Machine Vision and Applications
DOI 10.1007/s00138-012-0409-5

ORIGINAL PAPER

Visual tracking of hands, faces and facial features of multiple persons

Haris Baltzakis · Maria Pateraki · Panos Trahanias

Received: 17 November 2010 / Revised: 9 December 2011 / Accepted: 18 January 2012

Abstract This paper presents an integrated approach for tracking hands, faces and specific facial features (eyes, nose, and mouth) in image sequences. For hand and face tracking, we employ a state-of-the-art blob tracker which is specifically trained to track skin-colored regions. In this paper we extend the skin color tracker by proposing an incremental probabilistic classifier, which can be used to maintain and continuously update the belief about the class of each tracked blob, which can be left-hand, right hand or face as well as to associate hand blobs with their corresponding faces. An additional contribution of this paper is related to the employment of a novel method for the detection and tracking of specific facial features within each detected facial blob which consists of an appearance-based detector and a feature-based tracker. The proposed approach is intended to provide input for the analysis of hand gestures and facial expressions that humans utilize while engaged in various conversational states with robots that operate autonomously in public places. It has been integrated into a system which runs in real time on a conventional personal computer which is located on a mobile robot. Experimental results confirm its effectiveness for the specific task at hand.

H. Baltzakis (✉) · M. Pateraki · P. Trahanias
Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), P.O. Box 1385, 711 10 Heraklion, Crete, Greece
e-mail: xmpalt@ics.forth.gr

M. Pateraki
e-mail: pateraki@ics.forth.gr

P. Trahanias
e-mail: trahania@ics.forth.gr

P. Trahanias
Department of Computer Science, University of Crete, 714 09 Heraklion, Crete, Greece

Keywords Blob tracking · Hand and face tracking · Facial features · Incremental classification · Template matching

1 Introduction

Hands and faces play an important role for human communication. They are the main source of information to discriminate and identify people, to interpret communicative signs as hand and face gestures and to understand emotions and intentions based on facial expressions.

Applications involving human-robot interfaces with advanced interaction capabilities have started to receive considerable attention in the academic community in industrial laboratories and in the media. Some of the greatest scientific challenges towards such applications are related to the development of appropriate technologies and techniques for robots to perceive humans and track their activity. Tracking of hands movement provides information for hand-gesture recognition systems, whereas face and facial features encode critical information about facial expression and head movement.

Inline with the above, in this paper, we propose an integrated approach to identify and track human hands, human faces and specific facial features in image sequences. The proposed approach is mainly intended to support natural interaction with autonomously navigating robots that guide visitors in museums and exhibition centers and, more specifically, to provide input for the analysis of hand gestures and facial expressions that humans utilize while engaged in various conversational states with a robot. The operational requirements of such an application challenge existing approaches in that the visual perception system should operate effectively under difficult conditions regarding occlusions, variable illumination, moving cameras, and varying background. The proposed approach combines and integrates

Published online: 04 February 2012

a set of state-of-the-art techniques to solve three different but closely related problems: (a) identification and tracking of human hands and human faces which are detected as skin-colored blobs, (b) robust classification of the identified tracks to faces and hands, and, finally, (c) identification and tracking of specific facial features (eyes, nose and mouth) within each recognized facial blob.

For the first of the above-defined problems (identification and tracking of human hands and faces) a variety of approaches have been reported in the literature [1,2]. Several of them rely on the detection of skin-colored areas [3–6]. 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 blobs that correspond to human hands and/or human faces.

In contrast to blob tracking approaches, model-based ones [7–9] do not track objects on the image plane but, rather, in a hidden model-space. This is commonly facilitated by means of sequential Bayesian filters such as Kalman or particle filters. The state of each object is assumed to be an unobserved Markov process which evolves according to specific dynamics and which generates measurement predictions that can be evaluated by comparing them with the actual image measurements. Model-based approaches are computationally more expensive and often require the adoption of additional constraints for the dynamics of the system and for the plausibility of each pose but they inherently provide richer information regarding the actual pose of the tracked human as well as the correspondence of specific body parts with the observed image.

In this work, we employ and extend a blob-tracking approach which is based on our previous work [10]. According to this approach, foreground, skin-colored pixels are identified according to their color and grouped together into skin-colored blobs. Information about the location and shape of each tracked blob is maintained by means of a set of pixel hypotheses which are initially sampled from the observed blobs and are propagated from frame to frame according to linear object dynamics computed by a Kalman filter. The distribution of the propagated pixel hypotheses provides a representation for the uncertainty in both the position and the shape of the tracked object.

One of the most important contribution of this paper is related to the development of an incremental classifier which extends the above-described blob tracking approach and which is used to maintain and continuously update a belief about whether a tracked hypothesis corresponds to a facial region, a left hand or a right hand. For this purpose, we use a simple, yet robust feature set which conveys information about the shape of each tracked blob, its motion characteristics and its relative location with respect to other blobs. The class of each track is determined by incrementally improving a belief state based on the previous belief state and the likelihood of the currently observed feature set.

In the field of facial feature detection and tracking a number of approaches have already been presented in the existing literature [1]. Still, complexities arising from inter-personal variation (i.e. gender, race), intra-personal changes (i.e. pose, expression) and inconsistency of acquisition conditions render the task difficult and challenging. Related methods can be categorized on the basis of their inherent techniques. Color-based approaches were exploited in earlier systems by analyzing prior knowledge of color properties of facial features [11,12] and use them to verify that a candidate blob is a face, as in [13]. Although this category of approaches is sensitive to illuminations and head pose changes, it still gains attention in the literature [14], as it succeeds fast detection. Shape- or model-based approaches represent salient facial features via a model and its parameters are optimized to fit to the observations. Earlier examples included deformable templates [15], graph matching [16], active contours [17], Hough transformation [18] and Active appearance models (AAM) [19]. Later many derivatives based on AAM have been proposed [20–22] and although they may lead to accurate feature detection results, they may also converge to incorrect local minima due to improper initializations and feature variances and with a cost in time. Approaches based on machine learning techniques, like Principal Components Analysis [23], Neural Networks [24] and Adaboost Classifiers [25] are relative robust in illumination differences, but require a large number of images for training and are computationally less efficient in the case of high resolution video sequences.

For detecting and tracking the facial features within the detected facial blobs, we propose an approach which combines the boosted cascade detector of Viola and Jones [26] with a feature-based tracker. Therefore, both the advantages of appearance-based methods in detection (i.e. robustness in various lighting conditions) and the advantages of feature-based methods in tracking (i.e. computational speed and high accuracy when initial estimation is close to the real solution) are utilized. The resulting combined detector and tracker extends our previous work on facial feature localization [27] in that specific anthropometric constraints are imposed after the initial detection step to enforce the elimination of false positives and provide reliable initial values for tracking.

The purpose of the above-described approach for hand, face and facial features tracking is to support recognition of hand gestures and facial expressions for rich interaction with an autonomous mobile robot. It has been integrated into a system which runs in real time on a conventional personal computer which is located on the mobile robot itself.

Experimental results presented in this paper confirm its effectiveness for this demanding task.

To summarize, the main contribution of this paper regards a novel, unified approach for tracking hands, faces and specific facial features (eyes, nose, and mouth) in image sequences. For hand and face tracking, an existing blob tracking approach has been augmented with a novel incremental classifier which is used to maintain and continuously update a belief about whether a tracked blob corresponds to a facial region, a left hand or a right hand. An additional contribution of this paper is related to the combination of an appearance-based detector and a feature-based tracker to track facial features within the identified facial blobs. The developed facial feature tracker further extends our previous work by introducing anthropometric validation criteria at the track initiation process.

The outline of the paper is as follows. In the next section, we give an overview of our approach. The following two sections (Sects. 3 and 4) focus on the proposed approach for tracking skin-colored regions and classifying them as hands and faces. Our approach for tracking facial features within facial blobs is described in Sect. 5. Section 6 provides experimental results in laboratory and in real environments. The paper concludes with a brief overview in Sect. 7.

2 Methodology

A block diagram of the components that comprise the proposed approach is depicted in Fig. 1.

The first block in Fig. 1 is the hand and face tracker. This component is responsible for identifying and tracking hand and face blobs based on their color and on the information of whether they lay in the image foreground or not.

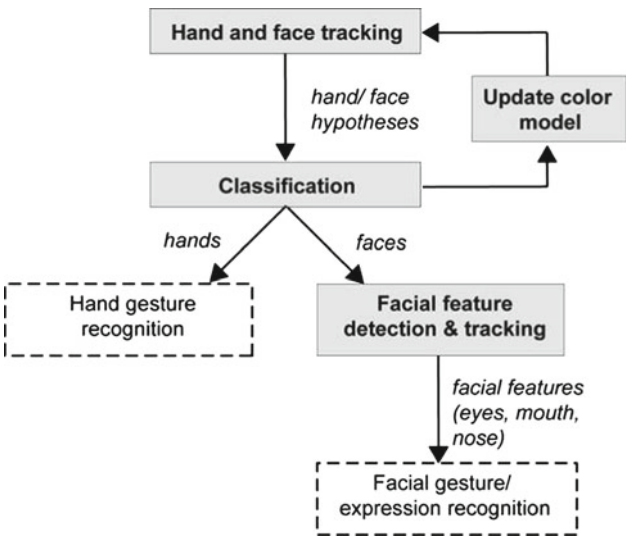

Fig. 1 Block diagram of the proposed system for hands and face tracking

The second step of the proposed system involves the classification of the resulting tracks into tracks that belong to facial blobs and tracks that belong to hands; left and right hands are also classified separately in this step.

Hand trajectories are forwarded to the hand-gesture recognition system (not described in this paper) while facial regions are further analyzed to detect and track-specific facial features (eyes, nose and mouth) and to facilitate facial gestures and expression recognition at a later processing stage of the system (not part of this paper).

Blobs classified as faces are also used to update the color distribution of skin-pixels, thus enabling the algorithm to quickly adapt to illumination changes.

In the following sections, we describe each of the above-mentioned components in detail.

3 Hand and face tracking

In this work, hand and face regions are detected as solid blobs of skin-colored, foreground pixels and they are tracked over time using the propagated pixel hypotheses algorithm[10]. This specific tracking algorithm allows the tracked regions to move in complex trajectories, change their shape, occlude each other in the field of view of the camera and vary in number over time.

3.1 Computation of skin color probabilities

The first step of our approach is to detect skin areas within the input images. For this purpose we use color a technique similar to the one described in [4] and [6]. Initially, the foreground area of the image is extracted by the use of a background subtraction algorithm [28]. Then, foreground pixels are characterized according to their probability to depict human skin and then grouped together into blobs using hysteresis thresholding and connected components labeling.

The result of this first step is the probability $P(s \mid c)$ for each pixel that this pixel belongs to a foreground skin-colored region (s), given its color c. Color c is assumed to be a 2D variable encoding the U and V components of the YUV color space but any other color space which provides a separate luminance component could have been used instead. In any case, the luminance (Y component in our case) is completely eliminated to provide robustness to light intensity changes.

$P(s \mid c)$ can be computed according to the Bayes rule as:

$$P(s \mid c) = \frac{P(s)}{P(c)} P(c \mid s) \tag{1}$$

where $P(s)$ and $P(c)$ are the prior probabilities of foreground skin pixels and foreground pixels having color c, respectively.

$P(c|s)$ is the likelihood of color c for skin-colored foreground regions. All three components in the right side of the above equation are computed off-line during training; $P(s)$ is a scalar variable while $P(c|s)$ and $P(c)$ are stored in 2D histograms.

3.2 From pixels to blobs

After probabilities have been assigned to each image pixel, hysteresis thresholding and connected components labeling are used to extract solid skin color blobs. These computed probabilities are initially thresholded by a "strong" threshold $T_{\max}$ to select all pixels with $P(s \mid c) > T_{\max}$. This yields high-confidence skin-colored pixels that constitute the seeds of potential hand or face blobs. A second thresholding step is applied, this time with a "weak" threshold $T_{\min}$, along with prior knowledge with respect to object connectivity to form the final skin-colored blobs. During this step, pixels with probability $P(s \mid c) > T_{\min}$ where $T_{\min} < T_{\max}$, that are immediate neighbors of skin-colored pixels, 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.

3.3 Tracking skin-colored blobs over time

To track skin-colored blobs over time we employ the pixel hypothesis propagation[10] algorithm. This specific tracking algorithm is able to maintain labeling of the tracked objects (be it hands of facial regions), even in cases of occlusions and shape deformations, without making explicit assumptions about the objects' motion, shapes and dynamics (i.e. how the shape changes over time).

More specifically, a linear model is used to model object trajectories and the uncertainty associated with them. Moreover, the shape of each object and the associated uncertainty are represented by a set of pixels (pixel hypotheses) which are propagated over time using the same linear dynamics as the ones used to model the object's trajectory. That is, for each tracked object, be it a hand or a facial region, the following two types of information are maintained:

- The location and the speed of the object's centroid, in image coordinates. This is encoded by means of a 4D vector $\mathbf{x}(t) = [c_x(t), c_y(t), u_x(t), u_y(t)]^{\mathrm{T}}$, where $c_x(t)$ and $c_y(t)$ are the image coordinates of the object's centroid at time t and $u_x(t)$ and $u_y(t)$ are the horizontal and vertical components of its speed. A Kalman filter is used to maintain a Gaussian estimate $\hat{\mathbf{x}}(t)$ of the above-described state vector and its associated 4×4 covariance matrix $\mathbf{P}(t)$.
- The spatial distribution of the object's pixels. This is encoded by means of a set $\mathbb{H} = \{(x_i, y_i) : i = 1 \ldots N\}$ of N pixel hypotheses that are sampled uniformly from the object's blob and propagated from frame to frame using the dynamics estimated by the Kalman filter.

The representation described above is further explained in Fig. 2. Figure 2a depicts the blob of a hypothetical object (a human hand in this example). Figure 2b–e depicts four possible states of the proposed tracker.

To propagate the state vector forward in time we have assumed a linear motion model for the state ($\mathbf{x}(t+1) = [c_x(t) + u_x(t), c_y(t) + u_y(t), u_x(t), u_y(t)]^{\mathrm{T}}$) and we have also assumed that we can directly observe $c_x(t), c_y(t)$ as the center of the mass of all pixels associated with a particular track. The state transition noise is a 4×4 diagonal matrix which may have particularly large elements in the diagonal to let the system compensate with the fact that the motion of the skin-colored blobs is not usually linear and that the speed greatly varies with time. The exact values can be found experimentally and depend on the camera's angular resolution, the frame rate as well as the distance of the person from the camera.

In contrast to the process noise, observation noise (a 2×2 covariance matrix) is set to contain particularly small values, assuming that in most cases we can very accurately measure the observed blob's centroid. In the case that two or more tracks share the same blob, the observation noise is set to contain larger elements to compensate with the unavoidable association errors.

The distribution of the propagated pixel hypotheses provides the metric used to associate measured evidence to existing object tracks. During the data association step, observed blob pixels are individually processed one-by-one to associate them with existing object tracks. The second step takes place only for blobs that are assigned to more than one existing tracks. In this step, the pixel of each such blob is individually processed and assigned to competing object tracks

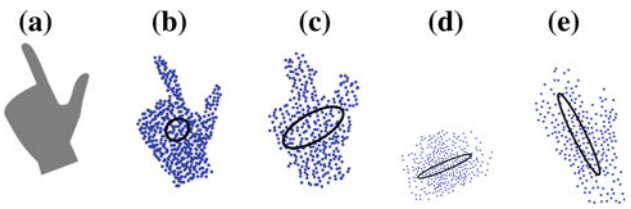

Fig. 2 Object's state representation. **a** Observed blob. **b–e** Examples of possible states. Ellipses represent iso-probability contours for the location of the object (i.e. the first two components of $\mathbf{x_t}$). *Dots* represent the pixel hypotheses

by a factor that is proportional to the density of propagated pixel hypotheses at the location of that specific pixel.

After skin-colored pixels have been associated with existing object tracks, the update phase follows in two steps: (a) the state-vector (centroid's location and speed) is updated using the Kalman filter's measurement-update equations and (b) pixel hypotheses are updated depending on how their track is associated with a blob.

More specifically, pixel hypotheses are updated depending on how tracks are associated with blobs:

- If a track is associated with a single blob and no other track competes for the same blob then the pixel hypotheses of this track are reinitialized by deleting them and sampling them again from their associated blob's pixels. This way we ensure that pixel hypotheses are uniformly distributed to their associated blob's region and we allow the pixel hypotheses to closely follow the blobs shape and size.
- If a track is not associated with any skin blob then we assume that this track is occluded by some other object and we keep propagating this track's pixel hypotheses from frame to frame so that when the hand or face re-appears, we can use the propagated pixel hypotheses associate the old track with the new one.
- If a track competes for the same blob with another track then we assume that the hand or face is occluded by (or it occludes or it gets merged with) another skin-colored object (e.g. another hand or face). In this case, propagated pixel hypotheses are also kept without any updating so that when the blob splits again, they can be used to correctly re-associate the two blobs with the old tracks.

Finally, track management techniques are employed to ensure that new objects are generated for blobs with pixels that are not assigned to any of the existing tracks and that objects which are not supported by observation are eventually removed from further consideration.

Figures 3 and 4 demonstrate the operation of the employed hand and face tracker on a test sequence which involves a man performing hand gestures in an office environment. Figure 3a shows a single frame from this sequence.

Figure 3b, c depicts foreground pixels and skin-colored pixels, respectively. White pixels are pixels with high probability to be foreground/skin-colored pixels and black pixels are non-skin pixels. Finally, Fig. 3d depicts the hand and face hypotheses as tracked by the proposed tracker.

The output of the tracking algorithm in two different frames from the same sequence is demonstrated in Fig. 4. As can be easily observed, this specific tracker succeeds in keeping track of all the three hypotheses, despite the occlusions and the blob merging events introduced at various fragments of the sequence.

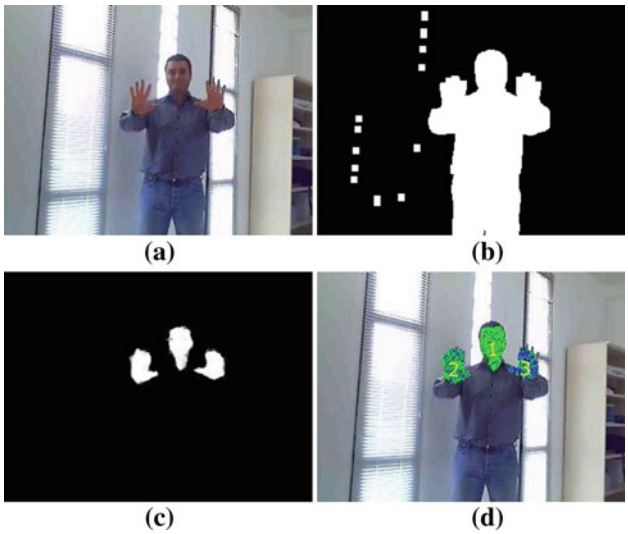

Fig. 3 The tracking approach. **a** Initial image, **b** background subtraction result, **c** pixel probabilities, **d** hand and face hypotheses

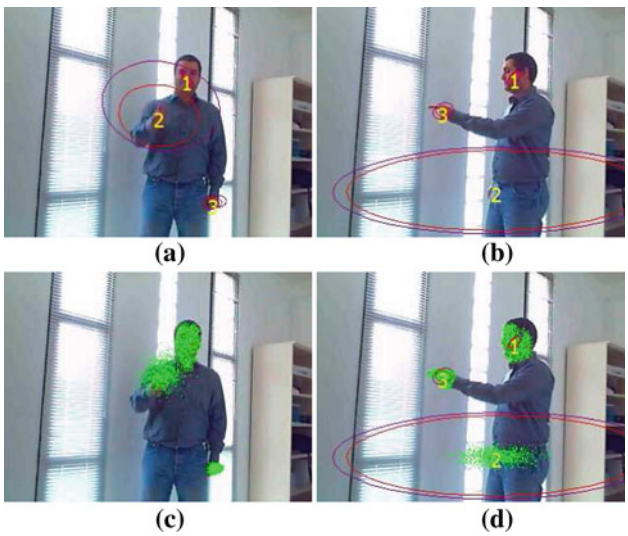

Fig. 4 Tracking hypotheses over time. **a**, **b** Uncertainty ellipses corresponding to predicted hypotheses locations and speed, **c**, **d** propagated pixel hypotheses

3.4 Adapting to illumination changes

When the illumination conditions vary, skin color detection may produce poor results, even if the employed color representation has illumination-independent characteristics (e.g the UV representation used in our case). Hence, a mechanism that adapts the employed representation according to the recent history of detected skin-colored points is required [4].

To solve this problem, skin color detection maintains two sets of prior probabilities: $P(s)$, $P(c)$, $P(c|s)$ corresponding to the off-line training set, and $P_h(s)$, $P_h(c)$, $P_h(c|s)$ corresponding to the evidence that the system gathers during

runtime. Clearly, the second set reflects more faithfully the "recent" appearance of hands and faces and is better adapted to the current illumination conditions. The probability used for skin color detection is given by:

$$P(s \mid c) = \gamma P(s \mid c) + (1 - \gamma) P_h(s \mid c) \quad (2)$$

where $P(s \mid c)$ and $P_h(s \mid c)$ are both given by (1) but involve prior probabilities that have been computed from the whole training set and from online training, respectively. In (2), γ is a sensitivity parameter that controls the influence of the training set in the detection process. We have experimentally set $\gamma = 0.5$, which gave good results in a series of experiments involving gradual variations of illumination.

To obtain $P_h(s)$, $P_h(c)$, and $P_h(c|s)$, we only utilize pixels that belong to tracks which are classified as facial tracks with high confidence (see Sect. 4). This is to completely eliminate false detections that could lead to deterioration of the color distributions.

4 Classifying between hands and faces

The hand and face tracker described in the previous section provides a set of blob tracks that correspond to the location of hands and faces of people that are in front of the robot. To proceed with higher level tasks, like hand gestures and facial expressions recognition, one has to distinguish between tracks that belong to hands and tracks that belong to faces. Moreover, for hand tracks, one has to know which tracks belong to left hands and which tracks belong to right hands.

Towards this goal, we have developed a technique that incrementally classifies a track into one of three classes: faces, left hands and right hands.

The input of the technique is a feature vector O_t which is extracted at each time instant t and is used to update the belief of the robot B_t regarding the class F of each track. The feature vector O_t consists of the following components:

- The periphery-to-area ratio r_t of the current track's blob. The ratio r_t is normalized to the corresponding ratio of a circle and provides a measure of the complexity of the blob's contour. It is expected that hands will generally have more complex contours than faces, i.e. larger values for r_t.
- The vertical and the horizontal components u_t and v_t of the speed of a tracked skin-colored blob. The intuition behind this choice is that hands are generally expected to move faster than faces. Moreover, faces are not expected to have large vertical components in their motion.
- The orientation θ_t of the blob. It is expected that faces will tend to have orientations close to $\pi/2$.
- The location l_t of the blob within the image. This location is relative to the location of each possible head hypothesis and it is normalized according to the radius of this head, as it will be explained later in this section.

We define the belief B_t of the robot at time instant t to be the probability that the track belongs to class f, given all observations O_i up to time instant t. That is:

$$B_t = P(F = f | O_1, \ldots, O_{t-1}, O_t) \quad (3)$$

By assuming the Markov property and the independence assumptions indicated by Fig. 5, the computation of B_t can be simplified as:

$$B_t = \alpha P(O_t | F = f) B_{t-1} \quad (4)$$

The above equation defines an incremental way to compute B_t, i.e. to classify the track by incrementally improving the belief B_t based on the previous belief B_{t-1} and the current observations. α is a normalization factor which ensures that the beliefs B_t for all possible values of F sum up to one.

To compute the term $P(O_t|F = f)$ in the right hand of Eq. (4), we assume the naive Bayes classifier depicted in the graph of Fig. 6, which gives:

$$P(O_t|F) = P(r_t|F) P(u_t|F) P(v_t|F) P(\theta_t|F) P(l_t|F) \quad (5)$$

All the probabilities in the right side of Eq. (5) can be estimated according to training data and encoded and stored in appropriate look-up tables that permit real-time computations.

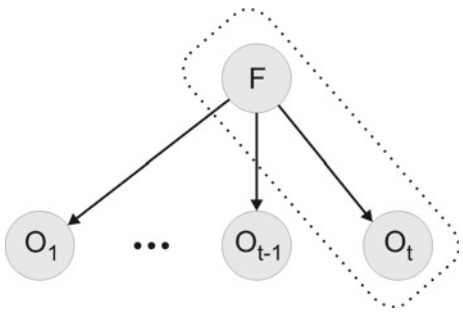

Fig. 5 Bayes graph encoding the independence assumptions of our approach

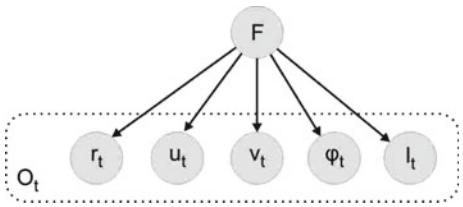

Fig. 6 The naive Bayes classifier used to compute the $P(O_t|F = f)$

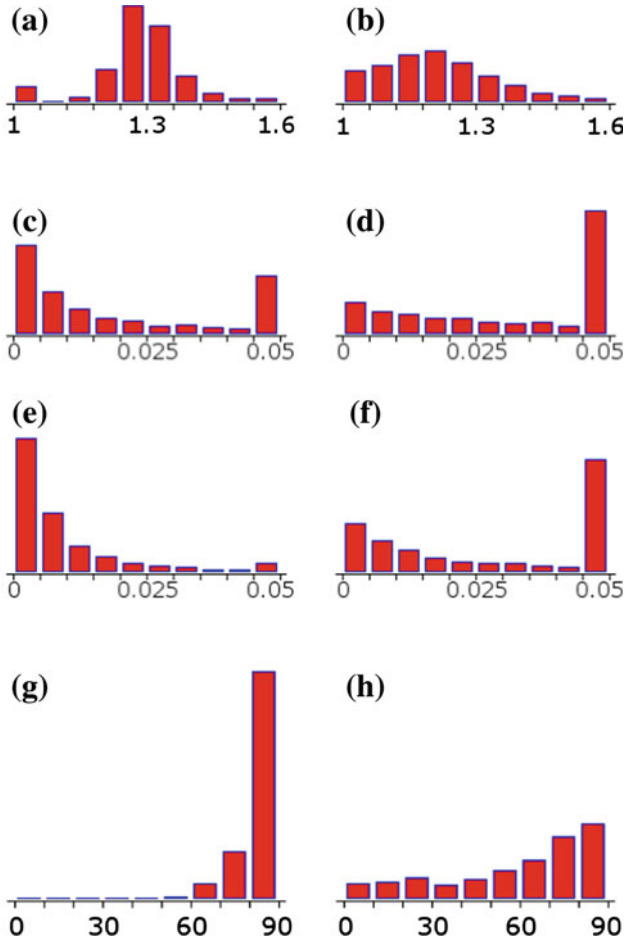

Fig. 7 1D Look-up tables used for the computation of Eq. (5). **a** $P(r_t|F = \text{face})$, **b** $P(r_t|F = \text{hand})$, **c** $P(u_t|F = \text{face})$, **d** $P(u_t|F = \text{hand})$, **e** $P(v_t|F = \text{face})$, **f** $P(v_t|F = \text{hand})$, **g** $P(\theta_t|F = \text{face})$, **h** $P(\theta_t|F = \text{hand})$

The lookup tables for $P(r_t|F)$, $P(u_t|F)$, $P(v_t|F)$ and $P(\theta_t|F)$ are depicted in Fig. 7. They are 1D lookup tables encoding the relevant quantity (r, u, v, or θ) with the probability of appearance of this quantity in the training set. These lookup tables are identical for left hands and right hands but they are different in the case of faces. This is because, the relevant quantities are not expected to vary significantly between left and right hands but, as can be easily observed in Fig. 7, they differ significantly in the case of faces.

$P(l_t|F)$, which is the probability of a blob being observed at location l_t given its class F, is computed and stored differently for faces and differently for hands.

For faces, $P(l_t|\text{face})$ is retrieved as the probability for a facial blob to be centered at this specific image location l_t. Obviously, the 2D lookup table for $P(l_t|\text{face})$ depends on the actual application at hand and involves assumptions about the pose of the camera and the relative location of the human(s) with respect to the camera. In our case, which involves a human-robot interaction application, we assumed a camera placement such that the field of view of the camera includes the upper body part of one or more humans standing at a convenient distance between 0.5 and 2 m in front of the robot. The actual lookup table that we compiled and used in our experiments is depicted in Fig. 8a.

For hands, $P(l_t|\text{left hand})$ and $P(l_t|\text{right hand})$ are computed relatively to the location of the corresponding person's face. Since we do not know which is the corresponding person's face, we marginalize over all possible face hypotheses.

That is, for $P(l_t|\text{right hand})$ we have:

$$P(l_t|\text{right hand}) = \sum_h P(l_t|\text{right hand}, h = \text{face})\, P(h = \text{face}) \tag{6}$$

and similarly for the left hand:

$$P(l_t|\text{left hand}) = \sum_h P(l_t|\text{left hand}, h = \text{face})\, P(h = \text{face}) \tag{7}$$

To compute the probabilities $P(l_t|\text{right hand}, h = \text{face})$ and $P(l_t|\text{left hand}, h = \text{face})$, in a way that is invariant to the location of the user with respect to the camera, we normalize the location of the current blob with respect to the location and the size of the face hypothesis h. This is achieved first by translating both the blob and the face hypothesis h in a way that h moves to the center of the image and second by scaling

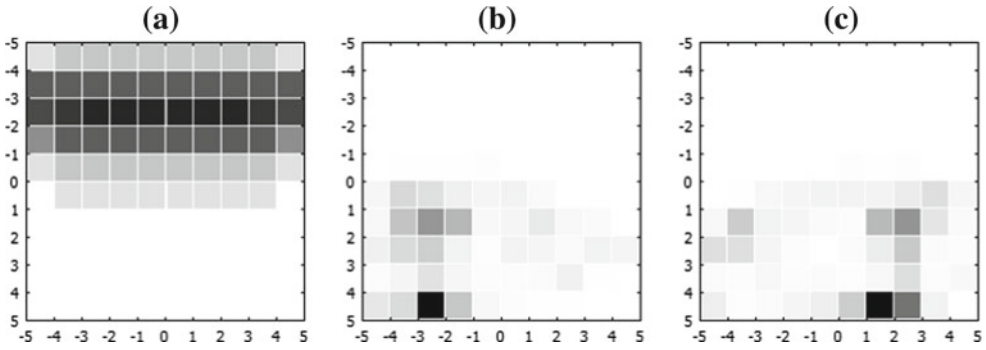

Fig. 8 2D Look-up tables used for the computation of $P(l_t|F)$ in Eq. (5). **a** For faces, **b** for left hands, **c** for right hands

the image (and, thus, moving the blob location) in a way that the area of the face hypothesis h is normalized to the average area of the faces used for training. Figure 8b, c depicts the resulting lookup tables for $P(l_t|\text{right hand}, h = \text{face})$ and $P(l_t|\text{left hand}, h = \text{face})$.

Evidently, the discriminative power of each of the above-described features is very related to the application scenario at hand. That is, different training data should be used for different applications, which is especially true for the speed components u_t and v_t and for the expected image location l_t of hand and facial blobs. In all experiments reported in this paper (with the exception of the database sequences reported in Sect. 6.3), we have trained our classifier assuming a human-robot interaction setup which involves human(s) standing at a distance of approximately 1 m from a camera which is placed at approximately 1.2 m above the ground (i.e. the robots chest).

5 Detection and tracking of facial features

For tracking individual facial features within each detected facial blob, we utilize a hybrid approach by integrating an appearance-based detector and a feature-based tracker for the eyes, the nose and and mouth. The combined approach inherits advantages from both approaches permitting robust identification of the facial features, correct maintenance of feature IDs among frames, as well as real-time computations.

The overview of the implemented approach is illustrated in Fig. 9 and is based on three steps: (a) initial detection of facial features using an appearance-based detector, (b) elimination of false positive detections via the application of anthropometric constraints, and, (c) real time tracking of the detected and filtered facial features using a feature-based method.

For the initial detection of facial features, we use the Boosted Cascade Detector of Viola and Jones [26] and the available implementation in the OpenCV open source library. This particular method combines four key concepts: (a) a set of Haar-like features,(b) an integral image for rapid feature detection, (c) the Adaboost machine-learning method, and, (d) a cascaded classifier used to efficiently combine the features. In our case, for the detection of the features within each face blob, individual sets of Haar-like features for eyes, nose and mouth are utilized and the method is initialized with frontal-view faces.

An important factor which affects both the reliability of detection and the tracking accuracy of facial features is the size of the detected face blob. According to Tian [29], facial features become hard to detect when the face region is smaller than a threshold of approximately 70×90 pixels. Therefore, the procedure of facial feature detection and tracking is only activated when the face blob satisfies the above size requirements.

After all features have been detected, specific anthropometric constraints are applied to cast out false positives. Motivated by the work of Sohail and Bhattacharya [30], we have

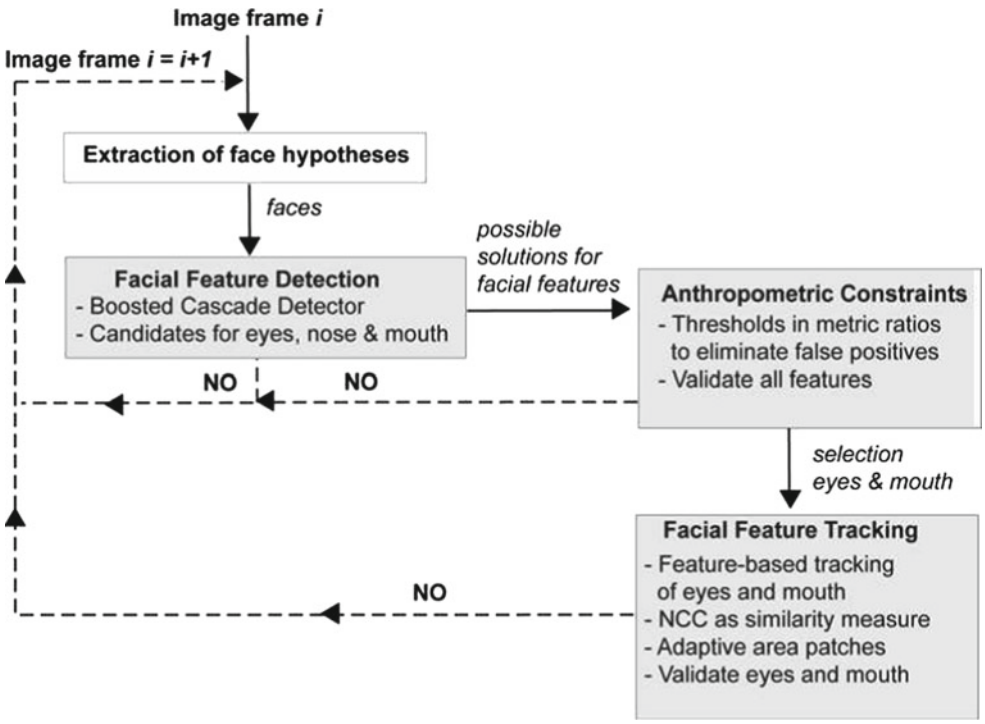

Fig. 9 Diagram of the proposed approach for detection and tracking of facial features

collected a large set of measurements from images depicting faces in frontal view. The collected measurements were used to built an anthropometric model of the human face and to define the necessary thresholds and validation gates used to filter out false positive detections. To define which measurements to use and which not in order to implement our validation criteria, we took into account both the quality of the provided information as well as the detection difficulty due to variations in appearance, high probability of occlusions, etc. The selected validation criteria involve the location and the size of the eyes, the nose and the mouth. Landmarks on other regions such as the eyebrows, used, for example, in [30], were not selected because they often proved to be occluded by hair, eyeglasses or, in some cases, they were entirely non-existent.

More specifically, considering a frontal face view, the eyes, the mouth and the nose regions are expected to be found in the upper half, lower half and in the central part of the face, respectively. Their respective widths as well as certain distances between them should obey certain rules. Furthermore, experiments have shown that the distance D_1 between the centers of the two eyes can serve as a normalization factor for measuring other dimensions on the face. That is, all measured distances and sizes are normalized to the distance D_1.

More specifically, we define the following criteria:

- All four selected features (eyes, nose, mouth) should be detected.
- The normalized sizes of the two eyes and mouth should be within certain bounds.
- The normalized distance between the midpoint of the eye centers and nose tip should be approximately 0.6. That is $D_2/D_1 \simeq 0.6$, where $D2$ is the distance between points P_3 and P_4 (see Fig. 10).
- The normalized distance between the midpoint of eye centers and mouth center should be approximately $\simeq 1.2$. That is $D_3/D_1 \simeq 1.2$, where D_3 is the distance between points (P_3 and P_5).

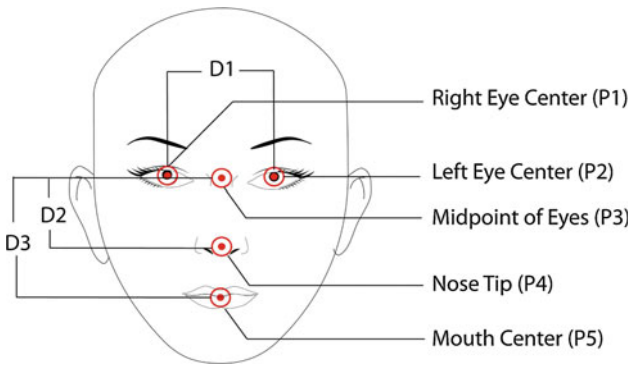

Fig. 10 Landmarks in the anthropometric face model

The above-defined criteria are applied for each facial blob, following the detection of facial features. Blobs that do not pass the above criteria are rejected by the system and a new re-initialization is attempted (by repeating the facial feature detection step) in the next frame.

For blobs that pass the above criteria, the tracking procedure is invoked. It is to be noted that tracking in only performed for the two eyes and for the mouth region. The nose region is not tracked because its actual location is not considered important for our target application, which is expression recognition and visual speech detection.

Tracking is based on template matching, using as eye and mouth templates the detected areas from each face. The normalized cross-correlation (NCC) measure of Eq. 8 is used as matching score/quality measure.

$$\mathrm{NCC} = \frac{(g_\mathrm{t} - \overline{g_t}) \cdot (g_\mathrm{s} - \overline{g_\mathrm{s}})}{\|g_\mathrm{t} - \overline{g_\mathrm{t}}\| \cdot \|g_\mathrm{s} - \overline{g_\mathrm{s}}\|} \tag{8}$$

The vectors g_t and g_s contain the grey levels of the pixels in the template and search image and their mean values are $\overline{g_\mathrm{t}}$ and $\overline{g_\mathrm{s}}$. NCC depicts a similarity measure in the variation interval of $[-1; 1]$.

The selection of NCC as quality measure is justified as only small deviations in the relative positions of the feature areas with respect to the position of the face blob in the image are expected. The search areas for the left and right eye are defined in the upper left and upper right half of the tracked face, whereas for the mouth in the lower half of the face area. The position with the maximum similarity score within each search area and above a certain lower threshold (i.e. 0.75) is selected as the new feature position. The size of templates is updated with a factor in every consecutive frame with the width and the height factor to be computed by the ratios of the template width and height to the respective face width and heights. The matching score is used to block results of low reliability and if it is below a certain threshold, detection is reinvoked. With this approach, there is a significant gain in processing time, allowing for real-time computations. For example for the case of images with size 640×480 pixels and a face of approximate size of 200×200 pixels, detection of each feature area is approximately at 200–400 ms whereas tracking is below 10 ms using a standard computer.

6 Experimental results

The proposed approach for combined tracking of hands, faces and facial features has been implemented on a system which is used to facilitate hand gesture and facial expressions recognition for a mobile robot with rich interaction capabilities.

In all reported experiments, the resolution of the camera was set to 640×480. Although the performance of the system greatly depended on the number of active hypotheses

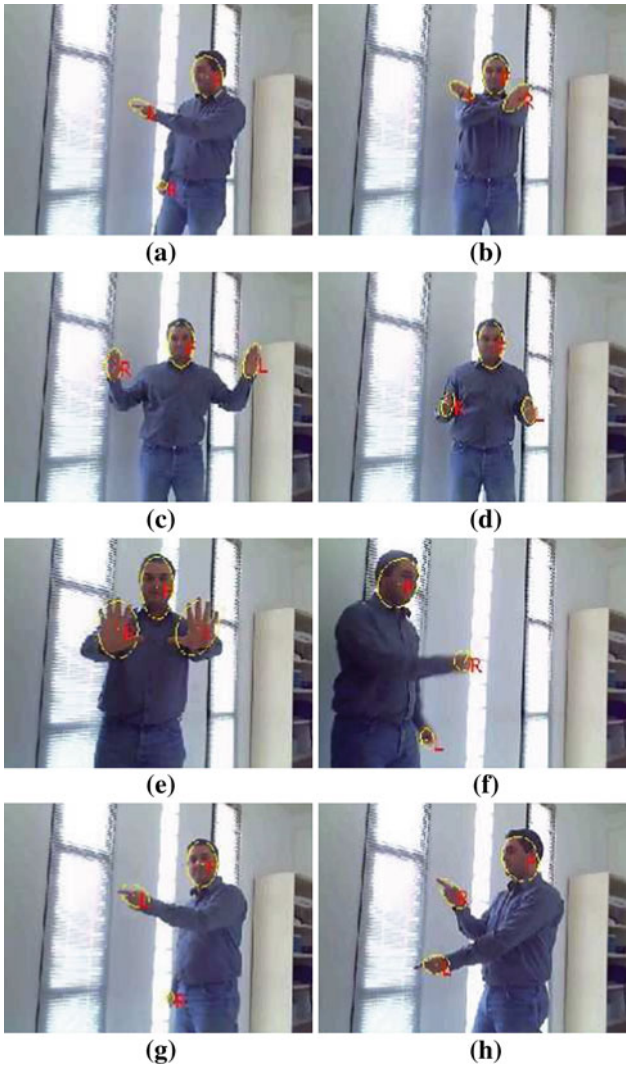

Fig. 11 Eight frames of a sequence depicting a person performing various hand gestures in an office environment

(especially on the number of facial hypotheses) in all cases, the algorithm was able to process the camera's input stream at a rate exceeding 16 frames per second on a standard personal computer. Most notably, in the case of the field trials reported in Sect. 6.2, all processing was done on a 2.8 GHz personal computer computer with 4 Gb of RAM, located on the robotic platform itself.

6.1 Evaluation in realistic environments

Figure 11 presents hand and face classification results for various frames of the office sequence of Fig. 4. Blobs classified as faces are marked with an "F", left hands are marked with an "L", and right hands are marked with an "R". The proposed approach has been successful in classifying the three observed tracks and it also managed to maintain its belief over the whole sequence.

During the whole sequence, which consists of 2,644 frames, five tracks are generated. The first two tracks correspond to the face and the right hand, respectively, which are successfully tracked, maintaining the same ID from their first appearance in the scene to the end of the sequence, despite a number of full and partial occlusions (in four different occasions the right hand's blob completely disappears behind the body of the performer, see for example Fig. 4d). The other three tracks correspond to the left hand. For the left hand the tracking algorithm could not maintain the same ID throughout the sequence because of heavy occlusions while the hand was moving rapidly towards the border of the image. In these cases the algorithm assumed that the left hand left the scene and deleted the corresponding track, causing a new track to be initiated when the hand re-appeared in the scene.

In all five cases, the algorithm succeeds in correctly classifying the track to one of the three classes with a certainty of 0.9 in less than 90 frames. Figure 12 depicts the belief for the first three tracks of the office sequence, as it evolves over the first 500 frames of this sequence. As can be easily observed, the belief of each track is initially uncertain but very soon it stabilizes to the correct class.

Figure 13 depicts additional results from a sequence with multiple persons. In this sequence, there are three performers which occasionally enter and leave the scene. Besides the occlusions which cause frequent disappearances/reappearances of both hand and face blobs there are also hands which get out of the field of view due to the fact that in some cases the performers are standing very close to the camera. Figure 14 depicts results from the same sequence with emphasis on facial feature localization within the provided face blobs.

Despite these difficulties, the tracking algorithm succeeds to correctly classifying all visible tracks given that the duration of the track is long enough (200 frames minimum) to facilitate classification. Facial feature tracking was also successful in that, tracks were created for all eye and mouth regions. Some quantitative tracking and classification results are presented in Table 1.

Figure 15 depicts frames from two additional sequences captured by the robot's camera in two different application environments within an exhibition center. In all our experiments, the algorithm successfully tracked the skin-colored blobs and very fast converged to the correct class for each track (i.e. left hands, right-hands and faces), following convergence curves which were very similar to the ones depicted in Fig. 12. Eyes, nose and mouth regions were also correctly localized and tracked, even in cases of usual off-plane head rotations and different facial expressions.

Figure 16 depicts facial feature tracking results from two additional, close-up, sequences captured at the same exhibition center. The first sequence comprised a total number of 1,100 frames, whereas the second sequence of 650 image frames. In all our experiments, the algorithm successfully

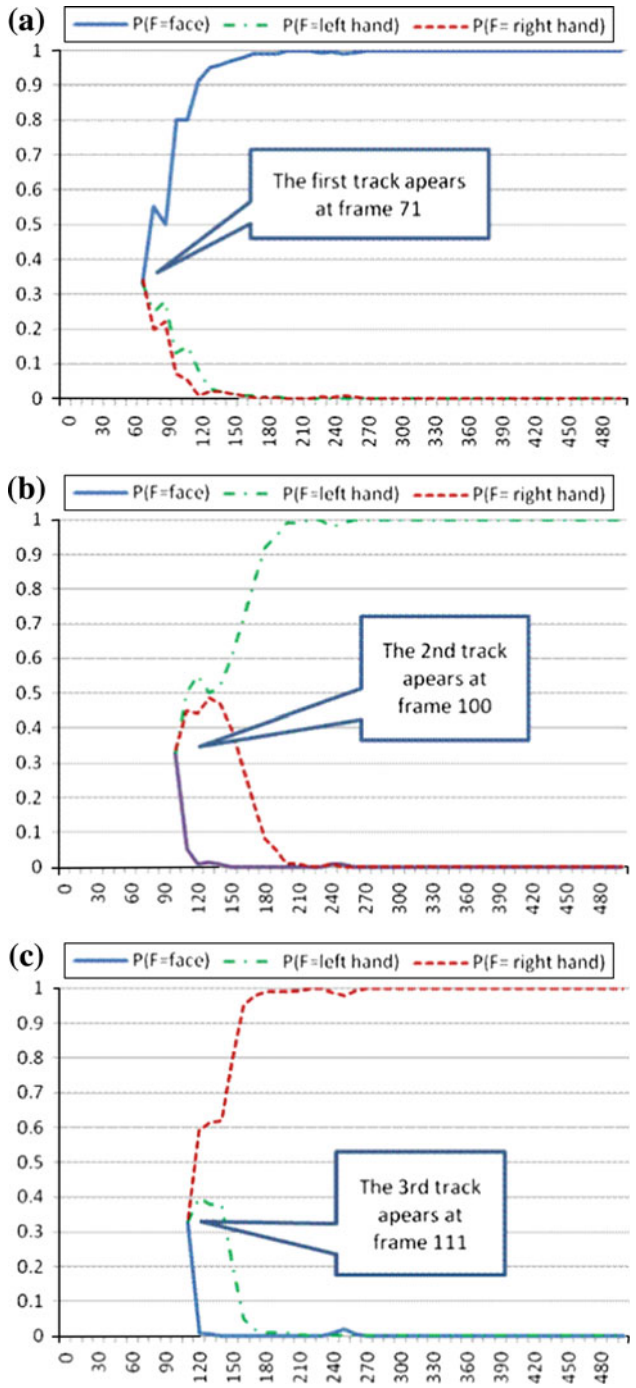

Fig. 12 The belief of each of the three tracks of the office sequence, as it evolves over the first 500 frames. The *solid blue lines* correspond to the probability of each blob being a face blob, the *dot-dashed green lines* correspond to left hands and the *dashed red lines* corresponds to right hands (color figure online)

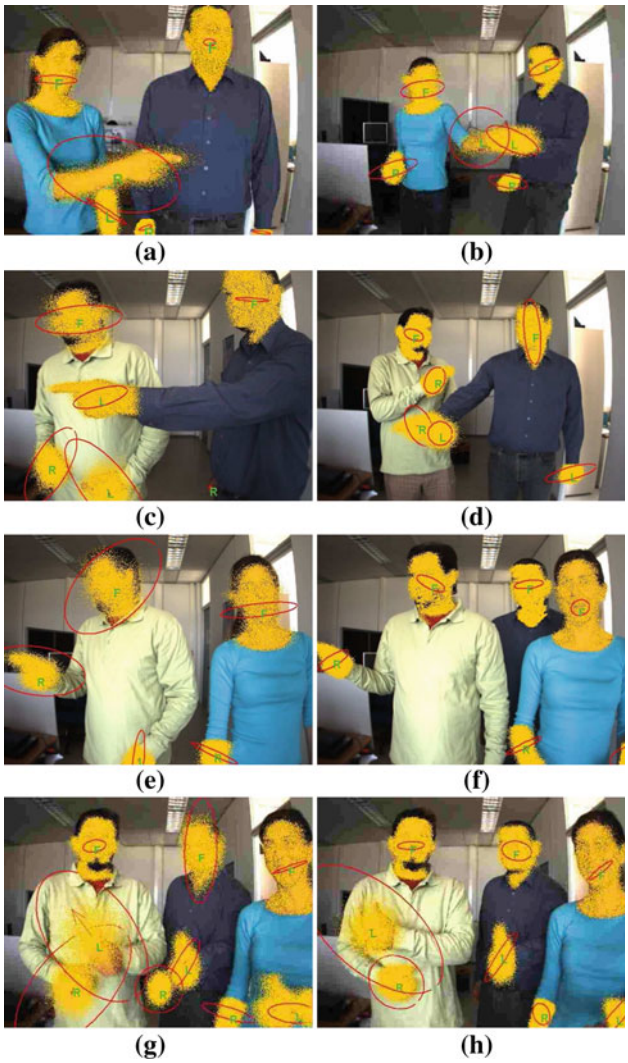

Fig. 13 Eight frames of a sequence depicting multiple persons performing hand and facial gestures in an office environment. Throughout the sequence, performers enter and leave the scene and there are frequent occlusions, which result in a varying number of hand and face hypotheses

tracked the skin-colored blobs corresponding to faces, following convergence curves which were very similar to the ones depicted in Fig. 12. As with the previous figure, facial features were correctly localized and tracked and Table 2 shows some indicative quantitative results, verified by a human supervisor, of the two sequences of Fig. 16. The high true positive (TP) percentages for each facial region indicate successful localization when the respective region is visible. The false negative (FN) percentages are generally low with the exception for the right eye of sequence 2 due to lower signal content in this area.

6.2 Field trials

The proposed system has been evaluated in a real human-robot interaction scenario which involved a tour-guide robot guiding users in an exhibition which consisted of seven screens depicting pictures of ancient temples and monuments.

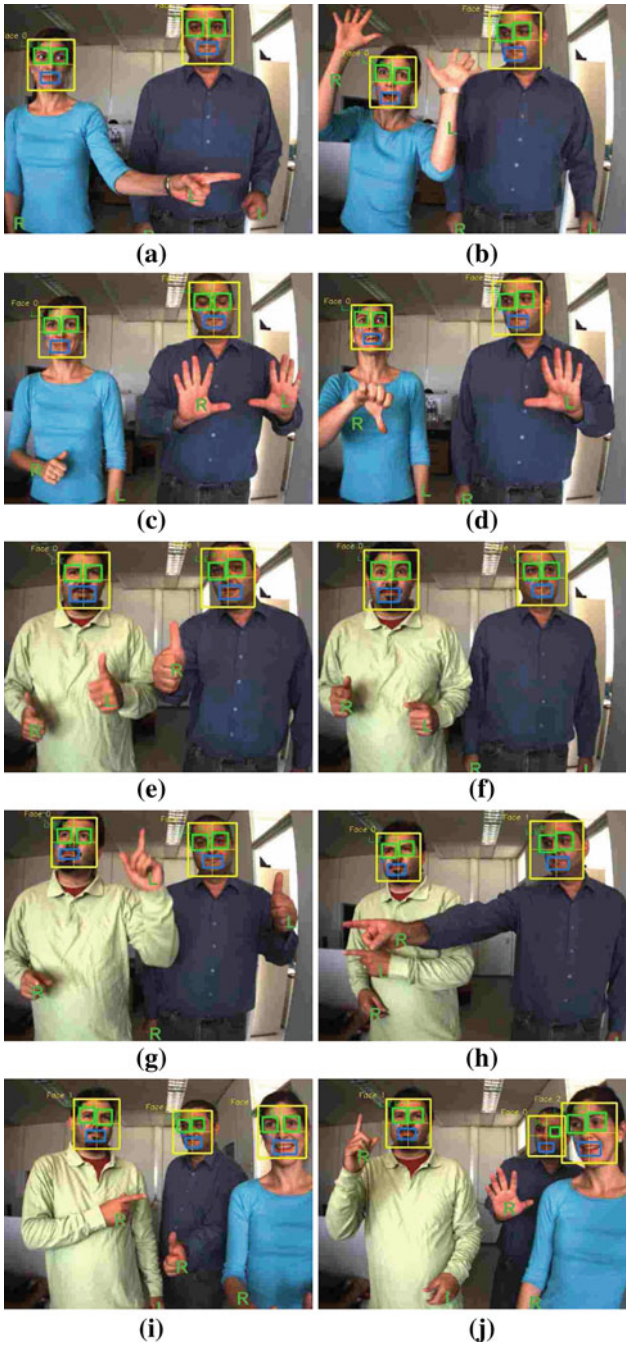

Fig. 14 Ten frames of a sequence depicting multiple persons performing hand and facial gestures in an office environment, focusing on the tracking of facial features. Throughout the sequence, performers enter and leave the scene, and perform gestures in various distances from the camera

The scenario involved the user(s) standing at a distance of approximately 1 m in front of the robot, interacting with it using natural language and hand gestures. The camera was properly mounted on the chest of the robot at a suitable height of approximately 1.2 m.

Table 1 Tracking and classification results for the sequence depicted in Figs. 13 and 14

Total number of frames	1,860
Number of times a person enters the scene	7
Total number of generated tracks	49
Generated tracks with more than 200 frames	28
Tracks correctly classified as faces	7 (100%)
Tracks correctly classified as left hands	12 (100%)
Tracks correctly classified as right hands	9 (100%)
Average number of frames needed for classification (for the certainty to become larger than 0.9)	53
Maximum number of frames needed for classification (for the certainty to become larger than 0.9)	140
Number of initialized eye tracks	14 (100%)
Number of initialized mouth tracks	7 (100%)
Average number of frames needed to initialize facial Feature tracks (to pass validation criteria)	3.2

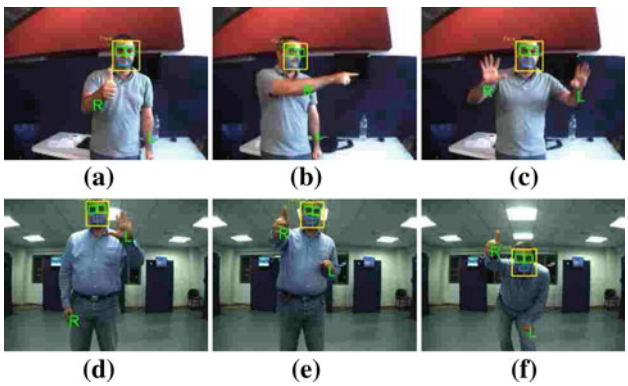

Fig. 15 Frames of two different sequences captured in an exhibition center that show results from hand, face and facial feature tracking

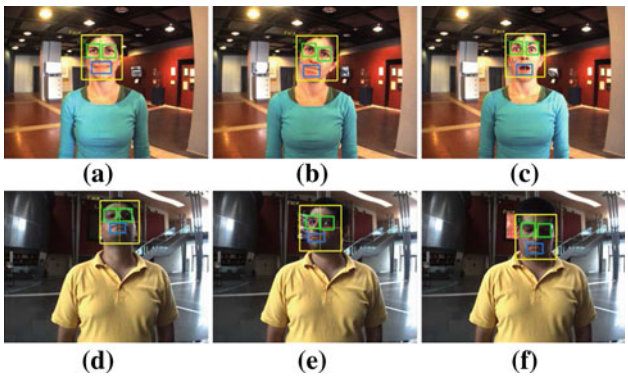

Fig. 16 Frames of different sequences captured in an exhibition center that show results from facial feature tracking of a person

Whenever the robot was asked by the user to present a different exhibit, the robot moved in front of the new exhibit with the user following the robot. Upon arriving at the new

Table 2 Percentages of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results for the two sequences of Fig. 16

	TP (%)	TN (%)	FP (%)	FN (%)
Sequence 1				
Mouth	95.09	0	0.22	4.69
Left eye	95.98	0	0	4.02
Right eye	93.30	2	0.22	4.64
Sequence 2				
Mouth	93.31	0	1.18	5.51
Left eye	94.49	4	0	3.94
Right eye	85.83	6	0	11.81

location, the robot auto adjusted the camera's white white balance and shutter speed parameters, reset the background model, and re-initialized the skin color histograms computed in Sect. 3.4. The robotic base and, thus, the camera remained still until the robot was asked to move to a new exhibit, which allowed the background subtraction and the skin color update algorithms to train new models for each location. For background subtraction, all involved parameters (learning rate and sensitivity of the background subtraction algorithm) were carefully chosen order to minimize false negatives, even if this meant that in some cases background pixels were falsely interpreted as foreground ones (they were subsequently rejected by the skin color detector).

Utilization of background subtraction in combination with skin color detection effectively eliminates static skin-colored regions that may exist in the background (e.g. in Figs. 12, 16, 17). It is to be noted however that skin-colored regions in the foreground (e.g. skin-colored clothes) cannot be eliminated in the current approach, with the exception of very large and very small skin-colored regions which are eliminated via appropriate blob size thresholds.

Figure 17 depicts some screen shots of a real interaction session involving a user interacting with the robot during a tour within the exhibition area. In all cases, the algorithm correctly classified hand and facial blobs in their correct classes. With the exception of very few frames which were very badly illuminated, where the user was standing very far away from the camera, facial features were correctly identified and tracked.

6.3 Evaluation of facial feature tracking

Different test data sets exist for evaluating algorithms mainly for facial expression or affect recognition. In our case, these data served as an alternative option to further assess individually the method for the detection and tracking of facial features, namely localization accuracy of the individual features within a given face area. The databases used in our

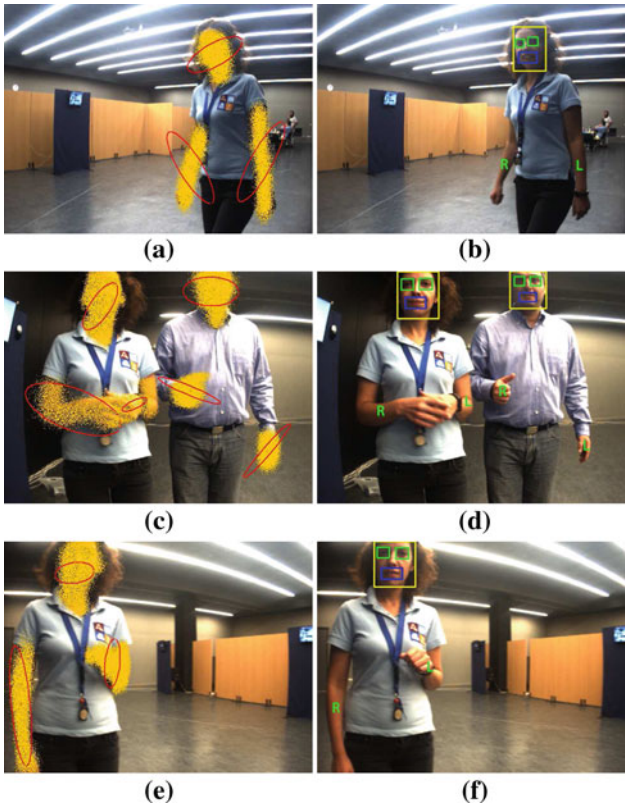

Fig. 17 Filed trial results. *First column* (**a**, **c**, **e**): hand and face tracking results. *Second column* (**b**, **d**, **f**): classification and facial feature tracking results

experiments are the Cohn-Kanade (CK) facial expression database [31], the FABO [32] and the BIOID database [33].

The CK database includes 486 greyscale image sequences from 97 posers exhibiting various facial expressions. Each sequence begins with a neutral expression and proceeds to a peak expression in the last frame. The FABO database contains videos in RGB mode (1,024 × 768 pixels) of face and body expressions of 23 subjects recorded by face and body cameras simultaneously. In our experiments, 1,010 videos from the face cameras were used for testing. Both databases were used for facial feature localization evaluation based on visual validation. Namely, the eye and mouth regions were detected and tracked in all image sequences and results of successful localization were visually verified by a human supervisor. High success rates were achieved for the localization of eyes and mouth in all image sequences of both databases. False positives did not occur in any image and both the left and right eye were correctly localized in a 93% of the images in the CK database and the mouth area in 98% of them. In the FABO database only visible facial features were considered. In 95 and 96% of the images, the left/right eye and the mouth were correctly localized, respectively. The fact that the images were recorded with a high frame rate, led to an increased success rate in feature tracking, since

differences in the relative position and shape of features between consecutive frames were not considerably large. Results from selected frames from the CK and FABO database are illustrated in Fig. 18.

The BIOID database, consisting of 1521 greyscale images (384 × 288 pixels) of frontal faces of 23 different persons, has been acquired in an office environment using a web camera and the output images exhibit illumination variability as well as background noise. Jesorsky et al. [33] first used the data set for face detection and eye localization. The average face size in the recorded images is approximately 140 × 150 pixels, allowing for facial processing tasks to be possible to apply. In the database, there are mainly recorded frontal-view faces and some faces with little in- and off-plane rotations, but there are also people that wear eye glasses and pose various expressions, therefore being a complete database for the evaluation of facial feature localization tasks. Ground truth data are also provided along with the images, namely 20 manually measured feature points per image. Below in this evaluation we focus only on eye localization, since also other authors [33,20] have restrict themselves in eye detection and hence comparative studies are possible.

To validate the performance of eye detection/localization over the BIOID database, we used the distance-based quality measure of Eq. 9, used also by [33,20], where d_l and d_r are distances of the estimated positions from the manually set positions C_l and C_r of the left and right eye center, respectively. The distance measure, hereafter called relative error, is computed from the maximum of the distances d_l and d_r normalized by the distance between the manually measured left and right eye centers.

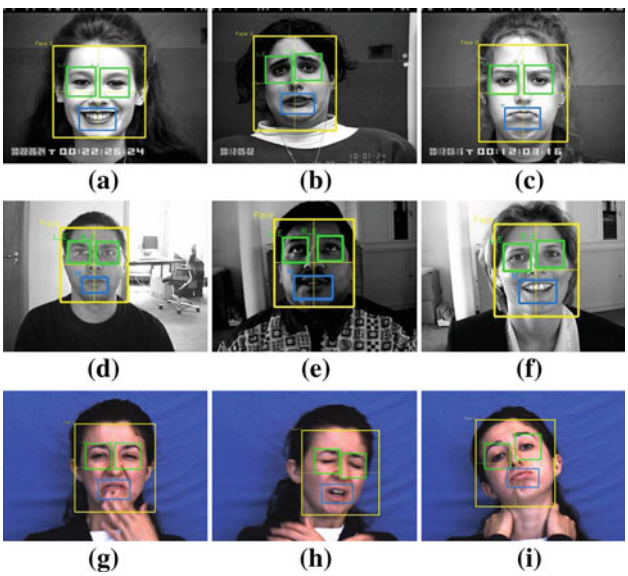

Fig. 18 Frames from the CK database (**a**–**c**), the BIOID database (**d**–**f**) and the FABO database (**g**–**i**) indicating facial feature tracking results

$$d_\mathrm{eye} = \frac{\max(d_\mathrm{l}, d_\mathrm{r})}{\|C_\mathrm{l} - C_\mathrm{r}\|} \tag{9}$$

The eye localization errors and the cumulative distribution have been computed using the two different methods: (a) the Boosted Cascade Detector of Viola-Jones (VJ) with individual sets of Haar-like features for eyes, nose and mouth and imposed anthropometric constraints applied in each frame and (b) the hybrid method for detection and tracking that has described in Sect. 5. We chose to compare our tracking method with the VJ detector, because the latter is commonly used in image sequences, also for tracking purposes. In the first case of the VJ detector (a), the images were processed as per their numbering in the database. Since detection does not depend on the sequence of frames, selection of a different first frame or a sequence of frames would not alter the final result. In the second case of our detection and tracking method (b), the images were initially grouped according to the person's identity, resulting in 23 image groups. The groups were separately processed in closed loop, namely tracking was restricted in images from each individual group following face and facial feature detection. Since the method considers the sequence of frames, we draw randomly different sequences of frames from each group. Results from selected frames from BIOID database are illustrated in Fig. 18.

Figure 19a plots the distribution and cumulative distribution of the relative errors for the VJ constrained method and Fig. 19b shows the respective errors for the hybrid method. From Fig. 19c is evident that the hybrid method outperforms the VJ constrained method, improving the localization accuracy. The method was successful in an average of 95% of cases, testing different order of frames from each group, whereas the VJ method succeeded in 85% of the cases. Out of the remainder 5% error cases of the hybrid method, in 1% the correlation score of a feature failed to pass the lower threshold , in 2% one of the criterions was not satisfied and in 2% the VJ detector failed.

The method could be further compared with previously published results. Jesorsky et al. [33] first introduced the BIOID data set and published results on eye localization accuracy of their algorithm, based on the Hausdorff distance for face matching followed by a Multi-Layer Perceptron eye finder. Cristinacce et al. [20] also presented eye finding results on the BIOID test set, combining the VJ detector, a shape constraint for feature detection also known as Pairwise Reinforcement of Feature Responses and a refinement of the predicted points using edge/corner Active Appearance Models. In the work of Cristinacce et al., a comparison with the results of Jesorsky et al. was given, using the same error measure, namely d_eye, described earlier in this section. Figure 20 shows that the Cristinacce method performs better over the Jesorsky method, however the hybrid method improves the localization accuracy overall. For example with

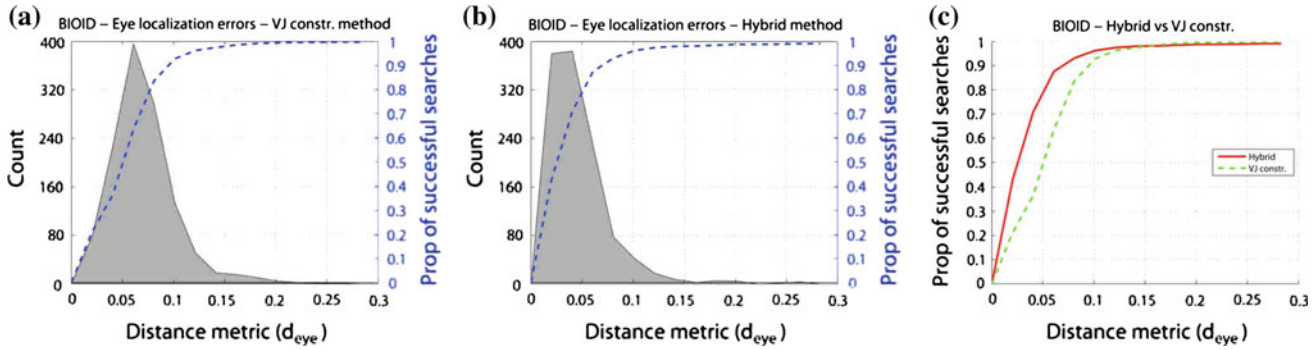

Fig. 19 Distributions and cumulative distributions of the accuracy (distance metric) of the VJ constrained and the hybrid method applied on the BIOID data set

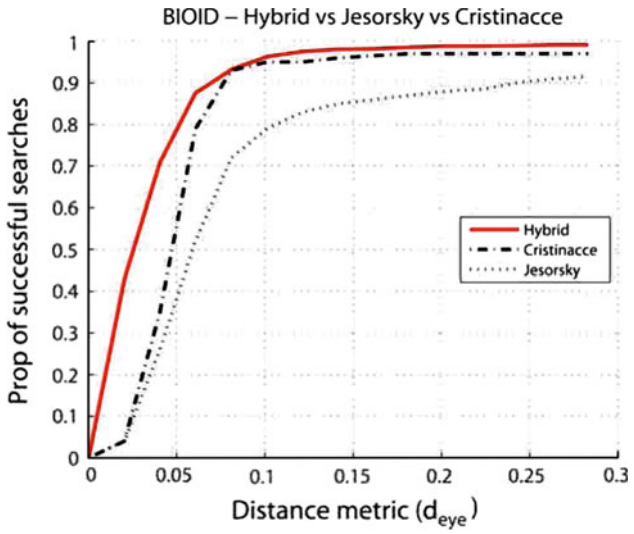

Fig. 20 Comparison with previously published results for eye localization on the BIOID test set

$d_{\text{eye}} = 0.05$, the hybrid search finds 80% of features successfully compared to 60 and 40% using the Cristinacce and the Jesorsky methods, respectively.

7 Conclusions

In this paper, we have presented an integrated approach for tracking of hands, faces and facial features in image sequences, intended to support natural interaction with autonomously navigating robots in public spaces and, more specifically, to provide input for the analysis of hand gestures and facial expressions that humans utilize while engaged in various conversational states with the robot.

For hand and face tracking, we employ a blob tracker which is specifically trained to track skin-colored regions. The skin color tracker was extended by incorporating an incremental probabilistic classifier which was used to maintain and continuously update the belief about the class of each tracked blob which can be a left-hand, a right hand or a face. Facial feature detection and tracking was performed via the employment of state-of-the-art appearance-based detection coupled with feature-based tracking, using a set of anthropometric constraints.

Experimental results have confirmed the effectiveness of the proposed approach proving that the individual advantages of all involved components are maintained, leading to implementations that combine accuracy, efficiency and robustness.

The purpose of the proposed tracking approach to facilitate human-robot interaction tasks but the methodology presented here possesses characteristics that constitutes it suitable for other tasks as well. Besides using it to give input for the analysis of hand gestures and facial expressions, we intend to use it for more general activity recognition tasks and tasks related to robot learning by demonstration.

Acknowledgments This work was partially supported by the European Commission under contract numbers FP6-045388 (INDIGO project) and FP7-248258 (First-MM project).

References

1. Yang, M.H., Kriegman, D., Ahuja, D.: Detecting faces in images: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **24**(1), 34–58 (2002)
2. Zabulis, X., Baltzakis, H., Argyros, A. : Vision-based hand gesture recognition for human-computer interaction. In: Stefanides, C. (ed.) The Universal Access Handbook, Human Factors and Ergonomics., pp. 34.1–34.30. Lawrence Erlbaum Associates, Inc. (LEA), New Jersey (2009)
3. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. Int. J. Comput. Vis. **46**(1), 81–96 (2002)
4. Argyros, A.A., Lourakis, M.I.A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, May 2004, pp. 368–379
5. Nickel, K., Seemann, E., Stiefelhagen, R.: 3D-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 2004, pp. 565–570

6. Baltzakis, H., Argyros, A., Lourakis, M., Trahanias, P.: Tracking of human hands and faces through probabilistic fusion of multiple visual cues. In: Proceedings of the International Conference on Computer Vision Systems (ICVS), Santorini, Greece, May 2008, pp. 33–42
7. Sigalas, M., Baltzakis, H., Trahanias, P.: Visual tracking of independently moving body and arms. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS '09), St. Louis, MO, USA, October 2009
8. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 126–133 (2000)
9. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9), 1372–1384 (2006)
10. Baltzakis, H., Argyros, A.: Propagation of pixel hypotheses for multiple objects tracking. In: Proceedings of the International Symposium on Visual Computating (ISVC), Las Vegas, Nevada, USA, November 2009
11. Zhang, X., Xu, Y., Du, L.: Locating facial features with color information. In: Proceedings of the Fourth International Conference on Signal Processing (ICSP), vol. 2, pp. 889–892 (1998)
12. Hsu, R.-L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 696–706 (2002)
13. Sobottka, K., Pitas, I.: A novel method for automatic face segmentation, facial feature extraction and tracking. Signal Process. Image Commun. **12**(3), 263–281 (1998)
14. Nguyen, T., Nguyen, V., Kim, H.: Robust feature extraction for facial image quality assessment. In: Chung, Y., Yung, M. (eds.) Information Security Applications. Lecture Notes in Computer Science, vol. 6513, pp. 292–306. Springer, Berlin (2011)
15. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. Int. J. Comput. Vis. **8**(2), 99–111 (1992)
16. Herpers, R., Sommer, G.: An attentive processing strategy for the analysis of facial features. In: Face recognition: From Theory to Applications, pp. 457–468 (1998)
17. Pardas, M., Losada, M.: Facial parameter extraction system based on active contours. In Proceedings of the International Conference on Image Processing (ICIP'01), Thessaloniki, Greece, October 2001, vol. 1, pp. 1058–1061
18. Kawaguchi, T., Rizon, M., Hidaka, D.: Detection of eyes from human faces by hough transform and separability filter. Electron. Commun. Japan **88**(5), 29–39 (2005)
19. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
20. Cristinacce, D., Cootes, T.F., Scott, I.: A multi-stage approach to facial feature detection. In: 15th British Machine Vision Conference, pp. 231–240 (2004)
21. Zuo, F., de With, P.H.: Facial feature extraction by a cascade of model-based algorithms. Signal Process Image Commun. **23**(3), 194–211 (2008)
22. Zhou, Y., Li, Y., Wu, Z., Ge, M.: Robust facial feature point extraction in color images. Eng. Appl. Artif. Intell. **24**, 195–200 (2011)
23. Kim, H.-C., Kim, D., Bang, S.-Y.: A pca mixture model with an efficient model selection method. In: Proceedings of the International Joint Conference on Neural Networks, 2001 (IJCNN '01), vol. 1, pp. 430–435 (2001)
24. Phimoltares, S., Lursinsap, C., Chamnongthai, K.: Locating essential facial features using neural visual model. In: Proceedings of the International Conference on Machine Learning and Cybernetics, 2002, vol. 4, pp. 1914–1919 (2002)
25. Wilson, P., Fernandez, J.: Facial feature detection using haar classifiers. J. Comput. Sci. Coll. **21**(4), 127–133 (2006)
26. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
27. Pateraki, M., Baltzakis, H., Kondaxakis, P., Trahanias, P.: Tracking of facial features to support human-robot interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '09), pp. 3755–3760 (2009)
28. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR),Ft. Collins, USA, June 1999, pp. 2246–2252
29. Tian, Y.: Evaluation of face resolution for expression analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), pp. 82–89. IEEE Computer Society, New York (2004)
30. Sohail, A.S.Md, Bhattacharya, P.: Detection of facial feature points using anthropometric face model. In: Damiani, E., Ytongnon, K., Schelkens, P., Dipanda, A., Legrand, L., Chbeir, R. (eds.) Signal Processing for Image Enhancement and Multimedia Processing. Multimedia Systems and Applications, vol. 31, pp. 189–200. Springer US, USA (2008)
31. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In Proceedings of the IEEE 4th International Conference on Automatic Face and Gesture Recognition (FG '00), Grenoble, France, pp. 46–53 (2000)
32. Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proceedings of ICPR 2006 the 18th International Conference on Pattern Recognition, Hong Kong, China, August 2006
33. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: Proceedings of the 3rd International Conference on Audio- and Video-based Biometric Person Authentication, Halmstad, Sweden, June 2001. Lecture Notes in Computer Science, vol. 2091, pp. 90–95. Springer, Berlin

Author Biographies

Haris Baltzakis is a principal researcher at the Institute of Computer Science of the Foundation for Research and Technology—Hellas (FORTH), Greece. He received a diploma in electrical engineering from the Democritus University of Thrace in 1997 and the M.Sc. and Ph.D. degrees in Computer Science from the University of Crete in 1999 and 2004, respectively. He has a broad background and research experience in robotics, with emphasis on autonomous robot navigation, and vision, with emphasis in vision based tracking of humans and hand/face gesture recognition. In these research fields he has published more than 30 papers in peer-reviewed scientific journals and conference proceedings.

Maria Pateraki received her Ph.D. in Photogrammetry from the Swiss Federal Institute of Technology in Zurich (ETHZ), Switzerland in 2005 and has been a research associate at the University of Melbourne and the Co-operative Research Center for Spatial Information (CRC-SI) in Melbourne, Australia. Since 2008 she is a member of the Computational Vision and Robotics Laboratory at the Institute of Computer Science of the Foundation for Research and Technology—Hellas (FORTH), Greece. Her research interests include topics related to robotic vision, motion tracking, registration matching and 3D reconstruction and she has published over 30 papers in peer-reviewed scientific journals and conference proceedings.

Panos Trahanias is a Professor with the Department of Computer Science, University of Crete, and the Institute of Computer Science, Foundation for Research and Technology—Hellas (FORTH). He

received his Ph.D. in Computer Science from the National Technical University of Athens, Greece, in 1988. Following that, he had positions as Research Associate at the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", Athens, Greece (1989–1991), and at the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada (1991–1993). He has participated in many research projects in image processing and analysis at the University of Toronto and has been a consultant to SPAR Aerospace Ltd., Toronto. Since 1993, he is with the University of Crete and FORTH. He has held the position of Director of Graduate Studies at the Department of Computer Science, University of Crete and currently he Chairs the same Department. At FORTH he Heads the Computational Vision and Robotics Laboratory, where he coordinates research and development activities in human-robot visual interaction, robot navigation, visual tracking, and brain-inspired robotic control. He has coordinated and participated in numerous research projects funded by the European Commission and Greek funding agencies. He has been General Chair of Eurographics 2008 (EG'08) and the European Conference of Computer Vision 2010 (ECCV'10). He has published over 110 papers in technical journals and conference proceedings.