Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 949–960
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Classification in an informative sample subspace

Guoping Qiu[a,b,*], Jianzhong Fang[a,c,1]

[a]School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, UK
[b]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[c]Endace Technology, Hamilton, New Zealand

Received 12 April 2006; received in revised form 19 June 2007; accepted 6 July 2007

Abstract

We have developed an informative sample subspace (ISS) method that is suitable for projecting high-dimensional data onto a low-dimensional subspace for classification purposes. In this paper, we present an ISS algorithm that uses a maximal mutual information criterion to search a labelled training data set directly for the subspace's projection base vectors. We evaluate the usefulness of the ISS method using synthetic data as well as real world problems. Experimental results demonstrate that the ISS algorithm is effective and can be used as a general method for representing high-dimensional data in a low-dimensional subspace for classification.
Published by Elsevier Ltd on behalf of Pattern Recognition Society.

Keywords: Information theory; Mutual information; Subspace methods; Representation; Classification; Object detection

1. Introduction

Representation plays a key role in the success of computer vision and pattern recognition algorithms. An effective representation method should be compact and discriminative. It is desired that the representation should have low dimensionality to combat the "curse of dimensionality" problem and to improve computational efficiency. The representation should also ideally be in a space where different classes of object/data are well separated.

A popular class of object representation technology is the linear subspace methods where high-dimensional inputs are projected onto a low-dimensional subspace in which object recognition can be carried out more efficiently. Depending on the ways in which the subspace's base vectors are chosen, the technology can be classified into supervised or unsupervised subspace methods. In unsupervised methods, the identity information of the inputs is not used in the derivation of the subspace base vectors, whilst in supervised methods, the identity information of the input is exploited in deriving the subspace base vectors. Examples of the former approach include the well-known classical technique of principal component analysis (PCA) whilst examples of the later include the linear discriminant analysis (LDA). In the context of pattern recognition/classification, PCA and LDA have been extensively studied in the past and there exists a huge body of literature, examples include [1–4].

Although unsupervised methods such as PCA can produce compact representation and is optimal in the sense of reconstruction errors under the L_2 norm, the eigen-subspace is not necessarily optimal in terms of discriminative power. In many classification applications, such as face or car detection [5,6], it is necessary to have labelled data to derive classification models. Because the label (identity) information of each training sample is available anyway, it makes sense and may also be advantageous to use the class label information to derive the base vectors of an input subspace. Using such supervised approaches to deriving object representations for classification has recently gained popularity and been demonstrated to be advantageous over unsupervised methods, see, e.g. Refs. [2–4,6,7]. A more recent work [8] combines the reconstructive capability of unsupervised approach and the discriminative power of supervised approach.

* Corresponding author. Tel.: +44 115 8466507; fax: +44 115 9514254.
E-mail address: qiu@cs.nott.ac.uk (G. Qiu).

[1] The work was done when J. Fang was with the School of Computer Science, University of Nottingham, UK.

0031-3203/$30.00 Published by Elsevier Ltd on behalf of Pattern Recognition Society.
doi:10.1016/j.patcog.2007.07.016

Projection-based supervised methods exist in the literature, such as those in Refs. [2,3] use Fisher's LDA to derive the projection subspace base vectors. However, LDA only makes use of the covariance of the inputs, it is only optimal for discriminating object classes having unimodal Gaussian density with well-separated means. Because most real world problems are not unimodal Gaussian nor well-separated, the discriminative power of LDA subspaces may also be limited.

In deriving a subspace in which, not only can the inputs be represented in low dimensions, but also the projections of the inputs in the subspace are well separated, it may be beneficial to exploit higher than second order statistical information. One way to exploit higher order statistics is to introduce kernel-based methods; kernel PCA [9] and other kernel based subspace methods, e.g. Refs. [10,11] have been proposed in recent years. Another way to exploit higher order statistics is to use information theory [12], e.g., use mutual information between the inputs and object class labels. Theoretically, information theoretic methods may have a number of advantages. For example, mutual information measures general statistical dependence between variables rather than their linear correlations. The mutual information is also invariant to monotonic transformations performed on the variables.

Mutual information has been successfully employed in deriving supervised but part-based representations for object classification [7], which has been demonstrated to be advantageous over non-supervised object representations. In this paper, we present a method that employs the maximal mutual information criterion to develop a supervised and projection-based method for object representation for classification purposes.

The organization of the paper is as follows. Section 2 gives a brief background overview on the Shannon information theory and Fano's inequality on the relationship between mutual information and a lower bound of misclassification error that motivates the use of mutual information for deriving projection-based object representation subspace methods. Section 3 describes a learning procedure for deriving a set of mutual information maximizing subspace base vectors. Section 4 presents experiments and results of applying the method to artificial data and real world problems. Section 5 concludes the paper. A preliminary version of this work has been published in Ref. [13].

2. Information theory background

Let ensemble X be a random variable x with a set of possible outcomes, $A_X = \{a_1, a_2, \ldots .a_n\}$, having probabilities $\{P(x = a_i)\}$, and ensemble Y be a random variable y with a set of possible outcomes, $A_Y = \{b_1, b_2, \ldots .b_m\}$, having probabilities $\{P(y = b_j)\}$. Let $p(x, y)$, $x \in A_X$, $y \in A_Y$ be the joint probability. The mutual information between X and Y can be defined as (other forms of definition also exist)

$$I(X; Y) = \sum_{x \in A_x} \sum_{y \in A_y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right). \quad (1)$$

The mutual information measures the average reduction in uncertainty of x as a result of learning the value of y, or vice versa. Another interpretation of the mutual information measure is that it measures the amount of information x conveys about y [12].

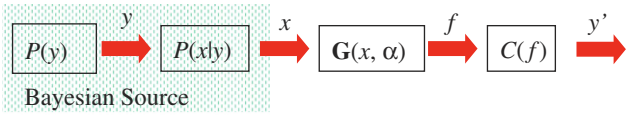

Fig. 1. Interpreting the classification process as a Markov chain [15,16], y is the object class random variable, x is the observation generated by the conditional probability density function $P(x|y)$. The observation is subjected to a transform G, which produces a new feature f from input x. The classifier C then estimates the class identity of input x as y' based on the transformed feature f.

In the context of object classification, Fano's inequality [14] gives a lower bound for the probability of error (an upper bound for the probability of correct classification). Our present application uses Fano's inequality in much the same way as it has been used by other authors [15,16]. The classification process can be interpreted as a Markov chain as illustrated in Fig. 1.

The probability of misclassification error in the setting of Fig. 1, $P_e = P(y \neq y')$, has the following bound [14]:

$$P(y \neq y') \geqslant \frac{H(Y) - I(Y, F) - 1}{\log(m)}, \quad (2)$$

where $H(Y)$ is the entropy of Y, F is the ensemble of random variable f, and m is the number of outputs of y (number of object classes). The form of the classifier, C, has not been specified. Eq. (2) quantifies at best how well we can classify the objects using the features f. However, an upper bound of the probability of misclassification error cannot be expressed in terms of Shannon's entropy. The best one can do is to minimize the lower bound to ensure an appropriately designed classification algorithm does well. Since both m and $H(Y)$ are constants in Eq. (2), we can maximize the mutual information $I(Y; F)$ to minimize the lower bound of the probability of misclassification error. The task now becomes that of finding the transform function that minimizes this lower bound. In the next section, we propose a practical solution.

3. The informative sample subspace algorithm (ISSA)

For the original d-dimensional input data $\boldsymbol{x}$, we would like to find its low-dimensional representation $\boldsymbol{f}$, by projecting it onto a k-dimensional ($k \ll d$) subspace using a $k \times d$ matrix $\boldsymbol{G}$

$$\boldsymbol{f} = \boldsymbol{G}\boldsymbol{x}. \quad (3)$$

The representational quality of $\boldsymbol{f}$ is directly dependent on the k row vectors of $\boldsymbol{G}$ and different subspace methods differ in the ways in which these vectors are computed. For example, in PCA, the first k eigenvectors of the covariance matrix of the input data form the projection matrix $\boldsymbol{G}$. In this paper, we are motivated by the relation between mutual information and classification errors, Eq. (2) in section 2, and we choose to use an information theoretic criterion to select the projection

The Informative Sample Subspace Algorithm (ISSA)

Step 1: Set $l = 0$, $G = 0$ (empty)

Step 2: Calculate the projections of all samples $\{\boldsymbol{x}_i\}$ onto each of the samples, $\boldsymbol{x}_j$, one at a time: $\boldsymbol{Z}_j = \{<\boldsymbol{x}_i, \boldsymbol{x}_j> \mid \forall\, \boldsymbol{x}_i\}, \forall\, \boldsymbol{x}_j$

Step 3: Calculate the mutual information $\boldsymbol{I}(\boldsymbol{Z}_j; \boldsymbol{Y}), \forall\, \boldsymbol{Z}_j$, according to (1).

Step 4: Find the sample, E, that generates the largest $\boldsymbol{I}(\boldsymbol{Z}_j; \boldsymbol{Y})$, i.e.

$$E = \arg\max_{X_j \quad \forall Z_j} \{I(Z_j; Y)\}$$

Step 5: Set $l = l + 1$

$$e_l = E \Big/ \|E\| \qquad \text{and} \qquad G_l = G_{l-1} \cup e_l \text{, i.e., } e_l \text{ used as the } l\text{th row of } \boldsymbol{G}$$

Step 6: If $(l \geq k)$ then stop, otherwise go to Step 7.

Step 7: $\{\boldsymbol{x}_i\} \leftarrow \{\boldsymbol{x}_i\} \setminus E$, and for all $\boldsymbol{x}_i$ do:

$$\boldsymbol{x}_i = \boldsymbol{x}_i - \langle \boldsymbol{x}_i, \boldsymbol{e}_l \rangle \boldsymbol{e}_l$$

Step 8: Go to Step 2

Fig. 2. The informative sample subspace (ISS) algorithm.

matrix $\boldsymbol{G}^*$:

$$G^* = \arg\max_{\forall \boldsymbol{G}} \boldsymbol{I}(\boldsymbol{GX}; \boldsymbol{Y}), \tag{4}$$

where $\boldsymbol{Y}$ is the identity variable of input variable $\boldsymbol{X}$, $\boldsymbol{I}(\boldsymbol{X}; \boldsymbol{Y})$ is the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Although Eq. (4) provides an elegant motivation for selecting an informative transform, it also presents huge computational challenges. To estimate the mutual information, it is necessary to estimate the probability density functions of the variables and calculate integration functions of these density functions, which leads to exponential complexity. A closed form solution is not practical (if not impossible). On the other hand, it should be recognized that because Eq. (2) is only a lower bound of the misclassification error, and an optimal solution to Eq. (4) does not necessarily guarantee optimal classification performance. Nevertheless, Eqs. (2) and (4) provide a motivation for developing a subspace in which data classification can be effectively carried out.

In many applications, we will have training samples available. Especially in the application domains we have in mind, image data, it is easy to obtain a large pool of data samples either by directly collecting from sensory input or by synthesis. At the core of our algorithm is a simplified assumption that the row vectors of the projection matrix $\boldsymbol{G}^*$ in Eq. (4) can be directly selected from the pool of training samples. This is similar to dimensionality reduction using random projection [17,18] where the projection matrix is generated randomly using a very simple probability distribution. Instead of using random vectors, we use a maximal mutual information criterion to select appropriate training samples and use them directly as the row vectors of the projection matrix. To simplify the procedure further, we select one row vector at a time and we make the row vectors unit length and orthogonal to each other. The procedure of our algorithm is summarized in Fig. 2.

In words, the algorithm works as follows. To find the first transform base (the first row vector of $\boldsymbol{G}$), we select one sample at a time, and project all other training samples onto that selected sample. The projections (a set of scalar numbers) and the samples' identities can be used to estimate the joint probability, which in turn can be used to estimate the mutual information of the projections and the class distribution. The sample with projection outputs that maximizes the mutual information is selected as the first transform base. This base is then removed from the training sample set. All remaining samples are then made orthogonal to the first base and used as training samples to find the second transform base. The process continues until all required k bases are found. From the procedure it is not difficult to see that all k bases are orthonormal.

Since projecting the samples onto a selected sample will produce a set of scalar values, which can be discretized to estimate the joint probability to compute the mutual information according to Eq. (1). In practice, we use the joint histogram to approximate the joint probability and this makes the method practicable with modern computers.

If we have a large enough pool of samples, it is reasonable to assume that it is possible to choose the representative ones which are most informative. In this scheme, the computational costs are proportional to the number of training samples. There may exist a trade-off between performance and computing costs. A larger pool of samples may enable us to find a better set of basis, but it will be computationally more costly. Using training pool sizes manageable by a modern PC, we will show that although such an approximated approach has no guarantee of finding the optimal solution, it is a practical solution and is effective.

4. Experiments

In this section, we evaluate the practical usefulness of the ISS algorithm. We have performed three experiments. In the first experiment, we use the method on simple synthetic data. In the second experiment, we apply the method to face detection. In the third experiment, we use the method in the detection of cars in photographs.

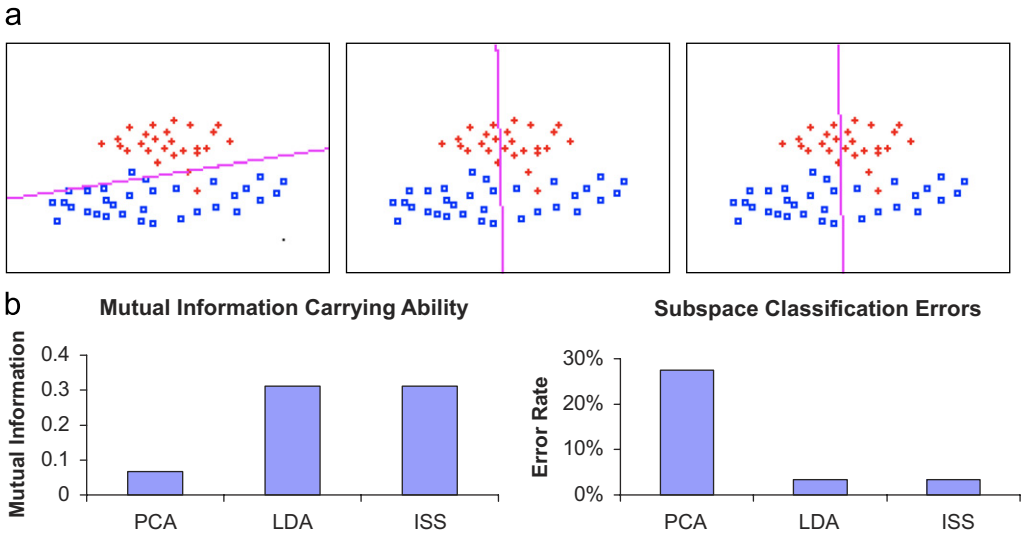

Fig. 3. Different subspaces and their classification ability. (a) Two-class 2-d features and the 1-d subspace base vector (the line) of different subspaces, from left to right: PCA, LDA, and ISS. (b) Mutual information carried in the 1-d feature of different subspaces and classification errors in the 1-d subspace using a naı¨ve Bayesian classifier.

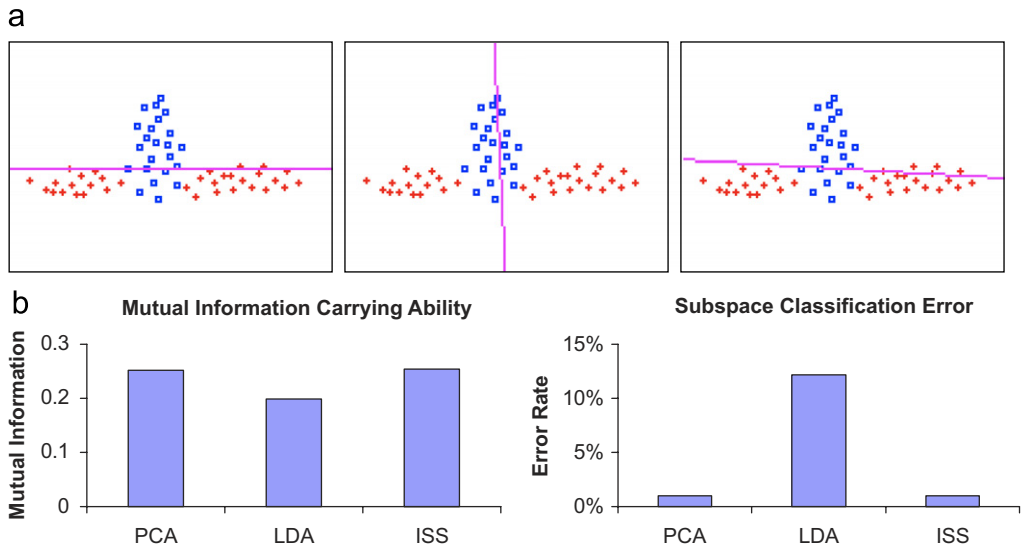

Fig. 4. Different subspaces and their classification ability. (a) Two-class 2-d features and the 1-d subspace base vector (the line) of different subspaces, from left to right: PCA, LDA, and ISS. (b) Mutual information carried in the 1-d feature of different subspaces and classification errors in the 1-d subspace using a naı¨ve Bayesian classifier.

4.1. Synthetic data

In this experiment, we have generated artificial two class problems in two-dimensional feature space for easy analysis and visualization. We then project the 2-d features in 1-d subspace and use a naı¨ve Bayesian classifier to classify the data in the projected 1-d space. Note there are 100 samples 50 from each class; 50% of randomly selected data were used for training and 50% for testing. The purpose of this experiment is to verify the soundness of our method and to demonstrate the possible advantages of our method over other subspace methods. We compare our method with two well-known subspace methods, PCA and Fisher's LDA. We also measure the mutual information between the projected 1-d features and the data's class labels to verify the relation between mutual information and classification error.

Fig. 3 shows a situation where projecting the data onto the 1-d PCA base will fail to separate the two classes while projecting the data onto LDA or ISS bases will make the classes much easier to separate. In this case, PCA found the maximal variance direction but failed to find the most discriminative direction. It is also shown that the mutual information and classification error has a direct relation. The higher the mutual information carried in the subspace the lower the classification errors and vice versus.

Fig. 4 shows a situation where projecting the data onto the LDA base will fail to separate the classes whilst projecting onto either PCA or ISS bases will make the separation much easier.

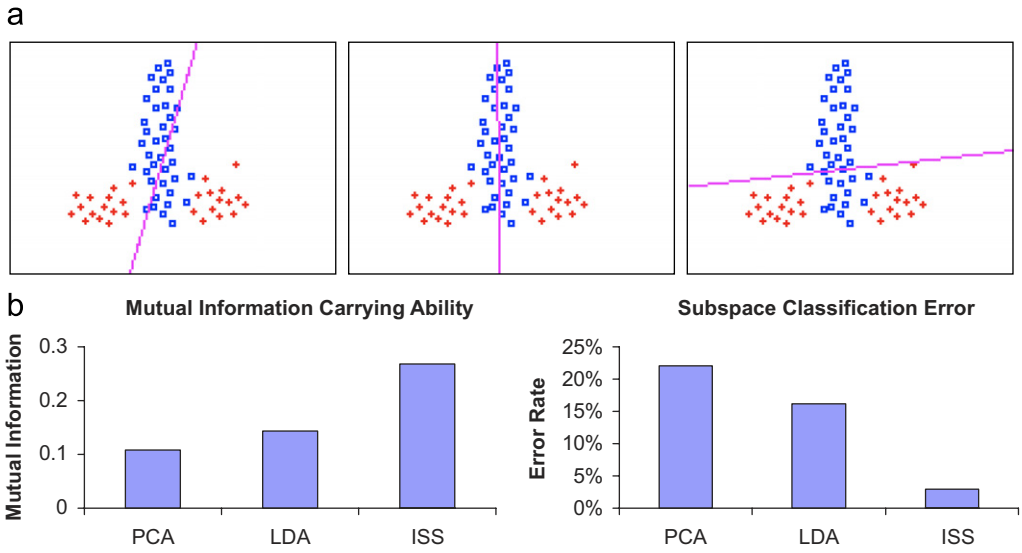

Fig. 5. Different subspaces and their classification ability. (a) Two-class 2-d features and the 1-d subspace base vector (the line) of different subspaces, from left to right: PCA, LDA, and ISS. (b) Mutual information carried in the 1-d feature of different subspaces and classification errors in the 1-d subspace using a naı¨ve Bayesian classifier.

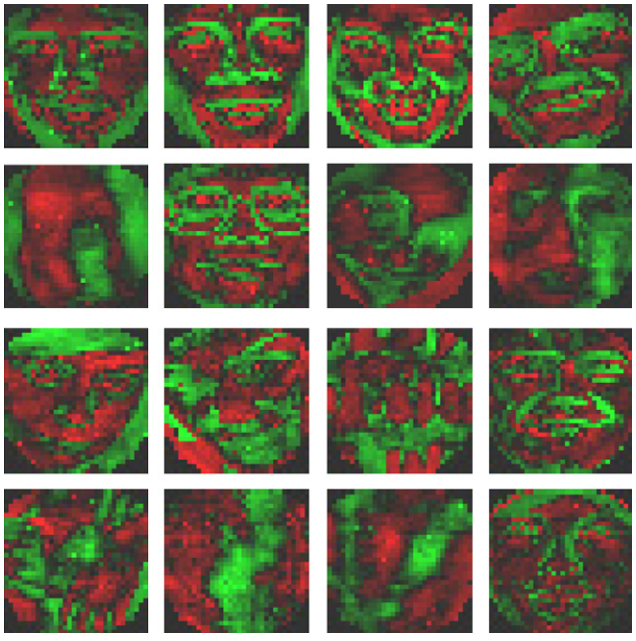

Fig. 6. Sixteen ISS subspace base vectors, red (darker) regions represent positive value; green (lighter) regions represent negative value. They are developed from 3153 face and 6237 non-face samples.

The reason that PCA performs well in this example is because the most discriminative direction happens to coincide with the largest variance direction of all the data points. It is seen that in this case, LDA found the direction of the maximal ratio of between-class scatter and within-class scatter [1–4] but failed to capture the maximal discriminative direction. This shows that LDA will not necessarily perform better than PCA even though LDA has used labelled data to develop the projection base. It also shows that when the decision boundaries are complex and the class distributions are not unimodal, LDA can easily suffer from failure. However, it is clearly seen that ISS projection is free from class distribution assumption and has successfully separated complicated class distributions. Again, it is also shown that the mutual information and classification error has a direct relation. The higher the mutual information carried in the subspace the lower the classification error and vice versa.

Fig. 5 shows a situation where projecting the data onto either PCA or LDA bases will fail to separate the classes whilst project them on the ISS base can successfully separate the classes. In the case of PCA, it captures the direction of the maximum variance, whilst LDA captures the direction of maximal ratio of between-class scatter and within-class scatter. However, both PCA and LDA fail to capture the most discriminative direction whilst ISS has successfully found the most discriminative direction. Once again, it is also shown that the mutual information and classification error has a direct relation. The higher the mutual information carried in the subspace the lower the classification errors and vice versa.

PCA is a widely used liner transform for dimensionality reduction. It is an optimal dimensional reduction technique in the mean square error sense. The eigen subspace captures the maximal variance directions of the data. However, as we have seen in Fig. 5, if the maximal variance directions and the maximal discriminative directions do not coincide (and there is no guarantee for this), then PCA subspace will not be suitable for representing the data for classification purpose.

At a first glance, our method is similar to Fisher's LDA [1–4] because both use labelled data. However, there are fundamental differences. Whilst LDA assumes unimodal equal variance Gaussian class distributions which are almost always not true for real world data, ISS does not make such a restricted assumption. As seen in Figs. 4 and 5, when this unimodal assumption is invalid, LDA fails to find the most discriminative subspace. Whilst LDA is not directly related to classification rates, maximizing mutual information directly minimizes a lower bound

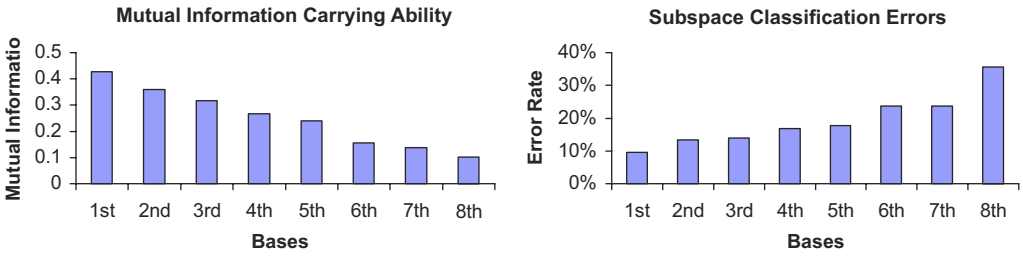

Fig. 7. Mutual information carried in the first eight projection vectors (left) and Bayesian classification error of each projection vector (right).

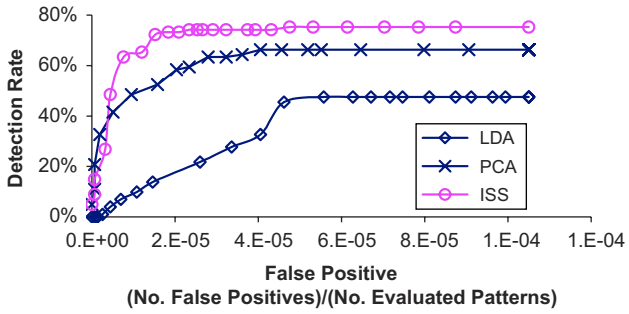

Fig. 8. Face detection performances in three different subspaces, The original 1024-d data is projected onto the first eight projection vectors in the subspaces to form 8-d data which is then used to train the SVMs. The SVMs are then used to detect faces in the 101 images randomly selected from the FERET database.

of classification errors. Whilst LDA only makes use of the covariance, ISS exploits higher than second order, more general statistical information. Although the experimental data used in this experiment is relatively simple, the results are nevertheless indicative. By analysing the way in which the subspaces are developed, we have reason to believe that the ISS method can overcome the limitations of LDA and PCA because ISS captures higher order, more general statistical information.

4.2. Application to face/non-face classification (face detection)

Human face detection has been a popular research subject in pattern recognition for many years, examples of face detection techniques include [19–24]. A comprehensive survey of technologies developed up to the year 2002 can be found in Ref. [20]. In this section, we apply our informative subspace method for representing face/non-face patterns in a low-dimensional space for face detection. To detect faces, we use a support vector machine (SVM) [19] which takes the representations in the low-dimensional informative subspace as inputs and classifies them into face/non-face. It is to be noted that here we essentially treat the problem as a classification problem. Our purpose is to test the usefulness and effectiveness of the ISS method for representing the face/non-face patterns for developing the classifier, rather than building the face detector, which, as has been pointed out in Ref. [25], is greatly affected by the face detection infrastructures other than the classifier.

For training the informative subspace and the SVM classifier, we have collected 3153 face and 6237 non-face samples. The original face samples are of various scales with different aspect ratios. These original samples are then scaled to a uniform size of 32×32 pixels. The non-face samples are 32×32 patches collected from various types of images. Therefore the original samples are 1024-d data and we make all samples to have zero mean. Obviously, the dimensionality of the data is too high to be directly used for classification. We therefore apply our informative sample subspace algorithm as described in Section 3 to reduce the dimensionality of the original samples before using them to train the SVM classifier. Examples of the first 16 maximal information sample subspace projection vectors are shown in Fig. 6. It is seen that some of the projection vectors come from face samples because they can be easily identified as faces whilst others come from non-face samples.

To study the amount of mutual information carried by the projection base vectors, we compute the mutual information between the projections of the training samples on a projection base vector and the sample labels. For the lth projection base vector, we compute the mutual information $I_l = I(\langle e_l, X\rangle; Y)$, where $\boldsymbol{X}$ is the training samples and $\boldsymbol{Y}$ is the sample's label. Fig. 7 shows the mutual information carried by each of the first 8 ISS projection vectors. It is seen that the amount of mutual information carried by the projection vectors is in decreasing order, i.e., the first projection vector contains the most mutual information, and so on. We also found that as more projection vectors are added, the difference in the amount of information carried by the consecutive projection vectors becomes smaller and smaller.

Again, as an exercise to study the relation between mutual information and classification error, we use every one of the first eight projections each independently to construct a Bayesian classifier to classify the face and non-face training samples. That is, we use the 1-d feature projected onto a projection vector to construct a Bayesian classifier. Classification errors of the first eight projection vectors are also shown in Fig. 7. It is seen that the higher the mutual information a projection vector carries, the lower the classification error.

Since the first subspace vector carries the highest mutual information and subsequent bases carry less and less mutual information, and high mutual information leads to low classification error, we therefore only need to use the first k bases of the ISS to perform classification. As a guide, we determine k in following manner: Following the ISS algorithm, we choose

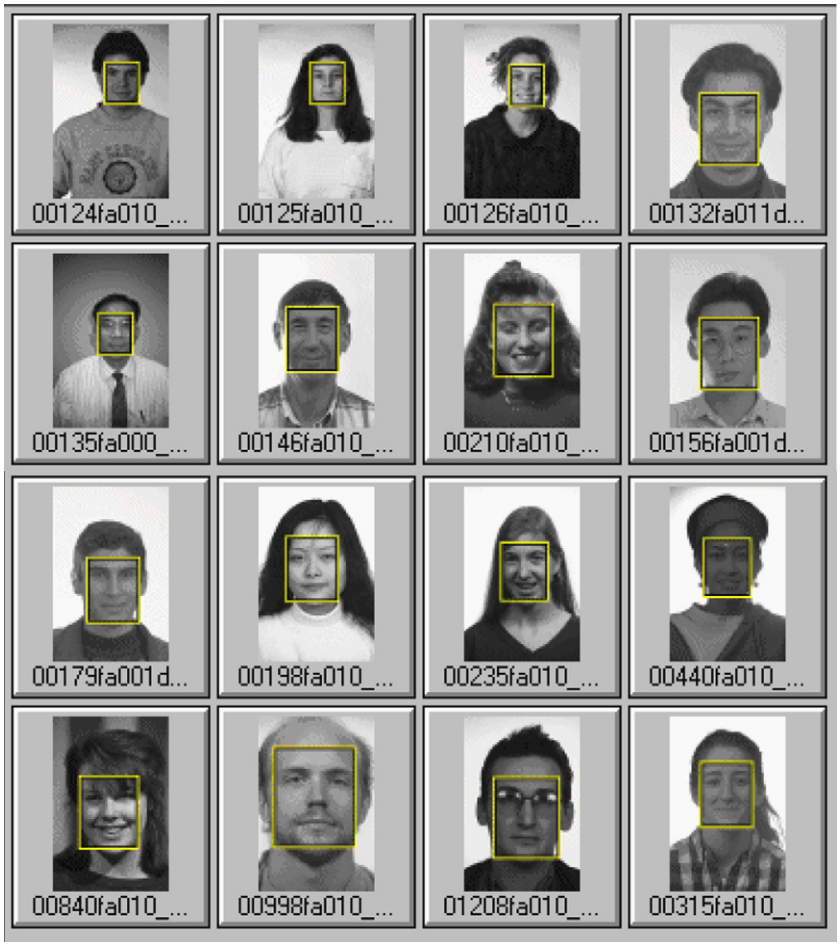

Fig. 9. Example results of face detection in the ISS subspace on the FERET data set.

Table 1
Comparison of different face detector on the CMU data set

Detectors	Test set (faces/images)	False positives	Detection rate
Li et al. [21]	481/125	31	90.2%
Schneiderman and Kanade [24]	481/125	65	94.4%
Osuna et al. [19]	155/23	20	74.2%
Rowley et al. [22]	507/130	95	89.2%
Viola & Jones [23][a]	507/130	95	90.8%
Our ISS/SVM Detector[b]	507/130	90	91.5%

[a]Note in Viola and Jones [23], several additional false positives/detection rates are listed, these are 78/90.1%, 110/91.1%, 167/91.8%.

[b]Note our result has been obtained by varying the detection threshold for different images. In the literature, it is not always clear whether different thresholds have been used for different images or a uniform threshold has been used for all images to obtain the reported results.

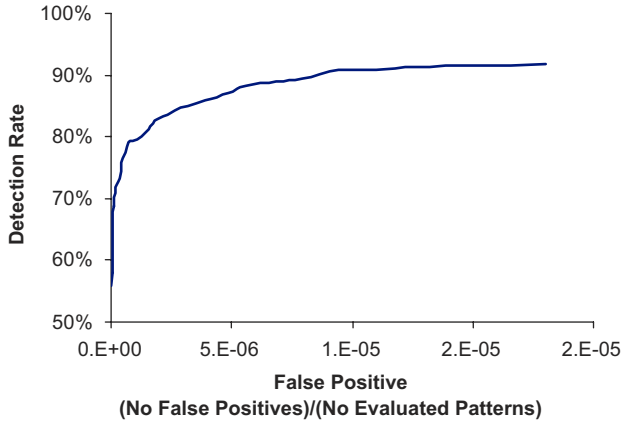

Fig. 10. Receiver operating curve of ISS subspace face detector performed on the CMU 130 images 507 faces data set.

the ISS base vector one at a time. For each base vector, we use the projection onto the base vector to build a naı¨ve Bayesian classifier to classify the training samples. If the classification error rate is $> 50\%$, then we stop adding more base vectors to the transform matrix $\boldsymbol{G}$. The rationale is that when the classification error is higher than 50%, it is no better than random guesses, therefore adding more bases will unlikely improve performances.

In the first testing, we compare the performances of SVM face detectors built in three subspaces, PCA, LDA and ISS. For the LDA subspace, the projection vectors are developed based on the technique of Ref. [2]. This testing dataset consists 101 images randomly selected from the FERET database [26]. Fig. 8 shows the receiver operating curves (ROC) of the SVM

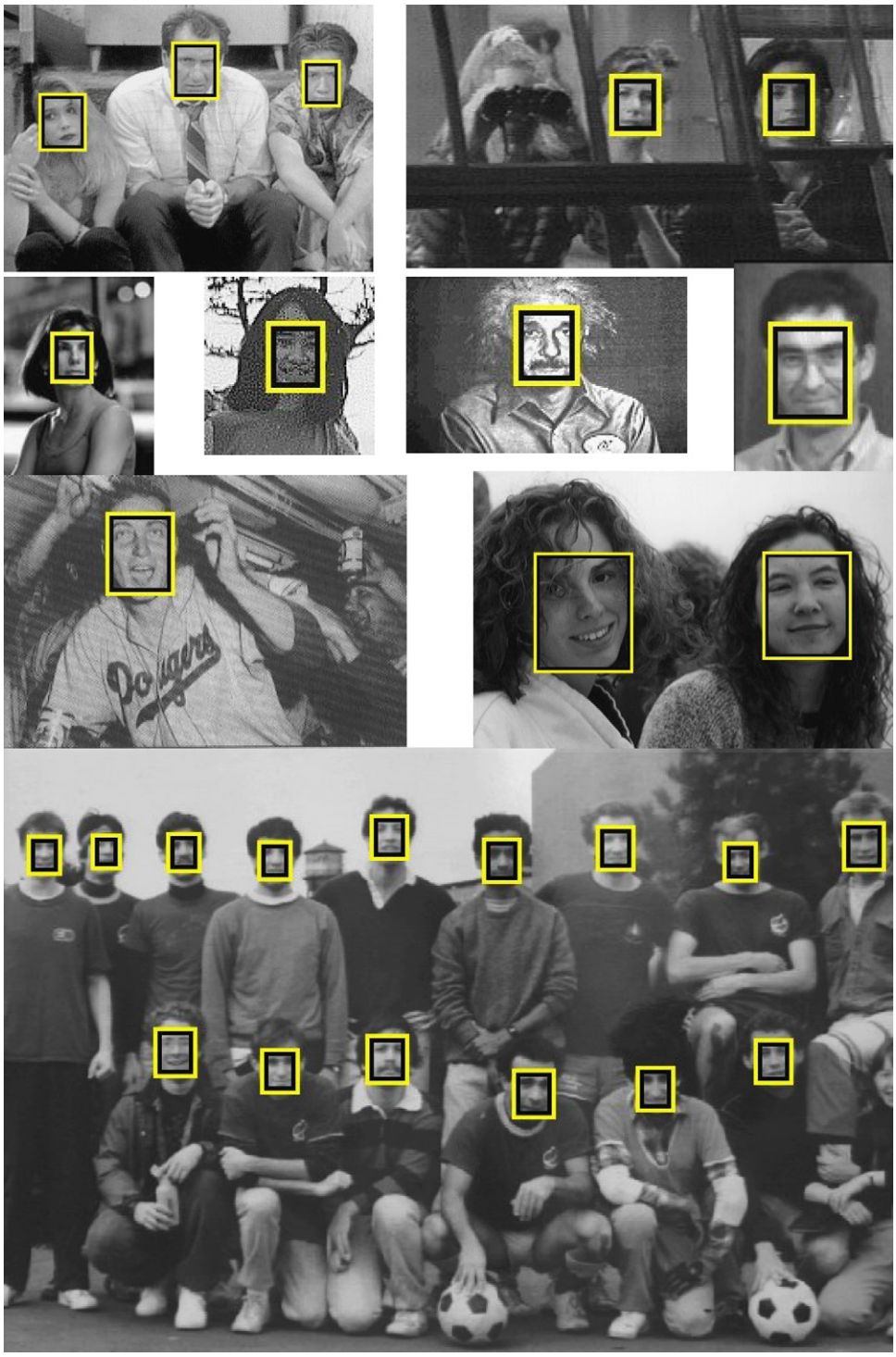

Fig. 11. Examples of face detection result in the ISS subspace for the CMU data set.

face detectors in the 8-d subspaces of PCA, LDA and ISS. It is seen that in these low dimensional spaces, the ISS subspace gives better performances. Some examples of face detection in the ISS subspace are shown in Fig. 9.

A popular face detection benchmark data set is that of CMU data set [27]. The data set contains 130 images and in them there are 507 faces of various sizes. We have implemented a face detector in a 64-d ISS subspace to detect faces in this data set. To train the detector, we first find the first 64 projection vectors in the ISS subspace according to the algorithm of Fig. 2. We then train a 64 input SVM to classify the input. For face detection in these images, we use 31 detection window sizes ranging from 18×22 pixels to 426×520 pixels. For each detection window, it is first rescaled to 32×32 pixels.

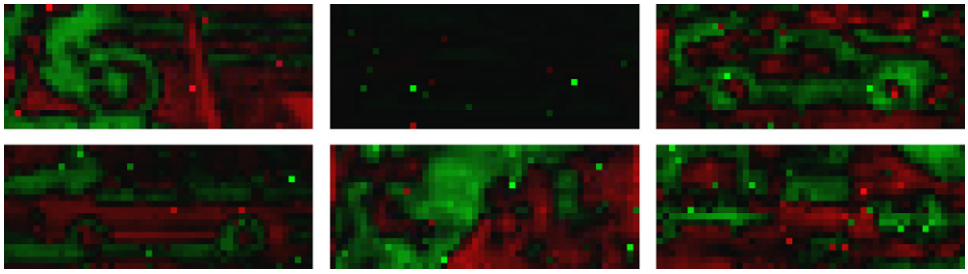

Fig. 12. Examples of the car/non-car sample ISS subspace projection vectors. Red (darker) colour represents positive values and green (lighter) colour represents negative values.

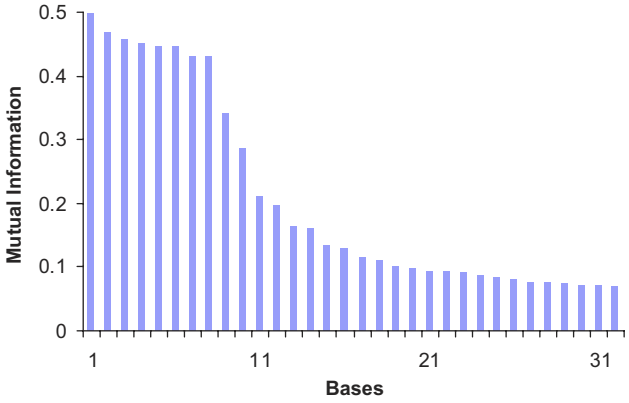

Fig. 13. Mutual information carried in the first 32 projection vectors of a car/non-car ISS subspace.

The 1024-d vector is then projected onto the first 64 projection vectors in the ISS subspace and then input to the SVM. In total the detector has evaluated 52,129,308 patterns. Results of our detector and several well-known face detectors that have used the dataset are shown in Table 1. By using the same detection threshold for all the images and varying this universal detection threshold, Fig. 10 plots the receiver operating curve (ROC) of our ISS subspace face detector for this test data set. Fig. 11 shows some examples of our face detection results. Overall, our detector achieves similar performance to state of the art technologies. Note that the first three detectors in the table did not use the full set whilst ours and the other two have used all 130 images in testing.

4.3. Application to car/non-car classification (car detection)

In this experiment, we apply our subspace method to another popular object detection application—car detection. We use the car and non-car training and testing images from the University of Illinois at Urbana-Champaign [6]. The advantage of using this database is that both the training data and testing data are publicly available which makes it easier to compare the performances of different methods. In this database, the training set contains 550 side-view car images and 500 non-car images with a size of 100 by 40 pixels stored in PGN raw data format. The testing set contains 170 images with 200 cars in them. The scales of the cars in the testing set are approximately the same as those in the training set. The testing images are different in size and the number of cars in them. Again, we treat the problem as a classification task and evaluate the usefulness of the ISS algorithm for representing car/non-car patterns for classification.

In our car detection system the input (detection window of 100×40 pixels) is first down-scaled by a factor of two in both dimensions. The purposes of down scaling are twofold. One is to reduce the dimensionality of the input and the other is to smooth the input (remove noise). We have tried a number of scaling factors and found that a down scaling factor of 2 gave the best results. We again make the samples to have zero mean. Even though down scaling the detection window from 4000-d to 1000-d is a significant reduction, the dimension of the input is still too high. We therefore apply our ISS subspace method to project the 1000-d sample vectors to 32-d representation. A SVM is then trained in the 32-d subspace to classify the samples.

Using the car/non-car training samples, we first use the algorithm in Fig. 2 to find the first 32 maximal mutual information projection vectors. Fig. 12 shows examples of six projection base vectors, some clearly come from car samples whilst others come from non-car samples. Fig. 13 shows the amount of mutual information carried in each individual projection vector.

We apply our car detector to the test set, which consists of 170 grey-scaled images with 200 cars in them. We employ exhaustive search method on the testing set. The 100 by 40 pixel detection window moves four pixels horizontally and two pixels vertically each time during the evaluation. The input image patch is then downscaled to 50 by 20 pixels before being projected down to 32-d vector in the maximal information sample subspace. We also define that a car is correctly detected if all parts of the car are enclosed in the 100 by 40 pixel window, which is a more strict criterion than the one defined in Ref. [6]. The car detector evaluated 176,792 patterns over the 170 test images, 166,856 out of which are negative outputs.

In order to make comprehensive comparison we adopt three different criteria as defined in Ref. [6] to characterise the car detector.

$$\textit{Correct detection rate (recall)} = \frac{\textit{Number of correct positives}}{\textit{Total number of cars in the data set}}.$$

Table 2
Performance of our ISS subspace car detector

Activation threshold	No. of correct detections, N	Recall $\frac{N}{200}$	No. of false detections, M	Precision $\frac{N}{N+M}$	False detection rate, $\frac{M}{166856}$
0.15	188	94.0%	45	80.6%	0.027%
0.25	184	92.0%	39	82.5%	0.023%
0.35	182	91.0%	34	84.3%	0.020%
0.45	180	90.0%	30	85.7%	0.018%
0.55	179	89.5%	25	87.7%	0.015%
0.65	174	87.0%	23	88.3%	0.014%

Table 3
Performance of the car detector in Ref. [6]

Activation threshold	No. of correct detections, N	Recall $\frac{N}{200}$	No. of false detections, M	Precision $\frac{N}{N+M}$	False detection rate, $\frac{M}{112000}$
0.55	181	90.5%	98	64.9%	0.09%
0.65	178	89.0%	92	65.9%	0.08%
0.75	171	85.5%	76	69.2%	0.07%
0.85	162	81.0%	48	77.1%	0.04%
0.90	154	77.0%	36	81.1%	0.03%
0.95	140	70.0%	29	82.8%	0.03%

$$\textit{False detection rate} = \frac{\textit{Number of false positives}}{\textit{Total number of negatives in the data set}}.$$

$$\textit{Precision} = \frac{\textit{Number of correct positives}}{\textit{Number of correct positives + Number of false positives}}.$$

The ideal detector should be of 100% correct detection rate, 0% false detection rate and 100% precision. We show our car detector performance in Table 2, and show the performance of UIUC car detector in Table 3. We evaluated more windows than Ref. [6], this may be one of the reasons that we have achieved better performances based on these measures. It is also to be noted that our purpose is not to build the best car detector as such but rather we evaluate the usefulness of our ISS subspace method for building good classifier. These results further demonstrate that ISS is an effective feature vector representation subspace method for classification. Examples of our car detection results are shown in Fig. 14.

5. Discussion and concluding remarks

A very popular class of nonlinear subspace method appears in recent pattern recognition literature is the kernel subspace method, see for example Refs. [8–11]. The underpinning theory of kernel subspace method is the VC (Vapnik–Chervonenkis) theory [28] which tells us that often mappings which take us into a higher dimensional space than the dimension of the input space provide us with greater classification power. However, the problem of high dimensionality is that this can seriously increase computation time. The kernel methods use the so-called "kernel trick" to take advantage of high-dimensional representations without actually having to work in the high-dimensional space. For example, kernel PCA [9] uses a Mercer kernel to nonlinearly map the input data onto a higher dimensional space and then perform PCA (using the kernel trick allows computation to be done in terms of dot products without having to do the computation in the high-dimensional space). It is important to distinguish linear PCA and kernel PCA. Suppose that the number of observations m exceeds the input dimensionality n; in linear PCA, we can find most n nonzero eigenvalues; however, kernel PCA can find up to m nonzero eigenvalues. Thus, this is not necessarily a dimensionality reduction method.

Put our present method in this context, the ISS is actually a linear subspace method. Unlike kernel subspace methods that first map the input onto a high-dimensional space and then perform feature extraction; our ISS method is similar to other linear subspace methods which extract linear features in the input space. The difference between our method and other methods such as PCA is that we use a maximum mutual information criterion to extract the projection directions. Since the objective of feature extraction is for classification, as explained in section 2, Fano's inequality, Eq. (2) tells us that features that maximize the mutual information will minimize the probability of misclassification error, which justifies and motivates this work to use maximum mutual information as a feature selection criterion and to seek a computationally practical solution.

One of the reasons that information theoretic methods such as mutual information (MI) have not been more widely used in real world applications is because of the exponential computational complexity in the estimation of MI. Our contribution in this paper has successfully introduced an MI based subspace method for representing high dimensional features in low-dimensional subspaces for classification applications.

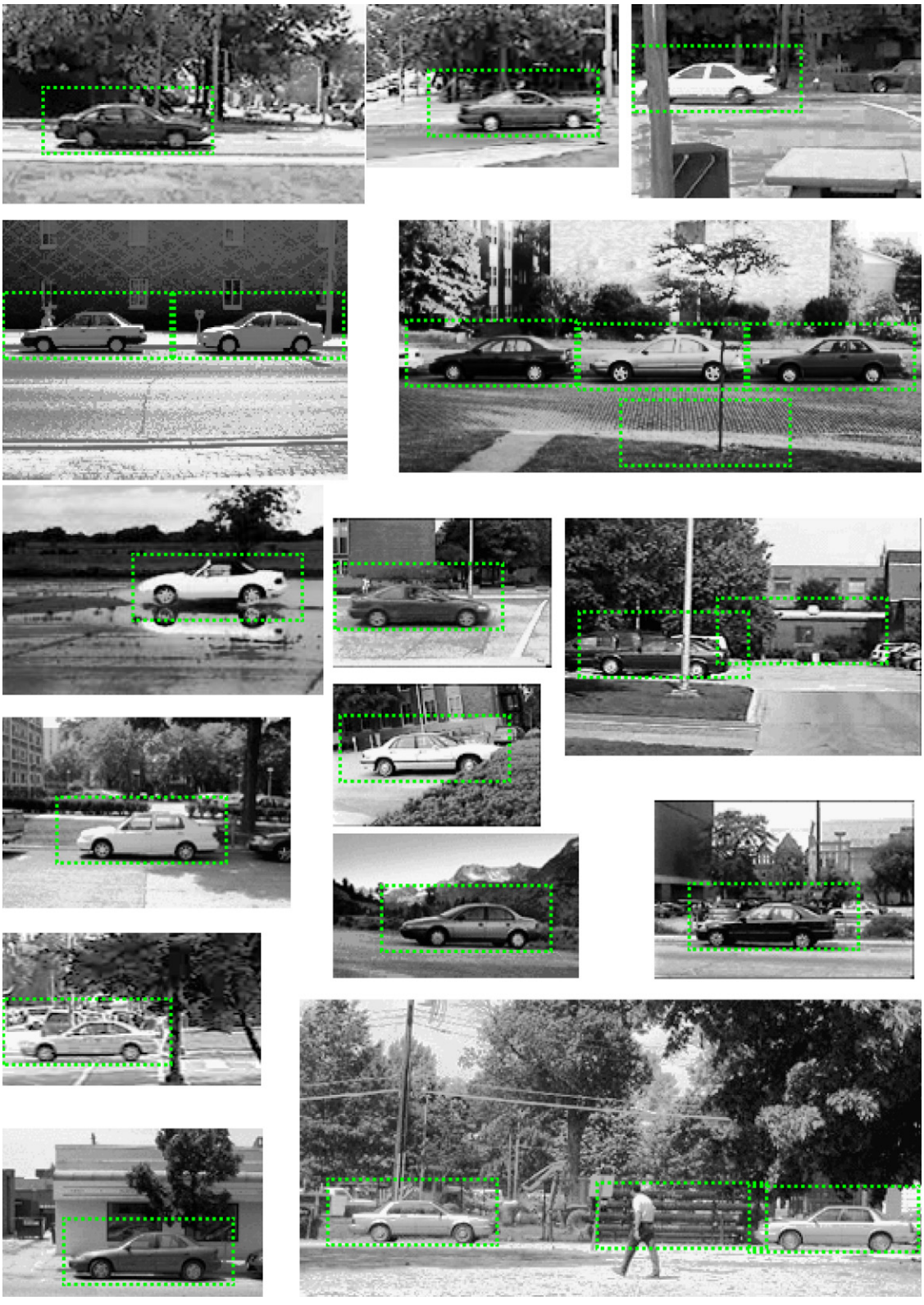

Fig. 14. Examples of our car detection results.

Our method is a generic one and can be applied to different application domains. Our solution is a pragmatic method for employing information theoretic criterion to find a compact and discriminative subspace for feature representation for classification. Empirical results have demonstrated the method is effective and useful. The drawbacks of our current solution are that it is a brute force technique. It is directly affected by the training samples and its computational costs are

proportional with sample size. However, as computers are getting more powerful, and it may also be possible to implement the ISS algorithm directly in hardware, the algorithm can be made computationally practicable. Our future work will investigate methods that can overcome these drawbacks. An interesting direction is to combine the random projection method [17,18] with our ISS technique.

References

[1] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed. Prentice-Hall, Englewood Cliffs, NJ.

[2] W. Zhao, Discriminant Component Analysis for Face Recognition, International Conference on Pattern Recognition, 2000.

[3] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. (1997) 711–720.

[4] M.H. Yang, N. Ahuja, D. Kriegman, Face detection using mixtures of linear subspaces, IEEE Conference on Automatic Face and Gesture Recognition 2000, pp. 70–76.

[5] M.-H. Yang, D. Kriegman, N. Ahuja, Detecting face in images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.

[6] S. Agarwal, D. Roth, Learning a sparse representation for object detection, in: Proceedings of 2002 ECCV2002, pp. 113–130.

[7] M. Vidal-Naquet, S. Ullman, Object recognition with informative features and linear classification, ICCV 2003, Nice, France.

[8] S. Fidler, D. Skocaj, A. Leonardis, Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling, IEEE Trans. Pattern Anal. Mach. Intell. 28 (3) (2006) 337–350.

[9] B. Scholkopf, A. Smola, K.-R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.

[10] P. Zhang, J. Peng, C. Domeniconi, Kernel pooled local subspaces for classification, IEEE Trans. Systems, Man Cybern. Part B: Cybernetics. Special Issue on Learning in Computer Vision and Pattern Recognition 35 (3) (2005) 489–502.

[11] S.-W. Kim, B. John Oommen, On using prototype reduction schemes and classifier fusion strategies to optimize kernel-based nonlinear subspace methods, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 455–460.

[12] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[13] J. Fang, G. Qiu, Learning an information theoretic transform for object detection in: A.C. Campilho, M.S. Kamel (Eds.), Image Analysis and Recognition: International Conference, ICIAR 2004, Porto, Portugal, September 29–October 1, 2004, Proceedings, Part I. Lecture Notes in Computer Science, vol. 3211, Springer, Berlin, 2004.

[14] R.M. Fano, Transmission of Information: A Statistical Theory of Communications, MIT Press, Cambridge, MA, 1961.

[15] J.W. Fisher III, J.C. Principe, A methodology for information theoretic feature extraction, World Congress on Computational Intelligence, March 1998.

[16] T. Butz, J.P. Thiran, Multi-modal signal processing: an information theoretical framework. Technical Report 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), 2002.

[17] D. Achlioptas, Database friendly random projections, in: Symposium on Principles of Database Systems (PODS), 2001, pp. 274–281.

[18] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), August 26–29, 2001, San Francisco, CA, USA, pp. 245–250.

[19] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.

[20] M.-H. Yang, D. Kriegman, N. Ahuja, Detecting face in images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (1) (2002) 34–58.

[21] Stan Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, H. Shum, Statistical Learning of Multi-View Face Detection. ECCV 2002, Lecture Notes in Computer Science, vol. 2353, 2002, pp. 67–81.

[22] H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1) (1998) 23–38.

[23] P. Viola, M. Jones, Robust Real-time Object Detection, in: Proceedings of Second International Workshop on Statistical and Computational Theories of Vision—Modeling, Learning, Computing and Sampling, Vancouver, Canada, July 2001.

[24] H. Schneiderman, T. Kanade, A statistical method for 3d object detection applied to faces and cars, in: IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 746–751.

[25] M. Alvira, R. Rifkin, An empirical comparison of SnoW and SVMs for face detection, AI Memo 2001-004, CBCL Memo 193, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, January 2001.

[26] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, The FERET database and evaluation procedure for face recognition algorithms, Image Vision Comput. J 16 (5) (1998) 295–306.

[27] CMU website: ⟨http://www.cs.cmu.edu/∼har/faces.html⟩.

[28] V. Vapnik, The Nature of Statistical Learning Theory, Springer, Berlin, 1999.

About the Author—GUOPING QIU received the B.Sc. degree in Electronic Measurement and Instrumentation from the University of Electronic Science and Technology of China, Chengdu, China, in July 1984 and the Ph.D. degree in Electrical and Electronic Engineering from the University of Central Lancashire, Preston, England, in 1993. He is currently a Reader in the School of Computer Science at the University of Nottingham, UK. Since September 2006, he is also a Visiting Scholar in the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong. He has research interests in the broad area of computational visual information processing and has published widely in this area. His current research is focused on two main areas—making better images and graphics and managing large image repositories. More about his research can be found in URL: http://www.cs.nott.ac.uk/∼qiu.

About the Author—JIANZHONG FANG received the Ph.D. degree in Computer Science from the University of Nottingham, UK in 2004 and the Bachelor's degree in Computer Engineering from the National University of Defense Technology, Changsha, China, in 1989. He is currently a freelance consultant and developer in visual information engineering. From 1989 to 1997, he was engaged in research and development of Central Processing Unit of mainframe computers at the East China Research Institute of Computer Technology, Shanghai, China. From 1997 to 2001, he worked in R & D department in Printroxin in Singapore.