

Research Article

A Real-Time Facial Expression Recognition System for Online Games

Ce Zhan, Wanqing Li, Philip Ogunbona, and Farzad Safaei

School of Computer Science and Software Engineering, University of Wollongong, NSW 2522, Australia

Correspondence should be addressed to Ce Zhan, cz847@uow.edu.au

Received 31 July 2007; Accepted 31 January 2008

Recommended by Kok Wai Wong

Multiplayer online games (MOGs) have become increasingly popular because of the opportunity they provide for collaboration, communication, and interaction. However, compared with ordinary human communication, MOG still has several limitations, especially in communication using facial expressions. Although detailed facial animation has already been achieved in a number of MOGs, players have to use text commands to control the expressions of avatars. In this paper, we propose an automatic expression recognition system that can be integrated into an MOG to control the facial expressions of avatars. To meet the specific requirements of such a system, a number of algorithms are studied, improved, and extended. In particular, Viola and Jones face-detection method is extended to detect small-scale key facial components; and fixed facial landmarks are used to reduce the computational load with little performance degradation in the recognition accuracy.

Copyright © 2008 Ce Zhan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Multiplayer online games (MOGs) have become popular over the last few years. The collaboration, communication, and interaction ability of MOGs enable players to cooperate or compete with each other on a large scale. Thus, players could experience relationships as real as those in the real world. The “real feeling” makes MOGs attractive to an increasing number of players, despite significant amounts of time and subscription fee required to play. Taking youths in China, for example, according to “Pacific Epoch’s 2006 Online Game Report” [1], China had 30.4 million online gamers by the end of 2006.

Despite the advances in interactive realism of MOGs, when compared with real-world human communication, the interfaces are still primitive. For example, in most of the existing MOGs, text-chat is used rather than real-time voice chatting during a conversation, avatars have no activities related to natural body gestures, facial expressions, and so forth.

Among the problems mentioned above, this paper focuses on facial communication in particular. In everyday life, the manifestation of facial expressions is a significant part

of our social communication. Our underlying emotions are conveyed by different facial expressions. To feel immersed and socially aware like in the real world, players must have an efficient method of conveying and observing changes in emotional states. Existing MOGs allow players to convey their expressions mainly through text-based commands augmented by facial and body expressions [2].

Although a number of existing MOGs have already achieved detailed animation, text commands do not offer an efficient way to control the avatar’s expressions easily and naturally. They are simple and straightforward, but not easy-to-use. First, players have to memorize all the commands. Thus the more sophisticated the facial system is, the harder it is to use. Second, humans convey emotions by expressions in real time. Players cannot type text commands every few seconds to update their current mood. Thirdly, facial communication should happen naturally and effortlessly; typing commands ruins the realism.

The goal of this paper is to automatically recognize the player’s facial expressions, so that the recognition results can be used to drive the “facial expression engine” of a multiplayer online game. While many facial recognition systems have been reported, MOGs pose unique requirements on the

system which have not been well addressed. In a summary, a facial expression recognition system for MOGs should meet the following requirements [2].

- (i) The recognition has to be performed automatically and in real time.
- (ii) The system should consume minimum system resources.
- (iii) The system should be robust under different lighting conditions and complex backgrounds.
- (iv) The system should be user-independent (e.g., the system should be able to handle users of different genders, ages, and ethnicities).
- (v) The input device should be easy to obtain and without any constraints, so only single regular web camera should be used.
- (vi) The system should be insensitive to distance of user to camera. (i.e., the system should be able to handle a relatively wide range of face resolutions).
- (vii) Players usually have to face the computer screen while playing game. Thus, input of the system should be user's frontal faces with certain degree of tolerance to head rotations.
- (viii) Due to entertainment purpose of the system, the recognition accuracy rate need not to be overly conservative.

In this paper, we propose a real-time automatic system that meets the requirements. It recognizes players' facial expressions, so that the recognition results can be used to control avatar's expressions by driving the MOG's "animation engine" instead of text commands. Section 2 provides a brief overview of existing technologies for facial expression recognition. Section 3 describes the proposed system and extension and improvement of several algorithms for an efficient implementation of the system. Section 4 presents the experimental results and Section 5 concludes the paper.

2. OVERVIEW OF FACIAL EXPRESSION RECOGNITION

In computer vision, a facial expression is usually considered as the deformations of facial components and their spatial relations, or changes in the pigmentation of the face. An automatic facial expression recognition system (AFERS) is a computer system that attempts to classify these changes or deformations into abstract classes automatically. A large number of approaches have been proposed since mid 1970s in the computer vision community. Early works have been surveyed by Samal and Iyengar [3] in 1992. Fasel and Luttin [4] and Pantic and Rothkrantz [5] published two comprehensive survey papers which summarized the facial expression recognition methods proposed before 1999. Recently, Tian et al. [6] presented the recent advances (before the year 2004) in facial expression recognition.

Generally, an AFERS consists of three processing stages: face detection, facial feature extraction and representation, and facial expression recognition. The face-detection stage seeks to automatically locate the face region in an input image or image sequences. As the first step of AFERS, its reliability has a major influence on the performance and usability of

the entire system. The face detector could detect faces frame by frame or just detect a face in the first frame and then track it in the subsequent images in a sequence.

After the face has been detected, the next step is to extract and represent the information about the facial expression to be recognized. The extraction process forms a high-level description of the expression as a function of the image pixel data. This description commonly referred to as "feature vector" is used for subsequent expression classification. Geometric features which present the shape and locations of facial components and spectral-transform-based features which are gained by applying image filters to face images are often used to represent the information of facial expressions. Irrespective of the kind of feature extraction approach employed, the essential information about the displayed expressions should be preserved. The extracted features should contain high discrimination power and high stability against different expressions.

Facial expression classification is the last stage of AFERS and it is decision procedure. The facial changes can be identified as facial action units (AUs) [7] or six prototypic emotional expressions [8]. Introduced by Ekman and Friesen, each of the six prototypic emotions possesses a distinctive content and can be uniquely characterized by a facial expression. These prototypic emotions are also referred to "basic emotions". They are claimed to be universal across human ethnicities and cultures and comprise happiness, sadness, fear, disgust, surprise, and anger. An AU is one of the 44 atomic elements of visible facial movement or its associated deformation. Ekman and Friesen first use AUs in their facial action coding system (FACS) [9] with the goal of describing all possible perceptible changes that may occur on the face. In applications, a facial expression is represented using a combination of AUs with respect to the locations and intensities of corresponding facial actions.

To attain successful performance, almost all the existing facial expression recognition approaches require some control over the imaging conditions, such as high-resolution faces, good lighting, and uncluttered backgrounds. With these constraints, the existing methods in the literature cannot be directly applied in most real-world applications, which always require operational flexibility. Although deployment of the existing methods in fully unconstrained environments is still in the relatively distant future, integrating and extending these algorithms to develop a facial expression recognition system for a certain application context such as MOG is achievable.

3. PROPOSED SYSTEM

Based on the specific requirements of MOGs, a facial expression recognition system is proposed in this section. The system categorizes each frame of user's facial video sequence into one of the six prototypic emotional expressions.

We hypothesize that recognition of the six prototypic emotional expressions would serve an MOG well in most cases, since players may not have enough time to perceive more subtle facial changes. Figure 1 shows the block diagram of the system, which consists of four components: face

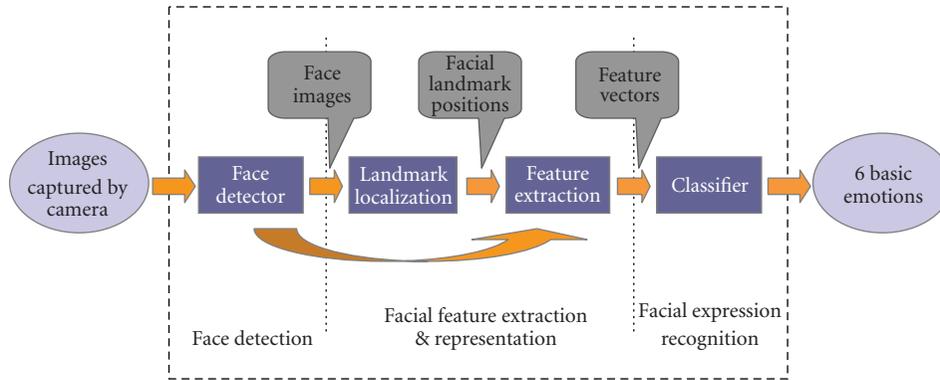


FIGURE 1: The proposed system for MOGs.

detection, facial landmark localization, feature extraction, and classification of the expressions.

3.1. Face detection

The face region is located in an input image by implementing one of the boosting methods proposed by Viola and Jones [10]. The method achieves real-time detection by using very simple and easily computable Haar-like features; and the good detection rate was obtained by the use of a fundamental boosting algorithm, AdaBoost [11], which selects the most representative features in a large set. As a machine-learning method, most of the time and computational expenditure are consumed during the offline training process. Thus, in the detection process, minimal system resources are needed. The trained face detector scans an image by a subwindow at different scales. Each subwindow is tested by a cascade classifier made of several stage classifiers. If the subwindow is clearly not a face, it will be rejected by one of the first stages in the cascade while more specific classifier will classify it, if it is more difficult to discriminate. For details on the Viola-Jones face-detection method, readers are referred to [10].

3.2. Facial landmark localization

To extract the facial feature automatically, facial landmarks need to be detected without manual efforts. Automatic facial landmark localization is a complex process. To find accurate position of landmarks, most of landmark detection methods involve multiple classification steps and a great number of training samples are required [4, 5]. Although coarse-to-fine localization is widely used to reduce the computational load, the detection process is still too complex and time-consuming for MOGs.

According to the results of the facial landmark location tolerance test conducted in our previous work [2], the facial landmark positions are relatively fixed after the normalization based on three key landmarks: mouth center and eye centers. Thus, it is reasonable to use fixed landmarks on normalized face images rather than performing traditional facial landmark detection; and in this way, only three key facial components are needed to be detected.

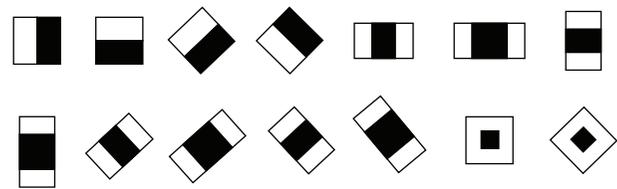


FIGURE 2: The extended Haar-like feature set.

To take advantage of the computational efficiency of Haar-like features and highly efficient cascade structure used in Viola-Jones Adaboost face-detection method, “AdaBoost” detection principle is still adopted to search the key facial components (the mouth and eyes) within the detected face area. However, low detection rate was observed when the conventional Viola-Jones method was trained with the facial components and employed in the detection process. This is probably due to the lack of structure information of the facial components (compared to the entire face). Especially, the structure of the facial components become less detectable when the detected face is at low resolution. Another possible cause of the low detection rate is the substantial variations of the component shape, especially, mouth, among the different expressions conveyed by the same or different people. This is also true for high-resolution face images. To solve these problems, we improved the “AdaBoost” detection method by employing extended Haar-like features, modified the training criteria, regional scanning, and probabilistic selection of candidate subwindow.

3.2.1. Extended Haar-like feature set

An extended feature set with 14 Haar-like features (Figure 2) based on [12] is used in the facial component detection. Besides the basic upright rectangle features employed in face detection, 45° rotated rectangle features and center-surround features are added into the feature pool. The additional features are more representative for different shapes than the original Haar-feature set, thus they would improve the detection performance.

3.2.2. High hit rate cascade training

In the conventional Viola-Jones method, the cascade classifier is trained based on the desirable hit rate and false-positive rate. Additional stage is added into the cascade classifier if the false positive is higher. However, when the false-positive rate decreases, the hit rate also decreases. In the case of facial components detection, hit rate will dramatically fall for low-resolution face images if the cascade classifier is trained for a target low false-positive rate.

To ensure that low-resolution facial components could be detected, a minimum overall hit rate is set before training. For each stage in the training, the training goal is set to achieve a high hit rate and an acceptable false-positive rate. The number of features used is then increased until the target hit rate and false-positive rate are met for the stage. If the overall hit rate is still greater than the minimum value, another stage is added to the cascade to reduce the overall false-positive rate. In this way, the trained detectors will detect the facial components at a guaranteed hit rate though some false positives will occur, which can be reduced or removed by the scanning scheme introduced below.

3.2.3. Regional scanning with a fixed classifier

Rather than rescaling the classifier as proposed by Viola and Jones, to achieve multiscale searching, input face images are resized to a range of predicted sizes and a fixed classifier is used for facial component detection. Due to the structure of face, we predict the face size according to the size of facial component used for training. In this way, the computation of the whole image pyramid is avoided. If the facial component size is larger than the training size, fewer false positives would be produced due to down sampling; when the component is smaller than the training sample, the input image is scaled up to match the training size.

In addition, prior knowledge of the face structure is used to partition the region of scanning. The top region of the face image is used for eye detection, and the mouth is searched in the lower region of the face. The regional scanning not only reduces the false positives, but also lowers the computation.

3.2.4. Candidate subwindow selection

To select the true subwindow which contains the facial component, it is assumed that the central position of the facial components among different persons follows a normal distribution. Thus, the probability that a candidate component at $\mathbf{k} = [x \ y]^T$ is the true position can be calculated as

$$P(\mathbf{k}) = \frac{1}{(2\pi)^{|s\Sigma|}^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{k} - \mathbf{sm})^T s\Sigma^{-1}(\mathbf{k} - \mathbf{sm})\right), \quad (1)$$

where the mean vector \mathbf{m} and the covariance matrix Σ are estimated from normalized face image dataset. The scale coefficient “ s ” can be computed as $s = w_d/w_n$; w_d is the width of detected face; and w_n is the width of normalized training faces. The candidate with maximum probability is selected as the true component.



FIGURE 3: The landmark localization process: (from left to right) detection of face and facial components, normalised face, and fixed set of facial landmarks on the normalised face.

3.2.5. Specialized classifiers

Two cascade classifiers are trained for mouth. One is for detecting closed mouths, and the other is for open mouths. During scanning, if the closed mouth detector failed to find a mouth, the open mouth detector is triggered. In addition, the left and right eye classifiers are trained separately.

After the area of key facial components, mouth and eyes, have been detected, face images are normalized based on the centers of the components; and finally, mean coordinates of facial landmarks obtained from the “location tolerance test” are used as landmarks. Figure 3 shows the landmark localization process.

3.3. Feature extraction

As stated previously, the extracted features should possess high discriminative power and high stability against different expressions. Among a number of feature extraction algorithms proposed in the literature, research has demonstrated that Gabor filters are more discriminative for facial expressions and robust against various types of noise than other methods [4]. However, applying Gabor filters to the whole face area is too costly for MOGs. In the proposed system, Gabor filters with different frequencies and orientations are applied only to a set of facial landmark positions. Thus, not only the real-time requirement can be met due to the reduced amount of data to be processed, but also the limited localization in space and frequency yields a certain amount of robustness against translation, distortion, rotation, and scaling of the images. At the same time, face cropping or alignment is not necessary in the whole recognition process since feature extraction is conducted at specific locations.

A 2D Gabor function is a plane wave with wave vector \mathbf{k} , restricted by a Gaussian envelope function with relative width σ :

$$\Psi(\mathbf{k}, \mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2} \exp\left(-\frac{\mathbf{k}^2 \mathbf{x}^2}{2\sigma^2}\right) \left[\exp(i\mathbf{k} \cdot \mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right) \right]. \quad (2)$$

In our implementation, we set $\sigma = \pi$ [13]. A set of Gabor kernels, which comprises 3 spatial frequencies ($\mathbf{k} = \pi/4, \pi/8, \pi/16$) and 6 different orientations ($\pi/6, 2\pi/6, 3\pi/6, 4\pi/6, 5\pi/6, \pi$) [14], is employed. The parameters of Gabor kernels are chosen based on a large number of experiments, so that the extracted feature vectors only contain the most important components with high discriminative power. Each image is convolved with both even and odd Gabor kernels

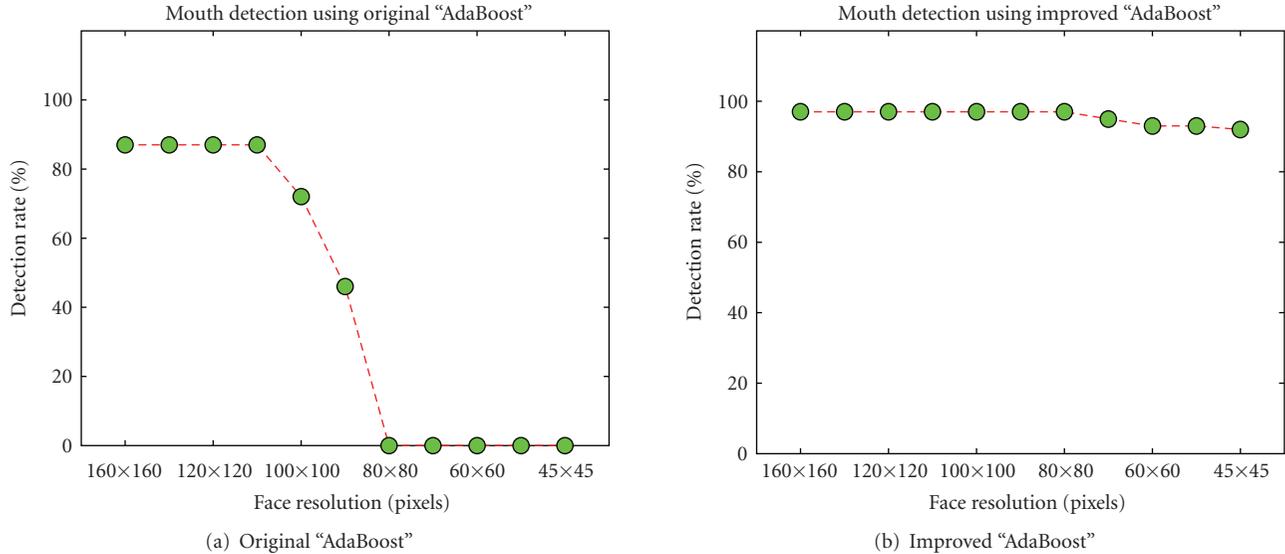


FIGURE 4: Mouth detection result. Both detectors are trained using same dataset.

at facial landmarks (as shown in Figure 3). Thus, 18 complex Gabor wavelet coefficients are obtained at each landmark. Since only magnitudes of these coefficients are used, each face image is represented by a vector of 360 ($3 \times 6 \times 20$) when 20 landmarks are used.

3.4. Classification

A wide range of classifiers in pattern recognition literature have been applied to expression classification. We evaluated a number of classification methods in [2]. In this paper, support vector machines (SVMs) [15] are employed.

SVMs belong to the class of kernel-based supervised learning machines and have been successfully employed in general-purpose pattern-recognition tasks. Based on statistical learning theory, SVMs find the biggest margin to separate different classes. The kernel functions employed in SVMs are used to efficiently map input data which may not be linearly separable to a high-dimensional feature space where linear methods can then be applied. Since there are often only subtle differences between different expressions posed by different people, for example, “anger” and “disgust” are very similar. The high discrimination ability of SVMs plays a major role in designing classifiers that can distinguish such expressions. SVMs also demonstrate relatively good performance when only a modest amount of training data is available, and this also makes SVMs suitable for the system under consideration. Furthermore, only inner products are involved in SVMs computation; the learning and prediction processes are much faster than some traditional classifiers such as a multilayer neural network.

In the implementation, classifiers are trained to identify Gabor coefficient vectors obtained from feature extraction process into one of the six basic emotional expressions or a neutral expression. Since SVMs are binary classifiers and there are 7 categories to distinguish, 21 SVMs are trained to

discriminate all pairs of expressions. A multiclass classifier is obtained by combining the SVM outputs through a voting principle. For example, if one SVM makes the decision that the input is “Happiness” and not “Sadness,” then happiness gets +1 and sadness gets -1. After all SVMs have made their decisions, votes for each category are summed together, and the expression with the highest score is considered to be the final decision.

4. EXPERIMENTAL RESULTS

4.1. Facial component detection

As introduced in Section 3.2, 4 cascade classifiers were trained to detect the key facial components, one for left eyes, one for right eyes, and two for mouths. Positive training samples of eyes and mouths and negative samples (nonfacial components) were cropped from AR database [16] and AT&T database [17]. To accommodate low-resolution facial components, the training samples were rescaled to small sizes: 10×6 for eyes and 16×8 for mouth. For each detector, about 1200 positive samples and 5000 negative samples were used for training.

The trained detectors were tested on BioID database [18]. To evaluate the performance on low-resolution input, the test images were down sampled to different resolutions to simulate low-resolution faces which are not included in the database. To show the improvement compared with the original detection method proposed by Viola and Jones, mouth detection results at different face resolutions are presented in Figure 4. The average left eye, right eye, and mouth detection rate for different face resolutions is 95.7%, 97.2%, and 95.6%, respectively. A few detection examples are shown in Figure 5.

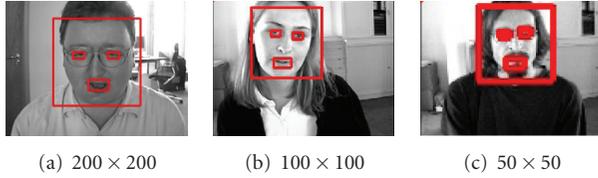


FIGURE 5: Facial component detection results for different resolution faces from BioID database.

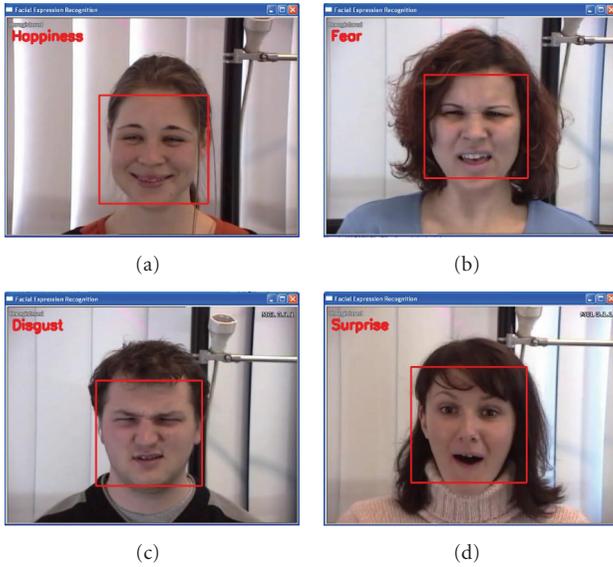


FIGURE 6: Recognition samples from FG-NET.

4.2. Expression recognition

FG-NET database [19] was used in the experiment. The database contains 399 video sequences of 6 prototypic emotional expressions and a neutral expression from 18 individuals. For each expression of each person, at least 3 sequences are provided. In the experiment, one sequence of each expression is left out for test, and the rest are used as the training samples. The recognition result is presented in Table 1 and some samples are shown in Figure 6. The results show that “Happiness,” “Surprise,” and “Neutral” are detected with relative high accuracy while other more subtle expressions were a little bit harder to recognize, especially for “Sadness”. During testing, we found that “Sadness,” “Anger,” “Fear,” and “Disgust” are confused with each other frequently, sometimes even human beings are not able to discriminate them, however, they are seldom confused with other expressions. Thus, if these four expressions are treated as one, together with “Happiness,” “Surprise,” and “Neutral,” we can estimate user’s emotional state more accurately on a higher level. Naming the new expression as unhappy, classification result for 4 expressions are presented in Table 2. In this way, the system is able to tell with an 85.5% accuracy if the user is in good mood, bad mood, or just surprised. We also tested the system in practical conditions, some samples are shown in Figure 7. The results show that the system is rel-

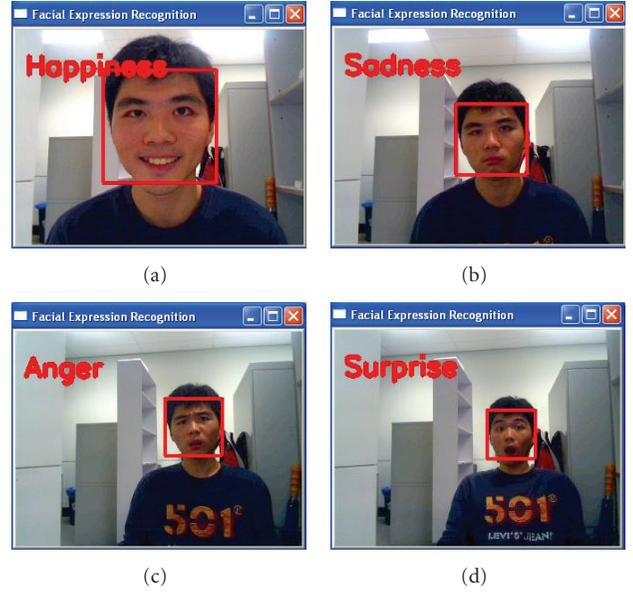


FIGURE 7: Recognition samples for real-time test.

TABLE 1: Recognition results for 7 expressions classification.

Expression	Recognition rate
Happiness	85.2%
Sadness	78.9%
Fear	80.7%
Disgust	81.6%
Surprise	86.3%
Anger	83.3%
Neutral	84.9%

TABLE 2: Recognition results for 4 expressions classification.

Expression	Recognition rate
Happy	85.2%
Unhappy	85.6%
Surprise	86.3%
Neutral	84.9%

atively robust against complex background and lighting conditions, furthermore, it works on the images taken from a practical range of distances from user to camera.

5. MOG IMPLEMENTATION ISSUES

In this section, we indicate the manner in which the proposed system can be incorporated in an MOG. A typical MOG is a complex distributed system connecting thousands of users. Two main types of network architecture are employed, namely, client-server and peer-to-peer [20]. We refrain from any comparative discussion about the two types of architecture since this paper is not about such considerations.

The system presented in this paper is implemented on the client side as it constitutes a user interface device enhancement. The system outputs a classification of the current emotion of the player and this is transmitted to the server. It is possible that an XML-based description of the emotions is employed. The game logic server running of the centralized server would incorporate a module that can parse the XML message and send the appropriate message to the game world module which in turn issues the necessary message that allows the correct view of the avatar to be generated. Thus, the facial expression recognition system allows a rendering of the appropriate avatar with the required emotion on clients' world views.

6. CONCLUSIONS

In this paper, we presented an automatic facial expression recognition system for MOGs. Several algorithms are improved and extended to meet the specific requirements. Despite recent advances in computer vision techniques for face detection, facial landmarks localization, and feature extraction, building a facial expression recognition system for real-life applications still remains challenging.

REFERENCES

- [1] <http://www.pacificepoch.com/>.
- [2] C. Zhan, W. Li, F. Safaei, and P. Ogunbona, "Facial expression recognition for multiplayer online games," in *Proceedings of the 3rd Australasian Conference on Interactive Entertainment*, vol. 207, pp. 52–58, Perth, Australia, December 2006.
- [3] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65–77, 1992.
- [4] B. Fasel and J. Luetin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.
- [5] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.
- [6] Y.-L. Tian, T. Kanade, and J. F. Cohn, *Hand Book of Face Recognition*, Springer, New York, NY, USA, 2004.
- [7] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.
- [8] P. Ekman, *Emotion in the Human Face*, Cambridge University Press, New York, NY, USA, 1982.
- [9] P. Ekman and W. Friesen, *Facial Action Coding System (FACS): Manual*, Consulting Psychologists Press, Palo Alto, Calif, USA, 1978.
- [10] P. Viola and M. J. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," in *Proceedings of the 2nd European Conference on Computational Learning Theory (EuroCOLT '95)*, pp. 23–37, Barcelona, Spain, March 1995.
- [12] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proceedings of the International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 900–903, Rochester, NY, USA, September 2002.
- [13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expression with gabor wavelets," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, Nara, Japan, April 1998.
- [14] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 454–459, Nara, Japan, April 1998.
- [15] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.
- [16] http://cobweb.ecn.purdue.edu/aleix/aleix_face_DB.html.
- [17] <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>.
- [18] <http://www.bioid.com/>.
- [19] <http://www.mmk.ei.tum.de/waf/fgnet/feedtum.html>.
- [20] S. Bogojevic and M. Kazemzadeh, "The architecture of massive multiplayer online games," M.S. thesis, Lund Institute of Technology, Lund University, Lund, Sweden, 2003.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

