รายงานวิจัยฉบับสมบูรณ์

โครงการ

การสร้างต้นแบบหน้าโดยอาศัยภาพตัวอย่าง เพื่อการค้นคืนฐานข้อมูลภาพล้อเลียน

โดย ดร.กิ่งกาญจน์ สุขคณาภิบาล

มกราคม 2554

สัญญาเลขที่ **MRG5080436**

รายงานวิจัยฉบับสมบูรณ์

โครงการ
การสร้างต้นแบบหน้าโดยอาศัยภาพตัวอย่าง เพื่อการค้นคืนฐานข้อมูลภาพล้อเลียน

ดร.กิ่งกาญจน์ สุขคณาภิบาล — **จุฬาลงกรณ์มหาวิทยาลัย**

Ritsumeikan University, Japan

อาจารย์ จักริน สมิตเวช — **มหาวิทยาลัยธุรกิจบัณฑิตย์**

ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์ — **จุฬาลงกรณ์มหาวิทยาลัย**

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกว.ไม่จำเป็นต้องเห็นด้วยเสมอไป)

บทคัดย่อ

รหัสโครงการ : MRG5080436
ชื่อโครงการ : การสร้างต้นแบบหน้าโดยอาศัยภาพตัวอย่างเพื่อนการค้นคืนข้อมูลภาพล้อเลียน
ชื่อนักวิจัย : **ดร.กิ่งกาญจน์ สุขคณาภิบาล**
ภาควิชาวิทยาศาสตร์ภาพถ่าย คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และ Intelligent Computer Entertainment Laboratory, Department of Human and Computer Intelligence, Ritsumeikan University, Japan
อาจารย์ จักริน สมิตเวช
คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์
Advanced Virtual and Intelligent Computing (AVIC) Center คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
E-mail Address : kingkarn@ice.ci.ritsumei.ac.jp
jackarin@it.dpu.ac.th
chidchanok.l@chula.ac.th
ระยะเวลาโครงการ : 2 ปี

ภาพล้อเลียนเป็นการนำเสนอภาพรวมของบุคคล หรือวัตถุโดยเน้นคุณสมบัติ ที่โดดเด่นที่สุด ให้มากเกินจริง และลดความซับซ้อนของคุณสมบัติทั่วไป เพื่อที่จะทำให้ภาพที่ได้แตกต่างจากภาพอื่นๆ และในขณะเดียวกันได้คงความเหมือนกันของภาพไว้ ปัจจุบันมีงานวิจัยจำนวนมาก รายงานผลการวิจัยว่า มนุษย์สามารถรู้จำ ใบ-หน้า จากภาพล้อเลียนได้รวดเร็วและแม่นยำกว่า การรู้จำใบหน้าจากภาพถ่าย โดยเฉพาะอย่างยิ่ง ในการรู้จำใบหน้าสำหรับคนที่ไม่คุ้นเคย

ในการศึกษานี้ เราได้ทำการทดลองสองประเด็นหลัก ที่เกี่ยวกับการใช้ภาพล้อเลียนในการรู้จำใบหน้า

- ประการที่หนึ่ง เราได้ทำการพิสูจน์ว่า อัตราการรู้จำภาพถ่ายใบหน้า นั้นสูงขึ้นด้วยการใช้ภาพล้อเลียน ร่วมกับภาพถ่ายจริงของใบหน้ามนุษย์
- ประการที่สอง เราได้ดำเนินการจัดการสืบค้นภาพล้อเลียน โดยใช้เซอร์ไนก์โมเมนต์ ซึ่งเป็นคุณสมบัติ ที่ไม่แปรเปลี่ยน

เซอร์ไนก์โมเมนต์มีคุณสมบัติที่ทนต่อการหมุนของภาพ และการแปรเปลี่ยนทางเรขาคณิตของภาพ การคำนวณความคล้ายกันของภาพล้อเลียน สามารถวัดได้โดยใช้ระยะทางแบบยุคลิด ระหว่าง เซอร์ไนก์-โมเมนต์ของภาพล้อเลียน

คำหลัก : การค้นคืนภาพ ภาพล้อเลียน โครงข่ายประสาทเทียม การรู้จำใบหน้ามนุษย์
กิตติกรรมประกาศ : ผู้วิจัยขอขอบคุณสำนักงานกองทุนสนับสนุนการวิจัย (สกว.)

Abstract

Project Code : MRG5080436

Project Title : Generating an example-guided prototype for facial caricature database retrieval

Investigators : **Dr. Kingkarn Sookhanaphibarn**
Department of Imaging Technology, Faculty of Sciences, Chulalongkorn University and Intelligent Computer Entertainment Laboratory, Department of Human and Computer Intelligence, Ritsumeikan University, Japan
Mr. Jackarin Smitaveja
Faculty of Information Technology, Dhurakij Pundit University
Professor Dr. Chidchanok Lursinsap
Advanced Virtual and Intelligent Computing (AVIC) Research Center, Faculty of Sciences, Chulalongkorn University

E-mail Address : kingkarn@ice.ci.ritsumei.ac.jp
jackarin@it.dpu.ac.th
chidchanok.l@chula.ac.th

Project Period : 2 years

Caricature is a pictorial representation of a person or subject in summarizing way by exaggerating the most distinctive features and simplifies the common features in order to make that subject different from others and at the same time, preserve the likeness of the subject. Currently, there have been a plenty number of researches on the recognition of facial caricatures outperforms that of their facial photographs, especially, on the recognition for unfamiliar people. In this study, we have had two main experiments on a use of caricatures for the facial recognition. First, we have examined the caricature advantage that the recognition rate will be increased with the extracted metrical features of facial caricatures that are combined with their facial photographs in the training set. Second, we have implemented a similarity retrieval of facial caricatures by using their invariant features extracted by Zernike moments. The Zernike moments are selected as feature extractor due to its robustness to image rotation, geometrical invariants property and orthogonal property. These features are used to identify the similarity between two facial caricatures by computing the Euclidean distance between feature vectors.

Keywords : Image retrieval; Caricature; Neural Network; Facial recognition;

Acknowledgement:

The authors are grateful to Thailand Research Fund (TRF) for their generous financial support.

I. INTRODUCTION

Today, people can easily find shops for drawing caricatures of their customers on a street fair, a carnival, or a shopping mall. It is a very interesting why people prefer to have their caricature, in a funny way, instead of their beautiful portraits. One reason is that human can memorize the caricatures one and a half faster than those of their portraits [Gooch et al., 2004]. Figure I-1 shows some caricatures posted in the Website and their source was taken at El Pasco Fair.

Figure I-1: Caricature pictures drawing on street [El Pasco Fair, 2007].

Initially, our motivation of this study is that not only portraits but also caricatures recognized as history artworks collected in a gallery; for example, a well-known gallery of a collection of caricatures, Sardi's is a restaurant in New York City's theater district at 234 West 44th Street in Manhattan. In 1979, Vincent Sardi, Jr. donated a collection of 227 caricatures from the restaurant to the Billy Rose Theatre Collection of The New York Public Library for the Performing Arts. The contributed caricatures date from the late 1920s through 1952.

In this study, we introduce a research on the caricatures that people have known them as insulting or complimentary tools in politic columns. Another aim of caricatures is for entertainment such as caricatures of super stars found in magazines. Caricatures also provide a powerful metaphor for illustrative visualization [Dror et al., 2008; Frowd, 2007; Rautek et al., 2006]. The caricature advantage demonstrates that performance is better when exaggerated stimuli are presented rather than a faithful

image. This can be understood with respect to a theoretical framework in which caricaturing maximizes the distinctiveness and thus minimizes any perceptual or representational confusion.

The caricatures have advantages over the portraits in terms of the human perception and cognition. Human faces share the same basic size and shape. They also share the same basic structure of facial features as two eyes above a central nose and above a central mouth. It is difficult that people must find the different facial features of one face from another. The perception of this distinctive facial information is crucial to people ability to process faces. Consequently, highly distinctive faces, in which the individuating cues are obvious, are processed with much more efficiency than typical faces in which distinguishing cures are very subtle. Caricatures as distinctive faces are more easily and more confidently recognized.

Over the last two decades or so, face recognition has become a popular area of research in computer vision and one of the most successful applications of image analysis and understanding. Because of the nature of the problem, not only computer science researchers are interested in it, but neuroscientists and psychologists also. It is the general opinion that advances in computer vision research will provide useful insights to neuroscientists and psychologists into how human brain works, and vice versa. We have briefly explained in the next section.

A general statement of the face recognition problem (in computer vision) can be formulated as follows: Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces. Nowadays, a facial recognition system has been developed as a computer application for automatically identifying or verifying a person from a digital image or a video frame from a video source. One of the ways to do this is by comparing selected facial features from the image and a facial database. This is typically used in security systems and can be compared to other biometrics such as fingerprint or eye iris recognition systems.

Nowadays, there have been a plenty number of researches investigated in the following problems [Zhao, 2003; Blanz, 2005;]:

- Recognition from outdoor facial images.
- Recognition from non-frontal facial images.
- Recognition at low false accept/alarm rates.

In this study, we initially introduce the facial identification guided by facial caricatures as well as the caricatures can retrieved by using the Zernike moments. Thus, the four objectives were as follow:

- To examine characteristics of portraits and caricatures of human faces.
- To test whether the characteristics of caricatures increase the recognition accuracy of facial recognition systems.

- To develop a similarity retrieval of facial caricatures.

To examine characteristics of photographs and caricatures of human faces, we have been interested in a use of different analytic techniques such as metrical analysis (measurements), morphological analysis (shape of the features) and superimposition. These techniques can be used for comparisons between two facial photographs, or between an actual face and a photograph. The dimensions and characteristics of the face on the two photographs are compared to investigate if it belongs to the same person, or if it can be excluded from being that person. A well-known application regarding this technique is access control systems.

To test whether the characteristics of caricatures increase the recognition accuracy of facial recognition systems, we have implemented the feed forward networks for recognition of human faces and compared two recognition systems that one uses only facial photographs and another one uses both facial photographs and caricatures. The implemented neural network is a simplest model with the following properties:

- Unidirectional flow of the input data
- Good at extracting patterns, generalisation and prediction
- Distributed representation of data
- Parallel processing of data
- Training by the Backpropagation

To develop a similarity retrieval of facial caricatures, Zernike moments are used as invariant descriptors of the image shape. Zernike moments are superior to other moment functions such as geometric moments in terms of their feature representation capabilities and robustness in the presence of image quantization error and noise. Their orthogonality property helps in achieving a near zero value of redundancy measure in a set of moment functions. Thus, moments of different orders correspond to independent characteristics of the image.

The rest of the study is structured as follows. Section II briefly describes the literature reviews relevant to the caricatures and the recognition systems. Section III examines the methodologies of our study and their experimental results. Finally, Section IV concludes the paper and gives suggestions for future research.

II. LITERATURE REVIEWS

In this section, we describe and discuss on the state-of-arts works relevant to the caricatures generated by computers, the facial recognition in aspects of both human perception and computer recognition.

2.1 Computer Generated Caricature

Recently, computer generated caricature becomes particularly interesting research topic due to the advantageous features of privacy, security, simplification, amusement and their explosive emergent real-world application such as in magazine, digital entertainment, Internet and mobile application. Software for generating caricatures have been developed in order to assist the user in producing caricature automatically or semi-automatically.

2.1.1. Human Facial Sketches

Starting with the human facial sketches, several researchers proposed a technique of generating human facial sketches. Their systems automatically generated a sketch from an input image, by learning from example sketches drawn with a particular style by an artist [Lee et al., 2007; Rautek et al., 2006; Liu et al., 2005; Yang, 2004; Liang, et al. 2002; Chen et al., 2001]. Given an input image pixel and its neighborhood, the conditional distribution of a sketch point was computed by querying the examples and finding all similar neighborhoods. Their work did not consider exaggeration of human faces.

2.1.2 Exaggerated Shape

Liang, et al. [Liang, et al. 2002] presented an example-based caricaturing system. They divided the caricature generation into two parts, i.e., shape exaggeration and texture style transferring. From example caricatures drawn by an artist, they captured the artist's understanding of what are distinctive features of a face and the exaggeration style. Combining this with sketches or images, they can automatically generate an effective caricature from an input image, with exaggeration.

The approach of how to capture the artist's understanding was to be done within two phases [Liang, et al. 2002]. At the training phase, they analyzed the correlation between images and caricatures from training examples and then construct a set of exaggeration prototypes. At the runtime phase, they classified it into one of the exaggeration prototypes for an input shape and then exaggerated the input shape by the selected prototype.

In [Liang, et al. 2002], they proposed an example-based method used to learn how to identify facial features and exaggerate them using the artist's style, albeit implicitly. However, key facial features were explicitly selected by their artist to exaggerate and to maintain consistent exaggeration styles throughout all examples.

2.1.3 Exaggerated Texture

Lee, et al. [Lee et al., 2007] proposed an approach to automatically exaggerate the distinctive features in extremely detailed 3D faces is discussed in this paper. They used two low polygon approximations of the detailed face. First, a working model was created by fitting a generic head model to the high-resolution data. Second, a simplified model was produced by any model simplification algorithm. The high-resolution detail was then captured in a two-step procedure by performing model parameterization.

They also employed the point-to-surface mapping to parameterize the dense model with respect to the simplified model and to parameterize the simplified model with respect to the working model. The working model's characteristics were exaggerated using a vector-based caricature algorithm that automatically enhanced the prominent features by comparing the working model to an average face.

At the final step, they used both the exaggerated working model and the simplified model parameterization in order to generate an exaggerated simplified model, which was in turn used with the dense model parameterization to produce the exaggerated version of the highly detailed face.

2.2 Face Recognition

2.2.1 Human Perception

What are the most important features of face playing in face recognition? There have been a number of researches in area of human perception trying to find answers the aforementioned question. In these studies, the principal features of face were divided into two aspects: the internal features of a face mean eyes, nose, mouth, etc. The external ones, by contrast, are hair, chin, face outline, and so on.

Young, et al. [Young et al., 1985] conducted an experiment of the influence of the internal and external features of a face to a human perception of familiar or unfamiliar faces. By conducting an experiment, subjects were asked to match a photograph of a complete face and a simultaneously presented photograph of internal or external features of a face, deciding whether or not the two photographs were pictures of the same person.

In the experiment, 'same' pairs were derived from different pictures of the same face, so that subjects had to match the faces and not the particular photographs used. The findings were reported as follows:

1) Matches based on internal features were found to be faster for familiar than for unfamiliar faces.
2) There was no difference in reaction time between matches based on the external features of familiar and unfamiliar faces.
3) Faster matching of internal features of familiar faces was found to hold equally for pairs of photographs that differed in orientation of the face or in facial expression.

In terms of neurosciences, some researchers [Andrew et al., 2010; Axelrod, 2010] have investigated the holistic processing of face recognition. It was a process that human beings perceive the face as a whole and not as a set of separate, independently processed features. For example, people have experienced the difficulty of recognizing a friend or a colleague who had changed her hairstyle or had shaved his beard. Such a change seems like not just a change in the facial hair, but rather the whole face looks different. This phenomenon of face processing was called holistic processing.

The studies [Akselrod-Ballin and Ulman, 2008; Axelrod, 2010; Andrew et al., 2010], who investigated holistic processing by manipulating external and internal facial features, concluded the following facts:

1) The response to isolated features of the face provided clear neurological support for the idea that the dominant role of internal features is in the neural representation of familiar faces.
2) The release from adaptation when viewing composite faces suggested that both the internal and external features are important in face perception.

Such as the aforementioned researches, the studies have indicated the significance of certain facial features, particularly internal ones such as the eyes, nose and mouth. People have known that eyebrows play an important role in emotional expression and nonverbal communication, as well as in facial aesthetics and sexual dimorphism. However, it is not clear that the role of eyebrows is in the identification of faces. Figure II-1 shows a comparison among complete faces, ones without eyebrows, ones without eyes.

Sadr, et al. [Sadr et al., 2003] concluded that eyebrows do indeed play an inherently important role in face recognition. They reported the experimental results that the eyebrows may be at least as influential as the eyes. Specifically, they found that the absence of eyebrows in familiar faces leads to a very large and significant disruption in recognition performance. In fact, a significantly greater decrement in face recognition is observed in the absence of eyebrows than in the absence of eyes. Figure II-1 shows a comparison among facial pictures with eyebrows or without ones.

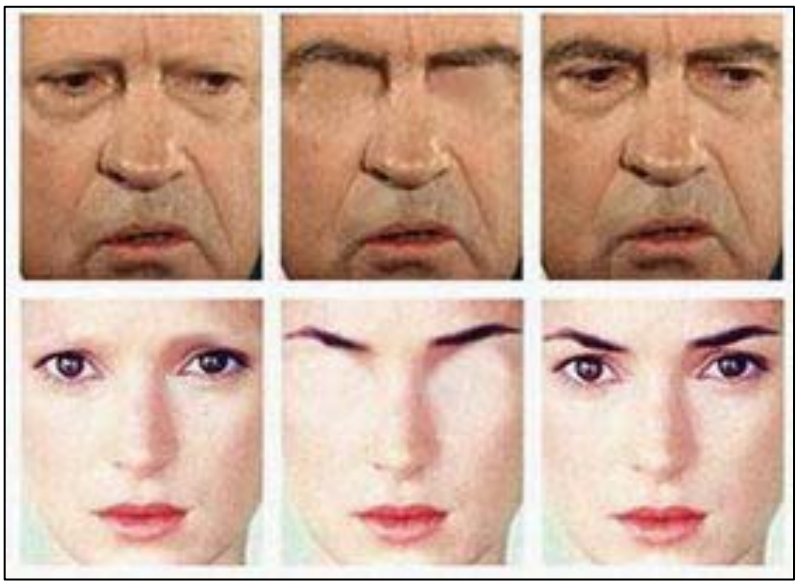

Figure II-1: The images on the right are original images of Richard Nixon and Winona Ryder; Most people have a better chance of recognizing a person with eyebrows and no eyes (center), rather than with eyes and no eyebrows (left) [Sinha, et al. 2006].

Studying the human perception, we gather these results having important implications for our understanding of the mechanisms of face recognition in humans as well as for the development of artificial face-recognition systems.

2.2.2 Computer Recognition System

Face Detection

Suphakant Phimoltares and et al. [Phimoltares et al., 2007] proposed a facial feature detection algorithm for all types of face images in the presence of several image conditions such as color images, gray images, binary images, wearing the sunglasses, wearing the scarf, lighting effect, noise and blurring images, color and sketch images from animated cartoons. Their face detection method consisted of two main stages. In the first stage, the faces were detected from an original image by using Canny edge detection and our proposed average face templates. Second, a neural network-based approach was used to recognize all possibilities of facial feature positions. Input parameters were obtained from the positions of facial features and the face characteristics that were low sensitive to intensity change.

Facial Feature

A wide rage of features, which is made of the image appearance, has been used directly for face recognition [Datta et al., 2005; Dauodi and Matusiak, 2007; Lew et al., 2006]. Consequently, their features are under various illuminations. These approaches use either global descriptions such as

Fisherfaces and Eigenfaces [Zhao et al., 2006; Gunturk et al., 2003] or the appearance of local face regions such as Laplacianfaces [He et al., 2005].

Facial Sketches

Alberto Del Bimbo and Pietro Pala [Bimbo and Pala, 1995] proposed a new matching of a user's sketches, called "Elastic Matching". They used elasting matching of sketched templates over the shapes in the images to evaluate similarity ranks. The degree of matching achieved and the elastic deformation energy spent by the sketch to achieve such a match were used to derive a measure of similarity between the sketches and the images in the database and to rank images to be displayed. The elastic matching was integrated with arrangements to provide scale invariance and take into account spatial relationships between objects in multi-object queries.

Stefan Muller and et al. [Muller et al., 1998] proposed an image retrieval system, which enabled users to search a gray-scale image database intuitively by presenting simple sketches. Their experiments were tested against on hand tool database, which each image is an isolated object. Their aim was to solve the problem of invariant feature extraction of isolated objects, i.e., rotation and scaling invariances.

Zhihua He and Maja Bystrom [He and Bystrom, 2005] as well as Qingshan Liu and et al. [Liu et al., 2005] proposed algorithms for sketch image retrieval. They introduced the approach of directional projection-based sketch image retrieval system. This system composed of efficient filter banks adopted to decompose the image into directional sub-bands in order to capture edges at different orientations. Through projecting of each sub-band into a pair of orthogonal profiles, the shape information was compactly captured by identifying the structure of each one-dimensional projection profile. By comparing the projection profiles of the sketched image with those in an image database, a set of images with shapes similar to the sketched images were retrieved.

Mohamed Daoudi and Stanislaw Matusiak [Dauoudi and Mutusiak, 2007] described an application allowing content-based retrieval, called "sketch-based database retrieval". This application allowed users to interact with the database by means of sketches. Users were able to draw his request with a pencil; the request image was then a binary image comprising a contour on a uniform bottom. The image-retrieval based on users' hand-draw sketches was emphasized in their research. Normally shape was subjective and widely various. They also proposed Curvature Scale Space (CSS) description, which defined an invariant distance in the shapes space. Their work did not focus on the problem of facial features, which were much different from general sketch such as tools, cars, machines, etc.

III. METHODOLOGIES

As shown in Figure III-1, caricatures can amplify the characteristics of intensive cues toward the details of interest. People can attract highly exaggerated features to aid in the recognition of differences.

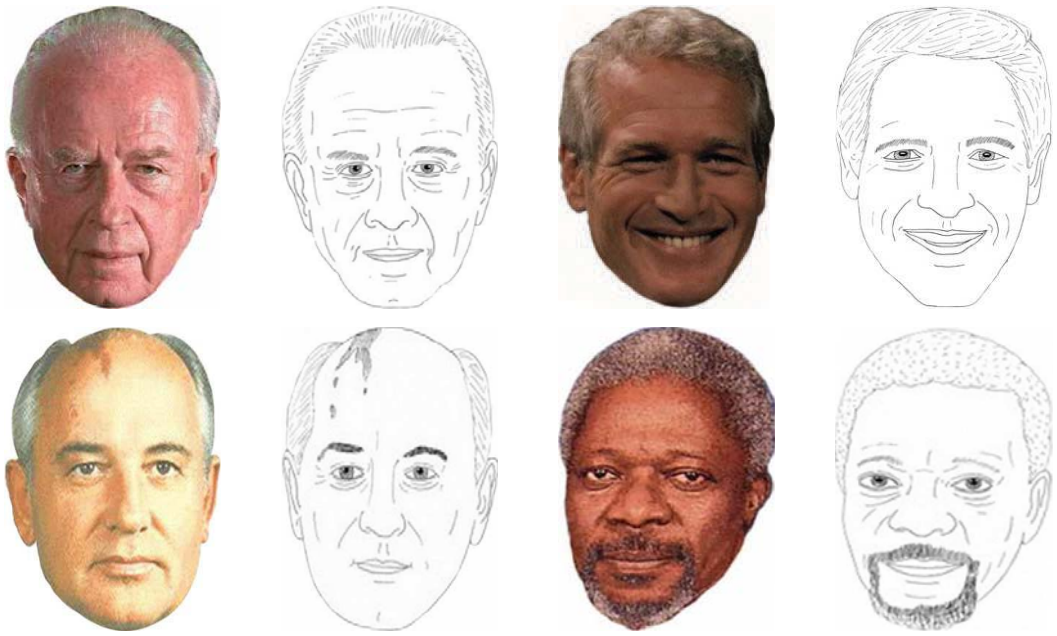

Figure III-1: Examples of face images and their corresponding caricatures.
First row: Yitzhak Rabin and Paul Newman.
Second row: Mikhail Gorbachev and Kofi Anan
(excerpted from [Newman, 2008; Alan, 2009; Prichett, 2009]).

Using metrical features such as forehead size, chin size, mouth breadth, vertical lip height, and nose breadth in comparisons between two facial photographs were investigated in [Roeofse et al. 2008] that if it belongs to the same person, or if it can be excluded from being that person. They emphasized the knowledge of common and rare facial characteristics seen in various populations. The metrical features extracted from biometric landmarks of the face reported in [Roelofse et al. 2008] are used in this study.

Neural network approaches are a utility to infer a function from observations. Unsurprisingly, neural networks yield effectiveness and efficiency in classification problems [Kussul et al., 2004; Xu and Ahmadi, 2007; Uglov et al., 2008]. Traditionally, training algorithm such as back-propagation is used in this study. In our experiments, we test the neural networks against training sets with caricatures and without caricatures. The results show that those with caricatures significantly outperform those without caricatures.

3.1 Facial Characteristics

To examine characteristics of portraits and caricatures of human face, we focus on researches of metric characteristics that have been done on various faces over the world. The use of these metrics minimizes the effects of distortion in photographs and differences in size between images, and measurement errors, as it is only necessary to reserve the individual in terms of the photographs and caricatures.

Biometric landmarks used for facial analysis [Roelofse et al. 2008] are equal to 22 points as shown in Figure III-2(a). All numbered landmarks are summarized in Table III-1.

Table III-1: Demonstrations of landmarks.

Points	Description
1,2	Endocanthion
3,4	Exocanthion
5,6	Iris
7	Vertex
8	Trichion
9	Glabella
10	Nasion
11	Subnasale
12	Labiale
13,14	Stomion
15	Labiale Inferius
16	Gnathion
17,18	Zygion
19,20	Alare
21,22	Cheilion

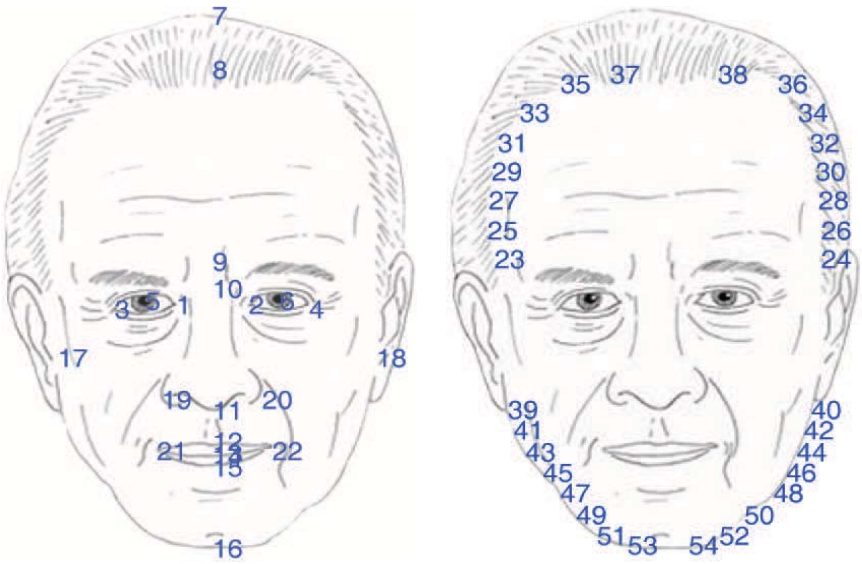

Figure III-2: (a) Point numbers 1-22 at standard biometric landmarks of a face.
(b) Point numbers 23-38 at upper facial contour and point numbers 39-54 at lower facial contour.

To achieve a facial identification, upper and lower facial contours as shown in Figure III-2(b) are also considered in this study. Those marked points are capable to be the representative of facial shape, jaw line, chin shape, upper lip notch, philtrum, septum tilt, nasolabial fold, and nose bridge height. All numbered facial contours are summarized in Table III-2.

Table III-2: Demonstrations of facial contours.

Facial contours	Descriptions
(23,24)	Both end of horizontal line through glabella (9)
(25,26), (27,28), (29,30), (31,32), (33,34), (35,36), (37,38)	Both end of horizontal line at 15%, 30%, 45%, 60%, 75%, 90%,95% of the vertical distance from glabella (9) to trichion (8)
(39,40)	Both end of horizontal line through subnasale (11)
(41,42), (43,44), (45,46), (47,48), (49,50), (51,52), (53,54)	Both end of horizontal lien at 15%, 30%, 45%, 60%, 75%, 90%, 95% of the vertical distance from subnasale (11) to gnathion (16)

In our approach, the features are normalized by the reference distance, which is a distance from left endocanthion to right endocanthion. Our facial features consist of 3 parts as illustrated in Figure III-3. All 29 facial features are called “distant index”, each of which is measured in Euclidean distances in vertical and horizontal directions.

On vertical line as shown in Figure III-3(a), 8 distant indices are drawn from each marked point to the facial center, which captures the vertical detail for facial parts. 21 horizontal distant indices cover remain regions, as shown in Figure III-3(b), composed of 2 distant indices and the reference distance that are drawn along the eye region, the other 3 distant indices are measured of facial width, nose width, and lip width. The remaining 16 distant indices are equally divided in the lower and upper regions as shown in Figure III-3(c).

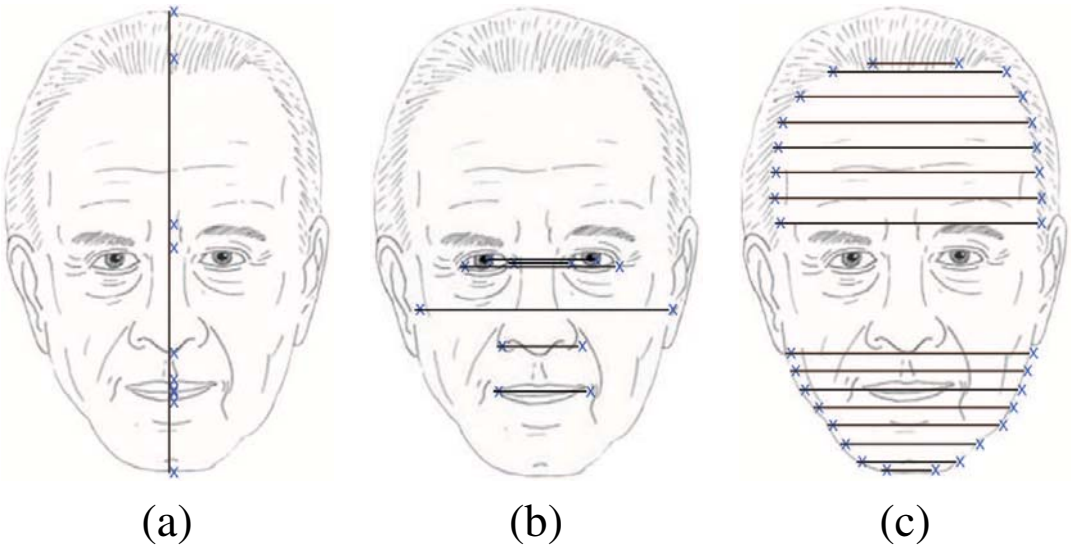

Figure III-3: (a) Vertical distant indices. (b) Horizontal distant indices in the middle region.
(c) Horizontal distant indices in the upper and lower regions.

3.2 Experiment I

3.2.1 Data sets

Our hypothesis is that face recognition systems with facial caricature images yield significantly outperformance of accuracy than those without facial-caricature images. We design the experiment into two parts: five-people classification and ten-people classification, because we also investigate the effect from the numbers of class (people) to the performance of our face recognition systems.

All face images and their caricatures using in our experiments are from the Internet resources. The corresponding caricature images are in pen and ink styles created by caricature artists [Alan, 200; Prichett, 2009].

In our experiments, there are two different data sets, the data set with photographic images will be called "facial photograph data set" and the other data set with caricature images will be called "facial caricature data set". The neural learning process based on only the facial photograph data set is called "Photo-Pattern-Based (PPB)" and the learning process based on facial caricature data set that add on to the facial photograph data set is called "Caricature- Pattern-Based (CPB)".

Let P_i^k be the i^{th} photographic image of a frontal face belonging to the k^{th} person. In our experiment, we gather 10 different image from the same person, therefore, we define the data set, A_i, consisting of distant indices of N different people for all images P_i^k, where $1 \leq k \leq N$, as shown in Eq. (1). In this equation $f()$ is the feature function mapping from an image to a distant index.

$$A_i = \{f(P_i^k) | 1 \leq k \leq N\} \tag{1}$$

For five-people and ten-people classification, N is set to five and ten, respectively. As the similar manner, we can define the caricature image of a frontal face belonging to the k^{th} person denoted by Q^k and there is only one caricature image for one person. Thus, the elements of the data set B are 10 caricature images.

$$B = \{f(Q^k) | 1 \leq k \leq N\} \tag{2}$$

In our study, we examine all the possible combinations of training data sets and testing data sets to avoid the problem of unknown data distribution and to increase the accuracy of the results. The super set of data sets selected for training and testing is computed from the simple binomial coefficient $C(n, r)$, that is the selection of r items from n items, where $n = 10$ and $1 \leq r \leq 9$.

For example, if the proportion of the number of data sets belonging to training and testing is 2:8, the training super set composes a set of the following members:

$$C(10,2) = \{\{1,2\}, \{1,3\}, \dots, \{9,10\}\} \tag{3}$$

where $|C(10,2)| = 45$. Reference with the set element in the super set, $\{1,2\} \in C(10,2)$ is a training set at particular run. Generally, $c_j(n, r)$ is the j^{th} element in the super set $C(n, r)$, hence, $C_1(10,2)$ defined as the training set of the 1^{st} run is equivalently $\{1,2\}$. We have two learning processes: PPB and CPB. Therefore, the training set of PPB and CPB at the j^{th} run is the following equations:

$$PPB_{Tr}(j) = \bigcup_{i \in c_j(n,r)} A_i \tag{4}$$

$$CPB_{Tr}(j) = PPB_{Tr}(j) \cup B \tag{5}$$

where $n = 10, 1 \leq r \leq 9$ and $1 \leq j \leq |C(n,r)|$.

In vice versa, the testing set of PPB and CPB at the j^{th} runtime is the following equations:

$$PPB_{Ts}(j) = \left(\bigcup_{1 \leq i \leq N} A_i \right) - PPB_{Tr}(j) \tag{6}$$

$$CPB_s(j) = PPB_{Ts}(j) \tag{7}$$

All the combination data sets and their data samples for both five-people (5 classes) and ten-people (10 classes) classifications are summarized in Table III-3. The proportion of the number of the data sets belonging to training and testing is denoted by $Tr:Ts$. Total ratios are listed as shown in the first column, i.e., 1:9, 2:8, ..., 9:1. In the second column, the number of runs is equal to $|C(10,r)|$.

3.2.2 Training and Testing

To compare the recognition performance between the CPB and the PPB, we setup the multilayer perceptron neural networks with different numbers of hidden neuron for various $Tr:Ts$ ratio.

Table III-3: Data set combinations and their data samples for five-people classification.

Five-people classification (5 classes)				
$\boldsymbol{Tr:Ts}$	# of runs	Train		Test
		Photographs	Caricatures	Photographs
1:9	10	5	5	45
2:8	45	10	5	40
3:7	120	15	5	35
4:6	210	20	5	30
5:5	255	25	5	25
6:4	210	30	5	20
7:3	120	35	5	15
8:2	45	40	5	10
9:1	10	45	5	5

The performance comparison between 5 classes and 10 classes recognition system were also investigated in our experiments. We train and test our face recognition system with various values of all combinations as shown in Tables III-3 and III-4. The parameter settings for these systems are also summarized in Table III-5.

Table III-4: Data set combinations and their data samples for ten-people classification.

Ten-people classification (10 classes)				
$Tr:Ts$	# of runs	Train		Test
		Photographs	Caricatures	Photographs
1:9	10	10	10	90
2:8	45	20	10	80
3:7	120	30	10	70
4:6	210	40	10	60
5:5	255	50	10	50
6:4	210	60	10	40
7:3	120	70	10	30
8:2	45	80	10	20
9:1	10	90	10	10

Table III-5: Parameter setting in the experiments.

Parameters	**Value**
Input nodes	29
Output nodes	5 (5 classes), 10 (10 classes)
Hidden neurons	15, 20, 25, 50
Momentum constant	0.8
Max. training epochs	10^6
No. of repeated runs	10,45,120,210,255,210,45,10

3.2.3 Experimental Results

In this section, the recognition performance of the CPB and the PPB with 15, 20, 25, and 50 hidden neurons and different $Tr:Ts$ varied from 1:9 to 9:1 will be evaluated in both 5 classes and 10 classes. The experimental results on 5 classes are summarized in Appendix A and depicted in Figure III-4, while the experimental results on 10 classes are summarized in Appendix A and depicted in Figure III-5. These

figures indicated that the performance comparison in terms of predictive accuracy of the CPB against 5 classes and 10 classes are not much different for both PPB and CPB.

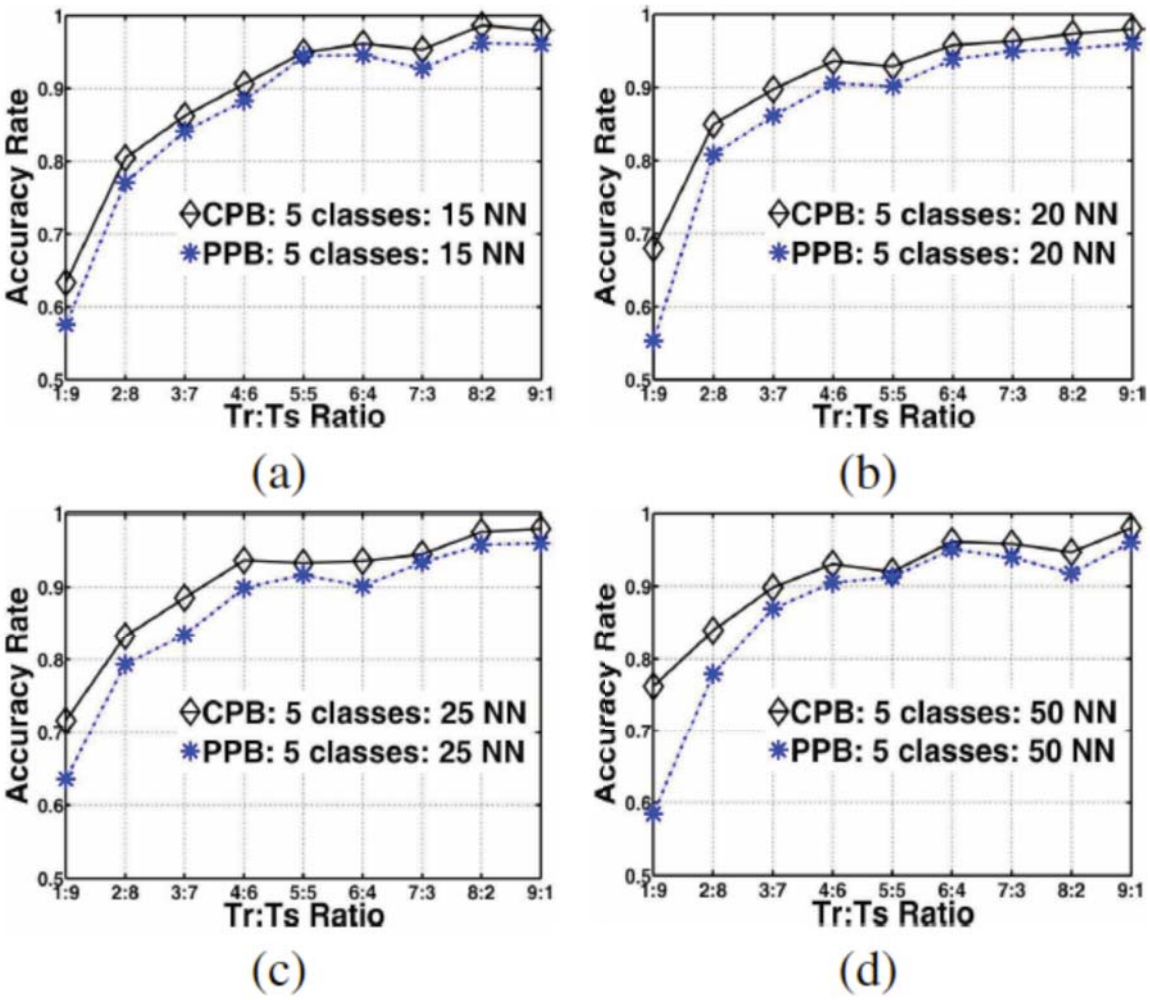

Figure III-4: Experimental results of CPB V.S. PPB on 5 classes; (a) 15 hidden neurons, (b) 20 hidden neurons, (c) 25 hidden neurons, (d) 50 hidden neurons.

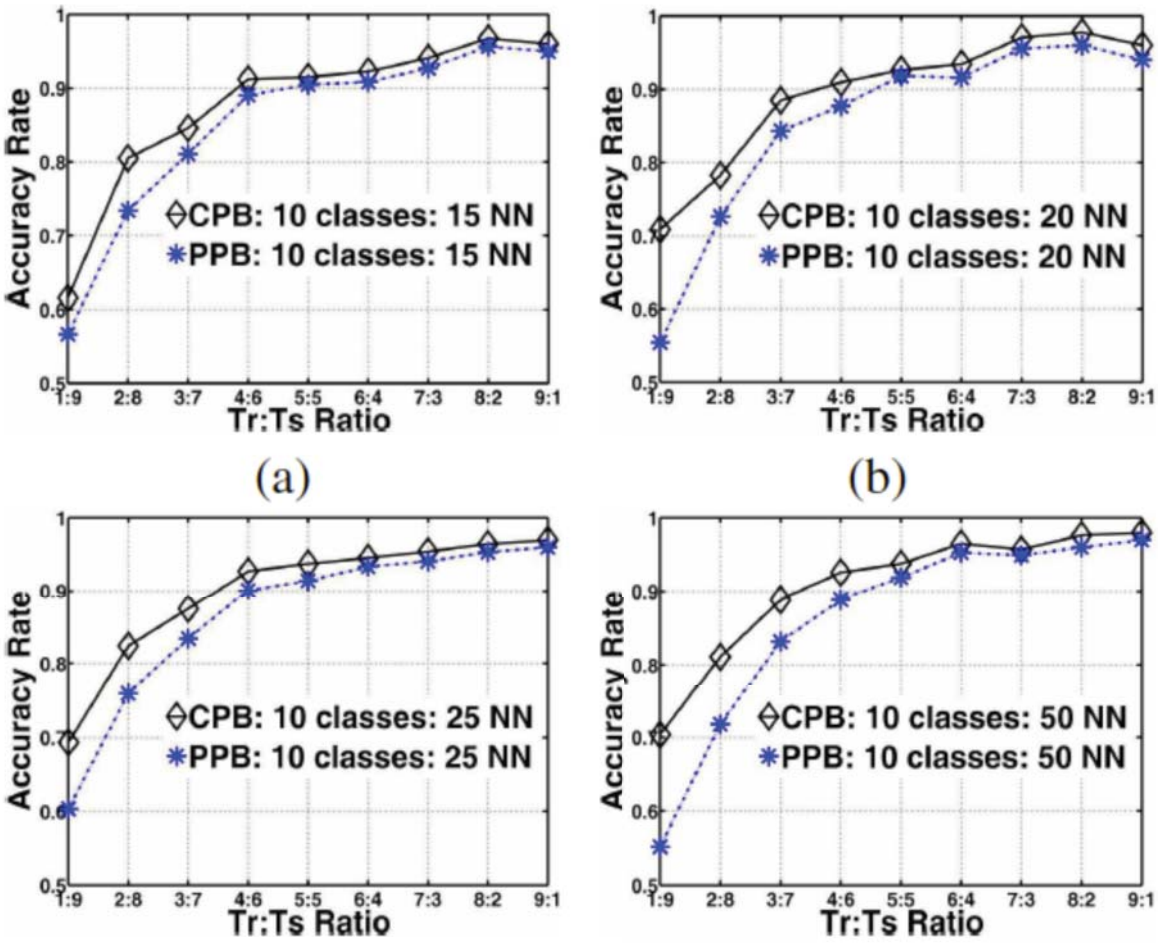

Figure III-5: Experimental results of CPB V.S. PPB on 10 classes; (a) 15 hidden neurons, (b) 20 hidden neurons, (c) 25 hidden neurons, (d) 50 hidden neurons.

This is because the numbers of classes, 5 classes versus 10 classes, were too little to make any effect to the performance of both CPB and PPB. We observe from the experimental results that the mean accuracy of CPB is shown to outperform the PPB in both 5 classes and 10 classes face recognition system.

The performance of the CPB also outperform of the PPB in every $Tr:Ts$ ratio. This ratio reflects the number of the distant indices that were used to train and test in our neural network systems. The $Tr:Ts$ ratio has the impact on how much distant indices (images) of each person were used to train and how much they were used to test. The similar methods were used also in other works [Kussul et al., 2004; Meng et al., 2002], but with the limited number of the repeated run and not done with all possible cases.

In our experiments, we investigated for every all-possible ratio of the 10 data sets (images) for each person from 1:9 to 9:1. In each case, the experiments were conducted with all possible number of repeated runs; those were 10, 45, 120, 210, 255, 210, 120, 45, and 10 runs, respectively. Again in our experiments, the CPB performance still outperform of the PPB in every cases.

To investigate the impact of the number of the hidden neurons to the CPB and the PPB, we implemented the hidden neurons in different ranges. Figure III-4(a) depicted that the best performance against 5 classes (98.67 %) is achieved with 15 hidden neurons. For the 10 classes, Figure III-5(b) depicted that the best performance (97.78 %) is achieved with 20 hidden neurons. The experimental results on the impact of the number of the hidden neurons to the CPB and the PPB, also shown the higher performance of the CPB to the PPB in every different numbers of hidden neurons.

Although the performance of the CPB is outperform the PPB in every cases, but both performance data are very similar. These performance similarities clearly observed in every $Tr:Ts$ ratio from 5:5 to 9:1. After applying a *t*-Test on each pair of the testing results of the CPB and the PPB, the results of the *t*-Test indicated that they are difference with a 95 % confidence level (α = 0.05) in every $Tr:Ts$ ratio from 1:9 to 8:2. Only in the 9:1 $Tr:Ts$ ratio, the *t*-Test implied that their results are not difference (*p*-value is more than the significance level).

We analyze that this is because of the few number of the caricature distant indices that mixed with the photo distant indices in the 9:1 $Tr:Ts$ ratio. Thus the ratio of the caricature samples per photo samples in this case is just only 11.11 %, and it is not effect to the learning process with different type of samples.

Conversely the spread of the performance expand when the $Tr:Ts$ ratio reduced from the equally ratio of 5:5 to 4:6, 3:7, 2:8, and 1:9 consecutively. The maximum spread on the comparison performance achieve in the 1:9 $Tr:Ts$ ratio. For example, in Table III-4 of 5 classes, 50 hidden neurons, the performance spread expand to 18 % (76.22 V.S. 58.44) and in Table III-5 of 10 classes, 50 hidden neurons the

performance spread became 15 % (70.33 V.S. 55.22). The reasons are that the mixing ratio of the caricature distant indices to the photo distant indices became maximum in the 1:9 $Tr:Ts$ experiment. In this case the train samples of the caricatures per photos in the CPB are equal and $Tr:Ts$ ratio is 1:1.

3.3 Binarization Techniques

Binarization is the task of converting a gray-scale image to a binary image by using threshold selection techniques to categorize the pixels of an image into either one of the two classes. Most of studies separated the binarization techniques into two main methods that are global thresholding and local adaptive thresholding techniques.

Global Thresholding Techniques: These techniques attempt to find a suitable single threshold value (τ) from the overall image. The pixels are separated into two classes: foreground and background. This can be expressed as follows:

$$f'(x,y) = \begin{cases} black & if \quad f(x,y) \leq \tau \\ white & if \quad f(x,y) > \tau \end{cases} \tag{8}$$

where $f(x,y)$ is the pixel of the input image from the noise reduction and $f'(x,y)$ is the pixel of the binarized image. Otus's algorithm [Otsu, 1979] is a popular global thresholding technique. Moreover, there are many popular thresholding techniques such as Kapur and et al. [Kapur, 1985], and Kittler and Illingworth [Kittler and Illingworth, 1986].

Local Thresholding Techniques: These techniques attempt to find a suitable single threshold value (τ) from the overall image. The pixels are separated into two classes: foreground and background. This can be expressed as follows:

$$f'(x,y) = \begin{cases} black & if \quad f(x,y) \leq \tau(x,y) \\ white & if \quad f(x,y) > \tau(x,y) \end{cases} \tag{8}$$

The conventional local thresholding techniques have been proposed by Niblack [Niblack, 1986] and Sauvola [Sauvola and Pietikainen, 2000].

Comparing the results from global thresholding techniques and local adaptive thresholding techniques, it was found that results from local adaptive thresholding techniques have adjusted the output among local areas so that the text document can be improved with a clarifying appearance of characters than those of global thresholding techniques. The results from [Chamchong et al., 2010] have shown that Otsu's algorithm gave the best performance. The binariazation techniques were used in their experiments as shown in Table III-6.

Table III-6: Collection of binarization techniques

Binarization Tecniques	Criteria
1.Otsu	$\eta(thr^*) = \sigma_B^2(thr^*)/\sigma_T^2$, $\sigma_B^2(thr^*) = \arg\max_{0\le thr<L-1} \sigma_B^2(thr)$ thr* is optimal threshold,η is separation criteria, $\sigma_B^2(thr)$ is variance between group of histogram at threshold thr and σ_T^2 is variance of histogram
2.Klittler and Illingworth	$T_{opt} = \arg\min\{P(T)\log\sigma_f(T) + [1-P(T)]\log\sigma_b(T) - P(T)\log P(T) - [1-P(T)]\log[1-P(T)]\}$ where $\sigma_f(T)$ and $\sigma_f(T)$ are foreground and background standard deviations.
3.Kapur	$T_{opt} = \arg\max[H_f(T) + H_b(T)]$ where $H_f(T) = -\sum_{l=0}^{T} \frac{p(l)}{P(T)} \log \frac{p(l)}{P(T)}$ and $H_b(T) = -\sum_{l=T+1}^{L} \frac{p(l)}{P(T)} \log \frac{p(l)}{P(T)}$
4.Yen and et al.	$T_{opt} = \arg\max[C_b(T) + C_f(T)]$ where $C_b(T) = -\log\left\{\sum_{l=0}^{T}\left[\frac{p(l)}{P(T)}\right]^2\right\}$ and $C_f(T) = -\log\left\{\sum_{l=T+1}^{L}\left[\frac{p(l)}{1-P(T)}\right]^2\right\}$
5.Huang	$T_{opt} = \arg\min\{-\frac{1}{N^2\log 2}\sum_{l=0}^{L}[\mu_f(l,T)\log(\mu_f(l,T))] + [1-\mu_f(l,T)\log(1-\mu_f(l,T))]p(l)\}$ where $\mu_f[l(i,j),T] = \frac{L}{L+\lvert I(i,j)-m_f(T)\rvert}$
6.Tsai	$T_{opt} = \arg\text{equal}[m_1 = b_1(T), m_2 = b_2(T), m_3 = b_3(T)]$ where $m_k = \sum_{l=0}^{T} p(l)l^k$ and $b_k = P_f m_f^k + P_b m_b^k$
7. Niblack	$T(x,y) = m(x,y)\cdot + k * s(x,y)$ k=-0.2, window size =20x20
8. Sauvola	$T(x,y) = m(x,y)\cdot\left[1 + k\cdot\left(\frac{s(x,y)}{R} - 1\right)\right]$ k=0.5, R=128, window size = 20x20
9.Bernsen	T(x, y) = (Zlow + Zhigh)/2, C(x, y) = (Z_{high} - Z_{low}<) e Window size (r x r) = 15x15 and e = 15

3.4 Zernike Moments

We extract the Zernike moments from an image for retrieval. Among various types of moments, Zernike moments are the most potential method for extracting the invariant features of images. To process Zernike moments method, a spatial coordinate $(x_1, x_2)$of an image is normalized on a unit disk region as follows:

$$D = \{(x_1, x_2) \in R^2 | x_1^2 + x_2^2 \leq 1\} \tag{9}$$

The region of interest can be normalized by scaling down their sizes until they fit into the unit disk. After the normalization, the centroid of the image should be located at the origin of the unit disk. In other words, a polar coordinate is replaced by a spatial coordinate as shown in Eq. (9).

$$D = \{((r, \theta)) | 0 \leq r \leq 1 \; and \; 0 \leq \theta \leq 2\pi\} \tag{10}$$

Then, the Zernike moments of order p with repetition q for the normalized image is defined in the form of polar coordinates in a unit disk becomes the following:

$$A_{p,q} = \frac{(p-1)}{\pi} \iint_D f(r, \theta) Z_{p,q}^*(r, \theta) r dr d\theta \tag{11}$$

where * denote the complex conjugate. $Z_{p,q}$ is the polynomial formed as

$$Z_{p,q}(r, \theta) = R_{p,q}(r) e^{jq\theta} \tag{12}$$

where p is a positive integer or zero, q is a positive or negative integer, subject to the constraints $(p - |q|)$ is even and $|q|$ is less than or equal to p. Variable $R_{p,q}$ is a radial polynomial defined by

$$R_{p,q}(r, \theta) = \sum_{s=0}^{(p-|q|)/2} \frac{(-1)^s (p-s)! \, r^{p-2s}}{s! \left(\frac{p+q}{2} - s\right)! \left(\frac{p-|q|}{2} - s\right)!} \tag{13}$$

In accordance with the above-defined formula, Zernike moments are the project of the image $f(x_1, x_2)$ onto the orthogonal basis functions except that the image $f(x_1, x_2)$ is outside the unit disk.

The Zernike moments of the image after rotation through an angle α denoted by $A'_{p,q}$ is calculated as shown in Eq.(13). In the same manner of the proper of the Fourier transform, it is obvious that the magnitude of Zernike moments do not change when the image is rotated.

$$A'_{p,q} = A_{p,q} e^{-jq\alpha} \tag{14}$$

Therefore, the image representative of an image can be shown as a vector containing the Zernike moments with different parameter either p or q. The vector of Zernike moments with the parameter p is defined as follows:

$$Z^p = \begin{cases} [A_{p,0}, A_{p,2}, \dots, A_{p,2i}]^T \; and \; i = 0,1, \dots \frac{p}{2}, if \; p \; is \; even \\ [A_{p,1}, A_{p,3}, \dots, A_{p,2i+1}]^T \; and \; i = 0,1, \dots \frac{p-1}{2}, if \; p \; is \; odd \end{cases} \quad (15)$$

3.5 Framework of Feature Extraction

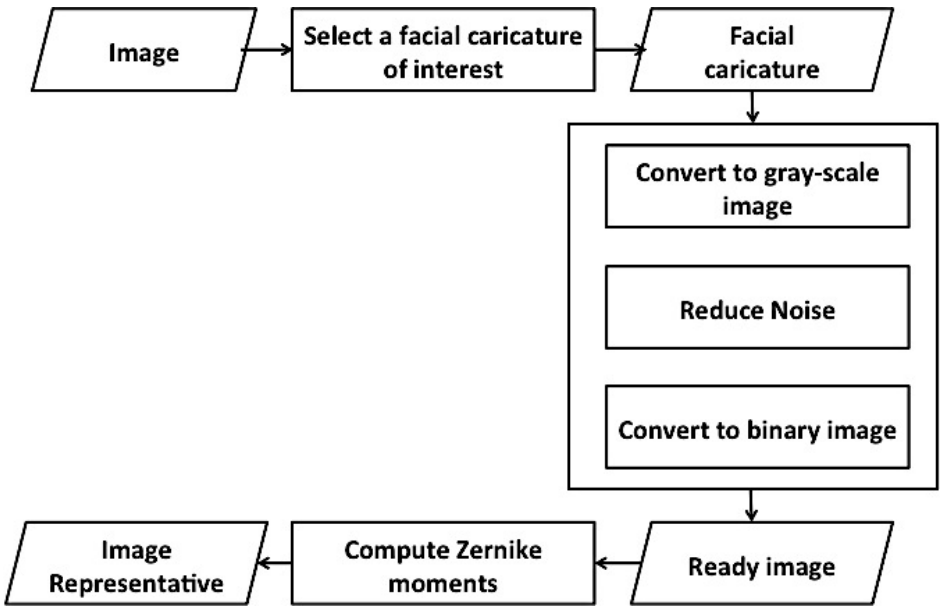

Figure III-6: Image processing and feature extraction for finding an image representative

Figure III-6 shows the framework of our feature extraction that starts with the selection of a facial caricature of interest, followed by the image preprocessing and the computation of Zernike moments. First, we are able to select a facial caricature of interest with graphic software. Second, we conduct the image preprocessing on the facial caricature in order to reduce noise and then convert to a binary image. The output of this preprocessing is called "ready image". Finally, we compute Zernike moments of the ready image. A set of Zernike moments is called "image representative" of a facial caricature.

3.6 Learning and Classification

Three different classification techniques have been evaluated. They are Support Vector Machines (SVM) [Vapnik, 1995], Minimum Mean Distance (MMD), and Nearest Neighbor (NN) [Khotazad, 2002]. The classifiers learn from the training set in which every example is represented with a multi-dimensional feature vector composed of extracted Zernike moments.

For the SVM multi-class classifier, we use a radial basis kernel and pair-wise classification [Schalkoff, 1992]. This results in $(N-1)N/2$ binary classifiers where N is the number of class. During classification, all classifiers are evaluated and the test example is classified to the class receiving the maximum number of votes. The training data is scaled to be in the range of [0, 1] in order to avoid numerical problems. The test data is also scaled according to the parameters obtained during the training stage.

In the minimum distance classifier, each symbol class, C_k, is represented with the sample means, μ^k, and standard deviations, σ^k learned from the training examples. When a new example is given, it is compared to each symbol class by calculating the normalized Euclidean distance. The normalization is done on each representative to account for the variance in that the dimension of image representative. The example is assigned to class *k* for which the distance is minimum. The equations are show below:

$$\mu_i^k = \frac{1}{n}\sum_{j=1}^{n} x_{i,j}^k \tag{16}$$

$$\sigma_i^k = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{i,j}^k - \mu_i^k)^2} \tag{17}$$

$$d(x, C_k) = \sum_{i=1}^{m}\left(\frac{x_i - \mu_i^k}{\sigma_i^k}\right)^2 \tag{18}$$

where

μ_i^k	=	the mean of the i^{th} representative in class k
σ_i^k	=	the standard deviation of the i^{th} representative in class k
$x_{i,j}^k$	=	the value of the i^{th} representative of example in class k
$d(x, C_k)$	=	the normalized distance between example x and class k
n	=	the number of examples in class k
m	=	the representative dimension

During training, the nearest neighbor classifier normalizes the representative vectors of the examples in the training classes using the corresponding, μ^k and σ^k. In the classification stage, the classifier extracts representatives from the test example and computes the normalized Euclidean distance, *d*, between the example and every training example. The test example has to be normalized using the parameters of the class under test. The training example of class *k*, C^k, with the smallest distance to the test example, a, is the nearest neighbor of a. Eq. (19) shows the normalization of an example to class *k*. Eq. (20) shows the Euclidean distance between two examples.

$$\hat{a}_i = \frac{a_i - \mu_i^k}{\sigma_i^k} \tag{19}$$

$$d(\hat{a}, \hat{c}^k) = \sum_{i=1}^{m} \left(\hat{a}_i - \hat{c}_i^k\right)^2 \tag{20}$$

The recognition can be made adaptive with each of these classification methods by updating the training parameters with added examples.

3.7 Experiment II

3.7.1 Data Sets

Our hypothesis is that the Zernike moments of a facial caricature can be used for a representative of the facial caricature since the Zernike moments will be tolerant to the rotation, scaling, shearing, and some degrees of prospective of an image. The data sets are separated into the training set and the test sets. The training set contains of 10 original facial caricatures (see in Appendix B) obtained from the Internet [Alan, 2009; Prichett, 2009]. The test sets are the corresponding transformation of the original images in terms of rotation, shearing, and prospective mapping. The transformed images as shown in Appendix C were done by using the GIMP software.

Our test sets of three transformation methods are shown as follows:

1) Rotation through the following angles: 5, 10, 15, 20, 25, 30, 35, 40, 45
2) Shear parallel to the x-axis with the following units: 5, 10, 15, 20, 25, 30, 35, 40, 45
3) Prospective transformation with nine ambiguous transformation matrices as shown in Figure III-7.

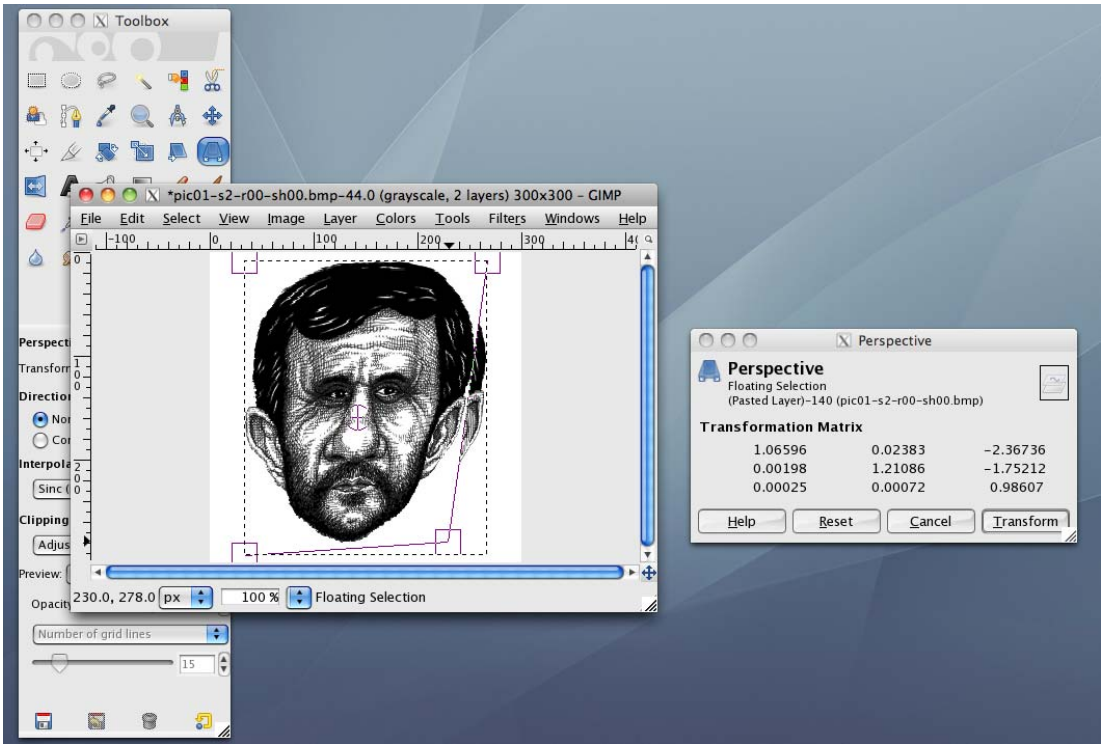

Figure III-7: Sample transformed image by using the GIMP software;
Prospective mapping with a particular transformation matrix (right panel)

3.7.2 Experimental Results

We tested whether the images representatives, which are obtained from the framework as shown in Figure III-6, are able to be invariant to rotation, shear mapping and perspective transformation. These transformations are common components of drawing caricatures. Training with a collection of caricatures, the classifiers can be used for a similarity retrieval system. We did the following experiments:

1) Statistical Analysis. The ANOVA and MANOVA were used to test on each moment and combination of moments, respectively.
2) Tests of invariance property by using NN. Tr:Ts = 1:27. This means that training with one original image while testing with other transformed images (9 rotated, 9 sheared, 9 perspective = 27 in total).
3) Recognition tests by using SVM, NN, and MMD. We designed the training with varying number of images with the following sets: Tr=3 (1 rotated, 1 sheared, 1 perspective), Tr=6 (2 rotated, 2 sheared, 2 perspective), Tr=9 (3 rotated, 3 sheared, 3 perspective) and so on.

Statistical Analysis

MANOVA is a generalized form of univariate analysis of variance (ANOVA). We used for analyzing two or more independent variables that are Zernike moments. Groups are defined as a set of images transformed by rotation, shear mapping and perspective transformation. In this analysis, there are ten groups each of which consists of 28 transformed images. The representative of each image is a set of Zernike moments, $[A_{20,18}, A_{19,17}, A_{18,16}, A_{17,15}, A_{16,14}, A_{15,13}, A_{14,12}, A_{13,11}, A_{12,10}, \dots, A_{3,1}]$.

We evaluated a hypothesis that includes not only equality among groups on the Zernike moments, but also equality among groups on linear combinations of these Zernike moments. By using MANOVA, we found that a significant difference at confidence 95% between groups of transformed images with a linear combination of at least six Zernike moments.

We also calculated an average distance between every pair of classes as shown in Table III-7. Each class consists of all transformed images of a person. Then, the tree diagram as shown in Figure III-8 illustrated the most similar class and the most dissimilar class and also the clusters of classes based on the similarity of their features. This tree diagram was computed by the hierarchical clustering.

Table III-7: Average distance between each pair of two groups (PEOPLE)

Avg. Dist	Pic_01	Pic_02	Pic_03	Pic_04	Pic_05	Pic_06	Pic_07	Pic_08	Pic_09	Pic_10
Pic_01	0.000	131.624	42.560	128.291	134.497	56.628	170.352	100.834	221.086	130.948
Pic_02	131.624	0.000	42.405	45.284	19.485	74.409	13.868	16.705	32.596	90.209
Pic_03	42.560	42.405	0.000	65.660	43.638	48.322	62.136	40.126	100.382	61.431
Pic_04	128.291	45.284	65.660	0.000	55.294	120.470	52.320	31.356	40.736	90.787
Pic_05	134.497	19.485	43.638	55.294	0.000	82.266	23.149	28.589	36.750	81.673
Pic_06	56.628	74.409	48.322	120.470	82.266	0.000	102.255	71.988	157.092	122.792
Pic_07	170.352	13.868	62.136	52.320	23.149	102.255	0.000	26.263	19.050	104.566
Pic_08	100.834	16.705	40.126	31.356	28.589	71.988	26.263	0.000	34.534	95.840
Pic_09	221.086	32.596	100.382	40.736	36.750	157.092	19.050	34.534	0.000	144.427
Pic_10	130.948	90.209	61.431	90.787	81.673	122.792	104.566	95.840	144.427	0.000