Probabilistic Video-Based Gesture Recognition Using Self-Organizing Feature Maps

George Caridakis, Christos Pateritsas, Athanasios Drosopoulos,
Andreas Stafylopatis and Stefanos Kollias

School of Electrical and Computer Engineering, National Technical University of Athens,
Politechnioupoli, Zographou, Greece
gcari@image.ece.ntua.gr, pater@softlab.ece.ntua.gr, ndroso@image.ece.ntua.gr
{andreas, stefanos}@cs.ece.ntua.gr

Abstract. Present work introduces a probabilistic recognition scheme for hand gestures. Self organizing feature maps are used to model spatiotemporal information extracted through image processing. Two models are built for each gesture category and, along with appropriate distance metrics, produce a validated classification mechanism that performs consistently during experiments on acted gestures video sequences.

Keywords: hand tracking, gesture recognition, gesture classification, self organizing feature map, markov processes

1 Introduction

Gesture recognition continuously receives abundant attention, especially throughout the research fields of sign language recognition, multimodal human computer interaction, cognitive systems and robotics. Renewed focus on interdisciplinary studies lead scientists to review and confront the questions raised when attempting to model and extract the information that a gesture conveys. Since hand gestures can be used for a wide variety of communicative purposes, classification becomes a significant problem, starting at the level of defining gesture taxonomy through psychological studies. Most commonly, gesturing behavior can be classified on a spectrum that ranges from highly structured languages (e.g. sign languages), through universal symbols, to natural and unconscious gesticulation [1]. Studies also show that gesture classification is, in general, a multimodal task that should make use of both hand movement trajectories and linguistic cues [2], [3].

In terms of computer vision, appropriate feature extraction and tracking is the focus of many researchers, in order to apply classification schemes for the hand trajectory and/or the hand shape. Depending on the scope of a study, approaches vary from multimodal interpretation (gesticulation, natural language, facial expression, domain knowledge, etc.) to gesture classification through a single modality. Present work deals with the classification of gestures from visual cues, focusing on robustness, performance and user independence. Aiming for naturalistic data, our intention is to localize and track hands, classifying their trajectories regardless of the hand shape.

An extensive review of several techniques is presented both in [4] and [5]. The first focuses mainly on SL recognition and classification issues, while examining closely hand localization and tracking, and on various feature extraction techniques related to automatic analysis of manual signing. In addition, it addresses the linguistic aspect of SL and non manual signals, along with methodologies to incorporate these in the SL recognition chain. On the other hand, Wu and Huang delve more into works related to hand modeling (shape analysis, kinematics chain and dynamics) and computer vision, and pattern recognition issues associated to hand localization and feature extraction from image sequences. Classification schemes involve several methods, depending on the features and the stages of the procedure. Methods used include neural networks and variants, hidden markov models and variants, principal component analysis, and numerous other machine learning methods or combinations (decision trees, template matching, etc.).

One of the most commonly proposed approaches involves feature extraction from the input signal and utilization of these features as input for a fine tuned HMM [6], [7]. In addition, variations of the previous group have been widely adopted [8], [9]. Other approaches employ alternate machine learning and artificial intelligence techniques such as recurrent fuzzy network [10], time delay neural network [11], finite state machines [12], Bayesian classifiers [13], etc. Finally, there have been several efforts combining more than one technique. Mantyla et al. [14] present a system for static gestures recognition using a self-organizing mapping scheme, while a hidden Markov model is used to recognize dynamic gestures. Black and Jepson [15] present an extension of the "condensation" algorithm, modeling gestures as temporal trajectories of the velocity of the tracked hands. Fang et al. [16] present an additional layer enhancing the HMM architecture with SOFM and improving their recognition rate by 5%, while introducing a fuzzy decision tree in an attempt to reduce the search space of recognized classes without loss of accuracy.

Present work introduces a novel approach for applying a combination of self organizing maps and markov models for gesture classification. The features extracted include the trajectory of the hand and the direction of motion in the various stages of the gesture. The classification scheme is based on the transformation of a gesture representation from a series of coordinates and movements to a symbolic form and on building probabilistic models using these transformed representations. Our study indicates that, although each of the two sets of features (trajectory and motion direction) can provide distinctive information in most cases, only an appropriate combination can result in robust and confident user independent gesture recognition.

2 System Overview

The steps of the introduced procedure, which is depicted in Fig. 1, begin with an image processing module. Taking into consideration computational cost and robustness, we employed an accurate, near real-time skin detection and tracking module [17] allowing a rate of around 12 fps (frames per second) on a usual PC configuration, which is adequate for continuous gesture tracking. The process involves the creation of moving skin masks and tracking their centroids to produce an estimate of the user's movements. The object correspondence heuristic makes it possible to individually track the hand segments correctly while the fusion of color and motion information eliminates any background noise or artifacts.

Following, each gesture instance is represented by a time series of points, representing the hand's location with respect to the head of the person performing the gesture. Consequently, a gesture G_i containing l points can de expressed as an ordered set of points:

$$G_i = \{(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)\} \ , \qquad \mathbf{(1)}$$

where l varies across different gesture instances. The system's input is a set of gestures D, assigned to c different categories.

The proposed modeling scheme is based on the transformation of a gesture representation from a series of coordinates and movements to a symbolic form which, in turn, is used to build the respective probabilistic models. The first transformation is based on the relative position of the hand during the gesture and is achieved using a self-organizing map model. Despite the fact that the map units are treated as symbols, the map's neighborhood function provides a distance metric between them, that is used during the classification of an unlabeled gesture. Additionally, this enables the use of the Levenshtein distance metric for the comparison between these sequences of symbols and the definition of a "mean" string of symbols representing e.g. the gestures included in a D_j set.

The second transformation is based on the optical flow of the gesture, aiming to describe the gesture's direction changes. The symbols generated from this transformation constitute the set of angles of the gesture's trajectory. This set is limited to quantized values that are treated as symbols in order to be used for the creation of an additional set of Markov models.

For the classification of an unlabeled gesture, the Markov models created from the first transformation play the primary role, while the models created from the second transformation are used for validation and decisions in cases of low confidence classification.

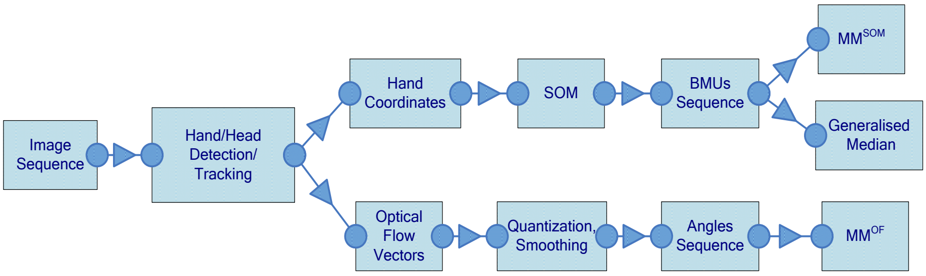


Fig. 1. Building gesture models from transformed gesture representations.

3 Probabilistic Models of Hand Movement

The coordinates of all the points from all the gestures are used to train a hexagonal, two-dimensional grid SOM with the batch mode learning procedure. The points are fed to the map in an unordered form, inconsequently to the gesture instance they belong to and to their ranking position into the gesture. Following training, each point is assigned to the respective best matching unit (BMU) on the map, i.e. the unit of the map closer to the point in the input data space, according to the Euclidean distance of the two vectors. Thus, a gesture G_{i} can be transformed from a series of points to a series of map units.

$$T(G_i) = (u_1, u_2, ..., u_l), \text{ where } u_i = BMU(x_i, y_i) \ . \qquad \mathbf{(2)}$$

Function $BMU(x_{\mathrm{i}},y_{\mathrm{i}})$ returns the index of the best-matching unit for point (x_{i},y_i) and $T(G_{\mathrm{i}})$ is the modified gesture representation. Given that u_{i} is the index of a map unit, this function can be is declared as BMU:R^2 ->S, where S is the set of the indices of all map units and can be treated as a set of symbols. In many cases, the u_{i} value of consequent points of a gesture remains the same since, although the continuous movement of the hand is represented by the distinct points, consequent points are generally close in the input data space. Replacing consequent equal values of u_{i} with a single value results in the following gesture definition,

$$G_i^{'} = N(T(G_i)) = \{u_1, u_2, ..., u_m,\} : m \le l, \forall t \in [2,l] \ u_t \ne u_{t-1} \ , \qquad \mathbf{(3)}$$

where N is a function that removes consecutive equal u_{i} values and $G_{\mathrm{i}}^{'}$ is the transformed gesture instance. The transformation of the gestures with the use of the SOM can be considered a transformation of the continuous trail to a sequence of m discrete symbols, different for every gesture class, that define the finite states to build first order Markov chain models.
Such a model, for each of the categories in the gestures' data set, is created. The sequence of the u_{i} values into the transformed gestures $G_{\mathrm{i}}^{'}$ of $D_{\mathrm{j}}^{'}$ set, will be used for the calculation of the transition probabilities of the model MM_j^{som} describing the j category and for the determination of the values of the function π_j^{som}, which is the first state probability function of this model. The result is a set MM^{som} of c Markov models.

$$MM^{som} = \{MM_1^{som}, MM_2^{som}, ..., MM_c^{som}\} : D_i^{'} = \{G_1^{'}, G_2^{'}, ..., G_n^{'}\} \to MM_i^{som} \qquad \mathbf{(4)}$$

These models are used to evaluate a new unlabeled gesture in order to be classified in one of the c categories. Fig. 2 depicts the above described transformation for a gesture instance.

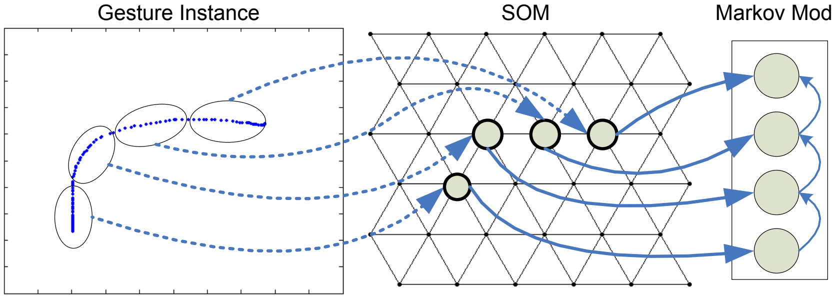


Fig. 2. Correspondence of gesture trajectory points to their respective BMUs on the SOM. These BMUs constitute the states of the Markov models.

With the purpose of providing a more descriptive representation of each gesture instance, an additional transformation is introduced, based on the optical flow of each gesture. This describes the different directions that the gesture trajectory presents instead of the spatial position of gesture points. In order to achieve such a representation, direction vectors are calculated from the consecutive gesture trajectory points. These angles are then quantized in 8 different symbolic values as depicted in Fig. 3. The segments of coordinates in Fig. 2 and Fig. 3 are considered to be a set of coordinates that belong to the same cluster (BMU and Quantized Angle for Fig. 2 and Fig. 3 respectively). In that sense, we define the transformation of a gesture instance G_i using the OF function as:

$$OF(G_i) = \{v_1, v_2, ..., v_m\} : v_i = W_r(Q(\arctan(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}))) \ , \qquad \textbf{(5)}$$

where v_i are the quantized values, Q the quantization function and W_r a median function applied to the values of a fixed length window around the input value. The purpose of the later is to smooth the quantized values against possible instabilities of the hand during the gesture. Applying the transformation function along with function N (eq. 3) for the removal of the equal consecutive values we get

$$G_i^{"} = N(OF(G_i)) = \{v_1, v_2, ..., v_m\} \qquad \textbf{(6)}$$

The v_i values define the states for a new set of Markov models MM^{of} that is built using the transformed set $D_\mathrm{j}^{''}$. The first state probability function π_j^{of} is also calculated using this set.

$$MM^{of} = \{MM_1^{of}, MM_2^{of}, ..., MM_c^{of}\} : D_i^{"} = \{G_1^{"}, G_2^{"}, ..., G_n^{"}\} \rightarrow MM_i^{of} \qquad \textbf{(7)}$$

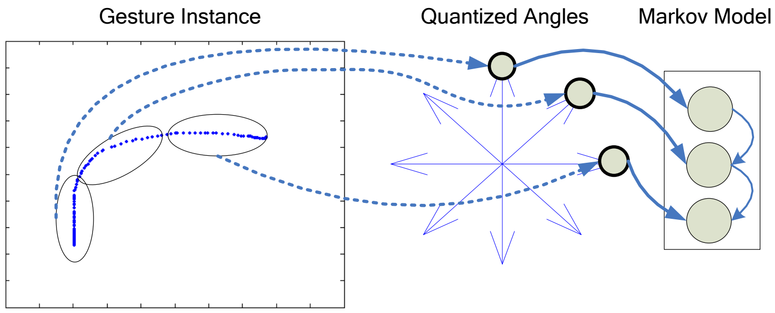


Fig.3. Building a Markov model for a gesture's optical flow

4 Classification of an Unlabeled Gesture

The classification of an input gesture will be based on the two sets of Markov models (eqs. 4 & 7). Let G_k be a gesture instance of unknown category, and $G_{\mathrm{k}}'$and G_{k}'' its transformed representations. Using the MM^{som} set of models, the probability of this gesture to belong in category j can be calculated as:

$$P(G_k' \mid MM_j^{som}) = \frac{\sum_{i=1}^{m} S_i^{som}}{m} \tag{8}$$

The above equation averages the values S_i^{som}, which represent an evaluation factor for each u_{i} value of the G_{k}' transformed gesture with respect to the MM_j^{som} Markov model. These values are calculated as:

$$S_i^{som} = \max_z (NF_{u_i}^{som}(z) P(z \mid u_i, MM_j^{som})) \tag{9}$$

$$u_i = \arg\max_z (S_i^{som}) \,, \tag{10}$$

where z is a variable that indexes the units of the trained map, $NF_{u_i}^{som}(z)$ is the distance of the unit z as defined by the self-organizing map Gaussian neighborhood function with the u_{i} unit as its center. In equation (9), the proximity between the state-unit z and the previous state-unit $u_{\mathrm{t-1}}$ of the gesture is multiplied with the probability of the transition from state-unit z to state-unit $u_{\mathrm{t-1}}$. As the z variable varies across all the units of the map, this product will provide the unit that combines a considerable transition probability from the previous state with a small distance onto the map grid from the current state. This unit will also be used as the previous state in the next step as defined by equation (10). The initial values used in the sum derive from the following equations.

$$S_1^{som} = \max_z(NF_{u_1}^{som}(z)\pi_j^{som}(z)), u_1 = \arg\max_z(S_1^{som}) \tag{11}$$

Using the MM^{of} set of models, the probability of this gesture to belong in category j can be calculated as:

$$P(G_k^{"} \mid MM_j^{of}) = \frac{\sum_{i=1}^{m} S_i^{of}}{m} \tag{12}$$

The values S_i^{of} are calculated from the following equations:

$$S_i^{of} = \max_z(NF_{v_{i-1}}^{of}(z)P(z \mid v_{i-1}, MM_j^{of})), v_i = \arg\max_z(S_i^{of}), \tag{13}$$

where z is a variable that indexes the different states-directions and $NF_{u_i}^{of}(z)$ a distance function between these states. These equations implement a search similar to the previous search on the map grid, but in this case the search is performed among the different possible gesture directions. The initial values are calculated in a similar way from the following equations.

$$S_1^{of} = \max_z(NF_{v_1}^{of}(z)\pi_j^{of}(z)), v_1 = \arg\max_z(S_1^{of}) \tag{14}$$

In order to compare the length of the unknown gesture with the length of the gestures included in each $D_j^{'}$ set, a distance metric for the comparison of symbol strings is necessary. From each set $D_j^{'}$, a *Generalized Median* gesture is calculated. Let S be a set of symbol strings s_i. We can then define m as a string that consists of a combination of all or some of the symbols used in the set and which minimizes the following expression.

$$\sum_{s_i} L(s_i, m), \forall s_i \in S \tag{15}$$

where $L(,)$ denotes the Levenshtein distance, one of the most widely used string distance metric. If the search for string m is restricted to the members of the set then m is the *set median*. But if m is a hypothetical string and the search is not restricted then m is the *Generalized Median* of the set. Using the above definition we calculate the Levenshtein distance $L_{kj} = L(G_k^{'} \mid M(D_j^{'}))$ between $G_k^{'}$ and the *Generalized median* $M(D_j^{'})$ of each $D_j^{'}$ set.

The category of the unknown gesture is primarily decided using the MM^{som} set of models. Subsequently, the category would be equal to:

$$\arg\max_j P(G_k^{'} \mid MM_j^{som}) \tag{16}$$

In order for the category of the unknown gesture to be decided by the above equation the three following conditions must be fulfilled.

$$\max_j(P(G_k^{'} \mid MM_j^{som})) \ge \alpha \qquad \textbf{(17)}$$

$$\max_j(P(G_k^{'} \mid MM_j^{som}) - 2^{nd} \max_j(P(G_k^{'} \mid MM_j^{som}) \ge \beta \qquad \textbf{(18)}$$

$$L_{k,\arg\max\limits_j(P(G_k^{'} \mid MM_j^{som})} \le \gamma LM(\arg\max_j(P(G_k^{'} \mid MM_j^{som})) \qquad \textbf{(19)}$$

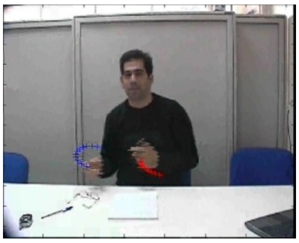
The two first conditions requires that the maximum probability calculated using position based models must exceed a threshold value *a* while the difference between the maximum probability and the second ranked ones must also exceed a threshold value *β*. These two values represent confidence thresholds. The last condition applied is that the Levenshtein distance between the gesture and the *Generalized Median* of the category with the maximum probability must be larger than the *LM* value of this category, multiplied by a user defined factor *γ*. This last comparison is made in order to assess the length of the unknown gesture with respect to the average length of the gestures of the category with the maximum probability. If one of these conditions is not fulfilled then the category of the unknown gesture is defined from a combination of values:

$$\arg\max_j(P(G_k^{'} \mid MM_j^{som})P(G_k^{''} \mid MM_j^{of})\frac{\frac{1}{L_{kj}}}{\left\| M(D_j) \right\|}) \qquad \textbf{(20)}$$

This classification rule combines the evaluation provided from both the MM^{som} and MM^{of} set of Markov models with the Levenshtein distance of the gesture and the *Generalized median* of the each category normalized by the length of the *Generalized median*.

5 Experimental Results

Fig. 4 shows a sample of frames from the input sequences that where used and indicative results of the hand localization and tracking module.



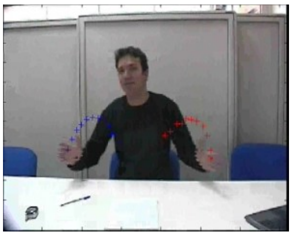


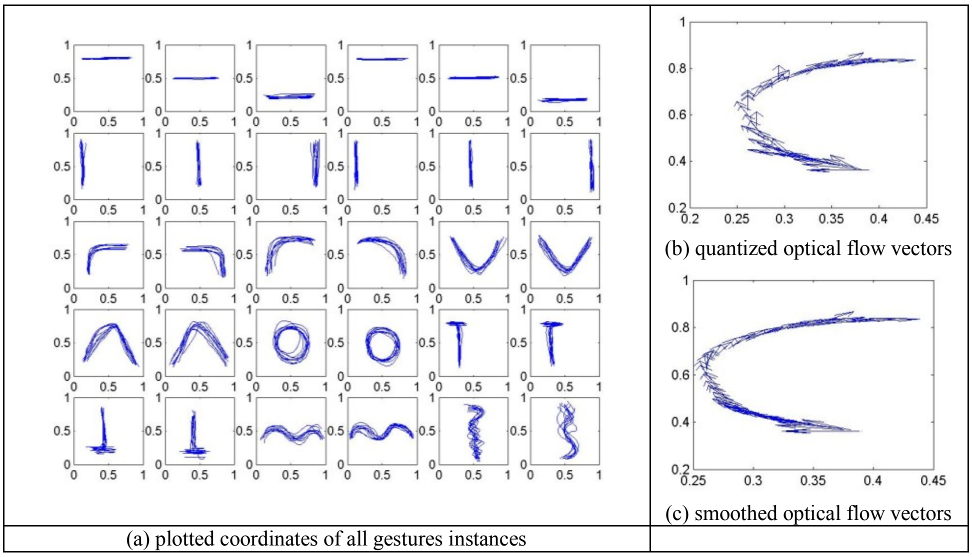
Fig.4 Sample of input sequences with hand tracking results

Experiments were conducted, with the above dataset, in order to evaluate the recognition performance of the proposed method. When all the gesture instances are used for both training and testing, the recognition rate is 100%. To evaluate the generalization capabilities of the proposed method the 10-fold cross validation strategy was used. In this case the average recognition rate was 93%. presents in detail the recognition percentages of each category.

In order to compare the results of our system with one of the most commonly used approaches in the literature we employed an HMM based classifier [7], training one HMM per gesture class. We used continuous left-to-right models and a mixture of 3 Gaussian probability density functions. During the decoding of a gesture it was tested against all models and the one with the highest log-likelihood value was selected as the winner. The above described process produced an average recognition rate of 85%.

Table 1. Proposed method's recognition rate per gesture category (93% average)

Category	%	Category	%	Category	%	Category	%	Category	%
1	100	7	100	13	80	19	100	25	90
2	100	8	100	14	80	20	90	26	90
3	100	9	100	15	100	21	50	27	90
4	100	10	100	16	90	22	70	28	100
5	100	11	100	17	100	23	100	29	100
6	100	12	100	18	100	24	60	30	100



(b) quantized optical flow vectors

(c) smoothed optical flow vectors

(a) plotted coordinates of all gestures instances

Fig. 5. The gesture dataset

6 Conclusions

Present work introduced a novel modeling scheme for gesture recognition from hand trajectories. The system builds models for gesture categories utilizing SOMs that are trained with features extracted through image processing. Experimental results indicate that the system is capable of performing robustly while also evaluating its results. Intended experiments on alternate gesture corpora will be used to assess the capabilities of the system in a broader spectrum of gesture based interaction. Through further research, we intend to address the classification strategy for gestures that present low confidence results, i.e they belong to unknown categories, as well as the evaluation of the system's gesture prediction capabilities.

7 Acknowledgements

Current work was funded by the 03ED375 research project, implemented within the framework of the "Reinforcement Programme of Human Research Manpower" (PENED) and co-financed by National and Community Funds (25% from the Greek Ministry of Development-General Secretariat of Research and Technology and 75% from E.U.-European Social Fund)

References

[1] Kendon, A.: Conducting Interaction. Cambridge, University Press (1990)

[2] Eisenstein, J., Davis, R.: Visual and Linguistic Information in Gesture Classification. Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04), USA, October 13 - 15, 2004. ACM Press, New York, NY (2004) 113-120

[3] Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., Tsapatsoulis, N.and Kollias, S., "Facial expression and gesture analysis for emotionally-rich man-machine interaction", N. Sarris, M. Strintzis, (eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques, pp. 175-200, Idea Group Publ., 2004

[4] Ong, S.C.W., Ranganath, S.: Automatic Sign Language Analysis: a Survey and the Future beyond Lexical Meaning. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.27, Iss.6, Jun, (2005) 873- 891

[5] Wu, Y., Huang, T.S.: Hand Modeling, Analysis and Recognition. Signal Processing Magazine, IEEE, Vol.18, Iss.3, May, (2001) 51-60

[6] Starner, T., Weaver, J., Pentland, A.: Real-time American Sign Language Recognition Using Desk and Wearable Computer-based Video. IEEE Trans. Pattern Analysis and Machine Intelligence, (1998)

[7] Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., Kollias, S., "Emotion Analysis in Man-Machine Interaction Systems", Samy Bengio, Hervé Bourlard (Eds.), Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, Vol. 3361, 2004, pp. 318 - 328, Springer-Verlag

[8] Ozer, I.B., Tiehan, Lu, Wolf, W.: Design of a Real-time Gesture Recognition System: High Performance through Algorithms and Software. Signal Processing Magazine, IEEE, Vol.22, Iss.3, May, (2005) 57- 64

[9] Wilson, Bobick, A.: Parametric Hidden Markov Models for Gesture Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 21(9), (1999)

[10] Juang, C.-F., Ku, K.C.: A Recurrent Fuzzy Network for Fuzzy Temporal Sequence Processing and Gesture Recognition. Systems, Man and Cybernetics, Part B, IEEE Transactions on, Vol.35, Iss.4, Aug., (2005) 646- 658

[11] Yang, M.H., Ahuja, N., Tabb, M.: Extraction of 2D Motion Trajectories and its Application to Hand Gesture Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.24, Iss.8, Aug, (2002) 1061- 1074

[12] Hong, P., Turk, M. and Huang, T.S., "Gesture modeling and recognition using finite state machines," Proc. Fourth IEEE International Conference and Gesture Recognition, March 2000, Grenoble, France.

[13] Wong, S. and Cipolla, R., Continuous Gesture Recognition using a Sparse Bayesian Classifier. In Proceedings of the 18th international Conference on Pattern Recognition - Volume 01 2006. ICPR. IEEE Computer Society, Washington, DC, 1084-1087

[14] Mantyla, V.-M., Mantyjarvi, J., Seppanen, T., Tuulari, E.: Hand Gesture Recognition of a Mobile Device User. Multimedia and Expo, 2000, ICME 2000, 2000 IEEE International Conference on, vol.1, (2000) 281-284

[15] Black, M. J., Jepson, A. D.: Recognizing Temporal Trajectories Using the Condensation Algorithm. Proceedings of the 3rd. international Conference on Face & Gesture Recognition FG. IEEE Computer Society, Washington, DC, (1998)

[16] Fang, G., Gao, W., Zhao, D.: Large Vocabulary Sign Language Recognition based on Fuzzy Decision Trees. Systems, Man and Cybernetics, Part A, IEEE Transactions on, Vol.34, Iss.3, May, (2004) 305- 314

[17] Martin, J. -C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S., "Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews", Personal and Ubiquitous Computing, Special issue on Emerging Multimodal Interfaces, Springer, 2007