Proceedings, ACM Multimedia 98, Bristol, England, Sept. 1998, pp. 427-434. See also: http://www.mic.atr.co.jp/~mlyons

Avatar Creation using Automatic Face Recognition

Michael Lyons, Andre Plante†, Sebastien Jehan, Seiki Inoue†, and Shigeru Akamatsu,
ATR Human Information Processing Research Laboratory
†ATR Media Integration and Communication Laboratory
2-2 Hikaridai, Seika-cho
Soraku-gun, Kyoto 619-02, Japan
mlyons@hip.atr.co.jp, jjap@mic.atr.co.jp

Abstract

In the multimedia context, an avatar is the visual representation of the self in a virtual world. It is desirable to incorporate personal information, such as an image of face, about the user into the avatar. To this end we have developed an algorithm which can automatically extract a face from an image, modify it, characterize it in terms of high-level properties, and apply it to the creation of a personalized avatar. The algorithm has been tested on several hundred facial images, including many taken under uncontrolled acquisition conditions, and found to exhibit satisfactory performance for immediate practical use.

Introduction

The term avatar, originally from Hindu mythology, refers to the temporary body a god inhabits while visiting earth. In virtual communities, it now describes the user's visual embodiment in cyberspace [1].

Cyberspace avatars can be as simple as an icon or as complex as a fully rendered and textured 3D model. Regardless, they play an identical role, which is to represent the user's likeness. Avatars may be distinguished from intelligent agents and other digital life forms in that they are specifically designed to convey a personal image to other users. Loosely speaking, they are similar to the nicknames used in text-based chat environments and their general function is to create a memorable impression of a particular user in a virtual community. In real life, one may recognize without hesitation a colleague by his/her external features such as body size, facial features, hair color or a particular style of prescription glasses. Ideally, recognition of a user's avatar by the other members of a specific virtual community should be similarly facile.

It seems clear that end-users must be able to customize their own avatar in order to truly differentiate it from those of other users. Currently, simple systems employ a library of generic bodies, body parts and accessories for avatar creation. While these systems do allow generation of interesting looking virtual creatures, the limited customization does not allow representation of the user likeness or personality. The face is one of the most easily recognizable, characteristic and expressive features of the human body and it seems natural to attempt to integrate it into the avatar itself.

Imagine for a moment, a virtual art exhibition in which whole classrooms of students could partake at a time with their teacher or a docent. The ability to recognize each participant by their face, just as in a real-life tour, could be advantageous to the virtual tour experience in many ways. For example, students could easily recognize their teacher or friends from the appearance of the avatar.

Egaokun: An Avatar Creation System

Photo-realistic self-representation is now easier with the advent of multimedia computers equipped with digital cameras and peripherals such as scanners. However, raw unedited images, with irrelevant backgrounds, are not convincing representations The system we propose here, which we call the Egaokun Automatic Avatar Building Tool, proposes to customize avatars by using face recognition technology to process raw images of the face. Egaokun finds the face within the digital image and positions a grid on specific facial markers which can then be used to creatively manipulate the facial features. The latter feature may find use in applications where a true likeness is not necessary or desired (for example, in playful situations) or where the user wishes to increase his distinctiveness by caricaturization, or accentuate emotions conveyed in the original picture.

A further design feature which we consider to be desirable is the ability to embody some intelligence about the face being processed. An automatic semantic analysis of the face into facial categories, empowers the system to generate intelligent suggestions of avatar body type. Important facial characteristics are, for example, male or female, child or adult, "race" (for want of a better word for what is essentially a fuzzy set) and emotional state at the time the picture was taken.

Egaokun not only attempts to convey a photo-realistic but also a "Kansei-realistic" image of the user. Kansei is a Japanese word which is somewhat difficult to translate. Kansei sometimes refers to emotions, feelings or senses. It can also define what in english is called "presence". Certain individuals make a greater impression on those around them than others. In Japanese such people are said to have strong "kansei" [3].

A important design feature of the Egaokun system is rapidity of use. In a few seconds a user's picture is filtered, the region containing the face is extracted and registered with an adaptable grid, facial attributes are classified and semantic labels attached to the face and finally the system suggests an interesting looking avatar body to the user. Swiftness and ease of operation would undoubtedly be of great advantage in an any virtual reality application requiring the timely acquisition of facial information.

The possibility to creatively manipulate the original image in order to create memorable picture of "strong kansei" is another advantage of this system. Through the use of spatial deformation and/or morphing with other images, users can create caricature, anti-caricature or fictionalize characters based on their feelings or simply depending on which virtual environment they intend to visit. To draw on the virtual art exhibition example again, the teacher, in this case, could morph in Salvador Dali while quoting him or explaining a particular work or aspect of his life and career. This would dramatically enhance the presentation and certainly create a memorable event for the students.

Automatic Face Processing System

To customize a generic avatar model, the user's specific personal information needs to be acquired. This paper is concerned with the addition of an individual face to a generic avatar body. However, the techniques we discuss here can in principle be extended to the aquisition and processing of other types of image information to the customization of an avatar model.

The acquisition and use of facial information in a multimedia application should be automatic, requiring as little user intervention as possible. Most current face processing software, such as morphing algorithms, for example, require laborious input of registration points on the image in order to establish a point-by-point correspondence mapping between the input face and a model. This procedure is impractical for programs where avatar generation is not the end goal, but only a part of a larger multimedia system. In order for the system we envisage here to achieve widespread popularity its ease of use much be such that a user should be able to sit down in front of a multimedia computer, select a few options, and see a personalized avatar appear in the application program. The main bottleneck in achieving this design goal is the problem of localizing the head and face and registering a grid with the facial features. A further design feature which we consider desirable is that an avatar creation tool should embody some degree of intelligence about the face being extracted from the input image. Is the face male or female face? Child or adult? Does the face display a pleasant mood or an unpleasant one? This semantic level information may be used to adapt properties or parameters of the avatar model to further customize it to the user. In this paper we describe automatic techniques for solving both the problem of localizing the face and its features in the image and extracting high-level semantic meaning about the face in a picture. An image is input to the video interface, the face is localized and cut from the

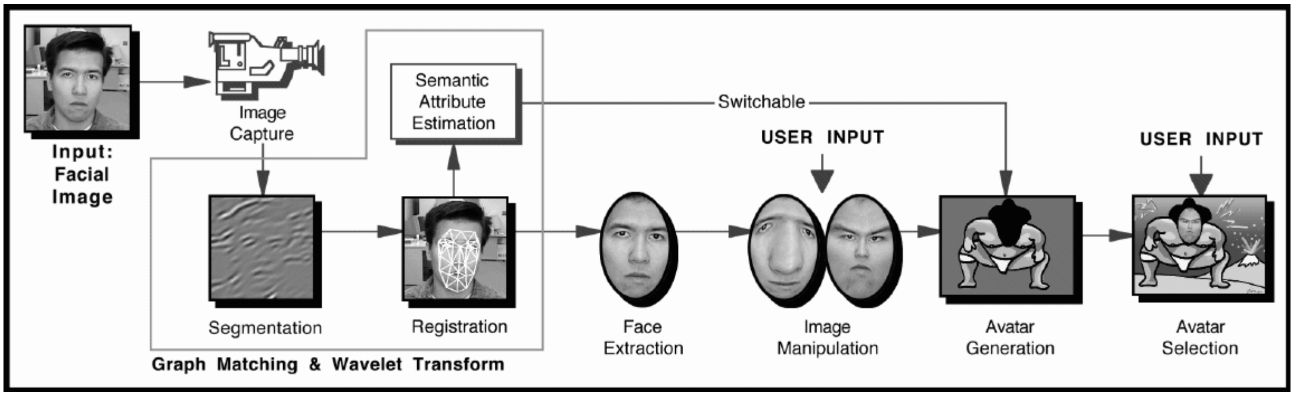

Figure 1. Major functional modules of the Egaokun avatar generation system.

image, using pattern recognition techniques, it is then semantically labelled in terms of several facial characteristics, and finally blended (after optional manipulation) with an avatar model whose properties may depend on the semantic labels and/or further input from the user. For a schematic of the Egaokun avatar generation system, see figure 1. Two stages of the face processing algorithm required to accomplish this may be distinguished. One stage localizes the head in the image and registers a grid with the face's internal features for accurate cutting and further image manipulations such as warping, morphing or caricaturization. In the second distinct stage of processing, the face is classified according to sex, "race" (for want of a better word to describe what is essentially a fuzzy set), and facial expression. A schematic of the Egaokun modules concerned with face processing may be found in figure 2. Both the positioning and classification stages make use of on an image representation based on the Gabor wavelet transform. The use of this representation in various image processing and computer vision applications was originally proposed by Daugman (for example, see [2]). The psychological plausibilty of the Gabor representation, which is important in its application to human-computer interface technology and content retrieval facial databases has been studied by Lyons et al. [6, 7]. To summarize briefly the advantages of the Gabor representation for vision applications: Gabor filters have minimum joint uncertainty in space and spatial frequency. This allows a good representation of spatial frequency structure, or texture, about an approximate location in the image.

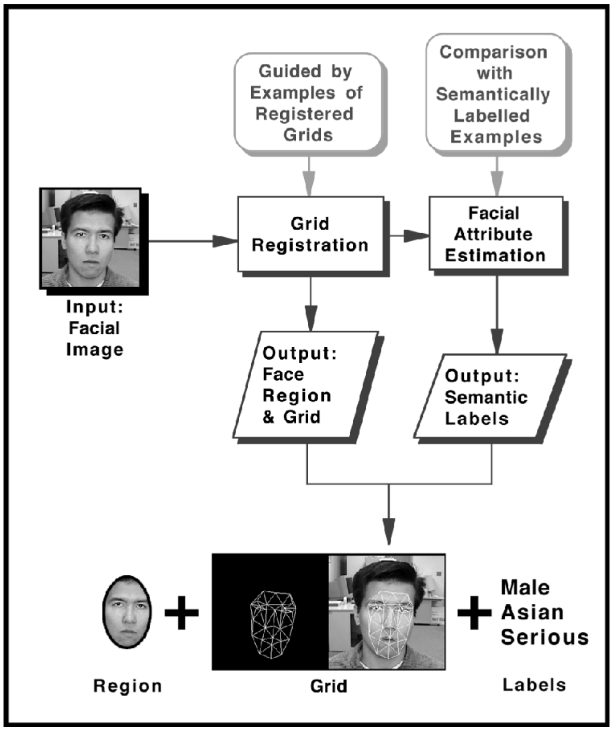
Guided by Examples of Registered Grids
Comparison with Semantically Labelled Examples
Input: Facial Image
Grid Registration
Facial Attribute Estimation
Output: Face Region & Grid
Output: Semantic Labels
Male
Asian
Serious
Region
Grid
Labels

Figure 2. Face processing modules of the Egaokun system.

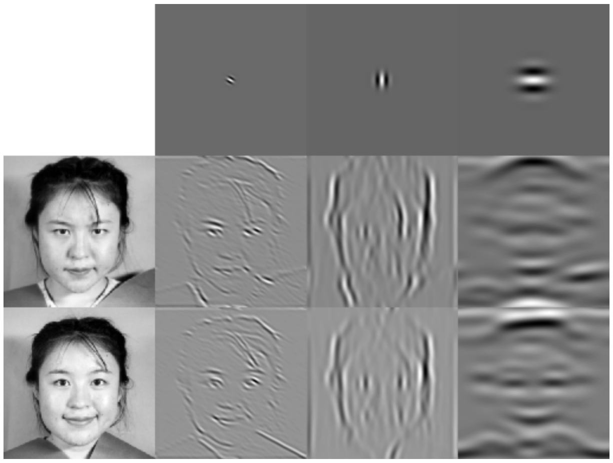

Figure 3. Examples of Gabor filter responses to two facial images for three of the filters used.

The following sections describe the functioning of the face processing algorithms used by the Egaokun avatar creation system. The first section, upon which the other sections depend, defines the representation in terms of Gabor filter responses. The second section discusses the labelled elastic graph matching algorithm (LEGM) for finding the head and registering a grid with the internal features of the face. The third section describes systems for the automatic semantic labelling of the face with a high-level description.

Gabor Representation of a Facial Image

In this section the Gabor wavelet representation of a facial image is described. A detailed discussion of the complex-valued two dimensional Gabor transform has been given by Daugman [2]. Each image is transformed using a bank of multiple spatial resolution, multiple angular orientation two-dimensional Gabor wavelet filters (examples are given in Fig. 3). Typically 3 or 4 spatial frequency resolution levels and 6 orientations are used. For each filter, the complex Gabor transform of the image is made up of sine and cosine parts corresponding to even- and odd-phased filtering functions. In the equations below, $G_{\vec{k},+}$ and $G_{\vec{k},-}$, are even and odd phase filters respectively. filter wave-vector, $\vec{k}$, determines the spatial frequency and orientation tuning. Responses of the filters to the input image I are calculated by convolution of the image with the Gabor functions: Responses of the filters to the image were combined into a vector, $\mathbf{R}$, with components given by:

$$R_{\vec{k},\pm}(\vec{r}_0) = \int G_{\vec{k},\pm}(\vec{r}_0,\vec{r}) I(\vec{r}) d\vec{r},$$

where,

$$G_{\vec{k},+}(\vec{r}) = \frac{k^2}{\sigma^2} e^{-k^2||\vec{r}-\vec{r}_0||^2/2\sigma^2} (cos(\vec{k}\cdot(\vec{r}-\vec{r}_0)) - e^{-\sigma^2/2}),$$

$$G_{\vec{k},-}(\vec{r}) = \frac{k^2}{\sigma^2} e^{-k^2||\vec{r}-\vec{r}_0||^2/2\sigma^2} sin(\vec{k}\cdot(\vec{r}-\vec{r}_0)).$$

In these equations, the integral of the cosine Gabor filter, $e^{-\sigma^2/2}$, is subtracted from the filter to render it insensitive to the local changes in the level of illumination. It is desirable for the representation to be fairly stable against changes in image conditions which are irrelevant to the purpose of the algorithm. The sine filter is spatially odd and does not depend on the absolute illumination level. Spatial frequencies are spaced at octaves, with the highest frequency set at half the Nyquist sampling frequency for the image. For example, when 3 spatial frequency levels were used the wavenumbers take values: $k = \{\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8}\}$, measured in inverse pixels. The parameter $\sigma = \pi$ was used in all calculations giving each filter a bandwidth approximately equal to an octave, independent of frequency level. The wave-vector orientations were spaced at equal intervals of $\frac{\pi}{6}$ from 0 to π. Though not it is not critical for the present discussion, these parameters are derived from experimentally measured properties of visual cortex receptive fields in primates and cats [4]. Hence this set of parameters may exhibit certain advantages for coding naturally occurring images, however discussion of this topic is beyond the scope of this article.

The filter responses as calculated with the above equations are sensitive to position changes on the scale of the wavelength of that filter. To smooth out rapid oscillations in the Gabor representation as image position is changed, the cosine and sine filter responses are combined into a single real number, the amplitude of the complex Gabor transform for a given filter at a spatial point in the image using the formula: $R_{\vec{k}} = \sqrt{R^2_{\vec{k},+} + R^2_{\vec{k},-}}$. This non-linear function is the analogue of complex cells of the primary visual cortex [8]

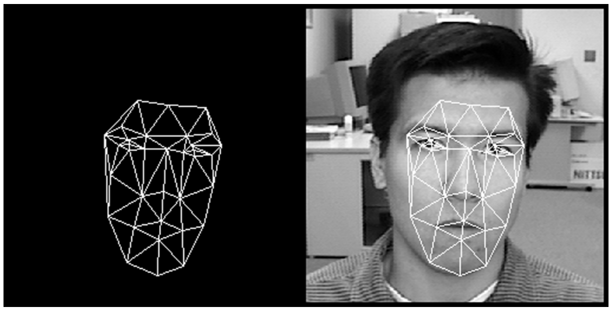

Figure 4. Sample input facial image with registration grid.

which constitute the majority of outputs of that structure. Thus in this representation, for each pixel in the image, there is a set of positive real numbers corresponding to the amplitudes of the responses of the filter bank, which characterize the multiscale texture around that point. For convenience we can think of these numbers as a list or response vector $\vec{R}$. The image is then replaced with a two dimensional array of lists. This representation is used both in localization of the face and its features, and in semantic classification of the face. For many applications the phase information in the transform is not used.

Facial Grid Registration

Our aim is to add or blend an image of the user's face with a generic avatar model. In addition, the user may wish to first modify the image of the face with various manipulations such as morphing, warping, caricaturization, anti-caricaturization or otherwise. The first requirement requires knowledge of the location of the head in the input image and an accurate positioning of the borders of the face so that it may be cut from the image and applied to the model. As for the second requirement, most image manipulations require registration of a grid with the internal features of the face with varying degrees of precision. For morphing or caricaturization, a detailed grid must be put into correspondence with the facial features, whereas for warping an image an approximately positioned rectangular grid often suffices. For head localization and facial grid registration, the Egaokun system uses an algorithm related to the labelled elastic graph matching (LEGM) approach taken by the Malsburg group [5]. Here we will review LEGM and introduce features specific to our implementation. A labelled graph is specified as the graph configuration given by the coordinates of the nodes of the facial grid, see Fig. 4, with the values of the response vector $\vec{R}$, defined above, for each node.

LEGM casts the problem of positioning and distorting the grid to match the face as an optimization problem. The function to be optimized is the similarity of the labelled graph at a particular position and configuration in the input image with a stored graph or set of stored graphs in which grids have been manually registered with example faces. Similarity between two response vectors $\vec{R}_1$ and $\vec{R}_2$ is given by the normed dot product between the two:

$$Sim(\vec{R}_1, \vec{R}_2) = \frac{\vec{R}_1 \dot{\vec{R}}_2}{\|\vec{R}_1\| \|\vec{R}_2\|}$$

Similarity between two labelled graphs is the average of this vector similarity over corresponding nodes of the facial graph. This is essentially a measure of the filter response amplitude at corresponding spatial frequency, orientation, and position on the grid.

For the implementation of LEGM used in Egaokun a set of five example faces were manually registered with facial graphs. To test a trial configuration of the graph, for each node of the input graph, the response vector with highest similarity is chosen from the five example labelled graphs at the same node. In this way a composite graph is built up, taking response vectors from different example graphs. The similarity is averaged over all nodes of the graph structure.

The space of graph positions and configurations can be searched in a fashion similar to gradient descent. At each trial configuration the maximum similarity with the example stack is computed in the fashion described above. The space is first searched for optimal position and scale (a 3 parameter search). Finally, when greater precision is needed for a given image manipulation, individual nodes are allowed to move locally distorting the grid so that it fits the particular face. For some applications a 3 parameter fit of a rectangular grid are sufficient for extracting a picture of the face, for simple manipulations such as warping, and for many classification problems (sex and "race"). Greater graph registration precision is required if the face is to morphed, caricatured or if detailed information about the facial expression is to be estimated.

The graph positioning system has been tested on hundreds of images under varying lighting, pose and background conditions. In approximately 98% of cases in our database, the positioning was accurate enough for the purposes of avatar creation.

Facial Semantic Attribute Estimation

Our second design goal for the Egaokun system is to embody in the system a degree of intelligence about what kind of face is being processed. Fortunately, for some facial attributes this is not as difficult as one may think. It turns out that in the Gabor face representation faces sharing certain characteristics are clustered together in the similarity space. This is illustrated in figures 5 and 6 which show clustering due to sex and "race" (Fig. 5) and facial expression (Fig. 6). These figures show a low dimensional projection of the similarity space between faces calculated using the labelled graph similarity measure defined above. This projection was found using multi-dimensional scaling (for example, see [9]) which embeds the objects in a euclidean space such that the distance between objects preserves, as much as possible, the rank ordering of the similarities between objects. These figures suggest that it should be possible to classify the input faces on the basis of their representation in the Gabor response vector labelled graph representation. Like the graph positioning scheme, the approach to semantic attribute estimation taken here is also example-based. Once a graph is registered with a face, the shape information contained in the filter response coefficients can be used to estimate properties of the face. Two types of classifier can be considered. In one type of classifier a multilayer perceptron network may be trained using supervised learning with example labelled graphs from faces which have been hand classified. The suitability of this scheme for expression recognition from Gabor coded facial images has been studied by Zhang et al. [11].

The method used by the Egaokun system is simpler but effective. An input facial graph is compared directly with example graphs which have been labelled with semantic attributes. This technique is derived from a procedure discussed by Wiskott [10]. For the Egaokun demonstration we constructed an exam-

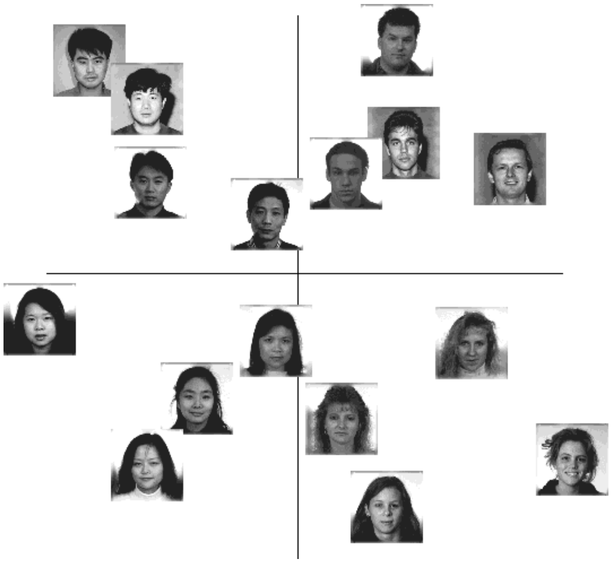

Figure 5. Face space showing clustering of sex and "race".

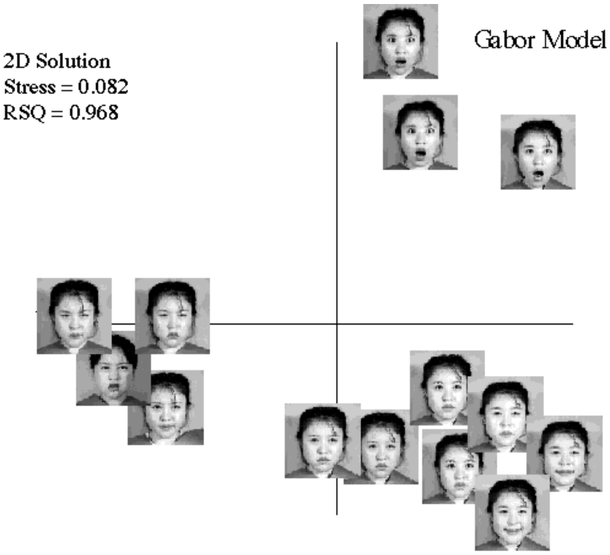

Figure 6. Facial space showing differential clustering of expressions.

ple stack with facial graphs semantically labelled with the binary attributes of sex (male or female), “race” (Japanese or Caucasian), and expression (smile or no smile). For each node, the three most similar response vectors are retrieved from the stack along with the corresponding semantic labels. A voting algorithm (the local, node based vote) is used to determine the estimate for the facial attribute at that node in the graph. The procedure is repeated for all nodes in the graph, yielding an attribute graph. The majority semantic label (the global, graph vote) is the estimate of the facial attribute for the input image.

Using a dataset consisting of 46 Japanese male and 46 Japanese female neutral faces taking under controlled lighting conditions, performance of the system on sex classification was 90 % correct responses.

As a further test of the system we demonstrated the Egaokun avatar creation system during an Open House event held at ATR. This represents a realistic, rather uncontrolled test of the system: pose was fairly uncontrolled; lighting variable; and the background open, changeing as people walked past the exhibit. Images had to be acquired and processed in a few seconds. More than 150 people tested the system with satisfactory results: the graph positioning failure rate was stable (at about 2 %) while the total classification performance on all categories was 80 %. In practice, a failure of the system on classification (eg. classification of a man as a woman or a Japanese as a Caucasian) produced humourous results which added to the entertainment value of the demo. Higher correct performance may be attainable with algorithm optimization, which we have not yet extensively explored.

In figure 7, we show an example output of the Egaokun avatar creation system. An image of a male face of Japanese background and neutral expression was input to the system. The head has been located, registered and applied to an avatar body automatically chosen to correspond to these three attribute categories: a Sumo wrestler’s body. The face has been distorted to further enhance the impression of the Sumo wrestler, with somewhat comical effect. The current implementation allows selection of eight avatar bodies (for each of the three binary attribute categories).

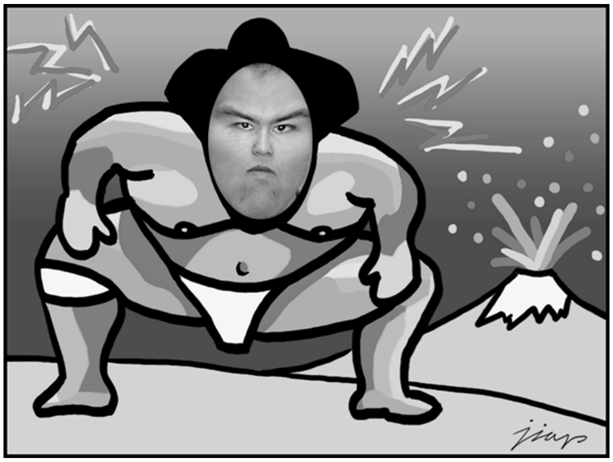

Figure 7. Sumo wrestler avatar combined with modified face from input image.

Concluding Remarks

Examination of the avatar concept in reference to virtual user interaction suggested to us that there be an increased use of the human face in the creation of an avatar since it is the single most important visual conveyor of non-verbal information about the individual user. Automatic face processing algorithms can fulfill several essential requirements of automating avatar creations systems. The face processing technology allows ease and rapidity of use and allow automation of such functionalities as caricaturization or morphing of the facial image. Use of pattern recognition techniques allows automatic semantic level labelling of the face. Indeed, systems that are not automatic and require laborious registration of image points will not find widespread use in avatar creation. In this paper we demonstrated the concept of automated avatar creation with a specific, but fully functional, implementation, the Egaokun avatar creation system. Egaokun draws on Gabor wavelet coding for facial image coding, labelled elastic graph matching for image search and pattern recognition for face classification.

Both in-house and public demonstrations of the system (during the 1997 ATR Open House) on hundreds of images, many taken under uncontrolled pose, lighting, and background conditions, showed a high degree of accuracy of the graph positioning and face extraction module (with approximately a 2 % failure rate). Accuracy of the classification algorithm is also sufficiently high for the purposes of the Egaokun demonstration system, but could be improved with further refinements of the classification algorithm.

As the interaction of users in virtual spaces through networked communities increases, it is inevitable that there will be a growing need to personalize avatars. The algorithm presented in this paper could find applications in educational systems (virtual museums or classrooms) and also in entertainment technology, for example interactive movies, multiple user role-playing communities and so on. The system we described fulfills the need for a system which can capture the facial image from any digital source (video or digital camera; scanned portraits or image files on a web page) and automatically personalize a generic avatar model. Future extensions of the Egaokun system could incorporate image manipulation techniques which also change texture and colour information of the face. In the longer term, the characteristics of body movement and gesture could be acquired to further individuate an avatar. Creative manipulation of this information could be used to hone the avatar's "kansei".

References

[1] B. Damer. Avatars! Exploring and Building Virtual Worlds on the Internet. Peachpit Press, Berkeley, 1997.

[2] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2: 1160-1169, 1985.

[3] S.Inoue, M. Ishiwaka, S. Tanaka. & J. Park. An image Expression Room. IEEE Proceedings of the International Conference on Virtual Systems and Multimedia VSMM 97 p.181, 1997.

[4] J.P. Jones & L.A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology* 58: 1233-1258, 1987.

[5] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, & W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture, *IEEE Transactions on Computers,* 42: 300-311, 1993.

[6] M. J. Lyons and K. Morikawa. A model based on V1 cell responses predicts human perception of facial similarity. *Investigative Ophthalmology and Visual Science,* 37: 910, 1996.

[7] M. J. Lyons, M. Kamachi, J. Gyoba, S. Akamatsu. Gabor Wavelet Representation of Facial Expression. *Technical Report of the IEICE.* HIP97-2: 9-16, 1997.

[8] D. A. Pollen & S. F. Ronner. Phase relationships between adjacent simple cells in the visual cortex. *Science,* 212: 1409-1411, 1981.

[9] Y. Takane, F. W. Young, & J. de Leeuw. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika* 42: 7-67, 1977.

[10] L. Wiskott, J. Fellous, N. Krüger, & C. von der Malsburg. Face Recognition and Gender Determination. In Bichsel, M., (Ed), *Proceedings of the International Workshop on Automatic Face and Gesture Recognition* (pp. 92-97), 1995.

[11] Z. Zhang, M. Lyons, M. Schuster, & S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. To appear in Proceedings of the Third IEEE Conference on Face and Gesture Recognition, Nara Japan, 1998.