# Temporal gesture recognition for human-robot interaction

Markos Sigalas[†‡], Haris Baltzakis[†] and Panos Trahanias[†‡]

† Institute of Computer Science,
Foundation for Research and Technology - Hellas,
Heraklion, Crete, Greece

‡ Department of Computer Science, University of Crete,
P.O.Box 1470, Heraklion, 714 09 Crete, Greece
{msigalas,xmpalt,trahania}@ics.forth.gr

***Abstract*— This paper describes a novel hand gesture recognition system intended to support natural interaction with autonomously navigating robots that guide visitors in museums and exhibition centers. The proposed system utilizes upper body part tracking and two neural network-based classifiers, one for each arm. Tracking is performed in a 9-DoF configuration space and it is facilitated by means of a probabilistic approach which combines particle filters with hidden Markov models in order to enable the simultaneous tracking of several hypotheses for the body orientation and the configuration of each of the two arms.**

**Given the arm trajectories in the configuration space, classification is facilitated separately for each arm by means of a combined MLP/RBF neural network structure. The MLP is trained as a standard classifier while the RBF neural network is trained as a predictor for the future state of the system. By feeding the output of the RBF back to the MLP classifier, we achieve temporal consistency and robustness to the classification results.**

## I. INTRODUCTION

Gesture recognition is an important, yet difficult task. It is important because it is a versatile and intuitive way to develop new, more natural and more human-centered forms of human-machine interaction. Moreover, it is difficult because it involves the solution of many challenging subtasks such as robust identification of hands and other body parts, motion modeling, tracking, pattern recognition and classification.

Early psycholinguistic studies [1], [2], initially targeting sign language gestures, revealed that gestures can be characterized based on four different aspects: shape, motion, position and orientation. All gesture recognition approaches try to approach the problem by concentrating one way or another on one or more of the above four aspects. Posture-based approaches, for example, utilize static images, concentrating only on the shape of the hand to extract features such as hand contours, fingertips and finger directions [3], [4], [5], [6]. Temporal approaches, on the other hand, not only make use of spacial features but also exploit temporal information such as the path followed by the hand, its speed, etc [7], [8], [9], [10].

A category of approaches utilize 3D hand models for the detection of hands in images. One of the advantages of these methods is that they can achieve view-independent detection. The employed 3D models should have enough degrees of freedom to adapt to the dimensions of the hand(s) present in an image. Different models require different image features to construct feature-model correspondences. Point and line features are employed in kinematic hand models to recover angles formed at the joints of the hand [11], [12]. In [13], a 3D model of the arm with 7 parameters is utilized. In [14], a 3D model with 22 degrees of freedom for the whole body with 4 degrees of freedom for each arm is proposed. In [15], the user's hand is modeled much more simply, as an articulated rigid object with three joints comprised by the first index finger and thumb.

In this paper we present a specific approach for vision-based hand gesture recognition, intended to support natural interaction with autonomously navigating robots that guide visitors in public places such as museums and exhibition centers. The operational requirements of such an application challenge existing approaches in that the visual perception system should operate efficiently under totally unconstrained conditions regarding occlusions, variable illumination, moving cameras, and varying background. Recognizing that the extraction of features related to hand shape may be very difficult task, we propose a gesture recognition system that emphasizes on the temporal aspects of the task. More specifically, the proposed approach takes into account information conveyed in the trajectory followed by user's arms while the user performs gestures in front of a robot.

The proposed gesture recognition system builds on our previous work in model-based visual tracking of human arms and body [16]. According to this tracking approach, a nine parameter model is employed to track both arms (4 parameters for each arm) as well as the orientation of the human torso (one additional parameter). In order to reduce the complexity of the problem and to achieve real-time performance, the model space is split into three different partitions and tracking is performed separately in each of them. More specifically, a Hidden Markov Model (HMM) is used to track the orientation of the human torso in the 1D space of all possible orientations and two different sets of particles are used to track the four Degrees of Freedom

(DoF) associated with each of the two hands, using a particle filter-based approach.

Given the arm trajectories in the configuration space, classification is facilitated separately for each arm by means of a combined Multi Layer Perceptron/Radial Basis Function (MLP/RBF) Neural Network structure. The MLP is trained as a standard classifier while the RBF neural network is trained as a predictor for the future state of the system. By feeding the output of the RBF back to the MLP classifier, we achieve temporal consistency and robustness in the classification results.

Sample experimental results presented in this paper, confirm the effectiveness and the efficiency of the proposed approach, meeting the robustness and performance requirements of this particular case of human-robot interaction.

## II. APPROACH OVERVIEW

A block diagram of the proposed gesture recognition system is illustrated in Figure 1.

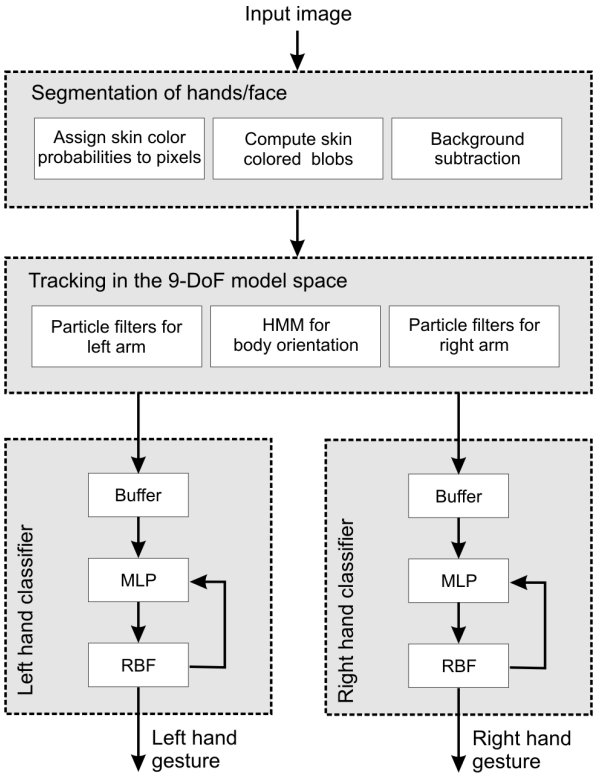


Fig. 1. Block diagram of the proposed approach for hand tracking and gesture recognition. Processing is organized into three layers.

The first step of the approach is to extract hand and face regions as skin-colored foreground blobs.

Assuming a 4 DoF kinematic model for each arm and one additional degree of freedom for the orientation $\phi$ of the user around the vertical axis (see Fig. 2), the pose of the user is tracked in a 9 DoF model space. The resulting 9-parameter tracking problem is tackled in realtime by fragmenting the 9-dimensional space into three sub-spaces; a 1D parameter space for body orientation angle and two 4D spaces, one for each hand. The body orientation angle $\phi$ is appropriately

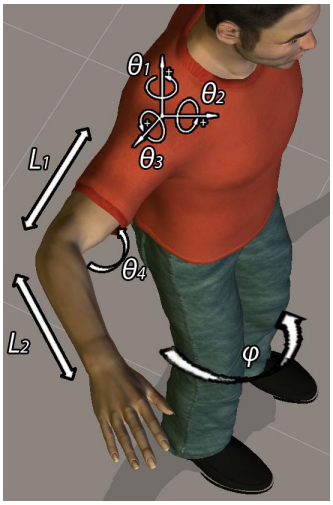


Fig. 2. The 9-parameter model used for the rotation of the body and the pose of the user's arms

quantized and tracked over time by means of an HMM. For every possible solution, a separate particle filter set is employed for each arm. The result of each particle filter is used to estimate the observation probability, which is subsequently employed to update the HMM.

Classification is achieved by buffering the trajectory of each arm (in its 4D configuration space) and feeding it to a feed-forward MLP Neural Network which is trained to recognize between five system states: idle (no gesture), preparation (hand moving towards a gesture), pointing gesture, hello (waiving) gesture, and retraction (hand retracting from a gesture). The output of the MLP is passed though an RBF which is trained as a predictor for the next state of the system and fed back to the MLP in order to improve temporal consistency and robustness of the achieved results.

More details regarding each of the above described modules are provided in the following sections.

## III. DETECTION OF HAND AND FACE BLOBS

The first step of the proposed approach is to detect skin-colored regions in the input images. For this purpose, a technique similar to [17], [18] is employed. Initially, background subtraction [19] is used to extract the foreground areas of the image. Then, for each pixel, $P(s \mid c)$ is computed, which is the probability that this pixel belongs to a skin-colored foreground region $s$, given its color $c$.

This 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. Color $c$ is assumed to be a 2D variable encoding the U and V components of the YUV color space. $P(c \mid s)$ is the prior probability of observing color $c$ in skin colored foreground regions. All three components in the right side of Eq.1 can be computed via offline training.

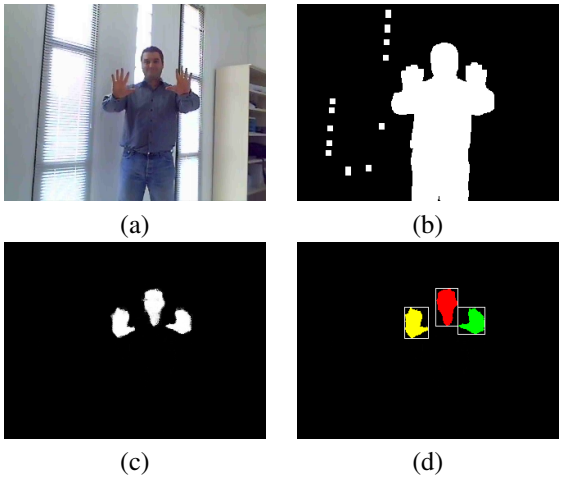

Fig. 3. Blob detection. (a) Initial image, (b) Foreground pixels, (c) skin-colored pixels, (d) resulting skin-colored blobs.

After probabilities have been assigned to each image pixel, hysteresis thresholding is usand connected components labeling are used to extract solid skin color blobs. Pixel 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 blobs. A second thresholding step, this time with a "weak" threshold $T_{min}$ is performed. During this step, pixels with probability $P(s \mid c) > T_{min}$ where $T_{min} < T_{max}$, that are immediate neighbors of already classified 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.

A set of simple heuristics based on location and size is used to characterize blobs as hand blobs and face blobs.

Results of the intermediate steps of this process are illustrated in Fig. 3. Figure 3(a) shows a single frame extracted out of a video sequence that shows a man performing various hand gestures in an office-like environment. Fig. 3(b) shows the result of the background subtraction algorithm and Fig. 3(c) shows skin-colored pixels after hysteresis thresholding. Finally, the resulting blobs (i.e. the result of the labeling algorithm) are shown in Fig. 3(d).

## IV. TRACKING IN THE MODEL SPACE

### A. Kinematic model

As already mentioned, for modeling the human body and arms, a nine-DOF model, has been employed. This model, which is similar to the one proposed in [20] is depicted in Figure 2. According to this model, the human body, with the exception of the arms, is assumed to be a rigid object with only one degree of freedom corresponding to its orientation $\phi$. Both arms are assumed to be attached to this rigid body at fixed locations (i.e. the shoulders) and they are modeled by a 4-DoF kinematic model each. The kinematics of each arm are defined as Denavit-Hartenberg parameters, shown in table I. $\theta_1$,$\theta_2$ and $\theta_3$, are the angles corresponding to the three DoFs of the human shoulder and $\theta_4$ corresponds to the angle of the elbow. $L_1$ and $L_2$ are the lengths of the upper arm and the forearm, respectively. They are assumed fixed in our implementation.

TABLE I

DENAVIT-HARTENBERG PARAMETERS FOR THE 4-DOF MODEL OF THE HUMAN ARM EMPLOYED IN OUR APPROACH.

| $i$ | $\alpha_{i-1}$ | $a_{i-1}$ | $d_i$ | $\theta_i$ |
|---|---|---|---|---|
| 1 | $+\pi/2$ | 0 | 0 | $\theta_1 - \pi/2$ |
| 2 | $-\pi/2$ | 0 | 0 | $\theta_2 + \pi/2$ |
| 3 | $+\pi/2$ | 0 | $L_1$ | $\theta_3 + \pi/2$ |
| 4 | $-\pi/2$ | 0 | 0 | $\theta_4 - \pi/2$ |
| 5 | 0 | $L_2$ | 0 | 0 |

### B. Model space partitioning and tracking

To track in the 9-DoF model space presented in the previous section, the approach presented in [16] has been assumed. According to this approach, in order to reduce the complexity of the problem and meet the increased computational requirements of the task at hand, the model space is split into three different partitions and tracking is performed separately in each of them. More specifically, a Hidden Markov Model (HMM) is used to track the orientation $\phi$ of the human body in the 1D space of all possible orientations and two different sets of particles are used to track the four DoFs associated with each of the two arms using a particle filtering approach.

To facilitate the implementation of the HMM, the body orientation angle $\phi$ is appropriately quantized (50 quantization levels were used in our implementation). For every possible solution, a separate particle filter set is employed for each arm. The result of each particle filter is used to estimate the observation probability, which is subsequently employed to update the HMM. This means that the weights of the particles are used to calculate the observation likelihood for a particular body orientation state.

To facilitate the implementation of likelihood function which is necessary in order to evaluate hypotheses in the particle filter-based trackers, the kinematic model defined in the previous section is used, along with the camera perspective transformations. More specifically, forward kinematic equations are used to transform the rotation of the human body and the angles of the arm joints to 3D coordinates for each joint (shoulder, elbow and hand). Accordingly, camera projection transformations are used to project the resulting 3D coordinates of the joints on the image frame. The projected joint locations are evaluated by comparing them with actual observations according to two different criteria: (a) Projected hand locations should be close to observed skin-colored blobs, and (b) projected elbows and shoulders should be within foreground segments of the image.

Figures 4(a) and 4(b) demonstrate the operation of the particle filter trackers that correspond to a specific value of the orientation angle ("0" in both cases). On the right

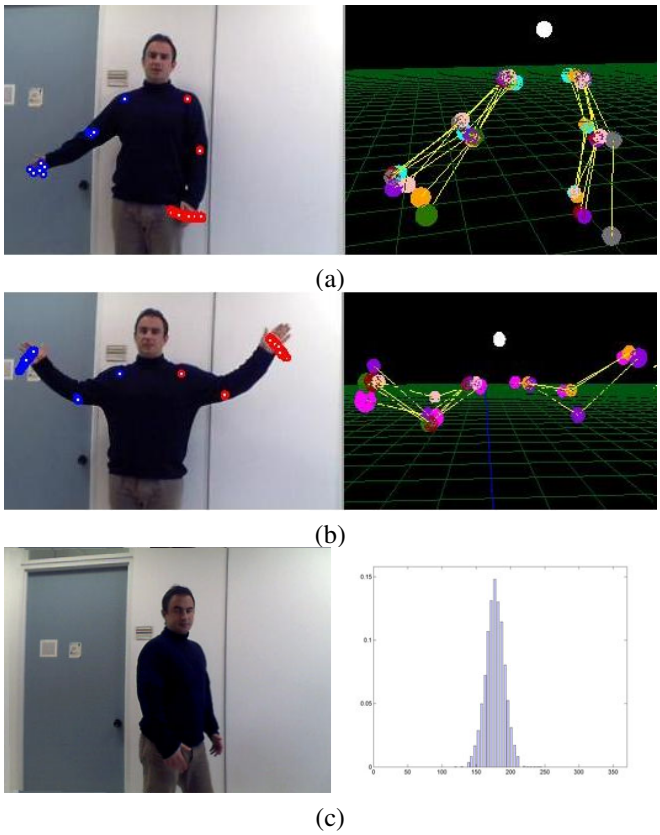

Fig. 4. Operation of the tracker; (a,b) Particle filter sets for a specific orientation angle, (c) A HMM histogram corresponding to a specific frame.

parts of the two images are the samples projected on the 3D space (using forward kinematics, as described above). The corresponding sample projections on the image plane are depicted on the left. Figure 4(c) depicts a sample orientation histogram as tracked by the HMM. The values of each histogram cell correspond to the probability of this specific orientation being the correct orientation.

## V. GESTURE CLASSIFICATION

As observed in [21], gestures are dynamic processes that typically consist of three phases: preparation, stroke and retraction. The preparation and retraction phases consist of arm movement from and towards the resting position, before and after the gesture, respectively. These phases have been found to be similar in content between many common gestures and therefore contribute little to the gesture recognition process. The stroke phase is the one that contains most of the information that characterizes a gesture.

Based on the above observations our system has been designed to recognize five different gesturing states:

- Idle. No gesture is performed,
- Preparation phase.
- Pointing gesture,
- Hello gesture. The user is waiving using his hand.
- Retraction phase.

The mentioned states correspond to two different strokes (pointing and hello gestures), the accompanying phases (preparation and retraction) and the idle phase. The transitions between the above-mentioned states are illustrated in Figure 5.

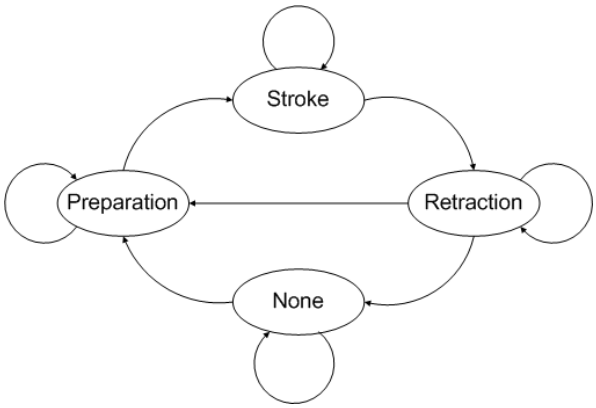


Fig. 5. Gesture state transitions.

Classification is achieved by buffering the trajectory of each arm (in its 4D configuration space) and feeding it to a feed-forward MLP Neural Network which is trained to recognize between the five system states. The output of the MLP is passed though an RBF neural network which is trained as a predictor for the next state of the system and fed back to the MLP in order to improve temporal consistency and robustness of the achieved results.

This classification structure is graphically illustrated in the lower part of Figure 1. As implied in Figure 1, the same structure is employed separately for gesture recognition in each arm.

### A. The MLP

The MLP neural network is mainly responsible for gesture recognition and classification, according to the trained patterns. It consists of the input layer, the output layer and two hidden layers. The output layer consists of 4 neurons encoding the five possible states of the system. The input layer consists of 44 neurons; 40 neurons are used to provide input about the trajectory of the arm (4 parameters per frame, 10 frames history) and the rest 4 neurons are used to provide the prediction for the next state of the system, which is fed back by the RBF neural network.

### B. The RBF

The input layer of the RBF network consists of four neurons which are connected to the output of the MLP. The output of the RBF Network also consists of four neurons and it is fed back to the MLP. Given the output of the MLP, the RBF is trained to provide a prediction for the next state of the system.

The intuition behind this is simple: one can think of a gesture as a state transition process with acceptable and unacceptable state transitions (Figure 5). However, due to possible discontinuities in the MLP input data (caused by erroneous tracking or lost frames in the video), the output (of the MLP) can itself present discontinuities, translated

to unacceptable state transitions, as well. The RBF network restricts unacceptable transitions and smooths out outliers at the output of the MLP.

### C. Network training

For training the proposed classifier, a dataset consisting of 12 sequences was used. This dataset contains six examples of each of the two cosidered gestures. In each of these sequences all three phases of a gesture appear, together with cases where none of the phases is performed or when both hands are acting simultaneously. The dataset was divided into two subsets, of 6 sequences each. The first subset contained 3 sequences from each of two gestures and it was used to train the MLP neural network while the second subset was used to train the RBF network. Using the two subsets, the training of the system was done in two steps.

Training of the MLP was performed by minimizing the mean of the squared error using the Levenberg-Marquardt algorithm. To train the RBF network, the sequences used for training the MLP cannot be used because they are known to the classifier. Thus, the second training set is used.

## VI. RESULTS

During our experiments, 3 sequences for each gesture, different from the ones used during the training step, have been tested. The examined scenarios contained both gestures performed by one arm only and by both hands simultaneously. Our main target was to study whether sequences of arm kinematic configurations contain enough information to describe a gesture, given the fact that no other information about the location of the arm has been used.

The proposed approach performed very well in all test cases. Four illustrative examples are depicted in Figure 6. In Figures 6(a) and 6(b) the user performs a right hand gesture that is correctly classified by the employed Neural Network structure. Figures 6(c) and 6(d) present two additional examples where the user gestures with both hands simultaneously.

Table II presents quantitative results obtained with the employed datasets. The TP figures shown in table II correspond to the percentage of correctly classified frames (True Positive classifications). Similarly, FP and FN figures correspond to percentages of false positive and false negative classified frames.

As can easily be observed, the successful recognition ratio does not drop below 86% while the false negative percentage remains in low levels as well. Further experiments have been conducted by eliminating the RBF neural network from the classification structure. In these cases the percentage of false positive decisions for the preparation and retraction phase was higher than 15%. Evidently, the utilization of the RBF network has greatly contributed to the robustness of the classifier by filtering out temporal inconsistencies in the output of the MLP.

## VII. CONCLUSION

In this paper, we have presented a novel temporal gesture recognition system intended for natural interaction with

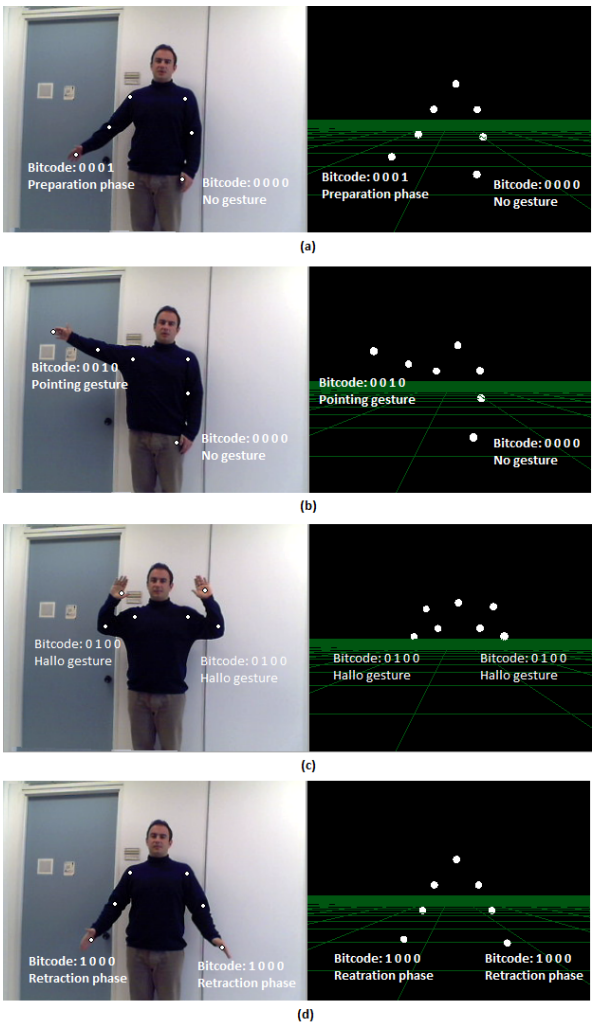


Fig. 6. Recognition and classification of gestures performed by one or both hands simultaneously. The left image depicts a 2D view from one camera of the stereo pair, while the right image shows the 3D representation (of the left image). The output of the classifier has been superimposed on the images for the sake of clarity. (a)The right arm prepares to perform a gesture. (b)The right hand performs a "pointing" gesture. (c)Both hands perform a "hello" gesture. (d)Both hands retract from the stroke phase.

TABLE II

GESTURE CLASSIFIER QUANTITATIVE RESULTS. TP:TRUE POSITIVES, FP: FALSE POSITIVES, FN: FALSE NEGATIVES.

| Preparation | | | Pointing | | |
|---|---|---|---|---|---|
| TP | FP | FN | TP | FP | FN |
| 88.46% | 11.54% | 6.47% | 86.48% | 13.52% | 2.08% |
| **Hello** | | | **Retraction** | | |
| TP | FP | FN | TP | FP | FN |
| 96.91% | 3.09% | 1.41% | 86.04% | 13.96% | 6.21% |

autonomous robots that guide visitors in museums and exhibition centers. The proposed gesture recognition system builds on our previous work on vision based tracking and more specifically on a probabilistic tracker capable to track both hands and the orientation of the human body on a nine-parameter configuration space.

Dependable tracking, combined a novel, two-stage neural network structure for classification, facilitates the definition of a small and simple hand gesture vocabulary that is both robustly interpretable and intuitive to humans. Experimental results presented in this paper, confirm the effectiveness and the efficiency of the proposed approach, meeting the run-time requirements of the task at hand.

Nevertheless, and despite the vast amount of relevant research efforts, the problem of efficient and robust vision-based recognition of natural hand gestures in unprepared environments still remains open and challenging, and is expected to remain of central importance in human-robot interaction in the forthcoming years. In this context we intend to continue our research efforts towards enhancing the current system. At first we plan to redesign the classification structure in order to take into account the multiple hypotheses provided by the employed tracker. This is expected to increase classification accuracy since errors in the early processing stages (tracking) are not propagated to later stages (classification). Additionally the training and test datasets will be expanded to include richer gesture vocabularies and larger intra-gesture variation. Finally, we intend to include a more sophisticated algorithm to classify skin colored blobs to hands and faces. This will allow our system to cope with more complex cases where multiple users simultaneously interact with the robot.

## VIII. 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] W. C. Stokoe, *Sign Language Structure*. Buffalo, NY: Buffalo Univ. Press, 1960.

[2] R. Battison, "Phonological deletion in american sign language," *Sign language studies*, vol. 5, no. 1, pp. 1–19, 1974.

[3] V. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[4] J. J. LaViola, "A survey of hand posture and gesture recognition techniques and technology," Department oc Computer Science, Brown University, Providence, Rhode Island., Tech. Rep. CS-99-11, 1999.

[5] Y. Wu and T. S. Huang, "Vision-based gesture recognition: A review," *Lecture Notes in Computer Science*, vol. 1739, pp. 103+, 1999.

[6] 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. Lawrence Erlbaum Associates, Inc. (LEA), to appear.

[7] A. Wilson and A. Bobick, "Parametric hidden markov models for gesture recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, Sept. 1999.

[8] S. Wang, A. Quattoni, L. Morency, and D. Demirdjian, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. II: 1521–1527.

[9] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[10] H. Suk, B. Sin, and S. Lee, "Robust modeling and recognition of hand gestures with dynamic bayesian network," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[11] J. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects," in *Proc. International Conference on Computer Vision (ICCV)*, 1995, pp. 612–617.

[12] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura, "Hand gesture estimation and model refinement using monocular camera - ambiguity limitation by inequality constraints," in *IEEE Int. Conf. on Face and Gesture Recognition*, Nara, Japan, 1998, pp. 268–273.

[13] L. Goncalves, E. di Bernardo, E. Ursella, and P. Perona, "Monocular tracking of the human arm in 3D," in *Proc. International Conference on Computer Vision (ICCV)*, Cambridge, 1995, pp. 764–770.

[14] D. Gavrila and L. Davis, "3-D model-based tracking of humans in action: a multi-view approach," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 1996, 1996, pp. 73–80.

[15] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *Proc. European Conference on Computer Vision*, 2000, pp. 3–19.

[16] M. Sigalas, H. Baltzakis, and P. Trahanias, "Visual tracking of independently moving body and arms," in *Proc. IEEE/RSJ Intlernational Conference on Intelligent Robotics and Systems (IROS)*, St. Louis, MO, USA, Oct. 2009.

[17] A. A. Argyros and M. I. A. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera." in *Proc. European Conference on Computer Vision*, Prague, Chech Republic, May 2004, pp. 368–379.

[18] 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.

[19] W. E. L. Grimson and C. Stauffer, "Adaptive background mixture models for real time tracking," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Ft. Collins, USA, June 1999, pp. 246–252.

[20] D. Tsetserukou, R. Tadakuma, H. Kajimoto, and N. Kawakami, "Development of a whole-sensitive teleoperated robot arm using torque sensing technique," in *WHC '07: Proceedings of the Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 476–481.

[21] A. Kendon, "Current issues in the study of gesture," *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, pp. 23–47, 1986.