Cephalomorphic interface for emotion-based music synthesis

Vassilios-Fivos A. Maniatakos
UPMC, LIMSI-CNRS
Paris, France
fivos.maniatakos@limsi.fr

***Abstract*—This article discusses to adapt 'Pogany', a tangible cephalomorphic interface designed and realized in LIMSI-CNRS laboratory, to use for music synthesis purposes. We are interested in methods for building an affective emotion-based system for gesture and posture identification, captured through a facial interface that understands variations of luminosity by distance or touch. After a brief discussion on related research, the article introduces issues in the development of a robust gesture learning and recognition tool based on HMMs. Results of the first gesture training and recognition system built are presented and evaluated. Explicit future work is described, as well as further possible improvements concerning the interface, the recognition system and its mapping with a music synthesis tool.**

I. Introduction

Research on natural communication between physical and virtual posed some new priorities for Human - Computer Interaction (HCI). Towards this direction, HCI research field find it useful to borrow models, theories and methodology from psychology, social sciences and art, keeping with them a bidirectional relationship. Furthermore, as described in [1]: *if for many years research was devoted to the investigation of more cognitive aspects, in the last ten years lot of studies emerged on emotional processes and social interaction.*

Consequently, during the last decade HCI has focussed on two very important subcategories of what we generally mean by the word 'interfaces': The *multimodal* and the *affective* interfaces. The first describes interfaces which employ information coming from several channels in order to build an application focused on the user and constrained by his/her need for ergonomic interaction[1]. The second refers to a user interface that appeals to the emotional state of users and allows to express themselves emotionally, as well as to receive information of emotional content.

The shift of interest on natural emotion-based interaction suggests *gesture interpretation* as a very important field of research for HCI. Gesture, *a movement of the body that contains information* [2], or an *expressive meaninful body motion with the intent to convey information or interact with the environment* [3], is regarded as a powerful carrier of emotional content and main channel of non-verbal communication. Additionally, new motion capture technologies and cost-effective powerful personal computers further encourage HCI to adopt gesture recognition methodology. In this framework, as a mean to form implicit messages of high-level emotional content, expressive gesture plays significant role in the development of innovative multimodal interactive systems able to provide users with natural expressive interfaces. Gesture is then regarded not only as a part of art performance (i.e definite gesture of fingers playing on a music instrument) but also as a statement of an emotional content to transform to artistic expression.

II. Related research and motivation

The first connection between unfettered gesture and electronic music control was firstly introduced by Theremin in the 1920s. Since then, one can meet numerous applications of different types of such type of interaction, such as direct control of computer synthesis through real or virtual instruments, device manipulation in a score or audio level control through real-time manipulation of music parameters, gestural control of audio effects, or interaction in the context of multimedia installations. Gestural control of music performance has been thoroughly reviewed, in works such as [4]. However, fewer works are coping with interpreting gesture data and extracting high-level information from gestures. Important work has been done in the framework of the augmented instrument projects such as the augmented violin [5], where the authors succeed in extracting expressive information from the violin player through a gesture-based real-time bowstroke analysis system.

Nowadays, control of sound/music based on movements is still considered as an outstanding challenge especially when treated in a more sophisticated basis. Such a basis, as far as multimodal affective interfaces are concerned, could be considered as:

- building the framework to *identify, describe* and *represent* an expressive gesture as a high-level information from a collection of cues,
- conceptualizing *the strategies* for expressive interaction, in other words to put the user into an affective music dialog between her/him and the multimodal interface.

In this article, we are discussing gesture recognition procedure and its methods, giving priority to the HMM approach as a mean for extraction of high-level expressive information. We then describe an implementation of HMM's for 'Pogany', a facial affective interface, as well

as mapping strategies for expressive interaction and future research plans.

III. RECOGNITION STRATEGIES

a) overview: Several methods have been used for gesture recognition. One of the earliest approaches was the one described in [7] in 1979 based on dictionary lookup methods. However, the last decade gesture recognition research was focused on statistic matching methods, usually employing statistics of feature vectors to derive classifiers. When applied, these methods either assume that the distribution of features follow one of the known distributions, or estimate the form of the distribution from the training data: the last avoids errors from assumption violations, however it requires a large amount of data to make proper estimation of the features distribution. Other approaches to gesture recognition employed template matching, linguistic matching, and neural networks technology. Some of the methods are suitable for only one type of feature representation, while others are more generally applicable.

b) Hidden Markov Models: Due to the widespread popularity of Hidden Markov Models in speech recognition and handwriting recognition, HMMs have begun to be applied in spatio-temporal pattern recognition and computer vision. The HMM approach to gesture recognition is motivated by the successful application of Hidden Markov modeling techniques to speech recognition problems, in the concept that techniques effective for one problem may be effective for the other as well. First, gestures, like spoken languages, vary according to location, time, and social factors. Second, body movements, like speech sounds, carry certain meanings. Third, regularities in gesture performances while speaking are similar to syntactic rules.

In general, the concept of HMM can be used in solving three basic problems: the evaluation problem, the decoding problem, and the learning problem. The reader can find a complete presentation in Rabiner's HMM tutorial [8]. The key idea of HMM-based gesture recognition is to use multi-dimensional HMM representing the defined gestures. The parameters of the model are determined by the training data. The trained models represent the most likely human performance and are used to evaluate new incoming gestures. According to Yang, Xu[6], the HMM-based gesture recognition approach can be described as follows:

1) Define meaningful gestures - To communicate with gestures, a vocabulary of meaningful gestures must first be specified.
2) Describe each gesture in term of an HMM - A multi-dimensional HMM is employed to model each gesture. A gesture is described by a set of N distinct hidden states and r dimensional M distinct observable symbols (Discrete HMM). An HMM is characterized by a transition matrix A and r discrete output distribution matrices $B_i, i = 1, ..., r$, where the values of elements in A and B are be estimated in the training process.
3) Collect training data - In the HMM-based approach, gestures are specified through the training data. It is essential that the training data is represented in a concise and invariant form.
4) Train the HMMs through training data - Training is one of the most important procedures in a HMM-based approach. The model parameters λ, are adjusted in such a way that they can maximize the likelihood $P(O|\lambda)$ for the given training data. That means the probability that the HMM gives an observation sequence O, given parameters λ. Training problem is usually solved by the Baum-Welch algorithm.
5) Evaluate gestures with the trained model - The trained model can be used to classify the incoming gestures. The Forward-Backward algorithm or the Viterbi algorithm can be used to classify isolated gestures, estimating the likelihood $P(O|\lambda)$ with reduced cost.

IV. POGANY: AN AFFECTIVE FACIAL INTERFACE

'Pogany' is a head-shaped tangible interface for the generation of facial expressions through intuitive contacts or proximity gestures [9]. The interface takes advantage of the existing non-expensive, integrated camera-capture technology. passing a video stream to a computer for processing. However, due to the design of the interface, only parts of the whole video image are worth processing. Hence the total amount of row video data to process is reduced.

A. Description of the interface

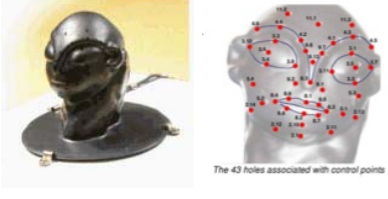

Fig. 1. physical interface and list of interactive keypoints-holes

Designers in [9], suggest a number of constraints for the physical and input part of the interface, which gave to the interface the form of figure 1, according to the interactive keypoints noted on the right. These keypoints, with the form of small holes on the surface of the interface, correspond exactly to a part of MPEG-4 compliant parameters (FAP's and FDP's). A camera is placed in the interior of the head. In a lit environment, passing over or covering the holes modifies the luminosity level captured by the camera. From each frame of the raw video we drag only the pixel blocks that correspond to the important keypoints. At the end, the normalized mean value of luminocity for every block is collected to a feature vector. The output of the front-end of the system consists of instantiations of a 43 element vector with a rate of 30 fr/sec. This direct vector cue is the data that feeds the HMM module.

B. experiments

The scope of the experiments was to build an HMM training module and recognizer for isolated gestures. For this reason we trained one seperate HMM for each isolated gesture to recognize.

1) First experiment: For the first experiment, we trained a 'left-to-right HMM' without any loop, with data from 6 gesture categories: *down-up*, *up-down*, *eyebrows-lifting*, *eye-closing*, *getting sad*, *smile* (figure 2). Concerning the names given to the types of the gestures, the categorization criterion was emotional content in the common sense, as each gesture described the intension of the user to express an emotion. However these common gestures can also be translated to Face Action Coding System (FACS)[10]. The user committed gestures that tended to animate certain face characteristics, in order to set the face from a neutral to a biased condition. Thus, the first two gestures correspond to the tangible contact throughout the face from the upper to the lower part and reversely. The $eyebrows - lifting$ is the tangible gesture tending to lift up the eybrows, the $eye - closing$ to close the eyes of the interface, and similarly *getting sad* and *smile* express gestures that tend to modify 'Pogany"s mouth-line analogously. Finally *static* corresponds to the motionless continuous touch on the mouth area of the interface (posture).

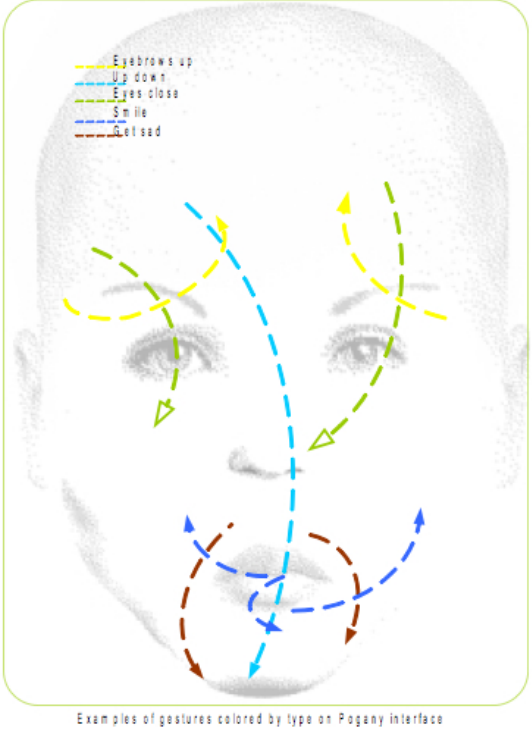

Fig. 2. examples of gestures used as data for training and testing during the experiments

On the first experiment we made use of 142 total gesture samples, taken under full, but slightly non-homogenous lighting conditions. Holdout validation method was used, so as 2/3 of the total samples were used for system training, and the other 1/3 for evaluating the system. The HMM to train was a 5-state left-to-right continuous multi-dimensionnal HMM, with observations vector of size 43. The interface had an orientation of zero angle with the user.

TABLE I
2ND EXPERIMENT , TESTING PHASE (DATA TAKEN CORRESPOND TO DIFFERENT LIGHT CONDITIONS): CONFUSION MATRIX

	eye. -cl.	eyb. -up.	get sad	sml	succ.
eye-cl.	11	2	1	0	78.57%
eyb-up	0	15	0	0	100%
g.-sad	0	0	26	2	92.85%
sml	0	0	5	7	58.33%
TOTAL					84.39 %

2) Second experiment: At the second experiment, the number of samples used was 90, but the gestures to recognize have been reduced to four: *eyebrows-lifting*, *eye-closing*, *getting sad*, *smile*. We used the same validation method to train and evaluate system performance. This time the experiment included an extra test-phase, under different lighting conditions. To specify, while all 90 samples were collected under homogenous artificial light conditions, test data(additional 70 samples of gestures) were taken under full physical light conditions, slightly brighter on the right of 'Pogany' than on the left. The HMM prototype used was similar to this of the first experiment, except for the number of states that this time was set to 4. Both HMM were strictly left to right without loops.

3) results: At both experiments the results we received were encouraging. System performed 96.77% and 100% at the two experiments, proving that both the interface and the HMM respond nearly perfectly during the validation procedure. Taking into consideration the random selection of the data samples through the holdout validation method, we dare to say that the isolated gesture recognition system exhibits a high-level performance for steady light conditions (validation results).

Table I reveals several weakness of the system to adapt in different lighting environments, its performance though remains tolerable 84.39%. Although fail in recognition can be attributed to known drawbacks of HMM structure, probabilities estimated through the testing process in second experiment prove a marginal victory of the wrong HMM estimation probabilities over the correct ones (mean difference: 4,7%), creating the expectation to improve results under careful modelisation.

Apart from successful validation results mentioned above, results presented in table I show that the problems the system encounters correspond to gestures sharing the similar area on the interface. Particularly for the recognition of 'smile' in the test phase of the second experiment, all wrong classified gestures were confused with 'sad'. An other type of confusion occurs between gestures with the same directivity (up to down common directivity between gestures 'eye-closing' - 'getting sad' at confusion matrix, Table I).

V. Strategies for expressive interaction

After a gesture is recognized, it can be used as a high-level emotional message from the user. The system, in order to create the conditions for an affective dialog, should then be able to produce a response containing information suitable with respect to the context and as much high-level as the users inputs. A multimodal system can analyze the user's gesture and create a semantic representation, prepare a semantic response to the input according to a set of rules, and then map it to a sound synthesizer. Thus, if the semantic content of the analyzed gesture is classified somewhere in the *expressive semantic space*, the system can then reply with an element of the same cluster, in order to emphasize the emotional content of the gesture, or select an other region in the semantic space (i.e a neutral or contradictory emotion from the one received) to smoothen or resist to the user's input. For 'Pogany', we employ a combination of such an approach with common direct mapping strategies; that means strategies that are often associated with the lower levels of the conceptual framework.

a) direct mapping strategies: The 43-feature vector used for the recognition module provides continuous information concerning activity in front of the interface. Apart from establishing one-to-one mapping connections among some feature elements and music parameters, other meaningful direct mapping techniques suggest a prior segmentation of the feature vector to smaller blocks with respect to the particular facial area they derive from (eyes, mouth, etc), followed by all-to-one mapping application. Useful additional features are obtained through estimation of the energy and velocity of the vector signal, which correspond to the horizontal axis, as well as the relative value of the energy among the facial areas. In a first approach, these values are metaphorically related with loudness and spectral modification respectively.

b) indirect mapping strategies: However, the HMM approach as a medium to capture semantic content, can also allow a higher-level mapping. Operating real-time HMM recognition on a layer above the direct mapping techniques can trigger more complex procedures of music parameter modification each time one or a series of gestures is recognized. Furthermore, with respect to the emotional content of gestures to the interface, it sounds inspiring to create correspondences between them and relevant or contradictory music procedures, that could establish an affective interaction. This can be achieved through concatenation of music structures, which are pre-classified according to perceptual criteria. However, the mapping of such information to music is a rather non-trivial task. First, because it is difficult to define a music emotional symbolic space that can be isomorphic to emotional gesture semantic space, due to the ambiguity of music emotional perception phenomenon. Second, in case of mapping to one of the known synthesis techniques, it is not straightforward how to conceptualize functions that transform emotional significance in terms of conventional parameters usually employed by standard synthesizers.

VI. Conclusions and future work

In this paper, we presented a system for posture and isolated gesture recognition for 'Pogany' interface. System performance for isolated gesture and posture recognition was outstanding in stable light conditions and remained promising in differentiable light conditions. Under careful vocabulary and grammar definition, both could result to high-performance recognition of more complex and continuous gestures separated by motionlessness postures (the hands of the user remain unmoved). Hence, the results of this first part of this research allow an optimistic approach to continuous gesture recognition. The last is currently being investigated. Low complexity in Viterbi calculation suggests real-time HMM recognizer feasible under appropriate hardware and software manipulations. Enhancing certain aspects of the 'Pogany' interface, could further improve general performance in both gesture recognition and precise, effective interactivity. Hence, providing the interface with a camera of better quality will improve image and speed up connection. Furthermore, a microphone capsule sticked on the surface of 'Pogany' is necessary in order for the system to be able to discriminate contact from minimum distance.

Results shown that HMM, through certain manipulations and enhancements, could provide 'Pogany' a gesture recognition module that decodes expressive information. In future research we will concentrate on the evaluation of a model for real-time recognition of complex gestures classified according to their emotional significance. High-level indirect expressive information will then be mapped to music synthesis tools. It would be of particular interest to map such information to emotional descriptors for data-driven concatenative music synthesis and evaluate the control of the evolution of implicit music events.

References

[1] Camurri, A., Mazzarino, B., Menocci, S., Roca, E., Vallone, I., Volpe, G., "Expressive gesture and multimodal interactive systems", InfoMus Lab,-Laboratorio di Informatica Musicale, DIST-University of Genova, AISB 2004 convention,2004.

[2] Kurtenbach, G., Hulteen, E. "Gestures in Human Computer Communication". In Brenda Laurel (Ed.) The Art and Science of Interface Design, Addison-Wesley, 1990.

[3] Turk, M., Robertson, G. ,"Perceptual User Interfaces", Communications of the ACM, vol. 43, no. 3, 2000.

[4] Wanderley, M., Battier, M., eds. "Trends in Gestural Control of Music". IRCAM Centre Pompidou 2000.

[5] Bevillacqua, F., Rasamimanana, N., Flety, E., Lemouton, S., Baschet, F., "The augmented violin project: research, composition and performance report", NIME 06, Paris, France, 2006.

[6] Yang,J., Xu, Y., "Hidden Markov Model for Gesture Recognition", report CMU-RI-TR-94-10, The Robotics Institute Carnegie Mellon University Pittsburgh, Pennsylvania 15213, May 1994.

[7] Newman, W., Sproull,R., Principles of Interactive Computer Graphics, McGraw-Hill, 1979.

[8] Rabiner,L., A tutorial on Hidden Markov Models and selected applications in Speech Recognition, Proceedings IEEE, pp. 257-284, February 1989.

[9] Jacquemin, C., "Pogany: A Tangible Cephalomorphic Interface for Expressive Facial Animation",LIMSI-CNRS and Univ. Paris 11,submission to second International Conference on Affective Computing and Intelligent Interaction ACII, 2007.

[10] Ekman, P., Friesen, W.V.: "Facial action coding system: A technique for the measurement of facial movement". Consulting Psychologists Press, Palo Alto, CA, USA,1978.