Towards Communicative Face Occlusions: Machine Detection of Hand-over-Face Gestures

Marwa Mahmoud[1], Rana El-Kaliouby[2], and Amr Goneid[1]

[1] The American University in Cairo, Cairo, Egypt
[2] Massachusetts Institute of Technology, USA

Abstract. Emotional body language constitutes an important channel of nonverbal information. Of this large set, hand-over-face gestures are treated as noise because they occlude facial expressions. In this paper, we propose an alternative facial processing framework where face occlusions instead of being removed, are detected, localized and eventually classified into communicative gestures. We present a video corpus of hand-over-face gestures and describe a multi-stage methodology for detecting and localizing these gestures. For pre-processing, we show that force fields form a better representation of images compared to edge detectors. For feature extraction, detection and localization, we show that Local Binary Patterns outperform Gabor filters in accuracy and speed. Our methodology yields an average detection rate of 97%, is robust to changes in facial expressions, hand shapes, and limited head motion, and preliminary testing with spontaneous videos suggests that it may generalize successfully to naturally evoked videos.

1 Introduction

Nonverbal communication plays a central role in how humans communicate and connect with each other. Ones ability to read nonverbal cues is essential to understanding, analyzing, and predicting the actions and intentions of others. As technology becomes more ubiquitous and ambient, machines will need to sense and respond to natural human behavior. Over the past few years, there has been an increased interest in machine understanding and recognition of people's affective and cognitive states, especially based on facial analysis. One of the main factors that limit the accuracy of facial analysis systems is occlusion.

The face can be occluded by many objects such as a pen or a mug, or by the hand. Hand-over-face occlusions are the most challenging to detect because the hand and face have the same color and texture, and hands can take many different shapes. Many facial analysis systems are based on facial feature point extraction and tracking. The motion of these facial points and the corresponding face-geometry changes are mapped into facial expressions, which in turn can be classified into affective or cognitive mental states e.g., [7]. As the face becomes occluded, facial feature points are lost or erroneously detected, resulting in an incorrect analysis of the person's facial expression. Similarly, in most facial analysis systems, face occlusions are mostly treated as noise.

In this paper, we argue that face occlusions, particularly hand-over-face ones, are *not* noise. To the contrary, these *gestures*—a subset of emotional body language—involve brain mechanisms similar to those used to process facial affect and are as important as

M. Kamel and A. Campilho (Eds.): ICIAR 2009, LNCS 5627, pp. 481–490, 2009.

Gesture	Meaning	Region Occluded	Gesture	Meaning	Region Occluded
hand holding face	boredom	cheeks	hand to cheek gesture	evaluation, interest	cheeks
chin stroking	evaluation, interest	chin	hands touching upper lips	evaluation, interest	lips
scratching head	evaluation, doubt	forehead	ruffling hair	evaluation, interest	cheeks / forehead
rubbing eyes	sleepiness	eyes	scratching in front of ear	doubt, suspicious	ears
rubbing nose	suspicious	nose	hand covering mouth	suspicious	mouth
biting nails	anxiety	mouth	hand over mouth	astonishment	mouth

Fig. 1. Hand-over-face gestures are an important channel of nonverbal communication: (clockwise) thinking, surprise, unsure, fatigue, concentration, sleepiness

the face in nonverbal communication [3]. Fig. 1 lists hand-over-face gestures and their meanings compiled from Ekman and Friesens [4] and Pease and Pease [10] classification of body movements; The position and shape of the hand carry different meanings. For example, rubbing one's eye indicates sleepiness or fatigue. Based on this literature, we propose an alternative face processing framework, where instead of being removed, face occlusions are detected, localized and classified into communicative gestures.

This paper makes three principal contributions: (1) to the best of our knowledge, our hand-over-face detection methodology is the first to apply and compare the performance of Local Binary Patterns (LBPs) and Gabor filters to the detection and localization of occluded areas of the face; (2) by using force field analysis followed by LBPs, we advance hand-over-face detection algorithms to perform in real-time and to be robust to changes in facial expressions, hand shapes and limited head motion; (3) we present the first online video corpus of meaningful hand-over-face gestures, which we have made available to the research community at `"http://web.media.mit.edu/kaliouby/handoverface"`. Our method serves as a first step toward classifying hand-over-face gestures and is well-suited to a system that responds in real-time to the person's affective and cognitive state.

The paper is organized as follows: section 2 surveys related work; section 3 overviews our methodology; sections 4, 5 and 6 present force fields for image pre-processing, compare LBPs and Gabor filters for feature extraction, and describe detection and localization; sections 7 and 8 present experimental results and conclude the paper.

2 Related Work

We have surveyed three research areas: face analysis, hand detection and tracking, and hand-over-face detection. Face analysis area considers the face as the main object of interest. Only a few facial analysis systems recognize facial expressions in the presence of partial face occlusion, either by estimation of lost facial points or by excluding the occluded face area from the classification process e.g., [1, 14].In all these systems, face occlusions are a nuisance and are mostly treated as noise.

Hand detection and tracking literature is very close to our problem domain, especially those that consider hand detection over skin-color backgrounds, the most complex background when detecting the hand. Table 1 compares several examples of hand

Table 1. Comparison between related work on method, dominant object, (A) real-time performance , (B) max. head rotation, (C) robustness to facial expressions, and (D) handling of articulated hands. Key: ✓ : yes, x: no, ?: not shown, -: not applicable.

Method	Dominant object	A	B	C	D	Method	Dominant object	A	B	C	D
Elastic graph matching [15]	Hand	?	-	-	✓	Particle filter and color [5]	Hand	✓	-	-	✓
Eigen-dynamics [17]	Hand	?	-	-	✓	Mean shift [2]	Non-rigid obj.	✓	-	-	-
Probabilistic reasoning [11]	Hand&Face	x	45^o	x	x	Bayesian Filters [13]	Hand	?	-	-	✓
Force field approach [12]	Hand-over-face	x	40^o	x	x						

detection and tracking with respect to method, dominant object in the video, real-time performance, robustness to head rotation, facial expressions and articulated hands. For a general survey of object detection and tracking, the reader is referred to [16].

Some hand tracking approaches use shape-based models while others use color/edge based models to represent the hand. Shape-based approaches, such as elastic graph matching [15], predefine a set of hand shapes that are tracked over time. The dynamics of the hand contours are defined manually or captured with a data glove [17]. Thus, shape-based models are often person-dependent and need to be trained for each new user of the system. Shape-based approaches have been applied extensively to articulated hand shapes, but not over a face that rotates or changes expressions. Color/edge-based approaches [2, 13] are simple but do not work well for hand-over-face detection since both the face and hand have the same color. Other approaches include Sherrah and Gong [11] who use probabilistic reasoning to track body parts but assume the two hands and face are always present in the video and do not handle articulated hand motion. Particle filters are used with color representations [5] for real-time tracking and handle cases where skin-like objects are occluded by the hand.

Unlike hand tracking, where the dominant object in the video is the hand—tracked as it occludes other objects—in our case, the face is the dominant object in the video and the hand is the occluding object: it may occlude the face partially, fully or not at all. Smith *et al.* [12] are the closest to our work because they track the hand as it occludes the face, assuming the hand is initially not present. We build on and extend their approach of using force fields to segment the hand over the face. Force fields are an excellent representation for hand-over-face occlusions as they represent the regional structure of an image, thereby avoiding local pixel-based analysis. We address the following limitations of their work: 1) the computational cost and non real-time performance of their algorithm; 2) the non-handling of facial expression changes as the hand moves over the face, and 3) the limitation on the set of hand gestures considered, namely only fully open hands occluding the face either vertically or horizontally.

3 Methodology

Our multi-stage methodology for handling hand-over-face occlusions detects the hand when it occludes the face in a video sequence and determines its position. Our approach consists of three stages: image pre-processing, feature extraction, and hand detection

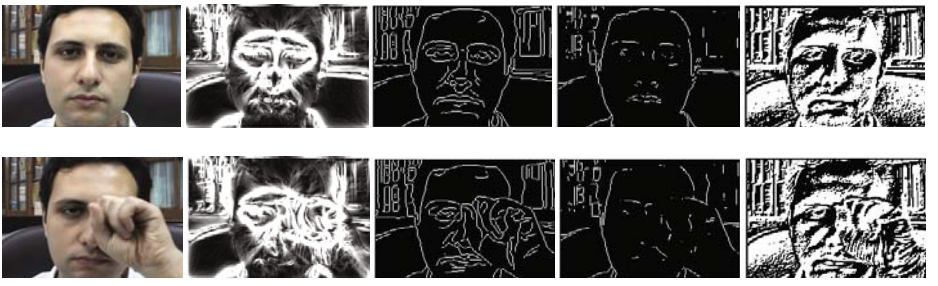

Fig. 2. Force fields encode the regional structure of an image, which is different for the face and hand: (left to right) raw image, force field, canny edge, sobel edge and binary representation

and localization. First, every frame I_t in a video of T frames is transformed into a representation that emphasizes the difference between hand and face. This stage is crucial since the hand and face have similar color and texture. Next, each frame is divided into $s = 9$ regions, assuming the face is centered; for comparison purposes, each region is encoded into a feature vector $H_{(s,t)}$ using LBPs or Gabor filters. Feature vectors at time t are compared to those at the initial frame, where larger differences $\delta H_{(s,t)}$ indicate a regional change in the structure of the image, which may be due to the hand. We describe two parameters for the detection and localization stage: magnitude threshold aK and number of occurrences above this threshold f to determine whether the difference $\delta H_{(s,t)}$ signals a hand. At each frame I_t, the output is a 3x3 matrix M_t which depicts the presence of a hand in each of the nine regions of the frame.

4 Image Pre-processing

The first stage of our methodology is image pre-processing, where the ideal filter would accentuate the difference between frames with a hand occluding the face and those with only a face. While color representations and edge detectors are simple image filters, they are not suitable here because the hand and the face have similar color and texture, resulting in similar representations (Fig. 2). Force fields, on the other hand, capture the regional structure of an image, which is substantially different for a hand and a face, and are therefore well-suited to our problem. By analyzing force field changes over time, we capture structural changes that are introduced as the hand occludes the face.

4.1 Force Field Analysis

Our implementation of force field is based on Hurley *et al.* [6]. Force fields describe the regional structure of an image by representing each pixel as a nonlinear combination of all other pixels in an image. Each pixel exerts a force on every other pixel in the image, directly proportional to the pixel's intensity, and inversely proportional to the square of the distance between the pixels. The force field exerted on a single pixel r in an mxn image is computed as follows:

$$FF(r) = \sum_{i=0}^{mxn} I(r_i) \frac{r_i - r}{|r_i - r|^3} \tag{1}$$

To compute an image's force field, Eq. 1 is repeated for every pixel. This computation is a convolution between the unit force field matrix and the image's intensity matrix. The unit force field matrix represents the force field that all pixels of unit intensity exert on a sample pixel. Note that this matrix is constant for all images of the same size because it depends only on the distance between any two pixels.

4.2 Discontinuity Detection

The resulting force field matrix has complex values, with a real component x and an imaginary component y. For each pixel, the angle $atan(y/x)$ is in the range of [-π,π] and yields the force direction at this location. An image's structure is described by changes or discontinuities in force direction, or *wells*. To extract well positions, the angles matrix is convolved with a Sobel operator. Fig. 3 shows how well positions change as the hand occludes the face in the force field representation.

Fig. 3. (left to right) raw image, angles representing force direction, lighter areas represent discontinuities or well positions. Note the change in well positions as the hand occludes the face.

5 Feature Extraction

The resulting force field image at time t is divided into $s = 9$ regions, and a feature vector $H_{(s,t)}$ is calculated for each region. Gabor filters, and more recently LBPs, have become popular feature descriptors of the face; we introduce them to the problem of hand-over-face detection and compare their performance.

5.1 Local Binary Patterns

LBPs are a simple, yet powerful method for texture analysis and description. The original LBP operator, introduced by Ojala *et al.* [9], is based on a texture unit that is represented by the eight elements in the surrounding 3x3 neighborhood. The eight pixel neighborhood is compared to the value of the center pixel: a pixel takes the value of 1 if it is greater than or equal to the center and 0 otherwise. The resulting binary number (or its decimal equivalent) gets assigned to the central pixel, so that each pixel is represented by a binary number. We then compute a histogram H of the frequency of each binary number. For a 3x3 neighborhood, pixel representations range from 0 to (2^8), so the resulting LBP feature vector size is 256. Assuming the hand is initially not present, the LBP vector $H_{(s,t)}$ for region s at time t is subtracted from the corresponding LBP vector in the initial frame. The difference $\delta H_{(s,t)}$ increases as a hand appears.

5.2 Gabor Filters

Gabor filters are based on a number of filters (which function as scale and orientation edge detectors) that are applied on an image, a force field representation in our case. The basic Gabor filter is a Gaussian function modulated by a complex sinusoid [8]. A bank of Gabor filters is then generated by dilating and rotating the above function for a number of scales n and orientations m. We use Gabor filters with $n = 5$ scales and $m = 4$ orientations. For a pre-processed image I_t, the 20 Gabor filters generate a 3-D matrix $G_{t_{mn}}$. The Gabor feature vector is constructed by getting the mean μ and standard deviation σ of the energy distribution of the transform coefficients. From correlation analysis, we found that μ and σ are strongly correlated. Therefore, we use only the mean μ values; thus, our Gabor feature vector has a length of 20.

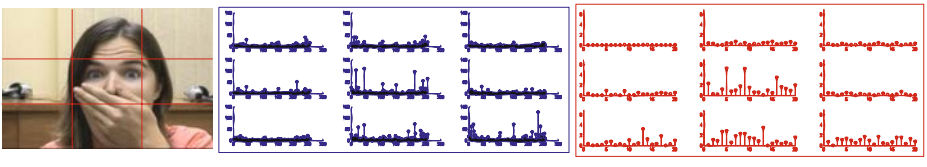

Fig. 4. (left to right) hand-over-face occlusion, LBP differences, Gabor differences. Note the magnitude and frequency of differences in regions 5, 8 and 9, indicating the presence of a hand.

6 Hand Detection and Localization

For each region s, the difference histogram $\delta H_{(s,t)}$ is compared to a magnitude threshold $a * K$ and a frequency threshold f. The two parameters $a >= 1$ and f are needed because facial expressions, head motion as well as hand-over-face occlusions all result in an observed difference in the feature vectors, but the latter results in a greater magnitude of difference. Threshold K is defined as $\delta H_{(s,t)}$ for the first $t = 10$ frames in a video, which have neutral, frontal and non-occluded faces. To select the best combination of parameters a and f, Receiver Operator Characteristic (ROC) curves were generated for sample videos to represent the true positive and false positive rates. We tried 24 parameter combinations: for LBPs, the best combination was $a = 0.5, f = 3$; for Gabor filters, $a = 1.5, f = 4$ worked best. Thus, The algorithm returns a hand when the frequency of differences above the threshold $a * K$ exceeds f. Fig. 4 shows how feature vectors differences increase as the hand occludes the face using LBPs and Gabor filters. The result is a 3x3 matrix M_t representing the nine regions at each frame I_t. The value of each cell is a one if a hand is detected, zero otherwise.

7 Experimental Evaluation

We present a comparative analysis of LBPs and Gabor filters for detection and localization of hand-over-face occlusions. For detection, true positive (TP) is computed as the number of frames where a hand was correctly detected divided the total number of

frames with a hand; false positive (FP) is computed as the number of frames where a hand was falsely identified, divided by the total number of frames without a hand. For localization, TP and FP rates are computed for each frame, for the nine regions, and then averaged for all frames in the video.

7.1 Video Corpus

To test our methodology, we constructed a video corpus of 138 videos of hand-over-face gestures. The videos feature 6 people (3 males and 3 females of different skin colors), were recorded at 30fps at a resolution of 352 x 288 and last about 450 frames. For labeling, each frame in the video is divided into nine regions and labeled with a 1 if a hand is present (i.e., covers more than 25% of a region), and 0 otherwise. To the best of our knowledge, this is the first publicly available video corpus of hand-over-face gestures. The corpus has been made available to the research community at `"http://web.media.mit.edu/kaliouby/handoverface"`. By sharing this corpus, we hope to encourage more researchers to address the problem of hand-over-face detection and provide a benchmark for comparing different approaches.

As shown in Fig. 5, the corpus is organized into five groups: (A) facial expressions {smile, face scrunch, surprise} without hand occlusions or head motion, 18 videos; (B) hand occlusions {hand over mouth, hand rubbing an eye, hand scratching cheeks, palms across the face} over a neutral static face, 24 videos; (C) hand occluding facial expressions {all combinations of A and B}, 42 videos; (D) head motion {pitch, yaw and roll up to 90 degrees} without hand occlusions or facial expressions, 18 videos; (E) hand occlusion with head motion {all combinations of B and D}, 36 videos.

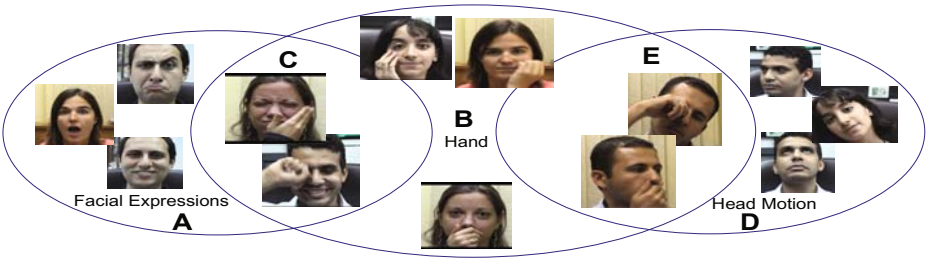

Fig. 5. Organization of video corpus: (A) facial expressions only; (B) hand over neutral, static face; (C) hand over facial expressions; (D) head motion only; (E) hand over moving head

7.2 Speed

Because our ultimate goal is to integrate our hand-over-face detection with systems that respond in real-time to a user's affective and cognitive state, real-time processing time is an important factor. For 256x256 video sampled at 30fps, using Matlab tested on Centrino 2.16GHz, 1Gb RAM, the preprocessing stage takes 0.25 sec/frame; feature extraction, localization and detection take 0.04 sec/frame using LBPs, and 1.02 sec/frame when using Gabor filters. Using LBPs, the overall performance is 0.29 sec/frame or 3.5fps, which achieves our real-time constraint.

7.3 Detection and Localization Results

We ran our system for 84 videos (37900 frames) using LBPs and Gabor filters. Table 2 summarizes the detection and localization results for groups A, B, and C, showing that LBPs outperform Gabor filters. Using LBPs, our methodology achieves real-time performance with average detection rate of: TP 97%, FP 12%, and localization rate of: TP 96%, FP 4%. For group B, average detection rate is: TP 96%, FP 8%, achieving localization rates per gesture as: hand over mouth 95%, eye rubbing 99%, hand scratching cheeks 98%, and palms across the face 91%. Groups A and C test methodology robustness to facial expressions. Group A average detection FP rate is 15%; since group A does not contain hands, detection TP and localization rates do not apply. Group C average detection rate is: TP 98%, FP 14%, with localization rate of: TP 96%, FP 4%; Fig. 6 presents results from group C. Note that even though participants were not asked to move their head for group C, many did so anyways at the onset of a hand gesture (as in frame 145 of Fig. 6), suggesting that head motion accompanies hand-over-face gestures and being able to handle it is crucial for dealing with natural videos.

Table 2. Detection and Localization results for Groups A:facial expressions, B:hand occlusions, C:hand occlusions with expressions changes. TP: True positive rate, FP: False positive rate.

	LBP				Gabor			
Corpus Group	A	B	C	All	A	B	C	All
Detection TP	-	96%	98%	97%	-	98%	98%	98%
Rate FP	15%	8%	14%	12%	50%	30%	47%	43%

	LBP				Gabor			
Corpus Group	A	B	C	All	A	B	C	All
Localization TP	-	95%	96%	96%	-	98%	98%	98%
Rate FP	-	3%	5%	4%	-	5%	9%	8%

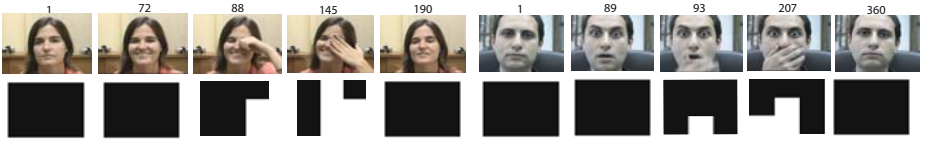

Fig. 6. Robustness to facial expressions(Group C). Bottom row: output 3x3 matrix (white if a hand is detected; black otherwise). Note the slight head motion in Frame 145.

Groups D and E test methodology robustness to head motion; Fig. 7 shows sample frames from group E, depicting the amount of pitch, yaw and roll tolerated by the system. Note the trade off between speed and precise localization: compared to [12], our localization is limited to the nine-region grid, which works well when the hand spans one or more regions, but not if the hand only partially occludes one region. One possibility to achieve more precise localization is to recursively divide occluded regions.

After our methodology using LBPs proved to be robust on a varied but posed corpus, we tested it with the most challenging corpus of spontaneous videos collected from a sipping study where participants were sampling beverages and answering question

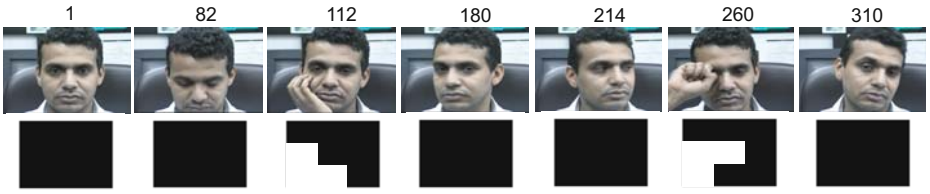

Fig. 7. Robustness to head rotation. Slight head motion (pitch, yaw, and roll) didn't cause false results; meanwhile, the hand is detected successfully. Note different subject skin color.

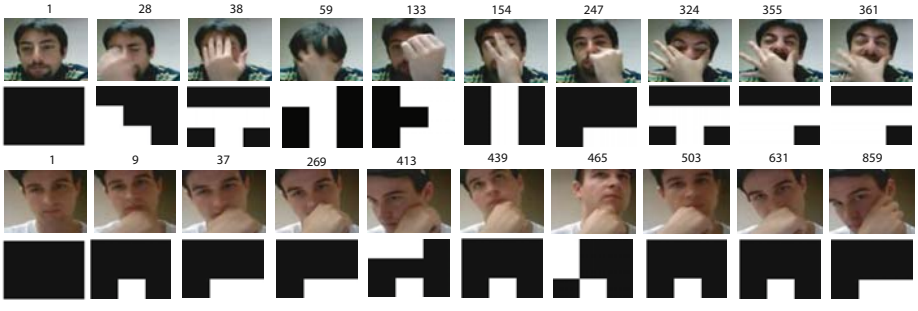

Fig. 8. The hand is detected successfully in the presence of different types of spontaneous facial expressions and head motion

online. Fig. 8 show examples of successful hand detection in the presence of different types of natural facial expressions and head motion.

8 Conclusion

This paper propose a face-analysis framework that emphasizes the meaning of hand-over-face occlusions, describes a multi-stage hand-over-face methodology to detect when and where the hand occludes the face, and present the first online video corpus of meaningful hand-over-face gestures. Our work is the first to apply and compare the performance of LBPs and Gabor filters to the detection and localization of occluded areas of the face. By using force field analysis followed by LBPs, we advance hand-over-face detection algorithms to perform in real-time and to be robust to changes in facial expressions, hand shapes and limited head motion. Preliminary testing with spontaneous videos yields promising results, suggesting that the methodology may generalize successfully to naturally evoked videos. Future work includes extending the methodology to be more robust to head motion by integrating with a head detector, using recursive LBPs for better hand localization, and classifying the gestures into affective or cognitive meaning. Ultimately, our goal is to combine hand-over-face gestures as a novel modality in facial analysis systems along with facial expressions and head gestures.

References

1. Buciu, I., Kotsia, I., Pitas, I.: Facial expression analysis under partial occlusion. Proc. of Acoustics, Speech, and Signal Processing 5, 453–456 (2005)
2. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using Mean Shift. In: Proc. of Computer Vision and Pattern Recognition, vol. 1, pp. 142–149 (2000)
3. de Gelder, B.: Towards the neurobiology of emotional body language. Nature Reviews Neuroscience 7, 242–249 (2006)
4. Ekman, P., Friesen, W.: The repertoire of nonverbal behavior. Categories, origins, usage, and coding. Mouton de Gruyter, Berlin (1969)
5. Fei, H., Reid, I.D.: Joint Bayes Filter: A Hybrid Tracker for Non-rigid Hand Motion Recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 497–508. Springer, Heidelberg (2004)
6. Hurley, D., Nixon, M., Carter, J.: Forcefield energy functionals for image feature extraction. Image and Vision Computing 20, 311–317 (2002)
7. el Kaliouby, R., Robinson, P.: Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In: Real-time vision for human computer interaction, pp. 181–200. Springer, Heidelberg (2005)
8. Ma, W., Manjunath, B.: Texture features and learning similarity. In: Proc. of Computer Vision and Pattern Recognition, pp. 425–430 (1996)
9. Ojala, T., Pietikainen, M., Harwood, D.: Comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29(1), 51–59 (1996)
10. Pease, A., Pease, B.: The definitive book of body language. Orion (2004)
11. Sherrah, J., Gong, S.: Resolving visual uncertainty and occlusion through probabilistic reasoning. In: Proc. of British Machine Vision Conference (BMVC), vol. 1, pp. 252–261 (2000)
12. Smith, P., da Vitoria Lobo, N., Shah, M.: Resolving hand over face occlusion. Image and Vision Computing 25(9), 1432–1448 (2007)
13. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(9), 1372–1384 (2006)
14. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(10), 1683–1699 (2007)
15. Triesch, J., von der Malsburg, C.: Classification of hand postures against complex backgrounds using elastic graph matching. Image and Vision Computing 20(13-14), 937–943 (2002)
16. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM computing surveys 38(4), 1–45 (2006)
17. Zhou, H., Huang, T.S.: Tracking articulated hand motion with eigen dynamics analysis. In: Proc. of International Conference on Computer Vision, pp. 1102–1109 (2003)