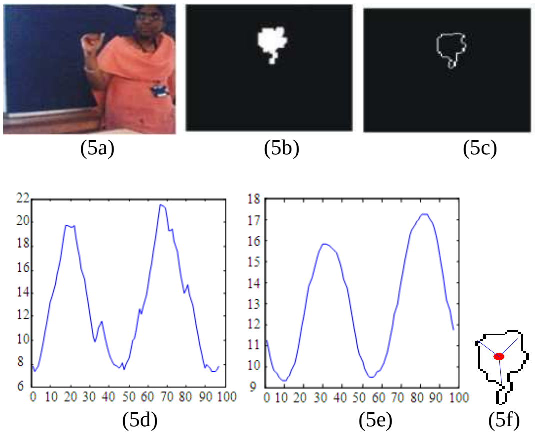
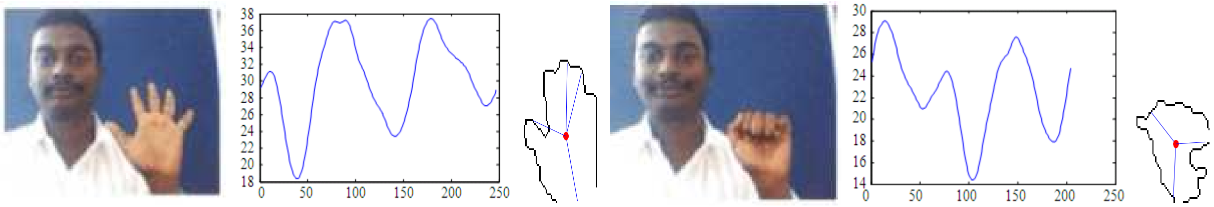
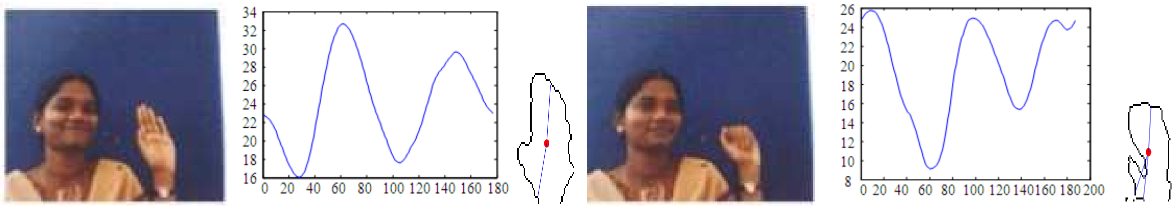


Fig. 5: (a) Input frame for close fist; (b) Segmented hand image; (c) Boundary extracted image; (d) Distance plot; (e) Smoothed distance plot; (f) Star skeleton for close fist



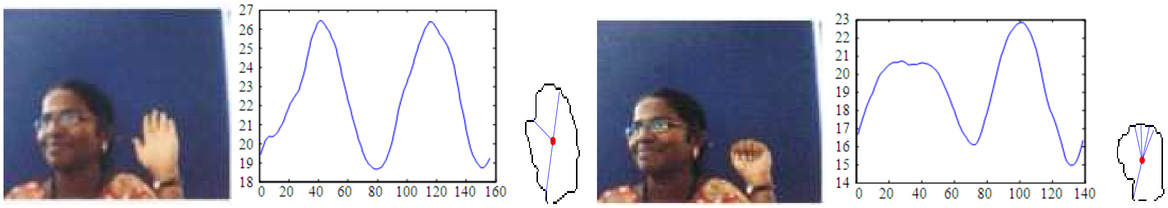
(6a) (6b) (6c) (6d) (6e) (6f)

Fig. 6: (a) Input frame for open fist (b) Smoothed distance plot (c) Star skeleton (d) Input frame for close fist (e) Smoothed distance plot (f) Star skeleton



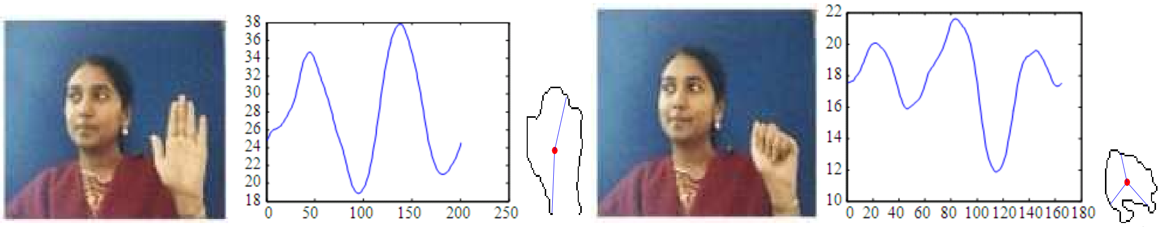
(7a) (7b) (7c) (7d) (7e) (7f)

Fig. 7: (a) Input frame for open fist (b) Smoothed distance plot (c) Star skeleton (d) Input frame for close fist (e) Smoothed distance plot (f) Star skeleton



(8a) (8b) (8c) (8d) (8e) (8f)

Fig. 8: (a) Input frame for open fist (b) Smoothed distance plot (c) Star skeleton (d) Input frame for close fist (e) Smoothed distance plot (f) Star skeleton



(9a) (9b) (9c) (9d) (9e) (9f)

Fig. 9: (a) Input frame for open fist (b) Smoothed distance plot (c) Star skeleton (d) Input frame for close fist (e) Smoothed distance plot (f) Star skeleton

By connecting these extrema points with the centroid, the skeleton of the object is obtained as shown in Fig. 4f and 5f.

The difference between the global maxima and minima of the distance signature is used to recognize the gestures. The proposed algorithm has been tested on various dataset and depicted in Fig. 6-9.

As an alternative effort for comparison, gesture recognition was implemented by extracting Multi-scale Fourier Shape Descriptors, (Direkoglu and Nixon, 2008) at various scales like σ1 = 15, σ2 = 11, σ3 = 8, σ4 = 5, σ5 = 3, σ6 = 1, on segmented hand images as in Fig 10. But this approach needs storage of pre-defined hand gesture templates leading to escalation in memory requirement.

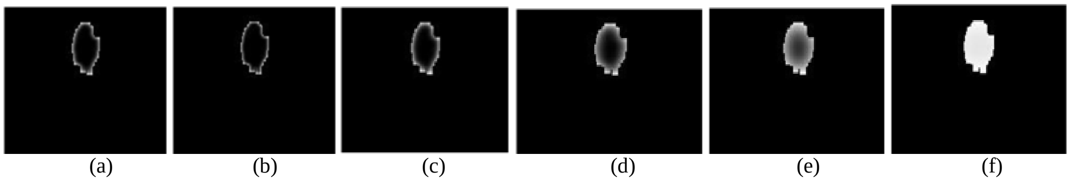


Fig. 10(a-f): Filtered images at different scales σ1 = 15, σ2 = 11, σ3 = 8, σ4 = 5, σ5 = 3 σ6 = 1 for open fist

On comparison with Sánchez-Nielsen *et al*. (2004) work, it is evident that the proposed method possesses scale invariance.

CONCLUSION

A hand gesture based recognition algorithm is proposed to control the PowerPoint application. In the proposed method, foreground is extracted through Gaussian Mixture Model. The extracted object is applied to Star Skeletonization process to detect the extreme points. The experimentation is tested on various dataset which justifies that the proposed solution outperforms the existing methods by being robust to scale variance and does not require any predefined templates for recognition.

REFERENCES

Cucchiara, R., C. Grana, M. Piccardi and A. Prati, 2003. Detecting moving objects, ghosts and shadows in video streams. IEEE Trans. Patt. Anal. Mach. Intell., 25: 1337-1342. DOI: 10.1109/TPAMI.2003.1233909

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B, 39: 1-38. http://www.jstor.org/stable/2984875

Direkoglu, C. and M.S. Nixon, 2008. Image-based multiscale shape description using Gaussian filter. Proceeding of the 6th Indian Conference on Computer Vision, Graphics and Image Processing, Dec. 16-19, IEEE Xplore Press, Bhubaneswar, pp: 673-678. DOI: 10.1109/ICVGIP.2008.40

Jones, M.J. and J.M. Rehg, 1999. Statistical color models with application to skin detection. Int. J. Comput. Vis., 46: 81-96. DOI: 10.1023/A:1013200319198

Lo, B.P.L. and S.A. Velastin, 2001. Automatic congestion detection system for underground platforms. Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, (IMVSP'01), IEEE Xplore Press, Hong Kong, pp: 158-161. DOI: 10.1109/ISIMP.2001.925356

Sánchez-Nielsen, E., L. Antón-Canalis and M. Hernández-Tejera, 2004. Hand gesture recognition for human-machine interaction. J. WSCG., 12: 1-8. DOI: 10.1.1.142.1205

Stauffer, C. and WE.L. Grimson, 1999. Adaptive background mixture models for real time tracking. Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 23-25, IEEE Xplore Press, Fort Collins, Co., USA., pp: 246-252. DOI: 10.1109/CVPR.1999.784637

Wren, C.R., A. Azarhayejani, T. Darrell and A.P. Pentland, 1997. Pfinder: Real-time tracking of the human body. IEEE Trans. Patt. Anal. Mach. Intell., 19: 780-785. DOI: 10.1109/34.598236

Zhang, Z., and C. Chen, J. Sun and K.L. Chan, 2003. Algorithms for Gaussian mixtures with split and merge operation. Patt. Recog., 36: 1973-1983. DOI: 10.1016/S0031-3203(03)00059-1