HAND TRAJECTORY BASED GESTURE RECOGNITION USING SELF-ORGANIZING FEATURE MAPS AND MARKOV MODELS

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

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

ABSTRACT
This work presents the design and experimental verification of an original system architecture aiming at recognizing gestures based solely on the hand trajectory. Self organizing feature maps are used to model spatial information while Markov models encode the temporal aspect of hand position within a trajectory. A validated classification mechanism is produced through a set of models and a committee machine setup ensures robustness as indicated by the experimental results performed.

Index Terms— gesture recognition, self organizing feature map, Markov processes

1. INTRODUCTION
Gesture recognition is a continuously growing research area receiving abundant attention, especially throughout the research fields of sign language recognition, multimodal human computer interaction, cognitive systems and robotics. 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. Studies also show that gesture classification is, in general, a multimodal task that should make use of both hand movement trajectories and linguistic cues.

An extensive review of several gesture recognition 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 and 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 or variations. Other approaches employ alternate machine learning and artificial intelligence techniques such as recurrent fuzzy network, time delay neural network, finite state machines, Bayesian classifiers, 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 to 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 accuracy loss.

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 resultant direction of motion during the gesture. The classification scheme is based on the transformation of a gesture representation from a series of coordinates and motion vectors to a symbolic form and on building probabilistic models using these transformed representations.
2. SYSTEM OVERVIEW

The steps of the introduced procedure begin with a quasi real time image processing module, which is described in detail in [17]. 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 be expressed as an ordered set of points

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

where $l$ varies across different gesture instances. The system’s input is a set of gestures $D_i$ 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 $D_i$ 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.

3. PROBABILISTIC HAND MOVEMENT MODELS

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.

$$T(G_i) = (u_1, u_2, \ldots, u_l), \text{ where } u_i = BMU(x_i, y_i) .$$  \hspace{1cm} (2)

Function $BMU(x_i, y_i)$ returns the index of the best-matching unit for point $(x_i, y_i)$ and $T(G_i)$ is the modified gesture representation. Given that $u_i$ is the index of a map unit, this function can be is declared as $BMU^r \rightarrow S$, 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,

$$G'_i = N(T(G_i)) = \{u_1, u_2, \ldots, u_m\} \hspace{1cm} (3)$$

where $N$ is a function that removes consecutive equal $u_i$ values and $G'_i$ 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 $G'_i$ of $D_i$ set, will be used for the calculation of the transition probabilities of the model $MM_{r}^{\text{seq}}$ describing the $j$ category and for the determination of the values of the function $\pi_{r}^{\text{seq}}$, which is the first state probability function of this model. The result is a set $MM_{r}^{\text{seq}}$ of $c$ Markov models.

$$MM_{r}^{\text{seq}} = \{MM_{r}^{\text{seq}}_1, MM_{r}^{\text{seq}}_2, \ldots, MM_{r}^{\text{seq}}_c\}$$

(4)

These models are used to evaluate a new unlabeled gesture in order to be classified in one of the $c$ categories.

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. In that sense, we define the transformation of a gesture instance $G_i$ using the OF function as

$$OF(G_i) = \{v_1, v_2, \ldots, v_m\} : v_i = W_i(Q(\arctan(\frac{y_i - y_{i-1}}{x_i - x_{i-1}})))$$  \hspace{1cm} (5)

where $v_i$ are the quantized values, $Q$ the quantization function and $W_i$ 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
\[
G'_i = N(OF(G_i)) = \{v_1, v_2, \ldots, v_m\}
\]  
(6)
The \( v_i \) values define the states for a new set of Markov models \( MM'_{ij} \) that is built using the transformed set \( D_j' \). The first state probability function \( \pi_j^{MM'} \) is also calculated using this set.
\[
MM'_{ij} = \{MM'_{1ij}, MM'_{2ij}, \ldots, MM'_{mij}\}
\]  
(7)

4. CLASSIFICATION OF A GESTURE
The classification of an input gesture will be based on the two sets of Markov models (eqs. 4 & 7). Let \( G_i \) be a gesture instance of unknown category, and \( G_k \) and \( G_m \) its transformed representations. Using the \( MM_{ij}^{som} \) set of models, the probability of this gesture to belong in category \( j \) can be calculated as
\[
P(G_i \mid MM_{ij}^{som}) = \frac{\sum_{z=1}^m S_{i}^{som}}{m}
\]  
(8)
The above equation averages the values \( S_{i}^{som} \), which represent an evaluation factor for each \( u_i \) value of the \( G_i \) transformed gesture with respect to the \( MM_{ij}^{som} \) Markov model. These values are calculated as
\[
S_{i}^{som} = \max_z (NF_{u_i}^{som}(z)P(z \mid u_i, MM_{j}^{som}))
\]  
(9)
\[
u_i = \arg\max_z (S_{i}^{som})
\]  
(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 \( u_{i,j} \) of the gesture is multiplied with the probability of the transition from state-unit \( z \) to state-unit \( u_{i,j} \). 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.
\[
S_{i}^{som} = \max_z (NF_{u_i}^{som}(z)\pi_{j}^{MM'}(z)), u_i = \arg\max_z (S_{i}^{som})
\]  
(11)
Using the \( MM'_{ij} \) set of models, the probability of this gesture to belong in category \( j \) can be calculated as
\[
P(G_i \mid MM'_{ij}) = \frac{\sum_{z=1}^m S_{i}^{MM'}}{m}
\]  
(12)
The values \( S_{i}^{MM'} \) are calculated from the following equations
\[
S_{i}^{MM'} = \max_z (NF_{u_i}^{MM'}(z)P(z \mid v_{t-1}, MM'_{ij})), v_i = \arg\max_z (S_{i}^{MM'})
\]  
(13)
where \( z \) is a variable that indexes the different state-directions and \( NF_{u_i}^{MM'}(z) \) 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.
\[
S_{i}^{MM'} = \max_z (NF_{u_i}^{MM'}(z)\pi_{j}^{MM'}(z)), v_i = \arg\max_z (S_{i}^{MM'})
\]  
(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 [18]. Let \( L_q = L(G_i \mid M(D_j)) \) denote the Levenshtein distance, one of the most widely used string distance metric, between \( G_i \) and the Generalized median \( M(D_j) \) of each \( D_j \) set.
\[
\sum_{q} L(s_m, m), \forall s_q \in S
\]  
(15)
The category of the unknown gesture is primarily decided using the \( MM_{ij}^{som} \) set of models. Subsequently, the category would be equal to
\[
\arg\max_j (P(G_i \mid MM_{ij}^{som}))
\]  
(16)
In order for the category of the unknown gesture to be decided by the above equation the three following conditions must be fulfilled.
\[
\max_j (P(G_i \mid MM_{ij}^{som})) \geq \alpha
\]  
(17)
\[
\max_j (P(G_i \mid MM_{ij}^{som})) - 2^{nd} \max_j (P(G_i \mid MM_{ij}^{som})) \geq \beta
\]  
(18)
\[
L_q \leq \gamma LM(\arg\max_j (P(G_i \mid MM_{ij}^{som})))
\]  
(19)
The two first conditions require that the maximum probability calculated using position based models must exceed a threshold value \( \alpha \), while the difference between the maximum probability and the second ranked ones must also exceed a threshold value \( \beta \). 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 arbitrary factor \( \gamma \). If one of these conditions is not satisfied then the category of the unknown gesture is defined by
\[
\arg\max_j (P(G_i \mid MM_{ij}^{som})P(G_i \mid MM'_{ij}) \frac{1}{L_q})
\]  
(20)
This classification rule (20) incorporates all three components described earlier.

5. EXPERIMENTAL RESULTS

Experiments were conducted 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%.

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 three 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%.

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

This work has been funded by the project PENED 2003. ISTERA (03E∆853). The project is cofinanced 75% of public expenditure through EC - European Social Fund, 25% of public expenditure through Ministry of Development - General Secretariat of Research and Technology and through private sector, under measure 8.3 of OPERATIONAL PROGRAMME "COMPETITIVENESS" in the 3rd Community Support Programme.

8. REFERENCES