투사에 기초한 얼굴 인식 알고리즘들의 통계적 분석

정회원 문 현 준*, 백 순 화**, 전 병 민***

Statistical Analysis of Projection-Based Face Recognition Algorithms

Hyeonjoon Moon*, Soon-Hwa Baek**, Byoung-Min Jun*** *Regular Members*

요 약

최근 수년간 얼굴인식에 관한 많은 알고리즘이 개발되었고 그 대다수가 view와 투사에 기초한 알고리즘이었다. 본 논문에서의 투사는 비단 직교 기저상에 영상을 투사하는 것으로 국한하지 않고 영상 화소값을 변환하는 일반적인 선형 변환으로써 상관관계, 주성분 분석, 클러스트링, gray scale 투사, 그리고 추적 필터매칭을 포함한다. 본 연구에서는 FERET 데이터베이스 상의 얼굴 영상을 평가한 알고리즘들을 세부적으로 분석하고자 한다. 투사에 기초한 알고리즘은 3단계로 구성된다. 첫 번째 단계는 off-line상에서 행하며 알고리즘 설계자에 의해 새로운 기저가 설정되거나 또는 학습을 통해 새로운 기저를 결정한다. 두 번째 단계는 on-line상에서 행해지며 영상을 설정된 새로운 기저상에 투사한다. 세 번째 단계는 on-line상에서 행해지며 영상내의 얼굴은 가장 인접한 이웃 분류자로 인식된다.

대부분의 평가 방법들은 단일 gallery 상에서의 성능 평가가 이루어짐으로써 알고리즘 성능을 충분히 측정하지 못하는 반면 본 연구에서는 독립된 galley들의 집합을 구성함으로써 각각의 알고리즘이 다른 galley상에서 가지는 변화와 이들의 상대적 성능을 평가한다.

ABSTRACT

Within the last several years, there has been a large number of algorithms developed for face recognition. The majority of these algorithms have been view- and projection-based algorithms.

Our definition of projection is not restricted to projecting the image onto an orthogonal basis; the definition is expansive and includes a general class of linear transformation of the image pixel values.

The class includes correlation, principal component analysis, clustering, gray scale projection, and matching pursuit filters. In this paper, we perform a detailed analysis of this class of algorithms by evaluating them on the FERET database of facial images. In our experiments, a projection-based algorithms consists of three steps. The first step is done off-line and determines the new basis for the images. The bases is either set by the algorithm designer or is learned from a training set. The last two steps are on-line and perform the recognition. The second step projects an image onto the new basis and the third step recognizes a face in an with a nearest neighbor classifier.

The classification is performed in the projection space. Most evaluation methods report algorithm performance on a single gallery. This does not fully capture algorithm performance. In our study, we construct set of independent galleries. This allows us to see how individual algorithm performance varies over different galleries. In addition, we report on the relative performance of the algorithms over the different galleries.

* Lau Technologies, Inc
** 천안외국어대학 컴퓨터정보과
*** 충북대학교 컴퓨터공학과
논문번호 : 99337-0823, 접수일자 : 1999년 8월 23일

Ⅰ. Introduction

The development of evaluation procedures for algorithms is starting to become an accepted practice in computer vision. One of the reasons is that evaluation procedures offer a way to assess competing performance claims. For one be able to make a fair assessment of these claims, it is necessary that the underlying assumptions of the evaluation procedure be clearly stated, and that the testing and scoring protocols are described. From this, a statistical model for comparing algorithms can be formulated.

For any given computer vision problem, there are numerous algorithms designed to solve it. The design of each algorithm is based on a set of decisions and assumptions. Because of these decisions and assumptions, it may not be appropriate to apply a particular test to an algorithm. The underlying test assumptions for scoring protocol are one of the criteria for determining if an evaluation procedure is appropriate for a particular algorithm.

The FERET database has fulfilled the data requirements for both development and testing, becoming the de facto standard for face recognition. [5,6] . The most recent of these procedures was the Sept96 FERET test, which provided a robust and comprehensive evaluation of face recognition algorithms. The success of the Sept96 FERET test was based on a few design assumptions from which a statistical model could be formulated. In this paper, we state these assumptions, present the resulting statistical model, and use it to assess the performance of projection-based algorithms. The primary goal of the Sept96 FERET test was to obtain an accurate assessment of the performance of face recognition algorithms on still images. To get this assessment, the evaluation methodology needed to be robust and comprehensive. To achieve this, scores had to be computed for a large range of galleries and probe sets. This led to the following design principles: (1) algorithms could not be trained during testing, (2) each facial image was treated as an unique face, and (3) the similarity score between a probe and a gallery image is a function of only those two images. The *gallery* is the set of known individuals. An image of an unknown face presented to the algorithm is called a *probe*, and the collection of probes is called the *probe set.*

Projection-based algorithms, the dominant approach to face recognition, include correlation, principal component analysis, clustering, gray-scale projection, and matching pursuit filters. The structure of these algorithms is amenable to a comprehensive evaluation procedure.

In this paper, an algorithm is projection-based if it projects the pixel intensity value onto a new basis, which need not be orthogonal. The original image is the primal space, where each pixel is a dimension; i.e., an m × n image is represented as a point $\Leftarrow R^{m \times n}$.

A projection-based algorithm consists of three steps. The first step is done off-line and determines the new basis for a facial image. The basis is either set by the algorithm designer or learned from a training set. The remaining steps are on-line and identify a face. The second step projects a facial image onto the new basis. The third step identifies a face using a nearest neighbor classifier in the projection space. (A more general definition of projection-based algorithms would allow a larger class of classifiers.)

Ⅱ. Testing Principles

To obtain a robust comparison of algorithms, it is necessary to calculate performance on a large number of galleries and probe sets. It is not practical to have an evaluation methodology that consists of a large number of separate galleries and probe sets. To allow scoring on multiple gallery and probe sets, we have to adopt an appropriate protocol. In the new protocol, an

algorithm is given two sets of images: the *target set* and the *query set*. We introduce this terminology to distinguish these sets from the gallery and probe sets that are used in computing performance statistics. The target set is given to the algorithm as the set of known facial images. The images in the query set are the unknown facial images to be identified. For each image q_i in the query set Q, an algorithm reports the similarity $s_i(k)$ between q_i and each image t_k in the target set T. The key property, which allows for greater flexibility in scoring, is that for any two images s_i and t_k, we know $s_i(k)$. (In fact, designations of which set is the target and which is the query are arbitrary. By reformatting the output, we can change the roles of the target and query sets.)

From the output files, algorithm performance can be computed for virtual galleries and probe sets. A gallery G is a virtual gallery if G is a proper subset of the target set, i.e., $G \subset T$. Similarly, P is a virtual probe set if $P \subset Q$. For a given gallery G and probe set P, the performance scores are computed by examination of the similarity measures $s_i(k)$ such that $q_i \in P$ and $t_k \in G$.

The score protocol allows for the implementation of the design assumptions and requirements. The first requirement is that training is before beginning the test. This forces each algorithm to have general representations for faces, not representations tuned to a specific gallery. Without this condition, virtual galleries would not be possible which allow for many different galleries to be constructed from a single target set.

For the algorithms to have general representations for faces, they must be gallery (class) insensitive, such as normalized correlation and principal component analysis (PCA). An algorithm is class sensitive if the representation is tuned to a specific gallery.

Virtual galleries and probe sets allow for an algorithm to be evaluated on different images of the same person, if more than one image of that person is placed in the target set. If such images are marked as the same person, then the algorithms being tested can use the information in the evaluation process. However, this would defeats the purpose of the virtual galleries. To avoid this happening, we require that each image in the target set be treated as a unique face. (In practice, this condition is enforced by giving every image in the target and query set a unique random identification.)

The remaining criterion for constructing virtual galleries and probe sets is that the similarity score between target and query images is a function of only those two images. If this is not the case, then all the images that contributed to $s_i(k)$ would have to be every gallery and probe set that contain t_k and q_i. This would place a significant restriction on the virtual galleries and probe sets that could be generated.

Ⅲ. Statistical Model

It was shown in Phillips *et al.* [5] that changing the gallery and probes changes absolute and relative performance of face recognition algorithms. Thus, it is necessary to measure the performance on a collection of galleries and probe sets. From these, score statistics can be computed. To be able to make informed statistical choices, it necessary to state the gallery population from the virtual galleries sampled.

Statistics are computed from a sample of virtual galleries G_i draw from a population of galleries G^*. The galleries in G^* all have a common property, e.g., all galleries of size 200. Statistical decisions and conclusions are made over G^*. To score an algorithm, both the gallery and probe set need to be specified. We assume that for each gallery, there is an unique probe set P_i, and we do not need to explicitly state P_i along with G_i. This is illustrated in the following example, which compares algorithms A_1 and A_2. Performance of an algorithm

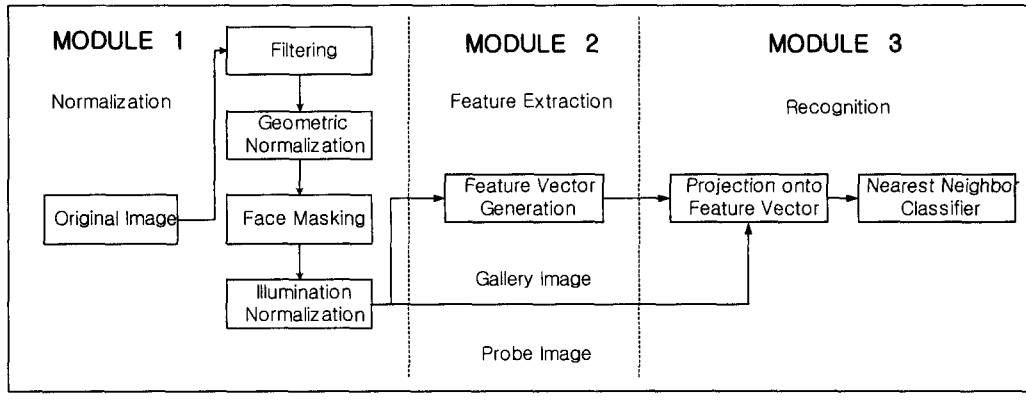


Figure 1. Block Diagram of Projetion-based Face Recognition System

A_j is measured with statistic h (j, G_i) on gallery G_i. The values of $h(j, \cdot)$ will have different values for different G_i and will have a distribution F_j over G^*. It is not practical to explicitly calculate F_j, therefore, we randomly sample G^* and estimate F_j by $\hat{F}_j$. From the empirical distribution $\hat{F}_j$, we estimate other statistics of h and perform hypothesis testing about the differences between algorithms.

Ⅳ. Experiments

We conducted two sets of experiments to show how the statistical model presented in section 3 can contribute to the design and evaluation of face recognition. In the first set, we performed hypothesis testing to estimate the relative performance of three algorithms. In the second set, we compared the effects of increasing the size of the gallery on algorithm performance.

4.1 System modules

Our face recognition system consists of three modules and each module is composed of a sequence of steps (figure 1). The first module normalize the input image. The goal of the normalization is to transform the facial image into a standard format that removes variations that can affect recognition performance. Figure 2 shows the input and output of some of the steps in the normalization module. The first step filters or compresses the original image. The image is filtered to remove high frequency noise in the image. An image is compressed to save storage space and reduce transmission time. The second step places the face in a standard geometric position by rotating, scaling, and translating the center of eyes to standard locations. The goal of this step is to remove variations in size, orientation, and location of the face. The third step masks out background pixels, hair, and clothes to remove unnecessary variations which can interfere identification process. The fourth module removes some of the variations in illumination between images. Changes in illumination are critical factors in algorithm performance.

The second module performs the feature extraction based on the training set which produces a set of feature vectors. The third module identifies the face from a normalized image, and consists of two steps. The first step projects the image onto the feature vectors. The critical parameter in this step is the subset of feature vectors that represent the face space. The second step recognizes faces using a nearest neighbor classifier. The critical design decision in this step is the similarity measure in the classifier. We have investigated performance results using L_1 distance, L_2 distance, angle between feature vectors, Mahalanobis distance [4]. Additionally,

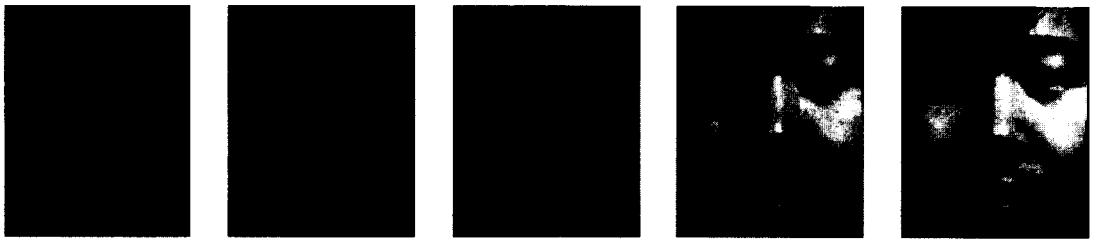


Figure 2. Input and output of several steps of the normalization module.

Mahalanobis distance was combined with L_1, L_2, and angle between feature vectors mentioned above.

All the algorithms were run on images from the FERET database of facial images using the Sept96 FERET evaluation protocol [2, 5]. The target set consisted of 3816 images and the query set consisted of 3323 images. In all the images, the eyes were manually located. Using the location of the eyes, the faces were translated, rotated, and scaled into a standard position. Once in the standard position, the background, hair, neck, and clothes were masked (figure 2).

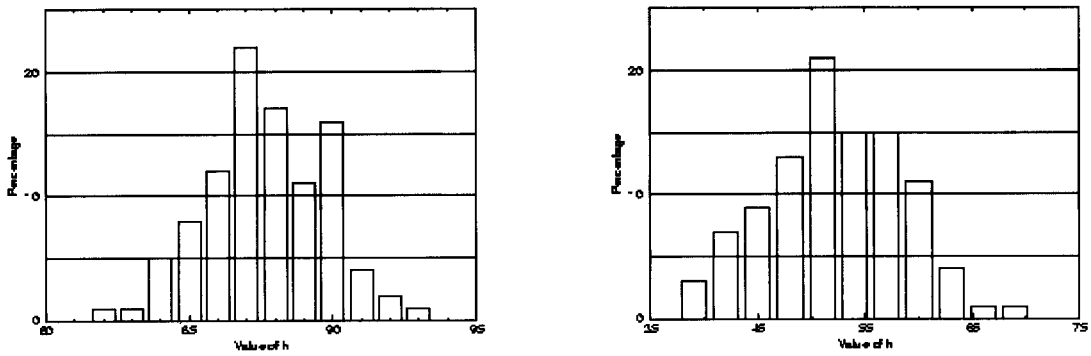
4.2 Test Design

From the target and query sets, we evaluated the algorithms of two categories of images. The first is the fb image. When the FERET database was acquired, two frontal images of each person were taken within five minutes under the same lighting conditions. One of these images is called the fa image and is in the gallery. The remaining image is the fb image, which is placed in the probe set. The fb test is a baseline test that evaluates the ability of an algorithm to recognize faces taken very close in time. The second category is the *duplicate* images. An image is a duplicate of a person in the gallery, if it was taken on a different day or under different circumstances than the gallery image. The FERET database contains images of people where the time between the first and most recent images is over a year and a half.

To demonstrate the statistical model, we ran the Sept96 FERET protocol on variations of normalized correlation and PCA algorithms [3]. After the images were transformed into a standard position, they went through a preprocessing step. We experimented with three types of preprocessing: (1) normalizing the images to have mean zero and unit variance, (2) histogram equalization, and (3) histogram equalization followed by normalization.

The recognition algorithms first projected the preprocessed image onto a new basis, then recognizes by the nearest neighbor classifier. PCA projects the image on an orthogonal basis that minimizes the variance in the training set [7]. In our implementation, the training set consisted of 500 images. We used the new basis with the first 200 eigenvectors. Normalized correlation performs recognition in the primal space (the images are not projected onto a new basis). For the nearest neighbor classifier, the L_2 and L_1 distances and the angle between vectors have been used for the PCA algorithm. For normalized correlation, the angle between vectors and the L_2 metric are equivalent.

The most common method of evaluating face recognition algorithms is to measure performance of algorithms on a single gallery and probe set. Since performance scores vary with the gallery and probe set, it is not possible to tell from a single score if the observed performance is average, optimist, or pessimistic. To address this issue, in the first set of experiments, we compared



(a) (b)

Figure 3. Histogram of h for 100 Galleries of size 200. The statistic h is the fraction of probes correctly identified. (a) Histogram for fb probe sets. (b) Histogram for duplicate probe sets.

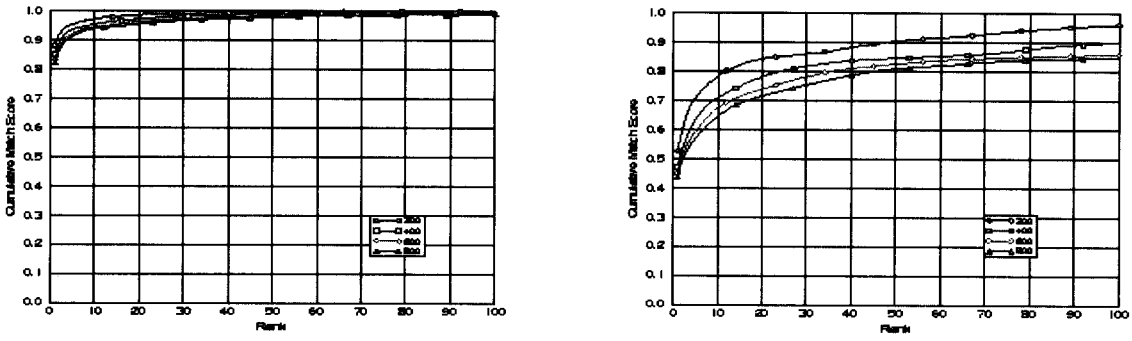
the performance of three algorithms over 100 galleries.

The gallery population G^* consisted of galleries of 200 individuals with one frontal image per person. From G^*, we randomly generated 100 galleries G_i and measured performance against two probe sets. The first probe set consisted of the fb images and the second consisted of duplicate probes. We measured performance as the fraction of probes that were correctly identified, the h statistic. Figure 3 shows a histogram of h over the G_i s for fb and duplicate probe sets. The statistic h was computed for the PCA algorithm with normalized preprocessing and the L_1 classifier.

The algorithms that we compare are (1) PCA with histogram equalization and normalized preprocessing, (2) PCA with histogram equalization preprocessing, and (3) normalized correlation (which implicitly includes normalized preprocessing). The classifier for all the algorithms was the L_1 classifier. We did a pair-wise comparison of all three methods. The problems were formulated as a hypothesis testing problem. The null hypothesis H_0 was that the empirical distributions were equal. We used the bootstrap permutation test (Efron and Tibshirani, chap. 15 [1]) comparing the differences of the sample means of h for each algorithm. Table 1 reports the sample means of h and the achieved significance level (ASL) of the tests. We can reject H_0 with a level a, usually 0.05 or 0.01, if the ASL is less than a. Based on the permutation test, we cannot reject that there is any performance difference between histogram equalization and combining normalization and histogram equalization. However, we can conclude that the performance of normalization by itself is different from the other two methods.

The above experiment examined algorithm performance for galleries of size 200. One of the concerns in face recognition is the effects on performance as the size of the gallery increases. To study this effect, we randomly generated galleries of size 200, 400, 600, and 800. For each gallery size, we generated 100 galleries and scored against the fb and duplicate probe sets. The scores were calculated using a PCA algorithm with histogram equalization preprocessing and the L_1 metric for recognition. The scores are tabulated as cumulative match scores for which computation is quite simple. Let p be the size of a probe set and R_k the number probes that the correct answer is in the top k. The fraction reported is $m_k = R_k/p$.

The results are presented in a graph, with k on the horizontal axis and m_k on the vertical axis. The m_k scores report performance for a single gallery and probe set. To summarize performance for a sample population, we average the



(a) (b)

Figure 4. Average cumulative rank scores for galleries of size 200, 400, 600, 800. The average was computed from 100 galleries for each size gallery. (a) Scores for fb probe sets. (b) Scores for duplicate probe sets.

cumulative match scores. If $m_k^i = R_k/p$ for G_i, then the average rank k score $\bar{m}_k = \frac{1}{N}\sum_i m_k^i$. The curve $(k, \bar{m}_k)$ is the average cumulative match scores for an algorithm. Figure 4 reports the average cumulative match scores for galleries of size 200, 400, 600, and 800, where N = 100.

The average cumulative score reports performance as a function of absolute rank. When comparing galleries of different sizes, it is insightful to compare relative performance of galleries. In relative performance, we measure the fraction of probes that are in the top ρ%. The average percent score for ρ% is $\tilde{m}_\rho = \bar{m}_{\lfloor \rho N/100 \rfloor}$. Figure 5 reports the results from figure 4 as average cumulative percent scores.

The graph of the average cumulative percent scores in figure 5 are very close in the value. This raises the hypothesis that distributions of h_ρ are the same for the different algorithms, where h_ρ is the fraction of probes that in the top ρ% . To test this hypothesis, we performed all pair-wise comparisons between different-sized galleries. We chose the null hypothesis to be that the distributions for h_{05} were equal for all gallery sizes. We performed the hypothesis testing with the bootstrap permutation test on the means of h_{05} . Table 2 reports ASL for all the tests conducted. The results show that in all cases, the null hypothesis cannot be rejected. We repeated the procedure for h_{10} , which produced the same conclusion, we cannot reject the null hypothesis.

Table 1. Hypothesis testing between h(a) for fb probe sets and (b) for duplicate probe sets, using the permutation test. The null hypothesis is that the distributions of h(j) are equal. We report ASL each pair-wise comparison of the algorithms and the mean of h(j) for each algorithm. EF = eigenface, Hist = Histogram Equalization, NMS = images have not been normalized. In all algorithms the L_1 metric is used.

	EF Hist	EF Hist(NMS)	Correlation
EF Hist	-	0.47	0.00
EF Hist(NMS)	-	-	0.00
Mean	0.96	0.96	0.98

(a)

	EF Hist	EF Hist(NMS)	Correlation
EF Hist	-	0.48	0.00
EF Hist(NMS)	-	-	0.00
Mean	0.74	0.73	0.67

(b)

V. Conclusions

In this paper we have used statistical methods to compare the performance of algorithms and the effects of using different galleries. This comparison allowed us to make intelligent and

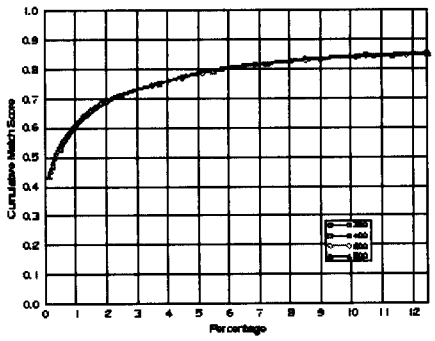


Figure 5. The cumulative match scores in figure 4(b) rescaled so that the rank is a percentage of the gallery.

informed decisions. We were able to perform the tests based on a statistical model, and the formulation was possible because we stated the assumptions underlying the testing methodology.

This is the critical step in applying testing methodologies to computer vision.

Table 2. Hypothesis testing comparing the distribution of galleries of different sizes. The statistic h selected was the percent of probes with ranked in the top 5%. The algorithm, galleries, and probe sets are the same as in figures 4 and 5.

	Gallery Size			
	200	400	600	800
200	-	0.44	0.46	0.20
400	-	-	0.43	0.31
600	-	-	-	0.32
Mean	0.78	0.78	0.78	0.79

References

[1] B. Efron and R. J. Tibshirani. An Introduction to the Bootstrap. *Chapman and Hall*, 1993.

[2] S. Gutta, J. Huang, D. Singh, I. Shah, B. Takacs, and H. Wechsler. Benchmark studies on face recognition. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 227--231, 1995.

[3] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In K. W. Bowyer and P. J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision. IEEE Computer Society Press*, Los Alamitos, CA, 1998.

[4] H. Moon and P. J. Phillips. Analysis of pca-based face recognition algorithms. *In 2nd Inter. Conf. on Audio-and video-based biometric person authentication*, pages 205--210, 1999.

[5] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings Computer Vision and Pattern Recognition 97*, pages 137--143, 1997.

[6] P. J. Phillips, P. Rauss, and S. Der. FERET (face recognition technology) recognition algorithm development and test report. Technical Report ARL-TR-995, *U.S. Army Research Laboratory*, 1996.

[7] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71--86, 1991.

문 현 준(Hyeonjoon Moon) 정회원

1990년 : Korea Univ., B.S., Electronics and Computer Engineering

1992년 : State Univ. of New York at Buffalo, NY. M.S., Electrical and Computer Engineering

1999년 : State Univ. of New York at Buffalo, NY. Ph.D., Electrical and Computer Engineering

1990년~1991년 : LG Information and Technology Co. Systems Engineer

1993년~1994년 : SDS Co., Systems Engineer

1994년~1995년 : State Univ. of New York at Buffalo, NY., Research Assistant

1996년~1999년 : U.S.Army Research Laboratory, Maryland, Research Associate

1999년~현 재 : Lau Technologies, Inc., Senior Research Scientist

<주관심 분야> 컴퓨터비전, 패턴인식, 영상처리, HCI(Human Computer Interaction)

백 순 화(Soon-Hwa Baek) 정회원

1985년 : 계명대학교 전자계산학과 졸업
1993년 : 호서대학교 전자계산학과 석사
1999년 : 충북대학교 컴퓨터공학과 박사과정수료

1999년~현재 : 천안외국어대학 컴퓨터정보과 전임강사
<주관심 분야> 영상처리, 패턴인식, 신호처리

전 병 민(Byoung-Min Jun) 정회원

1976년 : 한국항공대학교 전자공학과 졸업
1978년 : 연세대학교 전자공학과 석사
1988년 : 연세대학교 컴퓨터공학과 박사
1978년~1982년 : 공군사관학교 전자공학과 전임강사
1982년~1986년 : 동양공업전문대학 통신과 조교수
1992년~1993년 : 미시간대학교 교환교수
1986년~현 재 : 충북대학교 컴퓨터공학과 교수
<주관심 분야> 영상신호처리