Gesture recognition based on arm tracking for human-robot interaction

Markos Sigalas, Haris Baltzakis and Panos Trahanias
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH)
E-mail: {msigalas, xmpalt, trahania} @ ics.forth.gr

***Abstract*—In this paper we present a novel approach for hand gesture recognition. The proposed system utilizes upper body part tracking in a 9-dimensional configuration space and two Multi-Layer Perceptron/Radial Basis Function (MLP/RBF) neural network classifiers, one for each arm. Classification is achieved by buffering the trajectory of each arm and feeding it to the MLP Neural Network which is trained to recognize between five gesturing states. The RBF neural network is trained as a predictor for the future gesturing 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.**

The proposed approach has been assessed using several video sequences and the results obtained are presented in this paper.

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. At the same time, 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. Therefore, all gesture recognition approaches try to approach the problem by concentrating 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 spatial features but also exploit temporal information such as the path followed by the hand, its speed, etc [7], [8], [9], [10]. Additionally, there is strong neurobiological evidence which indicates that ongoing actions are interpreted by the human visual system into sequences of motor primitives (primitives of the human body motor control) [11]. Based on this idea, various gesture recognition approaches [12], [13], mostly within the *robotics* community, model and classify gestures according to the acquired motor primitives.

Depending on the application needs, various classification tools have been used for the recognition of gestures. Some approaches [14], [15] utilize Hidden Markov Models (HMMs) for the recognition process, as they offer rich mathematical structures and provide efficient spatio-temporal modeling. Particle filters have been also used in some gesture recognition works [16], [17], due to their ability of real-time estimation of nonlilnear, non-Gaussian dynamic systems. Furthermore, a set of approaches model gestures as ordered sequences of states in a spatio-temporal configuration and classify them with a finite-state machine (FSM) [18], while others exploit the adaptation ability of Neural Networks (NNs) for the recognition of gestures. Some of the the most commonly used NNs are Multi-Layer Perceptron (MLP) [19], Time-Delay NNs [20] and Radial Basis Function (RBF) NNs [21].

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 a very difficult task, we propose a gesture recognition system that emphasizes on the temporal aspects of the task. More specifically, the proposed approach exploits the extracted arm motor primitives conveyed in the trajectory followed by user's arms, while the user performs gestures in front of a robot.

The proposed gesture recognition system is built upon the human body tracker that we have proposed in [22]. 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, an 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.

In the current work we propose a new method to classify the arm trajectories, that is the sequences of motor primitives (joint angles), produced by the above mentioned tracker, into gestures 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 in the classification results.

II. Approach overview

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

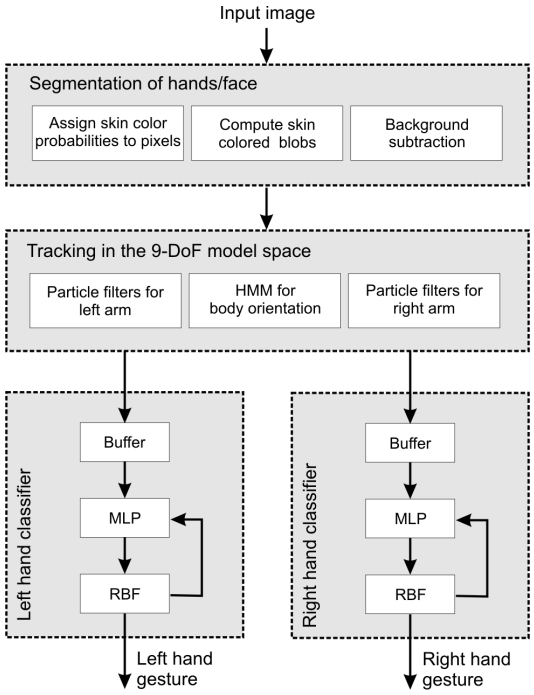


Figure 1. Block diagram of the proposed approach for hand tracking and gesture recognition. Processing is organized into three layers.

The first step of the approach is to extract hand and face regions as skin-colored foreground blobs. Then, assuming a 4 DoF kinematic model for each arm and one additional degree of freedom for the orientation ϕ of the user around the vertical axis (see Figure 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 ϕ is appropriately quantized and tracked over time by means of an HMM.

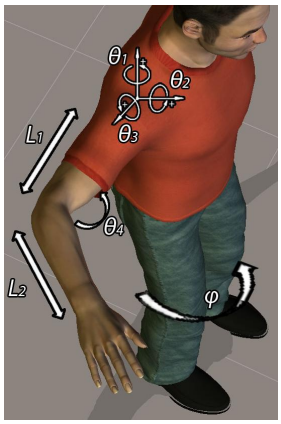


Figure 2. The 9-parameter model used for the rotation of the body and the pose of the user's arms.

Table I
Denavit-Hartenberg parameters for the 4-DoF model of the human arm employed in our approach.

i	α_{i-1}	a_{i-1}	d_i	θ_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

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. Specifically, due to the fact that the tracker has been presented in a previous work of ours [22], it will be briefly described in section III. A thorough presentation of the classifier structure and training follows in section IV, while section V contains some illustrative experimental results of the proposed gesture classification approach.

III. Upper body tracking

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 [23] 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 ϕ. 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. θ_1, θ_2 and θ_3, are the angles corresponding to the three DoFs of the shoulder and θ_4 corresponds to the angle of the elbow. L_1 and L_2 are the lengths of the upper arm and the forearm, respectively, and are assumed fixed in our implementation.

B. 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 [24], [25] is employed. Initially, background subtraction [26] is used to extract the foreground

areas of the image. Then, the probability of belonging to a skin-colored foreground area is computed for each pixel, while a connected components labeling algorithm is used to assign different labels to pixels that belong to different blobs. Finally, a set of simple heuristics based on location and size is used to characterize blobs as hand blobs and face blobs.

C. Model space partitioning and tracking

To track in the presented 9-DoF model space, the approach presented in [22] 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 ϕ of the human body in the 1D space and two different sets of particles are used to track the four DoFs associated with each of the two arms.

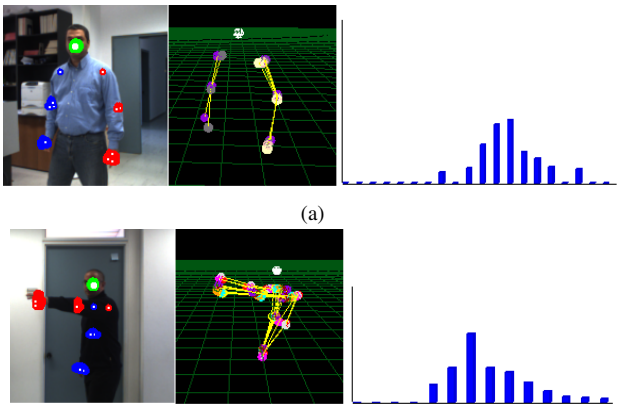
The body orientation angle ϕ is appropriately quantized and, for every possible solution, a separate particle filter set is employed for each arm. The result of each particle filter is used to estimate the observation probability, which is subsequently employed to update the HMM. In other words, the weights of the particles are used to calculate the observation likelihood for a particular body orientation state.

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

Figures 3(a) and 3(b) demonstrate the operation of the tracker at hand. Left images illustrate the operation of the particle filter trackers that correspond to a specific value of the orientation angle while the images in the center contain 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 images. Finally, sample orientation histograms, as tracked by the HMM, are depicted in the images on the right. The values of each histogram cell correspond to the probability of this specific orientation being the correct orientation.

IV. Gesture Classification

As observed in [27], 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.



(a)

(b)

Figure 3. Operation of the tracker; The first image of each row depicts the particle filter sets for a specific orienatation angle. The second image shows multiple hypotheses generated from the tracker, while the third one, illustrates a HMM histogram corresponding to a specific frame.

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/her arm.
- 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. As illustrated, apart from the normal preparation-stroke-retraction state transition, an arm may perform two sequential gestures without the need to pass from the resting position.

The motor primitives, provided by the above described tracker, are used in order to model each of the gesture states and, therefore, to train the gesture classifier. This implies that each gesture state is modeled as a sequence of arm joint angles. This representation facilitates the classification proccess, since it is not affected by factors such as the shape and size of the human and the arm, the body orientation, or the speed and duration of the performed gesture.

As described above, the employed tracker provides multiple configuration hypotheses. However, in this work we assume that the pose with the highest probability is the one closest to the actual pose of the human in the scene. Therefore, the sequence of joint angles of the most probable arm will be fed as input to the classifier.

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 through an RBF neural network which is trained as a predictor for the next state of the system

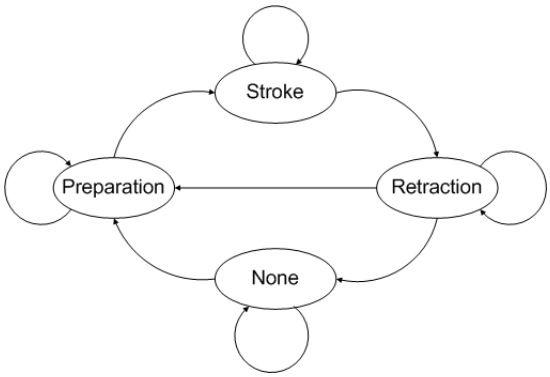


Figure 4. Gesture state transitions.

and fed back to the MLP in order to improve temporal consistency and robustness of the achieved results. In other words, the MLP is responsible for the initial classification of the gesture, while the RBF assures for valid state transitions. 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. Since gestures are dynamic procedures, i.e. having temporal context, a sort of history is needed to be provided to the classifier. In our experiments, a history of 10 frames proved to be adequate. Therefore, 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. Finally, the output layer consists of 4 neurons encoding the five possible states of the system.

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 4). 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 into 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 considered 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 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 has been 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.

V. Experimental results

We conducted a set of experiments, using 9 sequences for each gesture, different from the ones used during the training step. 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 training dataset is composed of a set of simple sequences depicting a single human performing various gestures, as illustrated in Figure 5. To evaluate the results of our approach we used a different dataset containing sequences of various humans in different scenes. The proposed approach performed very well in all test cases, even with the presence of large intra-gesture variations in duration, speed and/or arm trajectory. This provides a strong evidence that despite the differences on the way a gesture is performed, the sequence of motor primitives remains the same amongst the same gestures

Eight illustrative examples are depicted in Figure 6. The presented experiments, have been conducted upon 4 different sequences, containing various indoor scenes with different persons performing gestures with one or both arms simultaneously. As observed, the proposed classification approach was able to successfully cope with intra-person variabilities in shape, size and orientation, as well as with the aforementioned intra-gesture variations. This conclusion also implies that the classification is not affected by small tracking errors. For example in Figures 6(g) and 6(h) the arm performing the Pointing gesture is not completely extended, however the recognition was successful. That is because its not the accuracy in joint angle computations which is important but, rather, the abstract pattern of the arm's trajectory. For the sake of clarity, the output of the classifier (bitcode) has been superimposed on the images. This bitcode is a representation of the output of the neural network classifier.

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

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

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

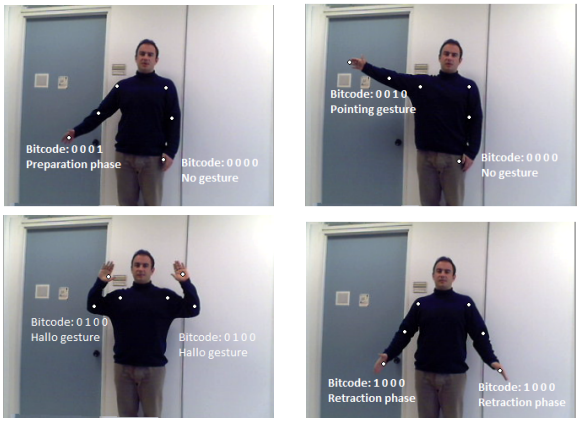


Figure 5. Sample images from the training dataset.

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%. This is justified by the fact that these phases are practically identical (with reverse trajectories) and the lack of state transition control would lead to misclassifications -this is not the case for the Hello gesture since its trajectory is easily distinguished. 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.

VI. CONCLUSIONS

In this paper we have presented a novel temporal gesture recognition system intended for natural interaction with autonomous robots that guide visitors in museums and exhibition centers. The proposed gesture recognition system builds on our previous work 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 with 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. Additionally, the use of motor primitives (joint angles), for the modeling and classification of each gesture, provides shape, size and orientation invariance as well as gesture speed and duration independability. 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.

VII. 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] R. Battison. Phonological deletion in American sign language. *Sign Language Studies*, 5:1–19, 1974.
[2] WC Stokoe. Sign language structure. *Annual Review of Anthropology*, 9(1):365–390, 1980.
[3] V.I. 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*, 19(7):677–695, 1997.
[4] J.J. LaViola Jr. A survey of hand posture and gesture recognition techniques and technology. Technical Report CS-99-11, Brown University Providence, RI, USA, 1999.
[5] Y. Wu, T.S. Huang, and N. Mathews. Vision-based Gesture Recognition: A Review. *Lecture Notes in Computer Science*, 1999.
[6] X. Zabulis, H. Baltzakis, and A. Argyros. Vision-based hand gesture recognition for human-computer interaction. In *The Universal Access Handbook*, Human Factors and Ergonomics. Lawrence Erlbaum Associates, Inc. (LEA).
[7] A.D. Wilson and A.F. Bobick. Parametric Hidden Markov Models for Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 884–900, 1999.
[8] S.B. Wang et al. Hidden conditional random fields for gesture recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2006.
[9] L.P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[10] H.I. Suk, B.K. Sin, and S.W. Lee. Robust Modeling and Recognition of Hand Gestures with Dynamic Bayesian Network. In *19th International Conference on Pattern Recognition (ICPR) 2008.*, pages 1–4, 2008.
[11] T.B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006.
[12] M.C. Lopes et al. Motor representations for hand gesture recognition and imitation. In *IROS Workshop on Robot Programming by Demonstration*. Citeseer, 2003.
[13] A.L. Edsinger. *A Gestural Language For A Humanoid Robot*. PhD thesis, Massachusetts Institute of Technology, 2001.

Figure 6. (a) Right hand performs a pointing gesture and then it is being retracted (b). (c)Left arm performs a pointing gesture while the human is slightly turned away from the camera. (d)Both arms prepare to perform a gesture. (e)Hallo gesture performed by the right arm. (f) Right arm performs a pointing gesture while the human is turned by 90 degrees from the camera and the left arm lies within the retraction state. (g) Right arm performs a pointing gesture while the left one prepares itself to perform a hallo gesture (h).

[14] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *Proc. Comp. Vis. and Pattern Rec*, pages 379–385, 1992.

[15] R. Bowden et al. A linguistic feature vector for the visual interpretation of sign language. *Computer Vision-ECCV 2004*, pages 390–401, 2004.

[16] M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. *Computer Vision - ECCV 98*, page 909, 1998.

[17] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002. Proceedings*, pages 423–428, 2002.

[18] P. Hong et al. Gesture Modeling and Recognition Using Finite State Machines. In *Procs. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*.

[19] Y.K.Ahn et al. Implementation of 3D gesture recognition system based on neural network. In *Procs. of the 9th WSEAS international conf. on Applied informatics and communications*, pages 84–87, 2009.

[20] M.H. Yang and N. Ahuja. Recognizing hand gesture using motion trajectories. In *CVPR 1999*.

[21] L.Gerlich et al. Gesture recognition for control of rehabilitation robots. *Cognition, Technology & Work*, 9(4):189–207, 2007.

[22] 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)*, 2009.

[23] D. Tsetserukou et al. Development of a Whole-Sensitive Teleoperated Robot Arm using Torque Sensing Technique. In *Procs. of World Haptics 2007*, pages 476–481. IEEE Computer Society, 2007.

[24] A.A. Argyros and M.I.A. Lourakis. Real-time tracking of multiple skin-colored objects with a possibly moving camera. *Lecture Notes in Computer Science*, pages 368–379, 2004.

[25] H. Baltzakis, A. Argyros, M. Lourakis, and P. Trahanias. Tracking of human hands and faces through probabilistic fusion of multiple visual cues. *Lecture Notes in Computer Science*, 5008:33–42, 2008.

[26] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[27] A. Kendon. Current issues in the study of gesture. *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, 1986.