Synthesizing Gesture Expressivity Based on Real Sequences

G. Caridakis, A. Raouzaiou, K. Karpouzis, S. Kollias
Image, Video and Multimedia Systems Laboratory, National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
{gcari, araouz, kkarpou}@image.ece.ntua.gr, stefanos@cs.ntua.gr
+302107723037

ABSTRACT
In this paper we describe an approach to synthesize gestures via the tools provided in the MPEG-4 standard, using the output of the analysis and taking into account the extracted values of expressivity parameters. We animate emotional gestures, using a symbolic representation of human emotion, based on real video sequences and we extract conclusions regarding the performance of every gesture. The results of the synthetic process can then be applied to emotional ECAs.

Author Keywords
Gesture analysis, MPEG-4, expressivity parameters

INTRODUCTION
Both analysis and synthesis of hand gestures constitute an important part of human computer interaction (HCI) [1]. Sometimes, a simple hand action, such as placing a person's hands over his ears, can pass on the message that he has had enough of what he is hearing; this is conveyed more expressively than with any other spoken phrase. To benefit from the use of gestures in HCI it is necessary to provide the means by which they can be interpreted by computers.

Since the processing of visual information provides strong cues in order to infer the states of a moving object through time, vision-based techniques provide at least adequate, alternatives to capture and interpret human hand motion. At the same time, applications can benefit from the fact that vision systems can be very cost efficient and do not affect the natural interaction with the user. Analyzing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies.

Our system uses as input image sequences and tracks the head and the hands of the actor. Following, we can estimate the MPEG-4 BAP (Body Animation Parameters) for every gesture and extract some important expressivity features. All the results are used to the synthetic and lifelike reconstruction of every gesture.

The presented system of the synthetic gesture reconstruction is illustrated in Figure 1:

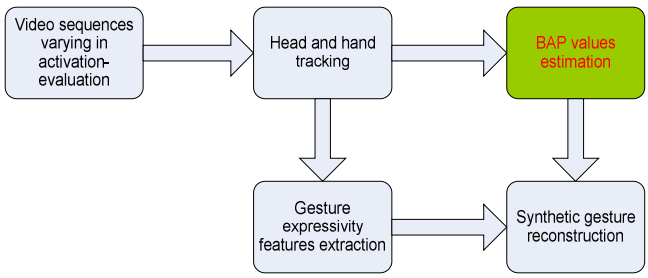


Figure 1: Synthetic Gesture Reconstruction

GESTURES ANALYSIS USING EXPRESSIVITY PARAMETERS

The System Input
The input image sequences of the presented system are videos captured at an acted session including 7 actors, every one of them performing 7 gestures. Each gesture was performed several times with the student-actor impersonating a different situation. Namely the gestures performed are: *"explain", "oh my god"* (both hands over head), *"leave me alone", "raise hand"* (draw attention), *"bored"* (one hand under chin), *"wave", "clap"*.

The different acted situations-emotions are illustrated in Table 1:

Gesture class	quadrant of Whissel's wheel [2]
explain	**(0,0), (+, +), (-, +), (-, -)**
oh my god	**(+, +), (-, +)**
leave me alone	**(-, +), (-, -)**
raise hand	**(0,0), (+, +), (-, -)**
bored	**(-, -)**
wave	**(0,0), (+, +), (-, +), (-, -)**
clap	**(0,0), (+, +), (-, +), (-, -)**

Table 1: Acted Emotions

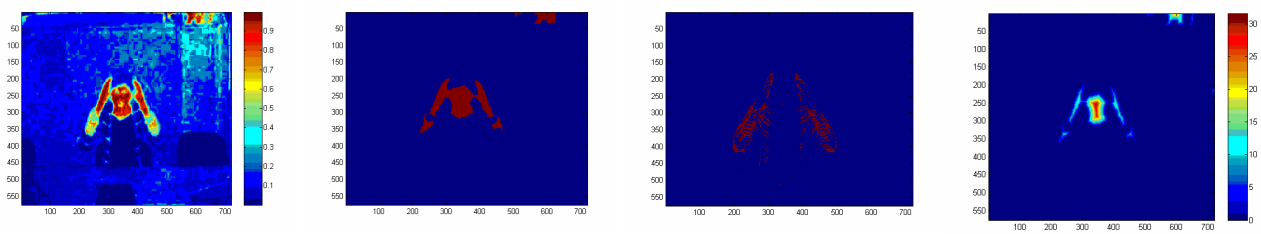
Some of the gesture-emotion combinations were not performed since it did not make much sense reproducing, for example, a "bored" gesture expressing joy. That led us to a whole of 7 actors x 20 variations (Table 1) of the 7 basic gestures =140 image sequences.

Head and Hand tracking

Several approaches have been reviewed for the head-hand tracking module. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames [9]. By tracking the centroid of those skin masks, we produce an estimate of the user's movements. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand). For each frame (*Figure 2*) a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values (*Figure 3a*). The skin color mask is then obtained from the skin probability matrix using thresholding (*Figure 3b*). Possible moving areas are found by thresholding the pixels' difference between the current frame and the next, resulting in the possible-motion mask (*Figure 3c*). This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated (*Figure 3d*) and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The described algorithm is lightweight, allowing a rate of around 12 fps on a usual PC during our experiments, which is enough for continuous gesture tracking. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions extracted from the described method. Skin region size, distance wrt the previous classified position of the region, flow alignment and spatial constraints. These criteria ensure that the next region selected to replace the current one is approximately the same size, close to the last position and moves along the same direction as the previous one as long as the instantaneous speed is above a certain threshold. As a result each candidate region is being awarded a bonus for satisfying these criteria or is being penalized for failing to comply with the restrictions applied. The winner region is appointed as the reference region for the next frame. The criteria don't have an eliminating effect, meaning that if a region fails to satisfy one of them is not being excluded from the process, and the bonus or penalty given to the region is relative to the score achieved in every criterion test. The finally selected region's score is thresholded so that poor scoring winning regions are excluded. In this case the position of the body part is unchanged wrt that in the previous frame. This feature is especially useful in occlusion cases when the position of the body part remains the same as just before occlusion occurs. After a certain number of frames the whole process is reinitialized so that a possible misclassification is not propagated.

Figure 2



(a) **(b)** **(c)** **(d)**

Figure 3

Gesture Expressivity Features Extraction

To define the expressivity parameters we searched through the literature of perception studies to see which parameters were investigated [3, 4]. Six dimensions representing behavior expressivity are defined. The expressivity dimensions have been designed for communicative behaviors only. Each dimension acts differently for each modality. For an arm gesture, expressivity works at the level of the phases of the gesture: for example the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated [5, 6]. We consider six dimensions of expressivity:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power/Energy
- Repetitivity

Overall activation is considered as the quantity of movement during a conversational turn. In our case it is computed as the sum of the motion vectors' norm: $OA = \sum_{i=0}^{n} |\vec{r}(i)| + |\vec{l}(i)|$. Spatial extent is modeled by expanding or condensing the entire space in front of the agent that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands: $SE = \max(|d(\vec{r}(i) - \vec{l}(i))|)$. The average spatial extent is also calculated for normalization reasons. The temporal parameter of the gesture determines the speed of the arm movement of a gesture's meaning carrying stroke phase and also signifies the duration of movements (e.g., quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. The power actually is identical with the first derivative of the motion vectors calculated in the first steps.

The testbed used for comparing the emotionally enriched gestures is GRETA [7]. The mechanisms employed to animate all the expressivity features described above are partly based on the attributes of the TCB Splines used to animate the virtual character. Details about the actual implementation can be found in [8].

EXPERIMENTAL RESULTS

Figure 4 illustrates the gesture *"oh!my god"*. The values for the six dimensions for two different subjects are presented in the diagram of Figure 5. The values shown are normalized.

The values of the results for the different gestures for a) overall activation, b) spatial extent, c) fluidity, d) power/energy are illustrated in Figures 6(a-d), while Figure 7 illustrates the mean values of the six expressivity parameters for three actors and Figures 8(a) and (b) illustrate respectively the mean values of *Overall Activation* and *Power* for positive and negative values of activation. As expected, the values of gestures lying in first and second quadrants (positive activation) are higher.

Some of the frames of the synthesized gesture are illustrated in Figure 9. The tool used for the synthesis is GretaPlayer [7].

Figure 4 Frames from the video of subject 21

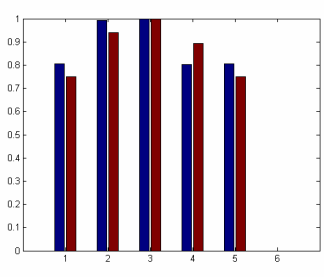
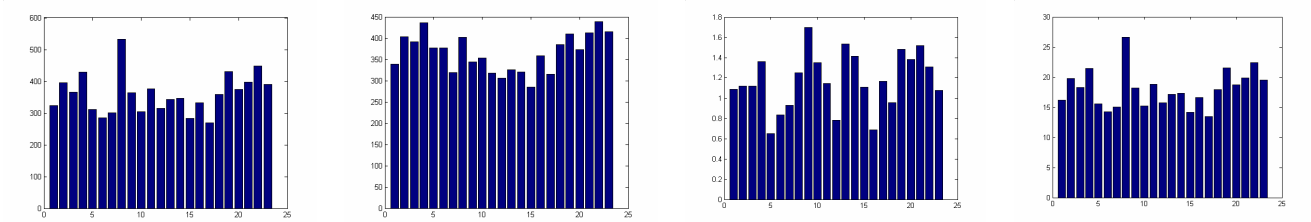


Figure 5



(a) overall activation **(b)** spatial extent **(c)** fluidity **(d)** power/energy

Figure 6

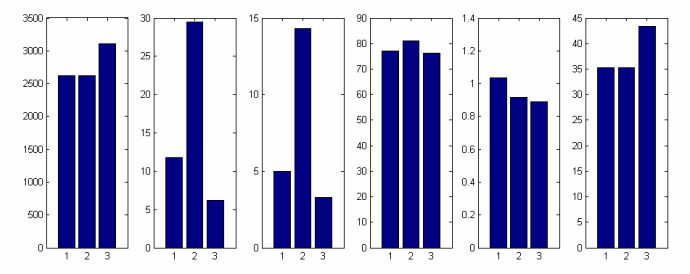
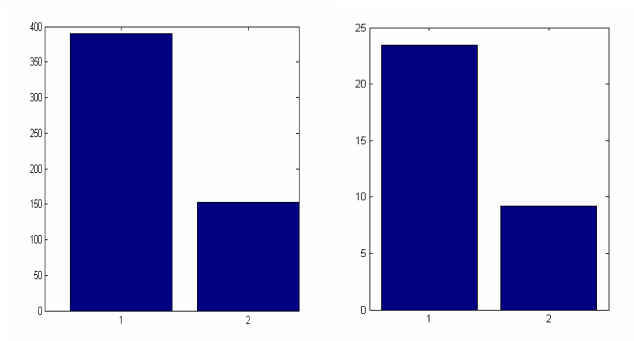


Figure 7: Mean values of the six expressivity parameters for three actors



(a) mean values of *Overall Activation*

(b) mean values of *Power*

Figure 8

Figure 9

CONCLUSIONS

Analysis and expressivity features extraction of a broader set of gestures are necessary in order to evaluate our results. The conclusions concerning the gestures belonging to different quadrants are very useful to further analysis but also to the synthesis of these gestures. The results of the synthetic process can then be applied to emotional ECAs and make the interaction more lifelike.

REFERENCES

1. Wu, Y. and Huang, T.S., "Hand modeling, analysis, and recognition for vision-based human computer interaction", *IEEE Signal Processing Magazine*, 18(3): 51-60, May 2001.
2. Whissel, C.M., The dictionary of affect in language, Emotion: Theory, Research and Experience: vol. 4, The Measurement of Emotions, R. Plutchik and H. Kellerman, Eds., New York: Academic, 1989.
3. Hartmann, B., Mancini, M. and Pelachaud, C., Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop (2005) Vannes
4. Wallbott, H.G, Bodily expression of emotion. European Journal of Social Psychology, 28:879–896, 1998.
5. Harrigan, J.A., Listener's body movements and speaking turns. Communication Research, 12(2):233–250, 1985.
6. Gallaher, P., Individual differences in nonverbal behavior: Dimensions of style. Journal of Personality and Social Psychology 63 (1992)
7. de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V. and De Carolis, B., From Greta's mind to her face: modeling the dynamics of affective states in a Conversational Embodied Agent. *International Journal of Human-Computer Studies*, 59, 81-118, 2003.
8. Maurizio Mancini, Bjoern Hartmann, Catherine Pelachaud, Non-verbal behaviors expressivity and their representation, PF-star report 3.
9. Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S., Manual Annotation and Image Processing of Multimodal Emotional Behaviours in TV Interviews, accepted for publication to LREC06.