Cartoon-like Avatar Generation Using Facial Component Matching

Chi-Hyoung Rhee and Chang Ha Lee*

**School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea*

kabapoo@vis.cau.ac.kr, chlee@cau.ac.kr
**Corresponding author: Chang Ha Lee, Ph.D. (chlee@cau.ac.kr)*

Abstract

Nowadays, avatars are widely used in games and Internet environments. Especially, video game consoles such as Wii (Nintendo) use avatars for representing the user's alter ego. There are several ways to generate avatars. Most existing games or Internet services provide manual systems for generating avatars. Many researchers have suggested automatic avatar generation methods, most of which generate avatars by simplifying images using non-photorealistic rendering techniques.

In this research, we suggest an example-based method that generates avatars by first matching the most similar avatar components for each facial feature and compositing them. We built a system that generates an avatar which is similar to the input front-view photograph by matching each facial feature to the corresponding feature of the avatar. The system first extracts six facial features from the input image: 2 eyes, 2 eyebrows, a mouth, and a face outline. Then it matches them to the corresponding avatar components using graph similarity and Hausdorff distance. Finally, the system generates an avatar by compositing the most similar components.

We have also experimented on the effectiveness of our approach, and the results show that our system generates avatars which successfully represent their corresponding photographs.

Key Words*: avatar generation, cartoon-like, feature extraction, face matching, facial component*

1. Introduction

Avatars are currently a crucial component of digital environments because they define how users can act and express themselves. Avatars are widely used on Internet forums, online games, and other communities. The purpose of avatars is to represent users, their actions, and personality. Users take time to customize their own avatars.

In many Internet services and games, a user can choose each facial component among a preset list. For example, Nintendo's Wii console provides avatars called "Miis" that can build a cartoon-like face images by compositing facial components chosen by users and can be used in games to represent players. Recently, many of the social network services and smart phones provide methods to represent individuals with images. Most of people use their photographs, but many people also want to hide their real faces because of privacy issues. Automated avatar generation would be useful in these cases as well.

Many approaches have been proposed to generate facial avatars automatically. There are two categories for those approaches. One is photo-realistic avatar generation and the other is non-photo-realistic (NPR) avatar generation. In the domain of the photo-realistic avatar generation, Hogue *et al.*, [1] proposed the system that generates the 3D textured and normal-mapped geometry of a personalized user avatar. NPR approaches include the work of Obaid *et al.*, [2] which proposed facial caricature generation using an Active Appearance Model (AAM) [3] and a quadratic deformation model representation of facial expressions.

Composite sketching of human portraits is proposed in Chen *et al.*, [4]. This example-based composite approach has an advantage if the training set contains a limited number of examples.

In this paper, we suggest an avatar system which automatically generates a cartoon-like image based on a human face photograph. To match realistic photograph components to the cartoon-like avatar components, we measure relative similarities with respect to the matching scores of face images in the database. Our system can generate qualified avatars because each component could be carefully designed and drawn by artists.

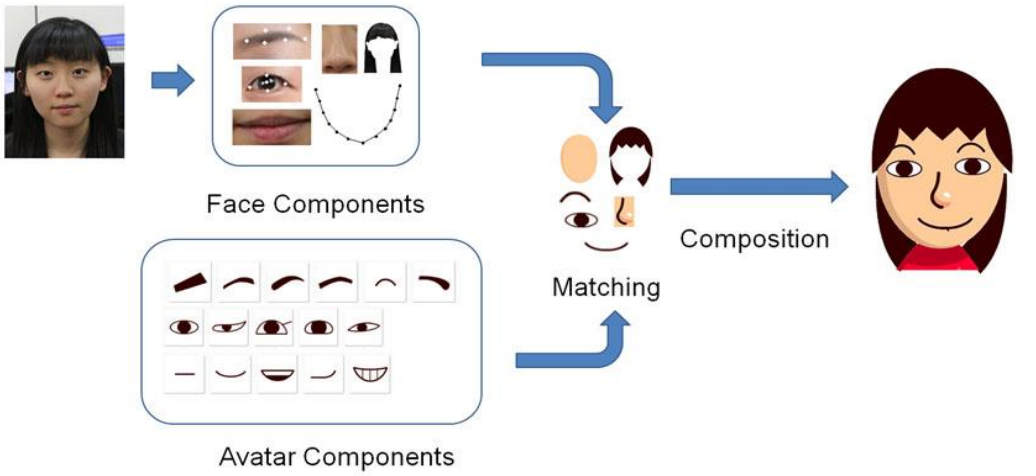

Figure 1. The Overall process of the Avatar Generation Method

2. Previous Works

2.1. Facial Feature Extraction

The importance of facial feature extraction for avatar creation cannot be overstated. Many works related to avatar creation, face recognition, or NPR-based caricature generation systems need to accurately locate facial features in a preceding step.

Three types of feature extraction method can be distinguished: (1) generic methods based on edges, lines, and curves; (2) feature-template-based methods; (3) structural matching methods which take into consideration geometrical constraints on the features. Hallinan presents a template-based approach focused on individual features [5]. For structural matching methods, the Active Shape Model (ASM) proposed by Cootes *et al.*, [6] enables flexible detection of facial features. The ASM is much more robust in terms of handling variations in image intensity and feature shape than earlier methods. Cootes *et al.*, also demonstrate a novel method of interpreting images using an Active Appearance Model (AMM) that combines a model of shape variation with a model of texture appearance variation.

2.2. Face Matching

One of the most widely used representations of the face region is eigen pictures [7] which are based on principal component analysis. An advantage of using such representations is their reduced sensitivity to statistical redundancies in natural images. Turk and Pentland [8] use eigen-pictures for face detection and identification. Every face in the database can be represented as a weight vector and when a new image is given, the image is also represented by its vector of weights. The identification of the new image involves locating the image in the database that has weights closest to the weights of the new image.

Belhumeur, *et al.*, [9], and Swets and Weng[10] use LDA/FLD for face recognition systems. LDA training is accomplished via scatter matrix analysis. Swets and Weng apply discriminant analysis of eigen features in an image retrieval system to determine the face class. To improve the performance of LDA-based systems, a unified method of PCA+LDA was proposed in Zhao [12].

In the structural matching category, the Elastic Bunch Graph Matching (EBGM) system is one of the most successful [13]. Gabor wavelets play a major role in facial representation in the graph matching methods. The EBGM is robust to illumination change, translation, distortion, rotation, and scaling.

2.3. Avatar Generation

Researches on avatar generation can be classified into two categories. One is photo-realistic 3D avatar generation and another is non-photo-realistic avatar generation involving facial sketch, caricature, and pen-and-ink illustration. Lee *et al.*, [14] presented the reconstruction of a 3D avatar for interactive mixed-reality in real-time. Ahmed *et al.*, [15] created a personalized avatar from multi-view video data of a moving person using a spatio-temporal approach. In the NPR domain, Obaid *et al.*, [2] proposed facial caricature generation using an Active Appearance Model (AAM) and a quadratic deformation model representation of facial expressions. Composite sketching of human portraits is proposed in Chen *et al.*, [4]. This example-based composite approach has an advantage if the training set contains a limited number of examples

3. Background

3.1. Active Shape Models

Active Shape Models (ASMs) are characterized by the use of the Mahalanobis distance on one-dimensional profiles at each landmark and a linear point distribution model. Landmarks are made by marking all images by hand. This is done before training begins. During training, the ASM determines the characteristics of the profile and point distribution models.

The ASM consists of two sub-models: the shape model and the profile model. A profile model describes the characteristics of the image around the landmark. During training, we sample the area around the landmarks across all training images in order to build a profile model for the landmark. During the search, we sample the area around each tentative landmark and move the landmark to the position that best matches its model profile.

A shape model defines the allowable relative position of the landmarks. During search, each landmark which follows the position suggested by profile model can lose its initial shape. The shape model adjusts the shape to conform to a legitimate shape.

3.2. Hausdorff Distance

The Hausdorff distance is a metric between two subsets of a metric space for measuring how far they are from each other. If every point of either set is close to some point of the other set, the two sets are close in terms of the Hausdorff distance. We used the HD as a similarity measure between an input face image and an avatar image.

Let A and B be two finite subsets of a metric space (M, d). Then the Hausdorff distance is defined as

$$H_{i,a} = \max\big(h(P_i, P_a), h(P_a, P_i)\big)$$
$$,where\ h(P_i, P_a) = \max_{a \in P_i} \min_{b \in P_a} |a - b| \quad (1)$$

4. Matching Algorithms

We first determine facial features in the input face photograph using the Active Shape Models. Next, we match each detected face component to the pre-generated cartoon-like avatar components, and find the most similar avatar component. Finally, we compose a cartoon-like avatar which looks similar to the input photograph by placing the matched components in appropriate positions. Figure 1 shows the overall procedure.

4.1. Graph Matching

We can intuitively assume that landmarks on the face image are the nodes of a graph. We can also define edges connecting pairs of landmarks (nodes) along the boundaries of the facial components. The edges have the value of both distance and slope. Since we need the graph similarity between a face image and an avatar, we also need a graph of an avatar.

Let us assume that two graphs have exactly the same number of nodes and the same edge structure. Graph similarity is very complicated problem in the general case but with the assumption above, it can be defined as:

$$G_{i,a} = \sum_{j=1}^{N} \frac{1}{N} \big(\omega_j |\delta_{i,j} - \delta_{a,j}|\big) \quad (2)$$

, where x denotes the coordinate of a node, and δ denotes the slope value of an edge. We define ω as weights for the node distances and the edge slope differences since the node distances are generally larger than the edge slope difference.

4.2. Image Matching

In the case of the eye matching, the number of landmarks from the ASM is only 5 for each eye. Those landmarks can barely represent the characteristic of eyes. So we need to match directly from the image components, and that is the area in which the modified Hausdorff distance is used.

The MHD may provide the proper similarity score between eye images and eye avatars, but there is a problem that one dominant avatar is the most likely to be selected if we directly score images by the MHD. For instance, the first eye image on the middle row in Figure 2 is obviously the most human-like eye. The other one may a represent slightly smaller and sharper eye, but in the real world, such eyes do not exist.

4.3. Relative Similarity

The graph matching or Hausdorff distance may provide a proper similarity score for exact

match. However, even very characterful features are slightly different from normal features in overall shape, and cartoon-like avatars normally exaggerate features. Therefore, one dominant avatar is the most likely to be selected with the exact match of cartoon-like avatar components.

For matching cartoon-like components, we compute a relative similarity. We collect face images to compute their graph similarities and Hausdorff distances to the avatar components. We normalize the graph similarity and the Hausdorff distance of the input image with respect to the average and the standard deviation of the similarity scores in the face database. Matching decision is made based on the normalized score.

We define the normalized graph similarity $\hat{G}_{i,a}$ and the normalized Hausdorff distance $\hat{H}_{i,a}$ as follows:

$$\hat{G}_{i,a} = \frac{G_{i,a} - \bar{G}_a}{\sigma_{G,a}}, \quad \hat{H}_{i,a} = \frac{H_{i,a} - \bar{H}_a}{\sigma_{H,a}} \tag{3}$$

,where $\bar{G}_a$ and $\sigma_{G,a}$ are the average and the standard deviation of the *a*-th avatar component's graph similarities to the face components in the database. Likewise, $\bar{H}_a$ and $\sigma_{H,a}$ are the average and the standard deviation of the *a*-th avatar component's Hausdorff distances to the face components in the database. These relative similarities denote how far the similarity locates from the average in fold of standard deviation.

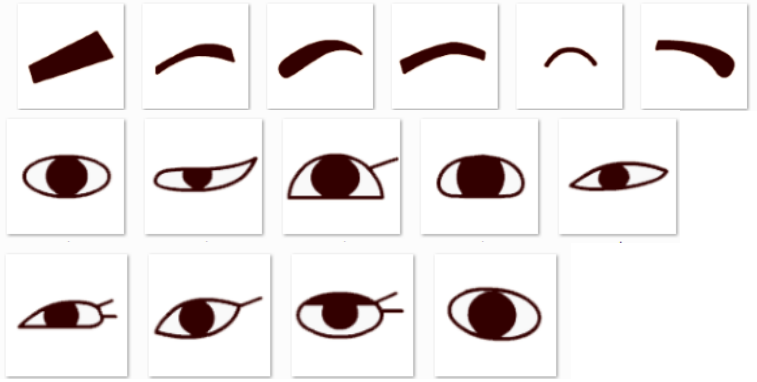

Figure 2. Samples of the Avatar Image

5. Component Matching

We have a total of 6 types of eyebrow, 9 types of eye, and 3 types of facial shape in the matching system, excluding extraordinary avatar components for the cartoon expression. And accessories such as glasses and mustaches are also excluded. Examples of sample avatar images are shown in Figure 2.

5.1. Eyebrow

Eyebrow landmarks directly represent the shape of the eyebrow and extracting the corresponding landmarks from the eyebrow avatar is simple. Evaluating the graph similarity is also quick, and this type of matching presents better performance than one based on the Hausdorff distance.

5.2. Eye

There are only 5 landmarks on each eye, left, right, top, bottom, and center. Those points can barely represent the characteristic of eyes, therefore graph matching is not appropriate. Before calculating the MHD, the eye image needs to be edge-detected since we define the distance between the black pixels in a pair of images.

Figure 3. Eye Images, Edge-deteced Eye Images and Matched Avatar Components

5.3. Face

A face has 15 landmarks on the boundary. Our 3 face avatars differ only in terms of the shape – round, normal and sharp. As we match the eyebrow image to the avatar, the face image is suitable for the graph method but slightly different from similarity matching. We select the 7-th landmark as the center and track down the first landmark to evaluate the slope values from the center to each landmark. Because the outline of the face is significantly distinguished in the vicinity of the chin, assigning a weight value on the edge near the center landmark provides an improved result.

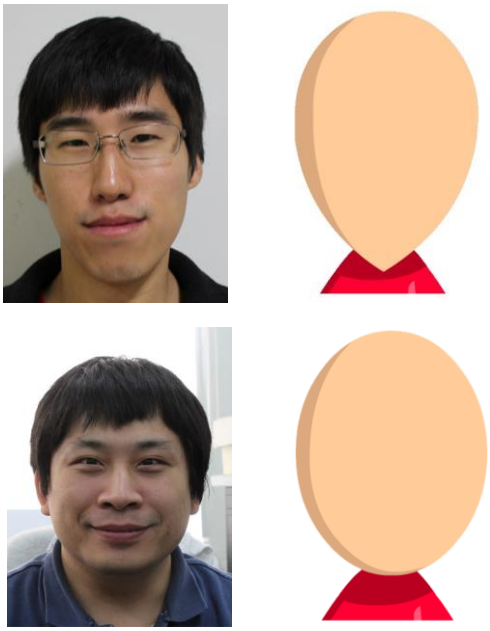

Figure 4. Result of the Face Matched Avatars

6. Results and Validation

6.1. Composite Avatars

Figure 5 shows sample face images and their corresponding result avatar images generated by our system. Our application only suggested eyes, eyebrows, a mouth, and face shape. Hair, a nose, and accessories were matched manually.

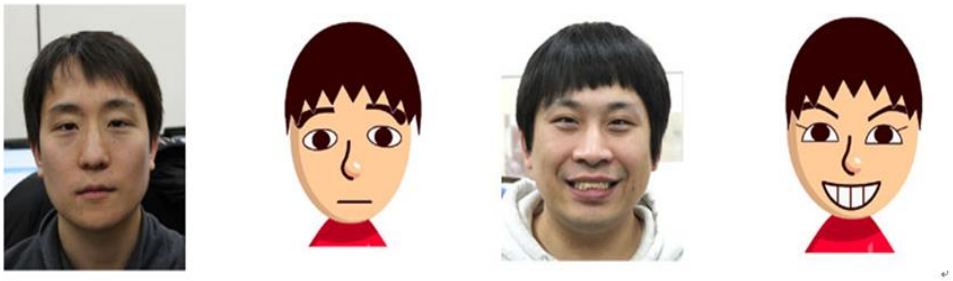

Figure 5. Avatars Matched by our Application (Right) and their Input Face Images (Left)

6.2. User Experiments

We performed two experiments for validating our approach. For both experiments, each subject is provided 10 face photographs and two avatars for each photograph where one of them is suggested by our system and the other is by the compared method. Subjects selected which avatar is more similar to the corresponding photograph. Subjects are 18 people with various backgrounds.

The first experiment is for showing that our system is better than a random system which randomly choose facial components and composite them. The result of the selection frequency distribution is shown in Table 1.

Table 1. Selection Frequency Distribution of Random and Suggested Avatar Set

Set No.	1	2	3	4	5	6	7	8	9
Random	3	2	2	1	0	1	3	1	2
Suggested	7	8	8	9	10	9	7	9	8
Set No.	10	11	12	13	14	15	16	17	18
Random	2	1	3	2	1	2	0	4	3
Suggested	8	9	7	8	9	8	10	6	7

We analyzed the result with one sample T-test. The null hypothesis is that our system is not better than the random system, and the result is $t=12.23$ and $p<0.001$. Therefore, we can reject the null hypothesis with 99% confidence level, which means that our system suggests better avatars than the random system in statistically significant manner.

Table 2. Selection Frequency Distribution of Manual and Suggested Avatar Set

Set No.	1	2	3	4	5	6	7	8	9
Manual	5	5	6	6	5	7	6	5	5
Suggested	5	5	4	4	5	3	4	5	5
Set No.	10	11	12	13	14	15	16	17	18
Manual	3	6	5	5	8	4	4	6	5
Suggested	7	4	5	5	2	6	6	4	5

The second experiment is for comparing our system to the manual avatar generation. There is no ground-truth for avatars since the quality of avatars is subjective aesthetic criteria. Avatars of the best quality would be avatars manually generated by ordinary people. The result of the selection frequency distribution is shown in Table 2. With one sample T-test, the result is $t=-1.24$ and $p=0.115$. Therefore, we cannot reject the null hypothesis with 99% confidence level.

7. Conclusions and Limitations

Our application composes each avatar component which best matches with input facial features such as eyes, eyebrows, a mouth and face shape. Each feature is independently chosen by various matching algorithms which are appropriate to explain its structure and characteristics. The overall components of the matched avatar are placed considering the original disposition of the facial feature. Images of the completely composed avatar are shown in Figure 5.

Works still needs to be done in the rough matching system. Designing avatars involves extensive research in various areas and the matching of even one facial feature needs intensive studies. Furthermore, regarding the characteristics of cartoon avatars, opinions about the best match depend on personal perspective.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology(grant number 2010-0011068). This work was also supported by Korea Science & Engineering Foundation through the NRL Program (Grant ROA-2008-000-20060-0).

References

[1] A. Hogue, M. Jenkins and S. Gill, "Automated Avatar Creation for 3D Games", Proceedings of the 2007 conference on Future Play, **(2007)**, pp. 174-180.
[2] M. Obaid, D. Lond, R. Mukundan and M. Billinghurst, "Facial Caricature Generation Using a Quadratic Deformation Model", Proceedings of the International Conference on Advances in Computer Entertainment, **(2009)**, pp. 285-288.
[3] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active Appearance Models", Proceedings of European Conference on Computer Vision, **(1998)**, pp. 484-498.
[4] H. Chen, Z. Liu, C. Rose, Y. Xu, H. Y. Shum and D. Salesin, "Example-based Composite Sketching of Human Portraits", Proceedings of the 3rd International Symposium on Non-photorealistic Animation and rendering, **(2004)**, pp. 95-153.
[5] P. W. Hallinan, "Recognizing human eyes", SPIE Proceedings, Geometric Methods in Computer Vision, vol. 1570, **(1991)**, pp. 214-226.
[6] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models – their training and their applications", Computer Vision and Image Understanding, vol. 61, **(1995)**, pp. 38-59.
[7] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human face", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 12, **(1990)**, pp. 103-108.
[8] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, **(1991)**, pp. 71-86.
[9] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "What is the set of images of an object under all possible light conditions?", IEEE Conference on Computer Vision and Pattern Recognition, **(1996)**, pp. 270-277.
[10] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrival", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, **(1996)**, pp. 831-836.
[11] K. Fukunaga, "Statistical Pattern Recognition", Academic Press, New York, **(1989)**.
[12] W. Zhao, "Robust Image Based 3D Face Recognition", Ph.D. dissertation, University of Maryland, College Park, MD, **(1999)**.
[13] L. Wiskott, J. -M. Fellous, N. Kruger and C. von der Malsburg, "Face recognition by elastic bunch graph matching", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, **(1997)**, pp. 775-779.
[14] S. Y. Lee, I. J. Kim, S. C. Ahn, H. Ko, M. T. Lim and H. G. Kim, "Realtime 3d avatar for interactive mixed reality", Proceedings of the ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry, **(2004)**, pp. 75-80.
[15] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor and H. P. Seidel, "Automatic generation of personalized human avatars from multiview video", Proceedings of the ACM symposium on Virtual reality software and technology, **(2005)**, pp. 257-260.