Gesture-Based Affective Computing on Motion Capture Data

Asha Kapur[2], Ajay Kapur[1], Naznin Virji-Babul[1], George Tzanetakis[1], and Peter F. Driessen[1]

[1] University of Victoria,
Victoria, British Columbia, Canada
{ajay, peter}@ece.uvic.ca, naznin@dsrf.org, gtzan@cs.uvic.ca
[2] Wake Forest University, School of Medicine,
North Carolina, United States
akapur@wfubmc.edu

Abstract. This paper presents research using full body skeletal movements captured using video-based sensor technology developed by Vicon Motion Systems, to train a machine to identify different human emotions. The Vicon system uses a series of 6 cameras to capture lightweight markers placed on various points of the body in 3D space, and digitizes movement into x, y, and z displacement data. Gestural data from five subjects was collected depicting four emotions: sadness, joy, anger, and fear. Experimental results with different machine learning techniques show that automatic classification of this data ranges from 84% to 92% depending on how it is calculated. In order to put these automatic classification results into perspective a user study on the human perception of the same data was conducted with average classification accuracy of 93%.

1 Introduction

Detecting and recognizing biological motion is an essential aspect of human evolutionary survival. The visual-perceptual system is extremely sensitive to the implicitly coherent structure revealed through biological movement. Humans have the ability to extract emotional content from non-verbal human interaction, facial expressions and body gestures. Training a machine to recognize human emotion is far more challenging and is an active field of research generally referred to as affective computing. Advances in this area will have significant impact on human-computer interactive interfaces and applications.

Imagine online learning systems which sense if a student is confused and re-explain a concept with further examples [1]. Imagine global positioning systems in cars re-routing drivers to less crowded, safer streets when they sense frustration or anger [2]. Imagine lawyers using laptops in the court room to analyze emotional behavior content from witnesses. Imagine audiovisual alarms activating when security guards, train conductors, surgeons or even nuclear power plant workers are bored or not paying attention [3]. These possible scenarios are indicative examples of what motivates researchers in this emerging field.

Currently there are two main approaches to affective computing: Audio-based techniques to determine emotion from spoken word are described for example in

J. Tao, T. Tan, and R.W. Picard (Eds.): ACII 2005, LNCS 3784, pp. 1 – 7, 2005.
© Springer-Verlag Berlin Heidelberg 2005

[4,5,6] and video-based techniques that examine and classify facial expressions are described in [7,8,9]. More advanced systems are multi-modal and use a variety of microphones, video cameras as well as other sensors to enlighten the machine with richer signals from the human [10,11,12]. The above list of references is representative of existing work and not exhaustive. For more details on the evolution and future of affective computing as well as more complete lists of references readers are pointed to papers [3,13].

In the review of the literature as briefly discussed above, almost all systems focus on emotion recognition based on audio or facial expression data. Most researchers do not analyze the full skeletal movements of the human body, with the exception of [14] that uses custom-built sensor systems such as a "Conductor's Jacket", glove, and respiratory sports bra for data acquisition of selected human body movements. Others have used motion capture systems for affective computing experiments with different methods to our own [15, 16]. Research by [17,18] present experiments which confirm that body movements and postures do contain emotional data. Our team has designed a system that uses the VICON[1] motion capturing system to obtain gestural data from the entire body to identify different types of emotion.

In this paper we will first describe the VICON motion capturing system and how it is used to collect data for our experiments. Using the collected data we show results of training automatic emotion classifiers using different machine learning algorithms. These results are compared with a user study of human perception of the same data.

2 Motion Capture System

In this section we will describe how the VICON motion system captures body movement and the method in which the data was collected for the experiments.

2.1 Vicon Motion Systems

The Vicon Motion System is designed to track human or other movement in a room-size space. Spheres covered with reflective tape, known as markers, are placed on visual reference points on different parts of the human body. The VICON system consists of 6 cameras and is designed to track and reconstruct these markers in 3-dimensional space. When a marker is seen by one of the cameras, it will appear in the camera's view as a series of highly illuminated pixels in comparison to the background. During capture the coordinates of all the markers in each camera's view are stored in a data-station. The VICON system then links the correct positions of each marker together to form continuous trajectories, which represent the paths that each marker has taken throughout the capture and thus how the subject has moved over time. At least three of the cameras must view a marker for the point to be captured. Therefore to obtain continuous signals interpolation is used to fill in the gaps [19].

2.2 Data Collection

Markers were placed at 14 reference points on five different subjects (2 of which were professional dancers). The subjects were asked to enact four basic emotions using

[1] http://www.vicon.com (May 2005).

their body movements. No specific instructions for how these emotions should be enacted were given resulting in a variety of different interpretations. The basic emotions used were sadness, joy, anger, and fear. The VICON system measured the trajectories of each subject's movement in 3D space at a sampling rate of 120 Hz. Each subject performed 25 times each emotion for a length of 10 seconds. We manually labeled the reference points of the body throughout the window of movement and filled missing data points by interpolation. A database of 500 raw data files with continuous x, y, and z-coordinates of each of the 14 reference points was created. This database was used to extract features for the machine learning analysis described in section 4. Figure 1 shows a screenshot of the data capturing process.

Data collection involving psychology and perception is challenging and its validity is frequently questioned. Although arguably in acting out these emotions the subject's cognitive processes might be different than the emotion depicted, it turns out that the data is consistently perceived correctly even when abstracted as described in the next section. In addition, since the choice of movements was done freely by the subjects we can stipulate that their motions are analogous to the actual display of these emotions. Even though this way of depicting emotions might be exaggerated it is perceptually salient and its variability provides an interesting challenge to affective computing.

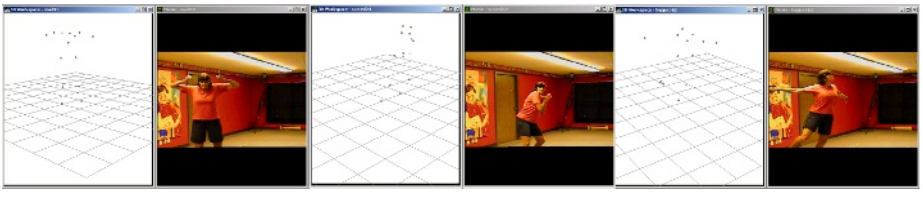

Fig. 1. Screenshot of the data capturing process. The dots on the screen correspond to the markers taped onto the human body.

3 Human Perception

A user study to examine human perception of the motion-capture data was performed in order to provide context for machine learning experiments, as well as to validate the collected data. A subset of 40 randomly ordered files from the database, with an equal proportion of each emotion and subject, were presented to each subject as point light displays. In these point light displays, only the 14 marker points are present (without stick figure lines) and the movement of the subject's emotion for a 10 second period is portrayed. Point light displays were used as they directly correspond to the data provided to the automatic classifiers and their perception is not affected by other semantic cues such as facial expressions.

A group of 10 subjects were tested in classification of these 40 point light displays. A confusion matrix from results of this experiment is shown in Table 1. An average recognition rate of 93% was achieved. It is worth noting that watching a series of 14 moving points humans can accurately identify different human emotions! This is probably achieved by looking at the dynamics and statistics of the motion parameters, which is what we use for features in the automatic system.

Table 1. Confusion matrix of human perception of 40 point light displays portraying 4 different emotions. Average recognition rate is 93%.

Sad	Joy	Anger	Fear	← Classified As
95	0	2	3	***Sad***
0	99	1	0	***Joy***
1	12	87	0	***Anger***
0	2	7	91	***Fear***

4 Machine Learning Experiments

From the human perception experiment described in section 3, it can be seen that motion-capturing preserves the information necessary for identifying emotional content. The next step was to see if machine learning algorithms could be trained on appropriate features to correctly classify the motion-capture data into the 4 emotions. This section describes the feature extraction process followed by experiments with a variety of machine learning algorithms.

4.1 Feature Extraction

After the raw data is exported from the VICON system, as described in section 2.2, feature extraction algorithms are run using a custom built MATLAB program for importing VICON data and extracting features. After experimentation the following dynamics of motion features were selected for training the classifiers. There were 14 markers, each represented as a point in 3D space, $\boldsymbol{v} = [x,y,z]$, where x, y, z are the Cartesian coordinates of the marker's position. In addition, for each point the velocity (first derivative of position) $d\boldsymbol{v}/dt$ and acceleration (second derivative) $d^2\boldsymbol{v}/dt^2$ were calculated. As we are mainly interested in the dynamics of the motion over larger time scales, we consider the mean values of velocity and acceleration and the standard deviation values of position, velocity and acceleration. The means and standard deviations are calculated over the length of 10-second duration of each emotion depicted. Although it is likely that alternative feature sets could be designed, the classification experiments described in the next section show that the proposed features provide enough information for quite accurate classification results.

4.2 Machine Emotion Recognition Experiments

Five different classifiers were used in the machine learning experiments: a *logistic regression,* a *naïve bayes* with a single multidimensional Gaussian distribution modeling each class, a *decision tree classifier* based on the C4.5 algorithm, a *multi-layer perceptron backpropogation artificial neural network,* and a *support vector machine* trained using the Sequential Minimal Optimization (SMO). More details about these classifiers can be found in [20]. Experiments were performed using *Weka* [20], a tool for data mining with a collection of various machine learning algorithms.

The column labeled "All" on Table 2 shows the classification accuracy obtained using 10-fold cross-validation on all the features from all the subjects and corresponds to a "subject-independent" emotion recognition system. The column labeled

"Subject" shows the means and standard deviations of classification accuracy for each subject separately using 10-fold cross-validation and corresponds to a "subject-specific" emotion recognition system. The last column labeled "Leave One Out" corresponds to the means and standard deviations of classification accuracy obtained by training using 4 subjects and leaving one out for testing.

Table 2. Recognition results for 5 different classifiers

Classifier	All	Subject	Leave One Out
Logistic	85.6 %	88.2%+-12.7%	72.8%+-12.9%
Naive Bayes	66.2 %	85.2% +- 8.8%	62.2%+-10.1%
Decision Tree (J48)	86.4 %	88.2% +- 9.7%	79.4%+-13.1%
Multilayer Perceptron	91.2 %	92.8%+-5.7%	84.6%+-12.1%
SMO	91.8 %	92.6%+-7.8%	83.6%+-15.4%

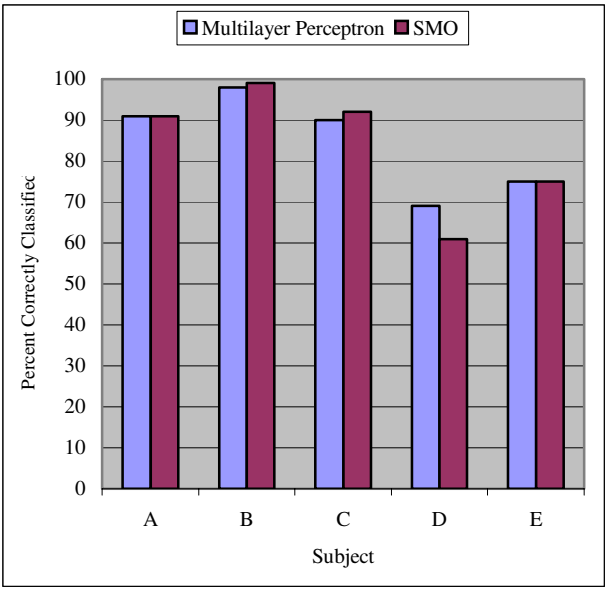

Fig. 2. Graph showing "Leave One Out" classification results for each subject using multi-player perceptron and support vector machine learning classifiers

As can be seen in Figure 2 there is considerable variation in classification accuracy based on which subject is left out. One observation is that the subjects who were professional dancers had a large repertoire of movements for each emotion making them good choices for the training set but poor for the testing set. As a consequence a professional dancer would be better if only one subject can be used to train a motion-based emotion recognition system.

Table 3. Confusion matrix for "subject independent" experiment using support vector machine classifier

Sad	Joy	Anger	Fear	← Classified As
114	0	2	9	***Sad***
0	120	4	1	***Joy***
2	3	117	3	***Anger***
10	3	4	108	***Fear***

Table 3 shows a confusion matrix for "subject independent" using the SMO classifier. As can be seen comparing the confusion matrix for human perception and automatic classification there is no correlation between the confusion errors indicating that even though computer algorithms are capable of detecting emotions they make different types of mistakes than humans.

In all the experiments the support vector machine and the multiplayer perceptron achieve the best classification results. It should be noted that training was significantly faster for the support vector machine.

5 Conclusions and Future Work

We have presented a system for machine emotion recognition using full body skeletal movements acquired by the VICON motion capture system. We validated our data by testing human perception of the point light displays. We found that humans achieved a recognition rate of 93% when shown a 10 second clip. From our machine learning experiments it is clear that a machine achieves a recognition rate of 84% to 92% depending on how it is calculated. SMO support vector machine and multiplayer perceptron neural network proved to be the most effective classifiers.

There are many directions for future work. We are exploring the use of different feature extraction techniques. We also are collecting larger databases of subjects including more intricate detail of facial expression and hand movements. Increasing the number of emotions our system classifies to include disgust, surprise, anticipation and confusion are planned upgrades in the near future. We are moving toward a real-time multimodal system that analyzes data from microphones, video cameras, and the VICON motion sensors and outputs a meaningful auditory response.

References

1. Kapoor, S. Mota, and R.W. Picard, 'Towards a Learning Companion that Recognizes Affect," *Proc. Emotional and Intelligent II: The Tangled Knot of Social Cognition, AAAI Fall Symposium*, North Falmouth, MA, November 2001.
2. R. Fernandez and R. W. Picard, "Modeling Driver's Speech under Stress," *Proc. ISCA Workshop on Speech and Emotions*, Belfast, 2000.
3. M. Pantic, "Toward an Affect-Sensitve Multimodal Human-Computer Interaction," *Proc of the IEEE.* vol. 91, no. 9, September 2003.
4. B.S. Kang, C.H. Han, S. T. Lee, D. H. Youn, and C. Lee, "Speaker dependent emotion recognition using speech signals," *Proc ICSLP,* pp. 383-386. 2000.

5. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Mag.,* vol 18, pp. 32-80, January 2001.
6. D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic Emotional Speech Classification," *Proc ICASSP,* pp. 593-596. 2004.
7. D. J. Schiano, S. M. Ehrlich, K. Rahardja, and K. Sheridan, "Face to interface: Facial affect in human and machine," *Proc CHI,* pp 193-200, 2000.
8. I.Essa and A. Pentland, "Coding analysis interpretation recognition of facial expressions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp 757-763, 1997.
9. M.J. Blackand Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *Int. J. Conmput. Vis.,* vol. 25, no. 1, pp 23-48, 1997.
10. Chen, L.S., Huang, T. S., Miyasato T., and Nakatsu R., "Multimodal Human Emotion/Expression Recognition," *Proc Third International Conference on Automatic Face and Gesture Recognition.* Nara, Japan, 1998.
11. L.C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multimodal information," *Proc FG,* pp. 332-335, 2000.
12. Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," *Proc. ROMAN,* pp. 178-183, 2000.
13. Picard, R. W. "Towards Computers that Recognize and Respond to User Emotions." *IBM System Journal*, vol 39, pp.705-719, 2001.
14. R. W. Picard and J. Healey, "Affective Wearables," *Personal Technologies,* vol. 1*,* no. 4, pp. 231-240, 1997.
15. F. E, Pollick, H. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement." *Cognition*, **82**. B51-B61, 2001.
16. Vines, M. M. Wanderley, C. Krumhansl, R. Nuzzo, and D. Levitin. "Performance Gestures of Musicians: What Structural and Emotional Information do they Convey?", *Gesture-Based Communication in Human-Computer Interaction - 5th International Gesture Workshop,* Genova, Italy. 2003.
17. H.G. Wallbott. "Bodily expression of emotion." *European Journal of Social Psychology* vol. 28, pp. 879-896, 1998.
18. M. DeMeijer. "The contribution of general features of body movement to the attribution of emotions." *Journal of Nonverbal Behavior.* vol. 13, pp. 247-268. 1989.
19. A. Woolard, *Vicon 512 User Manual*, Vicon Motion Systems, Tustin CA, January 1999
20. H. Ian, E. Frank, and M. Kaufmann, *Data Mining: Practical machine learning tools with Java implementations.* San Francisco, 2000.