SPATIAL HAND SEGMENTATION USING SKIN COLOUR AND BACKGROUND SUBTRACTION

Sergio Álvarez, David F. Llorca
Robesafe Research Group, Department of Automatics
Universidad de Alcalá, Madrid, Spain
email: {sergio.alvarez, llorca}@aut.uah.es

Gerard Lacey, Stefan Ameling
School of Computer Science and Statistics
Trinity College Dublin, Ireland
email: {gerard.lacey, stefan.Ameling}@cs.tcd.ie

Abstract

Despite advances in hand detection and hand tracking, robust hand segmentation remains a challenging task in many gesture recognition systems. Problems can be caused by a variety of factors, such as changing illumination and background clutter. We compare the most commonly used visual cues for hand segmentation, namely skin colour and background subtraction, applied both separately and combined. All three approaches are evaluated on video-data recorded with different backgrounds and under varying lighting conditions using a standard evaluation scheme based on overlapping masks. Additionally, we introduce a new evaluation scheme based on global histograms of oriented gradients.

1 Introduction

Hand gesture recognition is a crucial part in many systems for vision-based human computer interaction (e.g. [1]), sign language recognition (e.g. [2]) and other applications such as hand washing quality assessment (e.g. [3]).

To recognize hand gestures in a videostream, a computer vision system must perform both spatial and temporal gesture segmentation. Spatial gesture segmentation is the problem of determining where the gesturing hand(s) are located in each video frame. Temporal gesture segmentation is the problem of determining when the gesture starts and ends [4]. This paper addresses the problem of spatial gesture segmentation.

Hand segmentation remains a challenging task despite advances in hand detection and hand tracking (e.g. [5, 6]. Existing recognition methods often require as input the location of the hands which is unrealistic in most real-world scenarios. Problems for hand segmentation can be caused by a variety of factors, such as changing illumination, low video quality video and background clutter. The two most commonly used visual cues for hand segmentation are skin colour and background subtraction.

Surveys on techniques related to skin colour detection are provided in [7, 8]. A promising approach by Alon, et al. [4], combines skin colour histograms from [9] with motion cues to detect multiple candidate hand regions.

One of the first publications addressing the problem of hand gesture recognition [10] uses background subtraction to extract the moving objects. In [11] background subtraction is used to obtain hand gesture shape.

This paper is organized as follows. In Section 2 we present the methods based on skin colour and background subtraction. The captured video material used for the experiments and the evaluation schemes are described in Section 3. The results are presented and discussed in Section 4 and Section 5 concludes the paper and outlines the future work.

2 Method

In this section, the implemented methods are described. First the background subtraction and shadow detection, next the skin detection and finally a combination of both algorithms to make the most of the advantages of each one.

2.1 Background Subtraction and Shadow Detection

The basic idea of background subtraction is to subtract the current image from a reference image that models the background scene. Obviously the capturing system has to be fixed and the background static. Although the hands are the only objects which are moving in the field of view, the algorithm is susceptible to both global and local illumination changes such as shadows, so a detection and treatment of these problems is needed to achieve satisfying results.

Background Subtraction Rather than explicitly modeling the values of the pixels as one particular kind of distribution, like average, mean, etc., each pixel is modeled by a mixture of K Gaussian distributions [12], whose mean and variance is adapted over time. The probability that a certain pixel has a value X_t at time t can be written as:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \qquad (1)$$

where the mean $\mu_{i,t}$, the covariance $\Sigma_{i,t}$ and the weight $\omega_{i,t}$ (with $0 < \omega_{i,t} \leq 1$), are the parameters of the k^{th} gaussian component, and η is the gaussian probability density function:

$$\eta(X_t, \mu, \sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_t)^T \Sigma^{-1} (X_t-\mu_t)} \qquad (2)$$

For computational reasons the covariance matrices are isotropic so it can be expressed as:

$$\Sigma_{i,t} = \sigma_{i,t}^2 I \tag{3}$$

Given a new data sample X_t at time t, the recursive equations to update the model are [13]:

$$\omega_i = \omega_i + \alpha(\theta_i - \omega_i) \tag{4}$$

$$\mu_i = \mu_i + \theta_i(\frac{\alpha}{\mu_i})\delta_i \tag{5}$$

$$\sigma_i^2 = \sigma_i^2 + \theta_i(\frac{\alpha}{\mu_i})(\delta_i^T\delta_i - \sigma_i^2) \tag{6}$$

where α is the learning rate and $\delta_i = X_t - \mu_i$. For a new sample the ownership θ_i is set to 1 if the sample matches with a component of the mixture (sorted by the value of $\frac{\omega}{\sigma}$) and 0 for the remaining models. The matching is defined by the Mahalanobis distance between the sample and the gaussian component of the mixture and a threshold. If there is no matching, a new component is generated with $\omega_{i+1} = \alpha$, $\mu_{i+1} = X_t$ and $\sigma_{i+1} = \sigma_0$, where σ_0 is a predefined initial variance. If the maximum number of components has been reached, the component with the smallest weight is discarded. Figure 1 shows the result of this step.

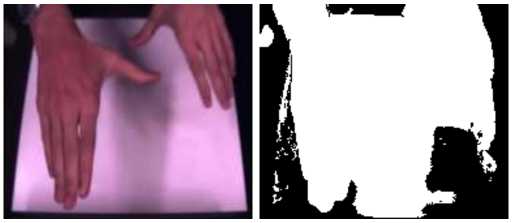


Figure 1: Original image (left) and background subtraction result (right).

Shadow Detection As can be seen in Figure 1 (right), the background subtraction step detects all the moving objects that do not belong to any component of the mixture, so cast shadows are also segmented as foreground. To remove these shadow pixels, the colour model proposed in [14] is used. It separates brightness from the chromaticity component using brightness distortion (b_t) and chromaticity distortion (CD_t). The method is based on the idea that a shadow decreases the brightness of the pixel but keeps the chromaticity. The model is defined as follows:

$$b_t = \frac{(\frac{X_{R,t}\mu_R}{\sigma_R^2} + \frac{X_{G,t}\mu_G}{\sigma_G^2} + \frac{X_{B,t}\mu_B}{\sigma_B^2})}{[\frac{\mu_R}{\sigma_R}]^2 + [\frac{\mu_G}{\sigma_G}]^2 + [\frac{\mu_B}{\sigma_B}]^2} \tag{7}$$

$$CD_t = \sqrt{[\frac{X_{R,t} - b_t\mu_R}{\sigma_R}]^2 + [\frac{X_{G,t} - b_t\mu_G}{\sigma_G}]^2 + [\frac{X_{B,t} - b_t\mu_B}{\sigma_B}]^2} \tag{8}$$

Where X is the current pixel of the image, μ is the mean intensity of the pixel in the background model (highest weighted component) and σ is the variance of the gaussian component, in the respective colour channels R, G, B. Finally a normalization of that parameters is used:

$$\widehat{b_t} = \frac{b_t - 1}{\sqrt{\frac{\sum_{i=t}^{N+t}(b_i-1)^2}{N}}} \tag{9}$$

$$\widehat{CD_t} = \frac{CD_t}{\sqrt{\frac{\sum_{i=t}^{N+t}(CD_i)^2}{N}}} \tag{10}$$

where N is the number of frames to evaluate the parameters. A pixel is classified to be part of a shadow if it has a small normalized chromaticity distortion and a lower brighness value than the background. More specifically, a pixel is labeled as a cast shadow if the following conditions hold:

$$\widehat{CD_t} < T_{CD} \tag{11}$$

$$b_{min} < \widehat{b_t} < 1 \tag{12}$$

Figure 2 shows the result after applying the shadow detection method and a neighbourhood filtering.

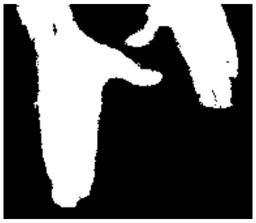


Figure 2: Result of the shadow detection step after background subtraction.

2.2 Skin Detection

The skin detection method mainly comprises three steps: Lighting compensation, the actual histogram-based skin detection and a mask refinement based on morfological operations.

Adaptive Light Compensation To deal with varying ambient lighting conditions, a lighting compensation is performed to achieve robust measurement of skin colour in the subsequent step. The grey-world approach presented in [15] is used, which is based on the assumption that the spatial average of surface reflectance in a scene is achromatic. Since the light reflected from an achromatic surface is changed equally at all wavelengths, it follows that the spatial average of the light leaving the scene will be of the colour of the incident illumination.

The grey-world algorithm is implemented using a scale factor s_i, $i \in \{R, G, B\}$ for each colour component of each pixel such that:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix}_{\text{new}} = \begin{bmatrix} s_R \\ s_G \\ s_B \end{bmatrix} \otimes \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{\text{old}} \tag{13}$$

where $\otimes$ means element wise multiplication. The scale factors are defined as:

$$s_R = \frac{C_{\text{std}}}{R_{\text{avg}}}, s_G = \frac{C_{\text{std}}}{G_{\text{avg}}}, s_B = \frac{C_{\text{std}}}{B_{\text{avg}}} \tag{14}$$

with

$$R_{\text{avg}} = \frac{\sum_i^m R_i}{n}, G_{\text{avg}} = \frac{\sum_i^m G_i}{n}, B_{\text{avg}} = \frac{\sum_i^m B_i}{n} \tag{15}$$

and

$$C_{\text{std}} = \frac{\sum_i^m (\max(R_i, G_i, B_i) + \min(R_i, G_i, B_i))}{2n} \tag{16}$$

Here m stands for the number of pixels in the image and n stands for the number of non-black pixels in the image to avoid over compensation in images with dark background.

Skin Detection The pre-processed frame with the aforementioned compensation is the input for the skin detection module. Skin and non-skin probability densities are estimated through histograms. The Bayes formula is employed to estimate the skin/non-skin probability given the RGB values of a pixel:

$$\begin{aligned} & P(\textsc{skin}|\textsc{rgb}) \\ & = \frac{p(\textsc{rgb}|\textsc{skin})p(\textsc{skin})}{p(\textsc{rgb}|\textsc{skin})p(\textsc{skin}) + p(\textsc{rgb}|\textsc{non-skin})p(\textsc{non-skin})} \end{aligned} \tag{17}$$

The probability densities in the right hand side of (17) is estimated using a training set. The skin detection results in a skin probability mask $M_{prob}(i,j)$ where each element represents the probability of the respecting pixel to be skin.

A binary skin mask $M_{bin}(i,j)$ is produced by applying a carefully selected threshold τ on the probability mask. A pixel (i,j) is classified as a skin pixel if $M_{prob}(i,j) > \tau$. Otherwise it is considered to be a non-skin pixel. The value of τ is experimentally determined to be 0.2.

Skin Mask Refinement The skin mask provided by the aforementioned skin detection step may contain errors originating from several sources. Therefore, to refine the skin mask, the steps followed are:

1. Dilation: To remove holes within the skin mask, a morphological dilation operator is applied on the binary skin mask.
2. Connected component analysis: To identify arms and hands and to get rid of small regions which were mistakenly classified as skin region, a connected component analysis, based on 8-neighborhood, is performed. The connected regions with area smaller than a certain threshold is not considered for further analysis.
3. Erosion: Small anomalies at the boundaries of the mask are removed by the morphological erosion operator.

In both erosion and dilation a 3×3 rectangular structuring element is used. After the refinement procedure the final skin mask is shown in Figure 3.

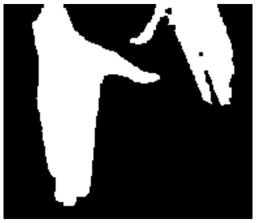
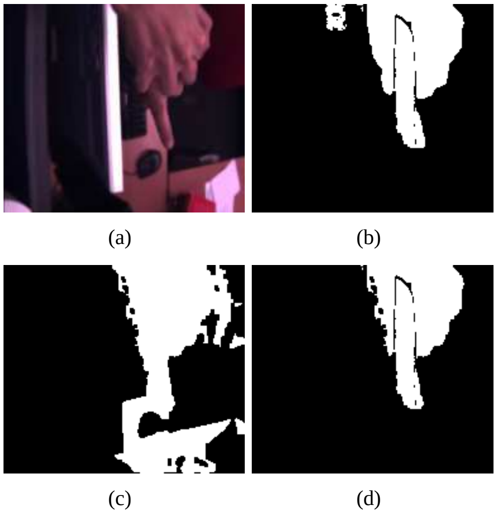


Figure 3: Result of the skin detection step.

2.3 Combined method

A combination of both methods was also implemented. Thus, hand movements are detected in the same way by background subtraction and shadow detection, and only the pixels with high probability to be skin are kept. For this purpose, the general threshold used by the skin detector is reduced to a minimum value, to increase the detection range. The results of the combined method can be seen in Figure 4 when the skin detector fails and in Figure 5 when the background subtraction fails.



(a) (b)

(c) (d)

Figure 4: Results of the three methods implemented. a) Original image. b) Background subtraction and shadow detection. c) Skin detection. d) Combined method.

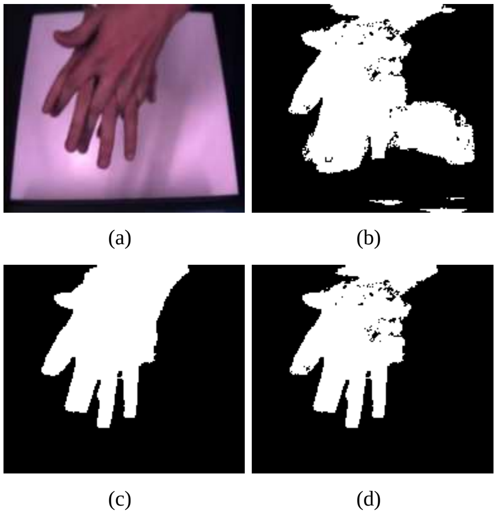


Figure 5: Results of the three methods implemented. a) Original image. b) Background subtraction and shadow detection. c) Skin detection. d) Combined method.

3 Evaluation

To evaluate the performance of the system, 8 videos have been captured under different illumination conditions and different backgrounds. The first 100 frames of each video contain only background and the following frames contain shadows and/or moving hands. More detailed characteristics of the videos are listed below:

- Video 1: White background with strong shadows.
- Video 2: Dark background.
- Video 3: Complex background containing wood.
- Video 4: High illuminated background (floor lights).
- Video 5: Mixed dark and white background.
- Video 6: White background and very dark illumination.
- Video 7: White background with highlights.
- Video 8: Complex background and dark illumination.

A ground-truth dataset for each video was manually created by labeling each frame with the two classes, hand and non-hand. In total, the ground-truth dataset comprises 2113 frames.

All three approaches are evaluated using a standard evaluation scheme based on overlapping masks, obtaining the specificity (18) and sensibility (19), which are defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{18}$$

where TN is the number of true negatives and FP the number of false positives, and

$$Sensivity = \frac{TP}{TP + FN} \tag{19}$$

where TP is the number of true positives and FN the number of false negatives.

Additionally, we introduce a new evaluation scheme based on global histograms of oriented gradients (HOG). While the traditional evaluation based on mask overlap evaluates the result only pixelwise, this evaluation scheme also takes the shape of the segment into account. To extract the HOG, the gradient image is computed and split into cells which are square regions with a predefined size. For each cell, histograms of gradients are computed by accumulating votes into bins for each orientation. These votes are weighted with the magnitude of the gradient vector. The cell histograms are also normalised and the final HOG is constructed by concatenating these cell histograms. Comparing the HOG of the ground-truth and the computed mask from the segmentation methods by using a distance metric allows to evaluate how accurate the segmentation is computed. We implemented the following distance metrics: Euclidean distance, χ^2-distance, Bhattacharyya distance and correlation – however, we only present results for the Euclidean distance and correlation, as the other distances did not show any remarkable differences to the Euclidean distance.

4 Results

In this section, the result of the three methods is presented.

In Figure 6 the specificity of each method for each video is shown. The value is near 1 in all cases, which means almost no negatives are incorrectly tagged as positive in the three methods. The sensitivity (Figure 7) shows how well the algorithms recognize the positive samples. In this case skin detection obtains a value lower than 0.8 in many videos, especially under dark or bright illumination conditions (videos 2, 4, 6 and 8) and complex backgrounds (videos 3 and 5). The other methods have a good performance except for video 6 due to the very dark illumination.

Figures 8 and 9 depict the results of the newly proposed evaluation method. The HOG correlation verifies the conclusion of the sensitivity, that skin detection is highly sensitive to illumination changes. Furthermore, the Eucludian distance shows that the conditions of the videos 3, 5 and 6 (corresponding to complex backgrounds and very dark illumination) are very challenging for skin detection, while the results for background subtraction are stable.

As a conclusion of the comparison of the three methods, figures 10 and 11 finally reveal that skin detection is not enough to cover all the possible cases of illumination conditions or backgrounds. This is shown by the low values in sensitivity and HOG correlation, and the high vector distance.

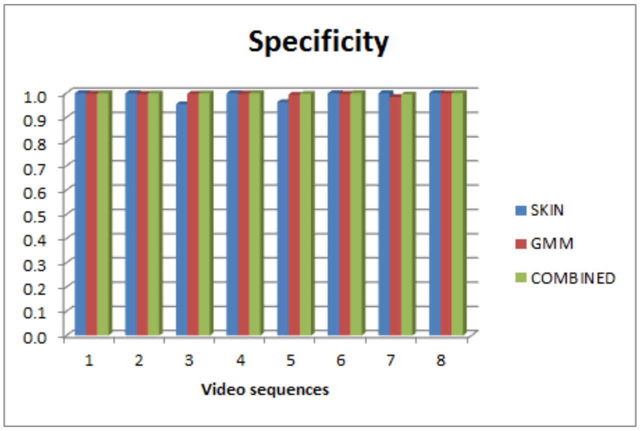


Figure 6: Specificity result for the 8 videos (higher is better).

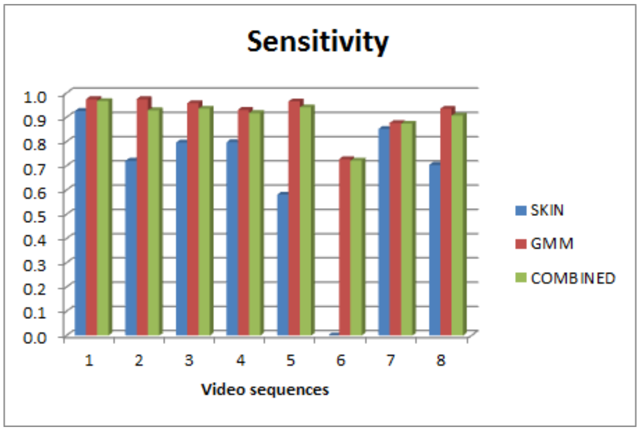


Figure 7: Sensitivity result for the 8 videos (higher is better).

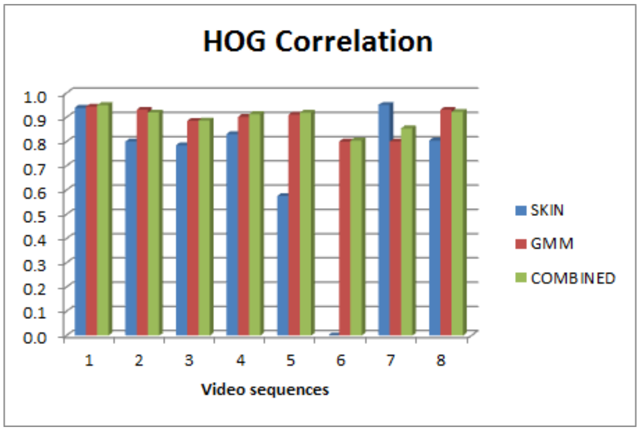


Figure 8: HOG Correlation result for the 8 videos (higher is better).

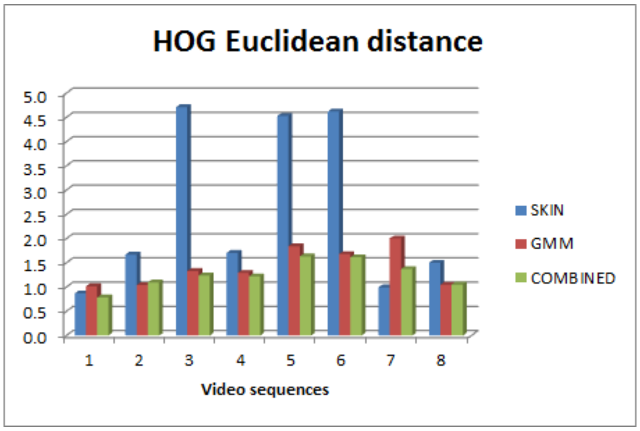


Figure 9: HOG Euclidean distance result for the 8 videos (lower is better).

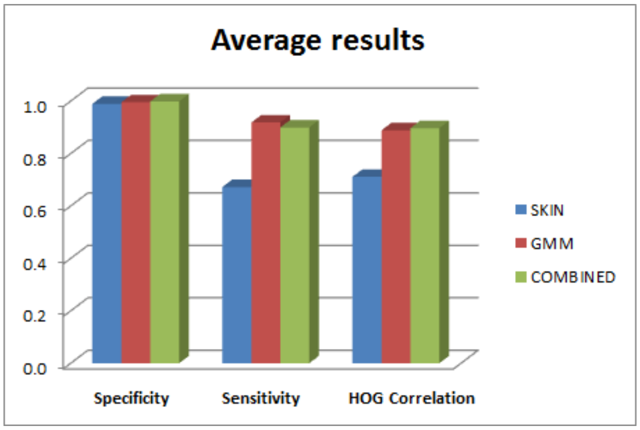


Figure 10: Average result of the evaluation methods for the 8 videos (higher is better).

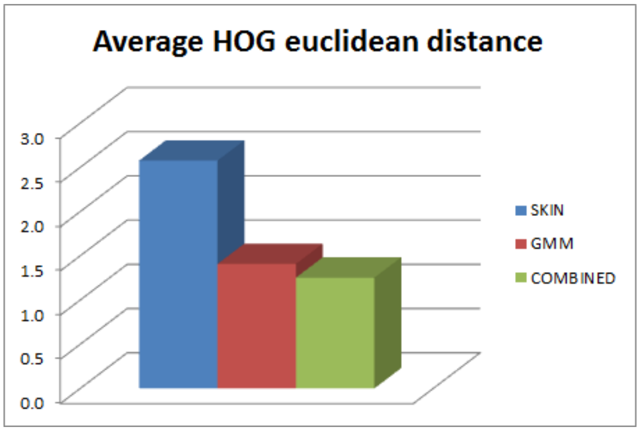


Figure 11: Average result of the HOG Euclidean distance for the 8 videos (lower is better).

5 Conclusion

In this paper, a segmentation method is presented for detecting hands from a static background scene. The method is shown to be accurate, robust, reliable and efficiently computed, under different illumination conditions, shadows, and complex backgrounds. A comparison of the most commonly used visual cues for hand segmentation has been done, namely skin colour and background subtraction, applied both separately and combined; using a standard evaluation scheme based on overlapping masks. Additionally, we introduce a new evaluation scheme based on global histograms of oriented gradients.

Although in general the results of the background subtraction method and the combined one do not show huge differences, it is assumed that in some cases the combined method will profit from the advantages of both methods. Thus, for future work, a large database containing more different backgrounds as well as moving objects (e.g arms or other parts of the body) will be used for further evaluation.

References

[1] V. I. Pavlovic, R. Sharma, and T. S. Huang, “Visual interpretation of hand gestures for human-computer interaction: A review,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, 1997.

[2] T. Starner, J. Weaver, and A. Pentland, “Real-time american sign language recognition using desk and wearable computer based video,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 12, pp. 1371 –1375, Dec. 1998.

[3] S. Ameling, J. Li, J. Zhou, A. Ghosh, G. Lacey, E. Creamer, and H. Humphreys, “A vision-based system for handwashing quality assessment with real-time feedback,” *The Eighth IASTED International Conference on Biomedical Engineering, Biomed 2011, Innsbruck, Austria*, 2011.

[4] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, “A unified framework for gesture recognition and spatiotemporal gesture segmentation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1685–1699, 2009.

[5] Y. Cui and J. Weng, “Appearance-based hand sign recognition from intensity image sequences,” *Computer Vision and Image Understanding*, vol. 78, no. 2, pp. 157 – 176, 2000.

[6] N. Stefanov, A. Galata, and R. Hubbold, “Real-time hand tracking with variable-length markov models of behaviour,” in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 2005, pp. 73 –73.

[7] V. Vezhnevets, V. Sazonov, and A. Andreeva, “A survey on pixel-based skin color detection techniques,” *Proceedings of Graphicon*, vol. 85, pp. 85–92, 2003.

[8] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A survey of skin-color modeling and detection methods,” *Pattern Recogn.*, vol. 40, pp. 1106–1122, March 2007.

[9] M. J. Jones and J. M. Rehg, “Statistical color models with application to skin detection,” *International Journal of Computer Vision*, vol. 46, pp. 81–96, 2002.

[10] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden markov model,” in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR ’92., 1992 IEEE Computer Society Conference on*, Jun. 1992, pp. 379 –385.

[11] F.-S. Chen, C.-M. Fu, and C.-L. Huang, “Hand gesture recognition using a real-time tracking method and hidden markov models,” *Image and Vision Computing*, vol. 21, no. 8, pp. 745 – 758, 2003.

[12] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking,” in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, 1999, p. 252 Vol. 2.

[13] Z. Zivkovic and F. Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[14] T. Horprasert, D. Harwood, and L. D. Davis, “A statistical approach for real-time robust background subtraction and shadow detection,” *Proceeding of IEEE International Conference on Computer Vision*, 1999.

[15] L. Chen and C. Grecos, “A fast skin region detector for colour images,” *IEE Conference Publications*, vol. 2005, no. CP509, pp. 195–201, 2005.