Video Image Based Multimodal Face Recognition System

Craig Belcher, Matt Terry, Sophia Vinci-Booher, Yingzi Du

Indiana Univ.-Purdue Univ. Indianapolis
Email: {csbelche, macterry, svincibo, yidu}@iupui.edu

Abstract

In this paper, a multimodal face recognition system based on video images is proposed to improve face recognition accuracy. The faces are first automatically extracted from the video image. Then each individual expression is extracted and grouped in categories using the difference images. Eigenface is calculated from each expression and is independently matched to our face database using principle components analysis (PCA). The Euclidian distance from the PCA method is used to obtain a matching score for each expression. The matching scores from all the expressions in the same video image will be fused together to generate an overall matching score. Our preliminary experimental results show that this method is effective.

1 Introduction

Biometrics[1] is the study of automatically identifying and verifying a person using physical, biological, and behavioral traits such as fingerprints, face, iris and voice. The goal of identification is to find out who a person is based on a database of known identities; whereas verification determines if someone is who they claim to be. The advantage of using biometrics for verification as opposed to traditional approaches like a password or keycard is that biometrics is more convenient for the users and can reduce fraud.

Face recognition is a very interesting biometric technique since it is very close to how humans identify one another. However, compared to other biometrics like fingerprint and iris, the accuracy of face recognition systems is often not high.[2] The challenges of face recognition are illumination and changing of face expression. In the literature, there have been a number of approaches for face recognition. They can be grouped in: local feature based methods, holistic approaches, and hybrid methods. Holistic approaches are more widely used due to the efficiency and high recognition rate. Popular holistic approaches include principle component analysis [3,4] (PCA), Linear Discriminant Analysis (LDA), and kernel-based approaches[5]. Among them, PCA is the simplest one. However, the accuracy from using PCA face recognition has been found to be low. In this paper, we propose a new approach to increase the accuracy of face recognition systems by using video image. Our method automatically extracts multiple face images from video of a subject and then classifies each image according to the expression. By comparing the

results of the PCA method between a set of test images and a set of images in a database, we are able to obtain a score for each expression. We then fuse each score to obtain an overall matching score.

2 Theory

2.1 Why face expression can help face recognition

To properly quantify the regularity at which a random person might choose to express their face, we divided the face into 7 different regions labeled A-G, Fig. 1, and developed ratios as to which parts of the face were used the most or least often. The first step was to closely examine the effects of the contraction of each individual muscle. The major 11 facial expressive muscles were used: frontalis, corrugator, levator palpebrae superioris, orbicularis oculi, levator labii superioris, mentalis, triangularis, incisivis labii superioris, orbicularis oris, risorus, zygomatic major. These major muscle groups are responsible for the majority of all facial expressions, and can be further divided into muscles of the upper face and lower face. The muscles of the upper face include the frontalis, corrugator, levator palpebrae superioris, orbicularis oculi, and the levator labii superioris. The muscles of the lower face include the mentalis, triangularis, incisivis labii superioris, orbicularis oris, risorus, and the zygomatic major[6].

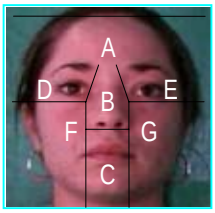


Figure 1. Defined Regions

With 44 different facial muscles able to produce over 5,000 different facial expressions[7], the odds of one person using the exact same facial muscles as another during any specific expression is very low. Furthermore, the odds of the same two people producing the same expression using the same muscles and in the same time sequence are even more unlikely. Therefore, a face recognition system that evaluates a live video feed of individual people making facial expressions should increase the accuracy of face recognition.

2.2 Face Segmentation

Face segmentation is necessary for comparison of face images and consists of finding a face in an image and cropping the surrounding area. We use the color space of each image to roughly detect the area containing a face. The major benefit from this method is that any shape or size of a face can be captured and detected. Skin tone no longer plays a huge roll in determining the accuracy of face detection since each image is projected into one color space.

The color space of an image refers to the projection of one, or more, of the red, green, or blue attributes of an image. The idea of using a color space to project its intensity can be used in a very powerful way. In this instance, the green space of each image is taken advantage of. As you can see in Eq. 1, the green space is calculated using the difference between the green and the

red; as well as, the difference between the green and the blue for each given point of an image. The equation takes a normal colored image, Fig. 2 (a), and creates an image that looks blue in color while surrounded by red, Fig. 2(b).

$$(((image(:,:,2) - image(:,:,1)) > 1 \,\&\, (image(:,:,2) - image(:,:,3)) > 1) = GreenSpace \quad (1)$$


(a) Color Video Frame


(b) Projected to Green Space


(c) Face Cropping


(d) Face Extraction

Figure 2. An Example of Face Extraction

The skin pigment and skin tone of a human face is different than the surroundings behind that person, and therefore the face can be isolated. In our approach, a threshold of a projected image is used to define the boundaries of that face. Each image goes through a process that starts on each of the four edges and scans inward toward the center. Once a change from red to blue is detected, that is considered an edge. A total of four edges are found and used to determine where the face is. This area is cropped from the image and used for the facial expression segmentation. This method of detecting the face is beneficial because it is dynamic enough to work on each frame of the video. It is also capable of detecting a face when the quality is poor, and if the person moves from side to side or up to down during the video. An example of a cropped image can be seen in Fig. 2(c).

To increase the accuracy of our system, we used an additional method of cropping developed in "When Holistic Processing is Not Enough: Local Features Save the Day[8]." This method requires the user to select the location of the eyes and mouth and the algorithm will then automatically crop and rotate the face, Fig. 2(d), so as to create face images that all have the eyes and mouth in the same orientation and location. In addition, each image is resized so that essentially the same areas of each face are directly compared which allows the face recognition method used to achieve the greatest accuracy. In future experiments, the selection of the location of the eyes and mouth will be done automatically as well.

2.3 Principle Component Analysis

PCA is a commonly used approach to face recognition because the algorithm is fast and does not take much space in a database. Normally, an image of each subject's face would need to be stored for comparison when identification or recognition is necessary. Using PCA, only a set of training weights is saved representing how much of each eigenface is present in each subject's image. An eigenface results from the projection of each face image in the database into eigenspace and is a combination of various significant features from each image.

To perform PCA, each training image needs to be the same size and converted to a single column vector of size N = (number of rows X number of columns). In Eq. 2, A contains columns of M images of size Nx1.

$$A = [T_1 \cdots T_M] \tag{2}$$

To remove the features common to all the training images, the average of all M columns is subtracted from each column, Eq. 3.

$$X_i = T_i - \frac{1}{M}\sum_{n=1}^{M} T_n \tag{3}$$

Eigenfaces are formed from the eigenvectors of the covariance matrix C, Eq. 4.

$$C = XX^T \tag{4}$$

Normally, the covariance matrix would be very computationally intensive since X is size NxM and X^T is size MxN. To reduce the amount of computation, singular value decomposition is used, Eq. 5, where U contains the eigenvectors of the covariance matrix C[9].

$$X = UEV^T \tag{5}$$

Using the eigenvectors in matrix U, each training image is projected into eigenspace and a weight Y results, Eq. 6, representing how much of each eigenvector or eigenface is used to create an individual's face.

$$Y_i = U^T X_i, i = 1,...,M \tag{6}$$

To find a match using PCA, the input image is reshaped to an Nx1 column, I, and the average of the columns of A are subtracted from it, Eq. 7.

$$B = I - \frac{1}{M}\sum_{n=1}^{M} T_n \tag{7}$$

Then the result is projected into eigenspace and a set of recognition weights is found from how much of each eigenface is contained in the test image, Eq. 8.

$$W = U^T B \tag{8}$$

To find the closest match, the Euclidean distance is found between the test image weight, W, and each training weight, Y, Eq. 9.

$$Z = \|W - Y_i\|, i = 1,...,M \tag{9}$$

The database image having the training weight with the smallest distance from the test weight is considered to be the best match. However, due to a difference in angle, lighting, face expression,

and facial hair, the lowest distance is not always to the correct image and a subject is incorrectly matched.

3 Method

Fig. 3 shows each step of our method and is described in the following sections.

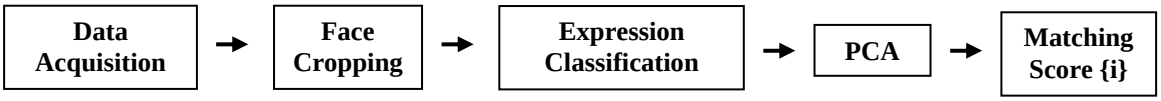


Figure 3. Multi-modal Face Recognition System

3.1 Data Acquisition

In our image acquisition system, a Logitech web camera interfaced with software designed in-house using Matlab, Fig. 4, to acquire videos with a resolution of 640X480, resulting in 100 to 150 frames per video. For this project, we asked each subject to start by making no expression, then smiling, then no expression, then a face expression of his/her choice, no expression, another face expression of his/her choice and finally no expression. We acquired a total of three videos of the same expressions each session and subjects were asked to come back for two more sessions with at least a week between each session.

Figure 4. Image Acquisition Software

In this way, we acquired a dataset that could be used for training data and test data with a known progression of expressions. At this time, we have useable videos from 38 subjects with only 27

of those subjects in more than one session. Since each subject repeated the same expressions in each video, the results of matching input images to training images would be based not only on common facial characteristics, but also on the expressions made. One video from each set of subject videos was used to manually form a training set consisting of the average of ten "no expression" images, five "smile" images, five "expression 1" images, and five "expression 2" images, Fig. 5. This resulted in a training set of 152 images with 38 images in each expression group mentioned. The remaining videos were used for testing our method.

Figure 5. Averaged Training Expressions

3.2 Face Cropping and Expression Classification

After images have been acquired, the above approach to face cropping in Section 2.2 is applied and the resulting images are compared for expression classification. When separating the facial expressions, the order of face expressions is already known. The order in the collected data consists of one "no expression" face between each other expression given. Knowing this, an "energy" can be taken from each image based on a "no expression" image. A graph of these energies can be seen in Fig. 6.

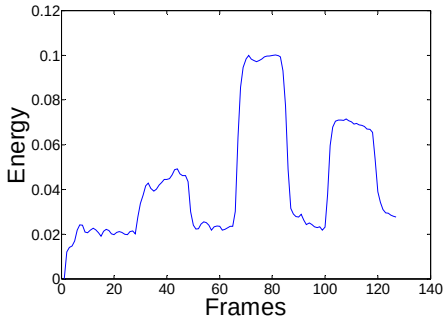


Figure 6. Graph of Expression Energies

The changes in energy in the above graph represent the changes in expression throughout a video. Each frame is compared to the first "no expression" frame and the peaks of the graph are the frames having expressions while the valleys are the frames having no expression. Since each expression is being compared to a database of images, the entire group of images per expression is then averaged together. This is done to reduce the number of comparisons performed on each subject when face recognition is done. In Fig. 7, the different expressions are displayed after the automated process is complete. Doing this reduces the test images down to about 8 images, versus the 110 or more frames from the entire video.

Figure 7. Averaged Expressions

3.3 PCA

Since the progression of expressions was known and the training set was divided into four different expressions, we created five different training sets; one for each expression and an additional set containing all 152 images. When finding the Euclidean distance for either the first or last image obtained in the Expression Classification module, it was assumed that the image would be "no expression" and only need be compared to the "no expression" training set. When finding the Euclidean distance for the second image, it was assumed the image was most likely either another "no expression" or a "smile" and only those two training sets were used. The same approach was used for the second to last image with the "no expression" and "expression 2" training sets. However, since the Expression Classification module would not divide each video into exactly 7 averaged expression images, any images not considered to be one of the previous four would be compared to the entire training set of 152 images.

3.4 Matching Score

For each image from the Expression Classification module, an image from the database was considered to be a match if the Euclidean distance between its recognition weight and the database image training weight was the lowest. Since each input image was an averaged image, not every image would match well and would have a very high Euclidean distance. To eliminate these images as valid results, a threshold was determined experimentally and applied to all of the results. If an image was the first, last, second or second to last image and was below a threshold of 300, it was accepted. If it was not one of those images and below a threshold of 200, it was also accepted. Then, the image most commonly found to be a match was decided to be the best match. In other words, if three of the four accepted images matched to subject A, then it was decided to be subject A. In the case of two subjects having the same number of accepted matches, the subject having the lowest Euclidean distance was decided to be the correct match.

4 Results

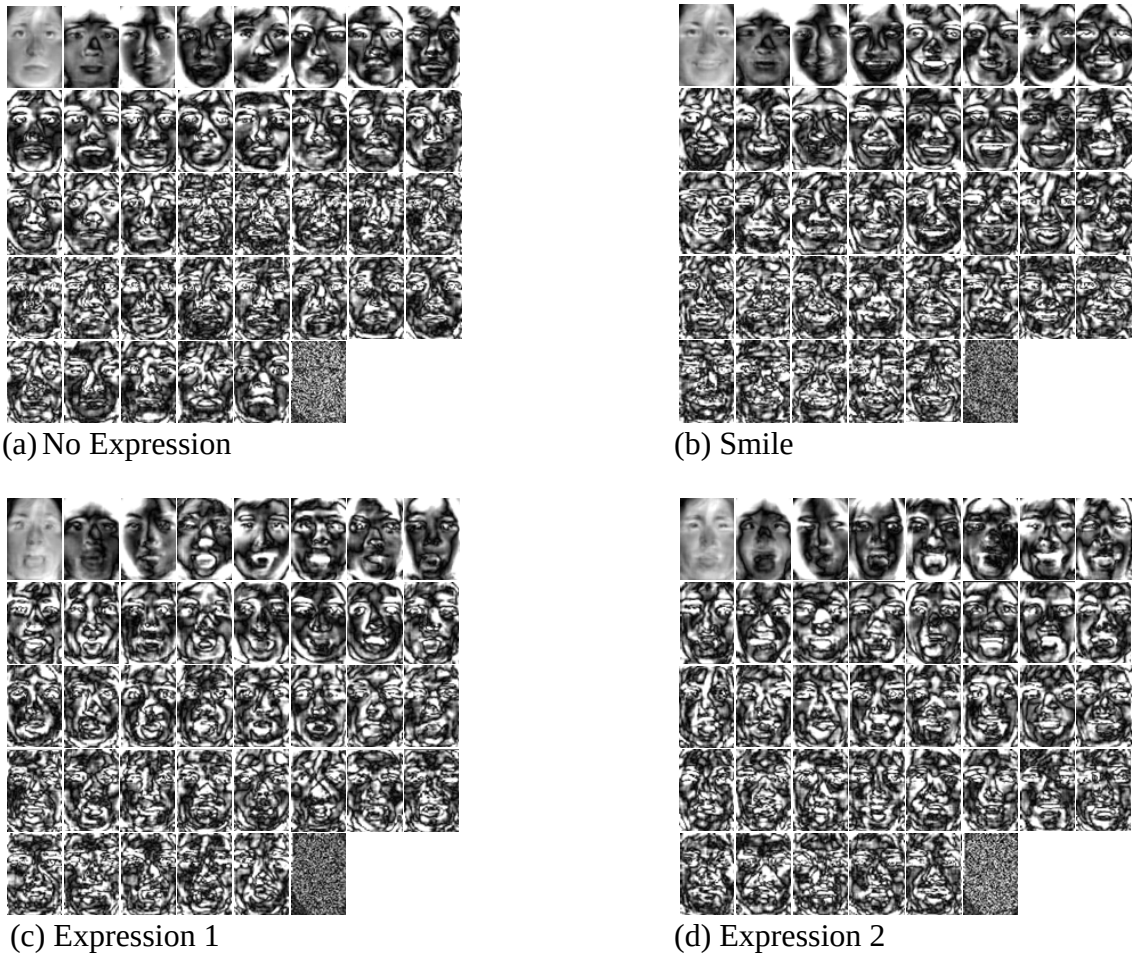
To create a control set of experimental results, we used the same video for the test set as for the training set for each subject. Since the training set was separated into classes manually and the test set was separated automatically, the images being compared would not be exactly the same. All 38 test images correctly matched the 38 different subjects in the database, as expected, with Euclidean distances ranging from 7 to 100. This shows that the Expression Classification module is capable of correctly dividing the input images into each expression class without enough overlap to create a large Euclidean distance. In other words, if the "no expression" and "expression 1" images had not been correctly separated and were averaged together; the resulting

image may have correctly matched but would not have had such a low Euclidean distance as was found.

The second experiment used the second video from the first session with a total of 38 videos. 37 of the 38 videos correctly matched with Euclidean distances ranging from 100 to 300. The third experiment used the third video from the first session and had the same results, although the subject that did not correctly match was different in each experiment. These results show that our method is capable of correctly matching videos of subjects to the stored still images in a database.

To have a comparison of our results to the results of the PCA using still images, we used a single frame from each subject video to match to the "no expression" images in our database. Running the same set of experiments, we found that the first experiment correctly matched all 38 images, the second experiment correctly matched only 36 images and the third experiment correctly matched 37 images.

The eigenfaces for all the training images used are shown in Fig. 8.



(a) No Expression (b) Smile

(c) Expression 1 (d) Expression 2

Figure 8. Eigenfaces

5 Conclusion

Based on our experimental results, our method has been shown to be slightly better than the normal PCA method of using only still images with a slight improvement in the second experiment. To determine if our method is a viable replacement for the normal use of PCA, we need a larger dataset. As the size of the dataset increases, the need for matching scores from multiple expressions will increase and our results should show that, even though PCA does not match 100% correctly, a fusion of matching scores can be used to decide more accurately. In addition, future experiments will use videos from the second and third sessions to compare results.

Acknowledgement

This project is sponsored by the Multidisciplinary Undergraduate Research Initiative Program of Indiana University-Purdue University Indianapolis. We thank graduate students Zhi Zhou and Emrah Arslanturk for collecting data from human subjects.

References

1 http://en.wikipedia.org/wiki/Biometrics January 2007.
2 http://www.itsc.org.sg/synthesis/2002/biometric.pdf January 2007.
3 M. Kirby and L. Sirovich, "Application of the Karhunen–Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.
4 M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
5 http://www.biometrics.gov/docs/facerec.pdf January 2007.
6 http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm January 2007.
7 http://www.psychologytoday.com/articles/pto-19980901-000024.html January 2007.
8 Lingyun Zhang and Garrison W. Cottrell, "When Holistic Processing is Not Enough: Local Features Save the Day," In Proceedings of the *Twenty-sixth Annual Cognitive Science Society Conference* 2004.
9 http://en.wikipedia.org/wiki/Singular_value_decomposition January 2007.

Biographies

CRAIG BELCHER is a senior undergraduate student from the Department of Electrical and Computer Engineering at IUPUI and will be attending graduate school with a focus on Biometrics. He is a student member of IEEE and Vice President of Eta Kappa Nu. His current work focuses on Iris Image Quality and Multimodal Face Recognition.

MATT TERRY is an alumni from the Department of Electrical and Computer Engineering at IUPUI. He is currently an IT specialist developing software for Ingersoll Rand.

SOPHIA VINCI-BOOHER is a junior undergraduate student from the Department of Biomedical Engineering at IUPUI. She is an active member of the Society for Women Engineers and the IUPUI Biomedical Engineering Club. Her current work focuses on Multimodal Face Recognition with a focus on facial structure.

DR. ELIZA DU is an Assistant Professor at the Department of Electrical and Computer Engineering, IUPUI. She got her Ph.D. in Electrical Engineering in 2003 and worked as an Assistant Research Professor at the U.S. Naval Academy from 2003 to 2005. She joined IUPUI in 2005. Her research interests include biometrics, image processing, pattern recognition, and digital signal processing. She is a member of Phi Kappa Phi and Tao Beta Pi honor societies, and a member of IEEE and SPIE.