Learning an Information Theoretic Transform for Object Detection

Jianzhong Fang and Guoping Qiu

School of Computer Science, The University of Nottingham
{jzf, qiu} @cs.nott.ac.uk

Abstract. We present an information theoretic approach for learning a linear dimension reduction transform for object classification. The theoretic guidance of the approach is that the transform should minimize the classification error, which, according to Fano's optimal classification bound, amounts to maximizing the mutual information between the object class and the transformed feature. We propose a three-stage learning process. First, we use a support vector machine to select a subset of the training samples that are near the class boundaries. Second, we search this subset for the most informative samples to be used as the initial transform bases. Third, we use hill-climbing to refine these initial bases one at a time to maximize the mutual information between the transform coefficients and the object class distribution. We have applied the technique to face detection and we present encouraging results.

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 objects are well separated.

Classical techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) [7] are well studied in the literature. Although PCA can produce compact representation, it cannot enhance the discriminative power. Since LDA only makes use of covariance, it is only optimal for classes having unimodal Gaussian density with well-separated means. In many applications, it may be beneficial to exploit higher than second order statistical information.

Theoretically, information theoretic approaches [8] have a number of advantages. For example, mutual information measures general statistical dependence between variables rather than the linear correlation. The mutual information is also invariant to monotonic transformations performed on the variables.

In this paper, we present a learning procedure for developing a dimension reduction linear transform based on the mutual information criterion, and apply it to object detection. The organization of the paper is as follows. Section 2 gives a brief back-

ground overview on the Shannon information theory and Fano's inequality on the relationship between mutual information and a lower bound of misclassification error [2]. Section 3 describes a 3-step learning procedure for deriving a mutual information maximizing linear dimension reduction transform. Section 4 presents experiments and results of applying the method to human face detection. Section 5 concludes the paper.

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. We can define the following Shannon information theory functions

The entropy of X is defined as

$$H(X) = -\sum_{x \in A_x} P(x) \log(P(x)) \tag{1}$$

The joint entropy of X and Y is defined as

$$H(X,Y) = -\sum_{x \in A_x} \sum_{y \in A_y} P(x,y) \log(P(x,y)) \tag{2}$$

The mutual information between X and Y can be defined as (other forms of definition also exist)

$$I(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{3}$$

The entropy measures the information content or uncertainty of the random variable. 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.

2.1 Fano's Mutual Information Bound

In the context of object classification, Fano's inequality [2] 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 is used by other authors [3, 4]. The classification process can be interpreted as a Markov chain as illustrated in Fig. 1.

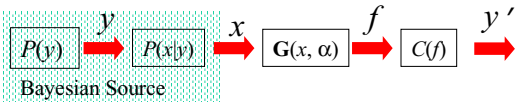

Fig. 1. Interpreting the classification process as a Markov chain [3, 4], y is the object class random variable, x are the observations generated by the conditional probability density function $P(x \mid y)$. The observations are subjected to a transform G, which produces a new feature f from input x. The classifier C then estimates the class identity of input x as y' based on the transformed feature f.

The probability of misclassification error in the setting of Fig. 1, $P_e = \mathrm{P}(\mathrm{y} \neq \mathrm{y}')$, has the following bound [2]

$$P(y \neq y') \geq \frac{H(Y) - I(Y,F) - 1}{\log(m)} \tag{4}$$

where 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. (4) 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 minimizes the lower bound to ensure an appropriately designed classification algorithm does well. Since both m and $H(Y)$ are constants in (4), 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 G that minimizes this lower bound. In the next section, we propose a three-stage solution.

3. Learning a Linear Informative Transform

Our objective is to find a dimension reduction linear transform G that minimizes the lower bound in (4). Because the observations x and the transformed feature f and class variable y are all normally multidimensional vectors, directly estimating an optimal G that maximizes $I(Y, F)$ is computationally extremely difficult. Assume x is an l-d column vector and f is a k-d column vector, ($k << l$), then $f = Gx$, G is a k (rows) by l (columns) transform matrix. In this section, we present an engineering solution to estimate G, which we will demonstrate in the next section it works satisfactorily.

The developed 3-stage process is as follows:

Stage 1: Use a support vector machine (SVM) [7] to select the training samples that are near the class boundaries

Stage 2: Use a maximum mutual information criterion to select an initial set of transform base vectors.

Stage 3: Use a hill-climbing algorithm to refine the transform base vectors one at a time.

Starting from labeled training samples, we use the raw data to train a support vector machine to classify the samples directly. After training, the "support vectors" of this SVM are those samples near the decision surfaces. Conceptually, this set of training samples (support vectors) are the most difficult to classify. We reason this set of samples is potentially the most "informative" because there is more uncertainty in relation to the class identity. We will therefore only use this set of samples to find the transform G. The rationale is that by reducing the number of samples, we can reduce the training time; also, if we can separate those samples near the decision boundaries, the rest of the samples can be separated easily. It is to be noted that the only use of the SVM at this stage is to select samples near the class boundaries.

Once the training samples have been selected, we use a constructive procedure to build an initial set of transform base vectors based on the maximum mutual information criterion. Let $X = \{x_1, x_2, \ldots x_N\}$ be the N labeled training samples that constitute the support vectors of the SVM, $Y = \{y_1, y_2, \ldots y_N\}$ their corresponding class labels, $G = [g_1, g_2, \ldots, g_k]^T$, and g_i be the ith transform base. We use the following procedure to find the initial value of G

Proc. Initialize$G(X,Y)$
  for $i = 1$ **to** k **do**
    for $m = 1$ **to** N **do**
      for $n = 1$ **to** N **do**
        $F(m, n) = \langle x_m, x_n \rangle$ //inner product
      End for
      Proc. Estimate joint probability $P(Y, F(m, \bullet))$
      Proc. Compute $I(m) = I(Y, F(m, \bullet))$
                //Compute mutual information (3)
    End for
    If $I(j) > I(m), \forall m$ **Then** $g_i = x_j / \|x_j\|$
    Remove x_j and y_j from X and Y respectively
    $N = N - 1$
    for $m = 1$ **to** N **do**
      $x_m = x_m - \langle x_m, g_i \rangle g_i$
    End for
  End for
End Proc.

To find the first transform base, we select one sample at a time, and project all other training samples onto that selected sample. The projection (a scalar) and the sample identity can be used to estimate the joint probability, which in turn can be used to estimate the mutual information of the projection and the class distribution. The sample with projection output 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 sam-

ples 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 initial bases are orthonormal.

It is clear that the bases are selected individually based on a maximum mutual information criterion. Ideally, these bases should be optimized jointly. However, estimating the joint probability of high dimensional vectors is computationally prohibitive. The mutual information function with respect to the transform G is non-differentiable. This makes a closed form optimization algorithm difficult to derive (if not impossible). Therefore some form of heuristic techniques have to be employed to refine the initial transform bases. We decided to use hill-climbing [10] to accomplish the task.

This is an iterative process. We refine the bases, g_i, one at a time. For each hill-climbing step, the criterion is the maximization of the mutual information between the projection of the training samples onto that base and the samples class distribution. Starting from the 1st base, hill climbing is used to refine the base such that the mutual information between the projections of the original data sample onto this base and the sample's class distribution reaches the highest possible value. Once a local maximum is reached, this base is normalized and fixed. All training samples are made orthogonal to the new base to form the new training samples to be used to refine the next base. The final transform bases all will have been made to have a unit length but are not necessarily orthogonal to each other.

The process first finds a single base, onto which the projection of the original signal will produce a scalar, whose distribution and the object class has maximum mutual information. We then make the signal orthogonal to the first base to form the residue signal. From this residue signal, we attempt to find another transform base that will maximize the mutual information between the projection of this residue signal onto the second base and the class distribution. This process repeats until a fixed number of bases are created.

An intuitive understanding of the method can be thought of as follows: from the original sample, we find a direction that conveys the maximum information about the object class distribution. We then remove what is already known by making the signal orthogonal to this base to form the residue signal. We then find a direction in the residue signal space that conveys the maximum information about the object classes. What is already known about this base is again removed by making the signal orthogonal to this base and the process continues.

Because each base (except the 1st one) is trained on the residue signals from the transform of the previous base, the transform should reflect this and the new maximum mutual information linear dimension reduction transform is illustrated in Fig. 2 as a neural network style diagram.

4. Experiments

In this section, we use human face detection [5] as an application example of the approach developed in section 3. We first collected 9390 face and nonface samples from

various sources (the numbers of face/nonface samples are roughly 1 : 2). The original samples are of various sizes and we normalize them to a uniform size of 32 x 32 pixels. We first use these 1024d vectors to train a support vector machine [1], when it is converged, there are 2217 support vectors, of which 986 are face and 1231 are non face samples. Using these support vector samples, we then follow the procedure in section 3 to develop the transform bases. Fig. 3 shows examples of 16 such bases.

In the following experiments, 64 transform bases are used, that is the input vector of 1024d is reduced to 64d for detection (16 : 1 compression). To search for faces in images, we use detection windows of 30 different sizes ranging from the smallest of 20 x 24 pixels to the largest of 426 x 520 pixels. For each of these windows, it is first resized to 32 x 32, then the 32 x 32 window is passed to the transform to reduce its dimension to 64. The 64 dimensional vector is then passed to a support vector machine, which has been trained to determine whether the current window is a face [1].

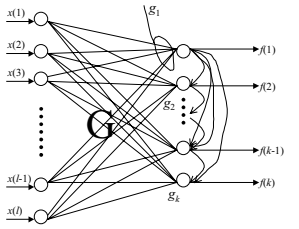

Fig. 2. (left) Schematic of the maximum mutual information dimension reduction linear transform. See eq. (5)**. Fig. 3.** (right) Examples of maximum mutual information linear dimension reduction transform bases for face/nonface objects

The transform output is defined as (5). In the next section, we use human face detection as an application example to evaluate the validity of the approach.

$$\begin{aligned} f(1) &= g_1 x \\ f(2) &= g_2(x - f(1)g_1) \\ &\vdots \\ f(i) &= g_i(x - f(i-1)g_{i-1} - f(i-2)g_{i-2} - \ldots - f(1)g_1) \\ &\vdots \\ f(k) &= g_k(x - f(k-1)g_{k-1} - f(k-2)g_{k-2} - \ldots - f(1)g_1) \end{aligned} \tag{5}$$

In the first experiment, we tested all 500 upright frontal view face images under the "/inp" directory in the FERET data set [9]. The detector correctly detected 495 faces from the data set achieving a detect rate of 99%. There were 22 false positives with a universal threshold. The testing result is comparable to recent work using this data, e.g. [11]. Fig. 4 shows some examples of detection results.

In a second experiment, we use 130 photographs from the CMU website [6]. This is the set of images used extensively by researchers in face detection. The 130 images contain 507 faces. For the 130 images, our experiment evaluated a total of 52,129,308 patterns. The receiver operating characteristic (ROC) curve of the detection is shown in Fig. 5. Detection examples are shown in Fig. 6. These results are quite good and are comparable to state of the art. This demonstrates that the new method is effective and its potential is very encouraging.

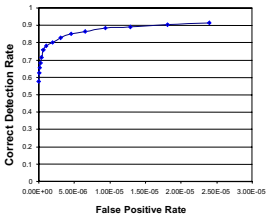

Fig. 4 (left). Examples of experimental results on the FERET database. **Fig. 5 (right).** Receiver operating characteristics of face detection using maximum mutual information dimension reduction linear transform (data representation) and support vector machine (decision making). 130 testing images with 507 faces and 52,129,308 evaluated patterns.

We have also performed some initial comparisons to other transform techniques, in particular principal component analysis (PCA). We found that at lower dimension (high compression), the new informative transform clearly outperforms PCA, the advantage of the new method is less pronounced when very high dimensions are used. Fig. 7 shows 3D plots of 1500 face samples and 1500 nonface samples in the first 3 dimensions of PCA and the new transform space. It is seen that the face/nonface patterns are better separated in the new maximum mutual information transform space. It is also seen that samples belonging to the same class are closer to each other in the new transform space.

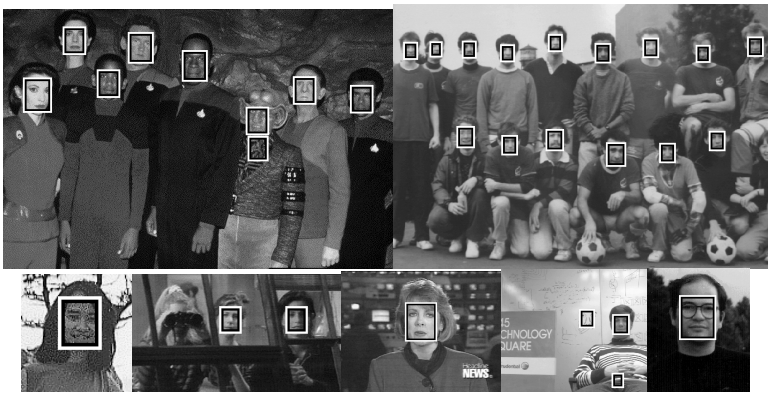

Fig. 6. Examples of face detection result performed on the CMU database

5. Concluding remarks

In this paper, we have presented a learning procedure to create a linear dimension reduction transform based on an information theoretic criterion. We have successfully applied the transform to face detection. Our initial results indicate that the new technique is very effective. Information theoretic approaches have many advantages com-

pared to other conventional methods. Our work here and recent work by others, e.g. [12], have clearly demonstrated the potential of information theoretic approaches to computer vision and pattern recognition problems.

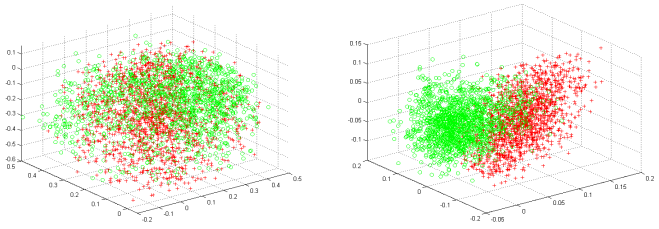

Fig. 7. 3D plots of 1500 face (green o) and 1500 nonface (red +) patterns in the first 3 dimensions of the PCA space (left) and the new maximum mutual information transform space (right).

References

1. E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: an application to face detection”. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Pages 130-136, 1997.
2. R. M. Fano, Transmission of Information: A Statistical Theory of Communications, MIT Press, Cambridge, MA, 1961.
3. J. W. Fisher III and J. C. Principe, “A methodology for information theoretic feature extraction”, World Cogress on Computational Intelligence, March 1998
4. T. Butz, J. P. Thiran, "Multi-modal signal processing: an information theoretical framework". Tech. Rep. 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Tech- nology (EPFL), 2002.
5. M-H Yang, D. Kriegman and N. Ahuja. "Detecting face in images: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, January, 2002.
6. CMU website: http://www.cs.cmu.edu/~har/faces.html
7. S. Haykin, Neural Networks: A Comprehensive Foundation (2nd Edition) , Englewood Cliffs, NJ: Prentice-Hall
8. T. M. Cover and J. A Thomas, Elements of Information Theory, Wiley, 1991
9. P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The feret database and evaluation procedure for face-recognition algorithms", Image and Vision Computing, 16(5):295-306, 1998. 27
10. D. H. Ackley. A Connectionist Machine for Genetic Hillclimbing. Boston: Kluwer Academic Publishers, 1987.
11. C. Liu, "A Bayesian discriminating features method for face detection". IEEE Transaction on PAMI, Vol. 25, No. 6, June 2003
12. M. Vidal-Naquet, S. Ullman, "Object recognition with informative features and linear classification", ICCV 2003, Nice, France