Visual tracking in dynamic scenes

M. Pateraki, P. Trahanias
Institute of Computer Science, Foundation for Research and Technology - Hellas
{pateraki,trahania}@ics.forth.gr

Abstract: Visual detection, recognition and tracking of specific objects in the scene is related to different applications ranging from visual surveillance for security purposes and navigation of autonomous vehicles/robots to human-computer and human-robot interfaces. 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 encodes critical information about head pose and gestures. Furthermore, both the head and torso pose support the derivation of the focus of attention, a key variable in the analysis of human behavior in human-robot interaction paradigms encompassing social aspects. The main challenge of visual tracking relates to the method's robustness and invariance in scene- and image-changing factors. Namely non-uniform illumination, background clutter, diversity in human appearance, shape, expression, gesturing, as well as occlusions are some of the factors which set the problem as non-trivial. Moreover, the inherent requirement of real-time computations increase the overall difficulty of the problem. Inline with the above, in this work, the proposed approach integrates and combines a number of state-of the art techniques to solve three different but closely related problems: (a) identification and tracking of human hands and faces which are detected as skin-colored blobs, (b) robust classification of the identified tracks to faces and hands, and, finally, (c) estimation of the head pose using each recognized facial blob as prior.
To detect and track faces and hands we employ and extend a blob-tracking approach (Baltzakis et al., 2008), according to which foreground, skin-coloured pixels are identified based on their colour and grouped together into skin-coloured blobs (Fig. 1). 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) (Fig. 2). In case we are interested in tracking objects in addition to hands and faces and these objects are characterized by a dominant color we may extend our method to track multiple color blobs based on a number of defined color classes $c_1, c_2, \dots, c_N$. 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.

Published online November 2015

The second step of the proposed approach 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 (Baltzakis et al., 2012) 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. Figure 3 shows results from the visual processing in a human-robot interaction sequence in a bartending environment where hand, face and object tracking results are illustrated along with recognition of specific communicative and manipulative gestures (Pateraki et al., 2013).

To further estimate the head pose from monocular RGB sequences extracted facial blobs are fed to a Least-Squares Matching (LSM) module (Gruen, A., 1985) which is extended to derive differential rotations via facial patch deformations across image frames as in (Pateraki et al, 2014). For example horizontal off-plane rotations (yaw angle) of the head mainly deform the facial patch in x-shift and x-scale. To derive the face rotations we perfom matching of the facial patch across image frames.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 (Pateraki et al, 2014). 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.

The proposed visual modules for combined tracking of hands, faces and for head pose estimation have been implemented on different robotic systems with rich interaction capabilities. Experimental results have confirmed the effectiveness of these methods 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 is to facilitate human-robot interaction tasks but the methodology presented here possesses characteristics that constitutes it suitable for other tasks as well and therefore can be used for more general activity recognition tasks.

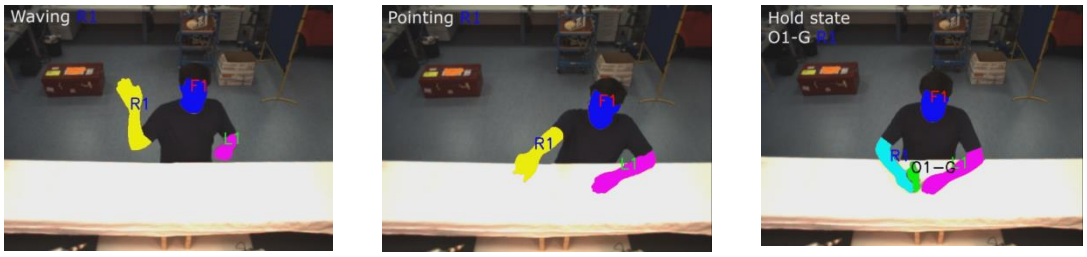
(a) (b) (c)

Figure 1. The tracking approach. (a) Initial image, (b) pixel probabilities, (c) hand and face hypotheses.

Published online November 2015

(a) (b) (c)

Figure 2. Tracking hypotheses over time. Indicative tracking results in three frames of the office image sequence used in the previous example. Green dots represent the pixel hypotheses.



(a) (b) (c)

Figure 3. Bartending environment, frames out of a human robot interaction sequence. Classification of hands and faces using the incremental classifier and detection on communicative and manipulative gestures.

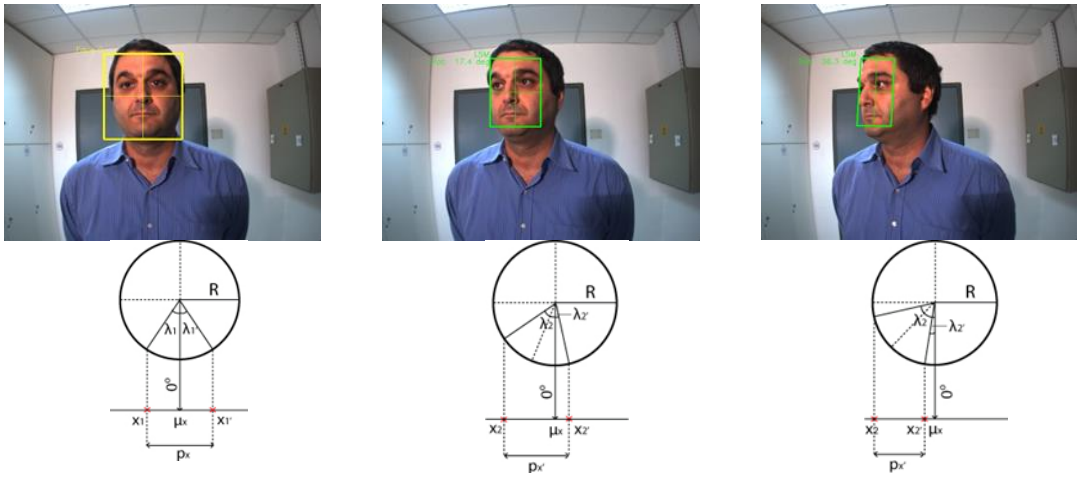


Figure 4. Yaw angle estimation via facial patch deformation.

Keywords: object tracking, pose estimation, image sequence analysis.

References

Baltzakis H., Argyros A., Lourakis M., Trahanias P., 2008, Tracking of human hands and faces through probabilistic fusion of multiple visual cues. In: Proc. International Conference on Computer Vision Systems (ICVS), Santorini, Greece, pp. 33-42.

Baltzakis H., Pateraki M., Trahanias P., 2012, Visual tracking of hands, faces and facial features of multiple persons. Machine Vision and Applications, pp. 1-17.

Published online November 2015

Gruen A. W., 1985, Adaptive least squares correlation: a powerful image matching technique. South African J. Photogramm. Remote Sens. Cartogr., Volume 14: 175-187.

Pateraki M., Sigalas M., Chliveros G., Trahanias P., 2013, Visual human-robot communication in social settings. In Proc. of the Workshop on Semantics, Identification and Control of Robot-Human-Environment Interaction, held within the IEEE International Conference on Robotics and Automation (ICRA), 10 May, Karlsruhe, Germany.

Pateraki M., Baltzakis H., Trahanias P., 2014, Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation. Computer Vision and Image Understanding, Volume 120: 1-13.

Published online November 2015