Visual human-robot communication in social settings

Maria Pateraki[1], Markos Sigalas[1,2], Georgios Chliveros[1] and Panos Trahanias[1,2]

Abstract— **Supporting human-robot interaction (HRI) in dynamic, multi-party social settings relies on a number of input and output modalities for visual human tracking, language processing, high-level reasoning, robot control, etc. Capturing visual human-centered information is a fundamental input source in HRI for effective and successful interaction. The current paper deals with visual processing in dynamic scenes and presents an integrated vision system that combines a number of different cues (such as color, depth, motion) to track and recognize human actions in challenging environments. The overall system comprises of a number of vision modules for human identification and tracking, extraction of pose-related information from body and face, identification of a specific set of communicative gestures (e.g. "waving, pointing") as well as tracking of objects towards identification of manipulative gestures that act on objects in the environment (e.g. "grab glass", "raise bottle"). Experimental results from a bartending scenario as well a comparative assessment of a subset of modules validate the effectiveness of the proposed system.**

I. Introduction

As robots become integrated into daily life, they must increasingly deal with situations in which effective human-robot interaction is characterized as continuous, natural and socially appropriate. In this framework, perception of humans and tracking of humans' actions and activities can be realized by an appropriate vision system that is able to operate in real-time in dynamic and cluttered scenes with variable illumination, capturing information from multiple users in the robot's environment. Towards this goal, the integrated vision system presented in this work combines a number of different cues, such as color, depth and motion extracted from RGB-D sequences to robustly track and recognize human actions in challenging environments. The proposed system combines different methods for identification and tracking of human hands and faces, extraction of pose-related information from body and face, the latter being of interest in the HRI domain as attentive cues of users in the robot's environment. Furthermore, methods for the identification of a specific set of communicative gestures, such as waving and pointing, as well as tracking of objects towards identification of manipulative gestures that act on objects in the environment (e.g. "grab glass", "raise bottle") are included in the overall vision system. In this paper the methods for the identification and tracking of human hands, faces and objects, as well as the methods for extracting pose-related information from body and face are discussed. These methods form the core components of a vision system utilized in a bartender robot [1].

For the identification and tracking of human hands and faces a variety of approaches have been reported in the literature [2], [3]. Several of them rely on the detection of skin-colored areas [4], [5]. 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. In contrast to blob tracking approaches, model-based ones [6], [7] do not track objects on the image plane but, rather, in a hidden model-space. 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.

With respect to the extraction of pose-related information from body and face, there is a large number of different methods in the current literature [8], [9]. With the emergence of real-time depth sensors, a few notable works have shown the usefulness of depth in solving pose estimation problems for body [10], [11], [12] and face [13], [14], [15]. Most of existing body pose estimation methods require an initialization phase for registering users (e.g. specific body pose by the user for a short time) and assume recovery of body pose parameters, which limits their efficiency in real life dynamic environments. With respect to face pose estimation, methods can be distinguished in appearance and feature-based using 2D images, depth data or combination of both. Appearance-based methods depend on a time-consuming training phase (e.g. [14]), and most feature-based methods are limited by the requirement to define pose-dependent features (e.g. [16]). Both methods on body- and face-pose tracking presented in this paper overcome requirements of large training data and initialization constraints.

In the next section, we give an overview of the integrated vision system and the individual methods for tracking skin-colored regions, classifying them as hands and faces, estimation of torso and face pose as well as tracking objects towards the identification of manipulative gestures (not part of this paper). Results of the vision system in real environments are presented in section III.

II. Methodology Overview

A block diagram of the components that comprise the proposed approach is depicted in Fig. 1 and the individual components are discussed in the following sections. Fig. 1

*This work has been partially supported by the EU Information Society Technologies research project James (FP7- 045388) and FIRST-MM (FP7-248258).

[1]Institute of Computer Science, Foundation for Research and Technology - Hellas, Heraklion, Crete, Greece

[2]Department of Computer Science, University of Crete, Greece

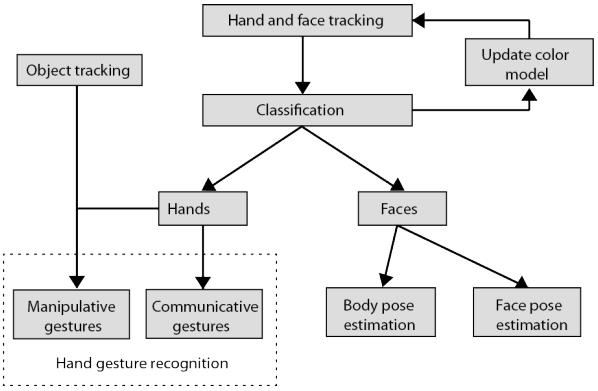


Fig. 1: Block diagram of the proposed vision system

comprises a number of modules, and notably the hand and face tracker and the hand/face classifier. The former 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. The latter 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. Blobs classified as faces are used to update the color distribution of skin-pixels, thus enabling the algorithm to quickly adapt to illumination changes. They are further used as input in the face and torso pose estimation modules. Hand trajectories are forwarded to the hand-gesture recognition system (not described in this paper) taking into consideration detected object tracks.

A. Hand and Face Tracking

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. To detect and track faces and hands we employ and extend a blob-tracking approach [17], according to which foreground, skin-coloured pixels are identified based on their colour and grouped together into skin-coloured 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. 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).

B. Object Tracking

In case we are interested in tracking objects in addition to hands and faces and these objects are characterized by a dominant color (e.g. coca cola, perrier, evian) we extend our technique described in section II-A to track multiple color blobs based on a number of defined color classes $c_1, ...c_N$. Following background subtraction, foreground pixels are characterized according to their probability to depict a color-class and then grouped together into blobs using hysteresis thresholding and connected components labeling. The algorithm handles the issue of assigning a pixel in more than one color classes by assigning it to the class with the highest probability. Then pixels are grouped together into blobs using hysteresis thresholding and connected components labelling as in [17].

The posterior probability for each pixel x_i with color c to belong to a color class c_N is computed according to Bayes rule as:

$$P(c_N \mid x_i) = \frac{P(c_N)}{P(x_i)} P(x_i \mid c_N) \tag{1}$$

where

$$P(x_i) = \sum_{j=1}^{N} P(xi \mid c_j) P(c_j) \tag{2}$$

$P(c_N)$ and $P(x_i)$ are the prior probabilities of foreground pixels of a specific color class and foreground pixels x_i having color c, respectively. $P(x_i|c_N)$ is the likelihood of color c for foreground regions of specific color class. All three components in the right side of the above equation are computed off-line during training. Tracking of object classes progresses in the same manner as with skin color classes (section II-A).

C. Hand and Face Classification

The second step of the proposed system involves the classification of the resulting skin-colored 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. An incremental classifier has been developed [18] which extends the above blob tracking approach and which is used to maintain and continuously update a belief about whether a tracked hypothesis of a skin blob 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.

D. Adapting to illumination changes

Blobs classified as faces are also used to update the color distribution of skin-colored pixels, thus enabling the algorithm to quickly adapt to illumination changes, which may deteriorate skin color detection. Hence, a mechanism that adapts the employed representation according to the recent history of detected skin-colored points is required [19]. To solve this problem, skin color detection maintains two sets of prior probabilities. The first set consists of $P(s), P(c), P(c|s)$, which are the prior probabilities of

foreground skin pixels and foreground pixels having color c and the likelihood of color c for skin-colored foreground regions and correspond to the off-line training set. The second set consists of $P_h(s), P_h(c), P_h(c|s)$, which correspond to the evidence that the system gathers during runtime from tracks classified as facial tracks with high confidence. 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:

$$P(s|c) = \gamma P(s|c) + (1-\gamma)P_h(s|c) \quad (3)$$

where $P(s|c)$ and $P_h(s|c)$ can both be derived from Bayes rule, but involve prior probabilities that have been computed from the whole training set and from online training, respectively. In 3, 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.

E. Torso pose estimation

A topic of particular interest in the HRI domain is the focus of attention of a person interacting with a robot, effectively conveying information on whether this person is seeking attention. Consequently, torso pose estimation and face pose estimation are identified as important attentive cues.

Although the hand and face (and object) tracking module takes into consideration color and motion cues derived from RGB image sequences, the torso pose estimation module utilizes an additional cue, that of depth, obtained from an RGB-D sensor, such as the Kinect™ [20]. A model-based approach is formulated, primarily focusing on overcoming the requirement of large training data and initialization constraints of other methods, while exhibiting robustness in dynamic settings. Face blobs extracted in the previous steps steer the detection of human body. Initially shoulder areas are extracted, based on illumination, scale and pose invariant features on the RGB silhouette. Depth point cloud information is further utilized to model hypotheses for the shoulder joints and the human torso based on a set of 3D geometric primitives. The final estimation of the 3D torso pose is derived via a global optimization scheme which is body pose and/or body morphology independent.

Below the major steps of the proposed approach are listed:

- **User detection and tracking**. Based on face detection and tracking, the human body silhouettes for the detected users in the scene are extracted by depth thresholding and refined via cubic spline fitting to secure piecewise continuity.
- **2D extraction of shoulder areas**. Given the location of the face, we select sets of points on the RGB silhouette, delineating possible shoulder areas. Selection is based on pose and scale invariant features satisfying certain geometric constraints. The shoulder areas on the silhouette are characterized each by two body parts, that of *acromial or shoulder point* and of *axillary or armpit*. The first is the upper part of the shoulder (red points in Fig. 2) and it's robustly detectable for all configurations where the elbow is below the shoulder. This silhouette region is characterized by high curvature, which is scale and pose invariant and the respective contour points posses certain geometric characteristics. The location of this area on the silhouette can be approximated via the tangent line. More specifically, we compute the tangent line for all silhouette points within certain bounds (estimated proportionally from camera position and face height) and check whether it intersects the face area (white segments in Fig. 2). The second is the area "below" the shoulder (light blue points in Fig. 2). Similarly, the *acromial* is visible in most of the shoulder-elbow configurations and is scale and pose invariant. For the set of points in this area the normal of the tangent line intersects the face area (green segment in Fig. 2).
- **Generation and 3D modeling of shoulder joints hypotheses**. Shoulder joints hypotheses are generated by projecting and proportionally expanding the 2D shoulder areas on the depth point cloud. Shoulder joints are approximated by least squares fitting of 3D spheres to the selected areas on the 3D point cloud. A set of quality criteria, such as spheres' radii and number of iterations are used to eliminate possible outliers. Figure 3 shows an example of shoulder joint modelling via sphere fitting.
- **Generation and 3D modelling of torso hypotheses**. Each pair of left and right shoulder joint-hypotheses along with a 3D point on the lower border of the torso with the pelvis area (extracted via anthropometric constraints) are used to select the area of 3D points on the torso surface and generate a torso hypothesis. This hypothesis is a 3D ellipsoid that is best fitted in a least squares sense to the selected 3D data on the point cloud.
- **Estimation of the 3D torso pose**. A global optimization scheme is adopted for the selection of the best set of shoulder joints and torso and to derive a refined torso pose per frame. Additionally, the best shoulder joints are separately tracked by means of an Extended Kalman Filter, to further refine shoulder detection and possibly handle partial occlusions.

F. Face pose estimation

Face pose estimation is extracted via Least-Squares Matching (LSM) on the RGB image and differential rotations are computed by analyzing the transformations of the facial patch across image frames [21]. The problem statement is finding the corresponding part of the template image patch, in our case the face path $f(x, y)$ in the search images $g_i(x, y), \quad i = 1, ...n-1$.

$$f(x, y) - e_i(x, y) = g_i(x, y) \quad (4)$$

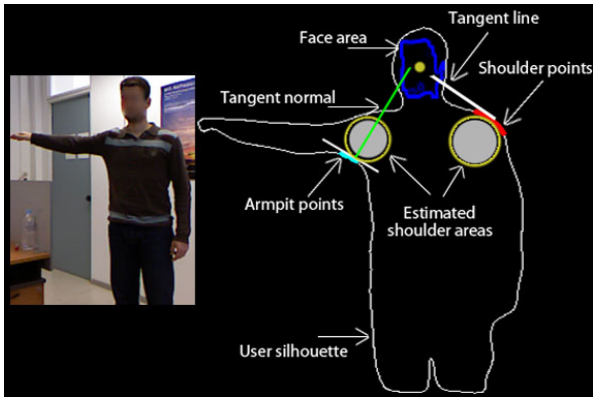


Fig. 2: Modelling shoulder areas. From a single RGB-D image, assuming the location of the face, 2D points along the user's silhouette are selected based on robust 2D features. Selected points are then used to determine the 3d area (fitted sphere) which constrains each shoulder.

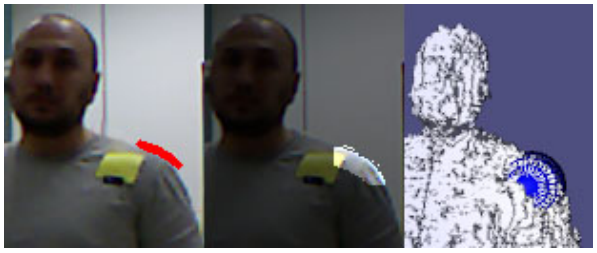


Fig. 3: Modelling of shoulder joints via sphere fitting. Delineated *acromial*(left), shoulder area selected for sphere fitting (middle), 3D sphere approximating shoulder joint.

Equation (4) gives the least squares grey level observation equations, which relate the $f(x, y)$ template and $g_i\ (x, y)$ image functions or image patches. The true error vector $e_i\ (x, y)$ is included to model errors that arise from radiometric and geometric differences in the images.

Assuming we have two images, in our case two consecutive frames, the $f(x, y)$ and $g(x, y)$, a set of transformation parameters need to be estimated from (4). Since (4) is nonlinear, it is linearized by expanding it into a Taylor series and keeping only zero and first order terms.

The estimation model should accommodate enough parameters in order to be able to model completely the underlying transformation. In the model only geometric parameters are included and radiometric corrections, e.g. equalization, for the compensation of different lighting conditions are applied prior to LSM in template and image. Assuming that the local surface patch of the face area is a plane to sufficient approximation (since depth variation exhibited by facial features are small enough) an affine transformation is used to model geometric differences between template or image frame n and search image or image frame $n+1$. The affine transformation (5) is applied with respect to an initial position (x_0, y_0):

$$\begin{aligned} x &= a_0 + a_1 \cdot x_0 + a_2 \cdot y_0 \\ y &= b_0 + b_1 \cdot x_0 + b_2 \cdot y_0 \end{aligned} \tag{5}$$

By differentiating (5) and the parameter vector being defined according to (6) the least squares solution of the system is derived.

$$x^T = (da_0, da_1, da_2, db_0, db_1, db_2) \tag{6}$$

The method requires that the change from frame to frame is small, considering the speed of the object and the framerate of the acquired image sequence, for the solution to converge. To improve performance and handle cases of fast motions we operate the algorithm at lower resolution levels.

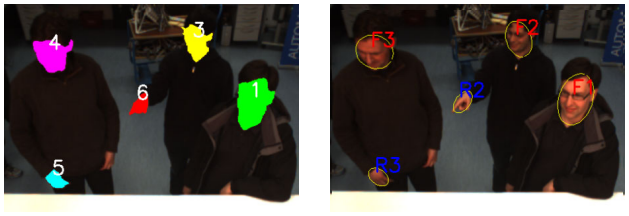
To derive the above-mentioned face rotations we employ LSM by initializing the template patch, at the center of the detected blob ellipse and updated the template in image frame $n + 1$ based on the estimated affine parameters and matched to the next image frame. The rotation between the initial position of the template and the final matched position is computed by accumulating the differential rotation angles derived by matching each consecutive template and patch. Under the assumption that the head approximates a spherical body and using the mapping equations of the vertical perspective projection we are able to compute the horizontal rotation of the face as in [21].

III. Results

The proposed methods form the core components of a vision system utilized in a bartender robot [1]. In all reported experiments, the resolution of the RGB camera was 640x480. Although the performance of the system greatly depended on the number of active hypotheses derived in the hand and face tracking module as well as in the torso pose estimation module in all cases, the algorithm was able to process the cameras input stream at a rate exceeding 20 frames per second on a standard computer.

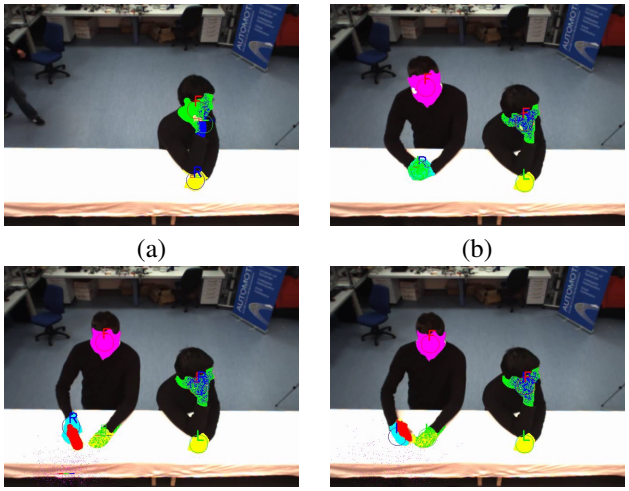
Fig. 4 shows results of skin-colored tracking and classification of hands and faces during fielded evaluation of the robot bartender. A number of sequences was captured with a maximum number of four users in the robots environment for up to 10 minutes each, enacting drink ordering variations. Blobs classified 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 up to twelve observed tracks and it also managed to maintain its belief over the whole sequence. There were a lot of cases in which hand and face hypotheses were partly occluded or merged as in the example of Fig. 5, where it can be seen that the face and hands hypotheses are maintained even in cases of merging to one blob. Similar results also were observed when object tracking (bottles) was invoked (Fig. 5), and face, hands and object hypotheses were successfully identified and tracked.

Torso pose estimation was also tested in the above bartending scenario with Fig. 6 illustrating indicative results of torso orientation. In all cases the results were more than promising, as the system could successfully recognize which user was seeking attention for all visited cases and without overlooking the fact that our method managed to cope with



(a) (b)

Fig. 4: Bartending environment. (a) Multiple hypothesis tracking of skin-colored blobs, the IDs of the different hypotheses is being shown and (b) classification of hands and faces using the incremental classifier of [18].



(a) (b)

(c) (d)

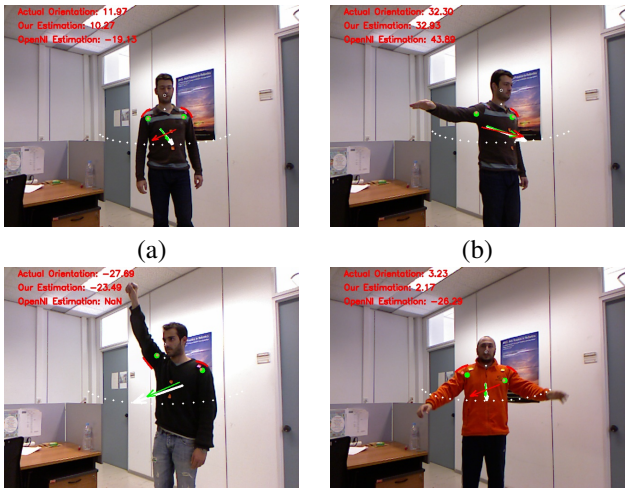
Fig. 5: Bartending environment, frames out of a sequence with humans and objects. Classification of hands and faces using the incremental classifier of [18].

the initialization problem, by recognizing the user and his torso orientation really fast.

(a) (b)

Fig. 6: Bartending environment, torso orientation estimation.

For a comparative assessment of our method on body pose estimation, constraint to torso orientation in the horizontal direction, against ground truth data we conducted a series of experiments in 3 sequences (summing to a total of 5000 frames of single users performing a variety of poses, in a controlled office environment. Lacking a motion capturing device, we used colored markers, on the user's clothing instead, in order to automatically detect the actual location of the shoulders, thus derive ground truth information. Additionally, we also tested the skeletonization module of the OpenNI [22] against the ground truth, compared the results with those of our method, and produced relative statistics of both approaches.



(a) (b)

(c) (d)

Fig. 7: Comparison with OpenNI skeletonization module. In each image, the orientation (in degrees) ground truth and the estimation of each method is shown at the upper left part of each image. The thick white arrow depicts the ground truth orientation, the green one depicts the estimated orientation of our methodology, and the red arrow depicts the orientation estimated by the OpenNI.

Fig.7 shows a variety of resulting images from the ground truth sequences. The user is roughly turned to the camera and performs a series of poses, by raising either or both hands (increasing the task difficulty) and rotating his body in various orientations. The actual (ground truth) orientation (in degrees) and the ones estimated by the two methods are superimposed on the images at the upper left part of each one. The thick white arrow depicts the actual orientation, while the green and red one illustrate the estimation of our methodology and OpenNI, respectively. It is interesting to note that there are cases where OpenNI estimation is far away from the truth Fig.7(a)and (d), whereas in others (as the ones of Fig.7(c)) OpenNI failed to derive an estimation. On the contrary, our method managed to robustly detect and track the shoulders in most of the cases and provide a very accurate estimation for the body orientation.

The face pose estimation method has been also tested in laboratory conditions as well in other environments with challenging illumination conditions. Fig. 8 shows results of face pose estimation in low light conditions, where off-plane and in-plane face rotations are robustly tracked. Moreover, a quantitative evaluation has been carried out using ground truth information [21]. The user was standing in front of the camera and turning his head in predefined directions, defined in the range of $0° \pm 180°$ with an angular step of

$10\,^{\circ}$. Results were derived from image sequences of a total of 7000 image frames and could be seen that the algorithm achieves high success rates for low angles (user looks in directions close to the direction of the camera) which are decreased for higher angles. The algorithm also maintains significant success rates (more than 50%) for angles up to $120\,^{\circ}$, where only a small part of the facial patch is visible.

To conclude it is also of interest to note that the vision system was used during a user evaluation study of the robot bartender, performed with a number of users enacted variations on drink-ordering scenarios (31 participants in 3 scenario variations), explained thoroughly in [1]. For the purposes of the above study, customers were considered to be seeking attention from the robot bartender if (a) they were close to the bar, and (b) their torso was oriented towards the bartender. This simple rule was based on a study of customer behaviour in natural bar settings [23], where nearly all customers were found to exhibit those behaviours when initiating a drink-ordering transaction. The vision system was running over long sequences and was able to extract the required information to initiate the interaction of users with the robot, resulting to many successful drink-ordering transactions.

Fig. 8: Head pose estimation in real-world environment with challenging illumination conditions.

IV. CONCLUSIONS

The paper presented an integrated vision system targeted for HRI scenarios. The system has been designed to work in dynamic, multi-party social settings combining different color, depth and motion cues, towards tracking and recognition of human actions engaged in a robotic environment. The system has been tested in a bartending scenario with overall success. Moreover, the torso estimation module was tested against OpenNI using ground truth data, exhibiting superior performance.

REFERENCES

[1] M. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. Petrick, ""two people walk into a bar: Dynamic multi-party social interaction with a robot agent," in *In Proc. of the 14th ACM International Conference on Multimodal Interaction*, Santa Monica, CA, USA, 22-26 October 2012.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," vol. 24, no. 1, pp. 34–58, 2002.

[3] X. Zabulis, H. Baltzakis, and A. Argyros, "Vision-based hand gesture recognition for human-computer interaction," in *The Universal Access Handbook*, ser. Human Factors and Ergonomics, C. Stefanides, Ed. LEA Inc., Jun. 2009, pp. 34.1 – 34.30.

[4] M. Jones and J. Rehg, "Statistical color models with application to skin detection," in *International Journal of Computer Vision*, vol. 46, no. 1, 2002, pp. 81–96.

[5] H. Baltzakis, A. Argyros, M. Lourakis, and P. Trahanias, "Tracking of human hands and faces through probabilistic fusion of multiple visual cues," in *Proc. International Conference on Computer Vision Systems (ICVS)*, Santorini, Greece, May 2008, pp. 33–42.

[6] M. Sigalas, H. Baltzakis, and P. Trahanias, "Visual tracking of independently moving body and arms," St. Louis, MO, USA, Oct. 2009.

[7] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter." *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 9, p. 13721384, 2006.

[8] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 4–18, 2007.

[9] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2008.106

[10] S. Escalera, "Human behavior analysis from depth maps," in *Articulated Motion and Deformable Objects*, ser. Lecture Notes in Computer Science, F. Perales, R. Fisher, and T. Moeslund, Eds. Springer Berlin / Heidelberg, 2012, vol. 7378, pp. 282–292.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *In Proc. Computer Vision and Pattern Recognition*, 2011.

[12] B. Holt, E.-J. Ong, H. Cooper, and R. Bowden, "Putting the pieces together: Connected poselets for human pose estimation." in *Proc. International Conference on Computer Vision*, 2011.

[13] M. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[14] G. Fanelli, J. Gall, and L. V. Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 617–624.

[15] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3d deformable face tracking with a commodity depth camera," in *Proceedings of the 11th European conference on computer vision conference on Computer vision: Part III*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 229–242. [Online]. Available: http://dl.acm.org/citation.cfm?id=1927006.1927026

[16] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *In Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2001, pp. 255–260.

[17] H. Baltzakis and A. Argyros, "Propagation of pixel hypotheses for multiple objects tracking," *Advances in Visual Computing*, pp. 140–149, 2009.

[18] H. Baltzakis, M. Pateraki, and P. Trahanias, "Visual tracking of hands, faces and facial features of multiple persons," *Machine Vision and Applications*, pp. 1–17, 2012, 10.1007/s00138-012-0409-5. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0409-5

[19] A. A. Argyros and M. I. A. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera," in *In: ECCV*, 2004, pp. 368–379.

[20] M. Corp. Kinect for xbox 360. Redmond WA. [Online]. Available: http://www.xbox.com/en-US/kinect

[21] M. Pateraki, H. Baltzakis, and P. Trahanias., "Using dempsters rule of combination to robustly estimate pointed targets." in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2012.

[22] OpenNI. [Online]. Available: http://openni.org

[23] K. Huth, "Wie man ein bier bestellt," Master's thesis, Universitaet Bielefeld, 2011.