FEATURE SELECTION ON COMBINATIONS FOR EFFICIENT LEARNING FROM IMAGES

Rong Xiao, Lei Zhang, and Hong-Jiang Zhang

Microsoft Research Asia, Beijing 100080, P.R. China
{t-rxiao, leizhang, hjzhang}@microsoft.com

ABSTRACT

Due to the high computation complexity and intra-class variance in the area of image pattern recognition, feature extraction for image pattern recognition has been the focus of interest for quite some time. In this paper, a novel feature extraction framework is presented, which first constructs an over-complete feature combination set, and then selects effective combinations by using feature selection algorithm. Experimental results show that this structure can do pattern recognition on images more efficiently in both accuracy and speed.

1. INTRODUCTION

Pattern recognition on images has been studied for many years. Although lots of works have been published to improve performance in this area, there is still a long distance from human's vision skill in both speed and accuracy. To make further investigation, we first take a close look at the two essential ingredients for building a good recognition system.

The first one is feature. Single pixel carries little information of given patterns, and has large intra-class variance. It is very hard to recognize patterns from images on those features. A variety of algorithms have been proposed to transform raw features, such as PCA, LDA, wavelets, ISOMAP, kernel PCA and etc. Based on these features, successive classification task is greatly simplified. Turk. M[11] used Eigenface to do face recognition. C. Papageorgiou et al. [2] developed a system to detect pedestrians in still images that uses SVM and 1326 wavelet features.

Another one is learning algorithm. Due to the lighting, pose and background variations, the pattern intra-class structure turns to be very complicated. To address this problem, generalization ability and capacity of the learning algorithm should be especially considered. Therefore, large margin algorithms, such as SVM, boosting and etc. are widely used in recent years.

Moreover, as the computation over the whole image is always time consuming, the evaluation speed of learning algorithm is very important. Boosting algorithm and cascade structure began popular. In these framework, weak classifiers based on decision stump, linear method and etc., can be easily applied to improve evaluation speed. But, due to the poor learning capacity of these weak classifiers, how to find the effective feature set is critical.

Generally, optimal feature set is very hard to be determined manually by prior-knowledge. An over-completed feature set together with a feature selection algorithm is widely used for this problem. For example, P. Viola [9] use a boosting cascade to build a face detector based on 45891 haar-like features and each haar-like feature can be a linear combination of image pixels. According to Viola's experiment, the classifier with only two combinations can be adjusted to detect 100% of the faces with a false alarm rate of 40%. Compared with raw pixel feature, these feature combinations carry more class information and tend to be more robust.

Motivated by viola's approach, we extended the idea of feature combination to more general area. The remainder of this article is organized as follows; section 2 presents the framework of feature combinations for classification, section 3 presents some experimental results on different pattern recognition problems on images, section 4 concludes the article.

2. FEATURE COMBINATIONS FOR PATTERN RECOGNITION

2.1 Feature Combinations

In pattern recognition problem, suppose we are given a training data set X : $x_i \in R^n$, $i = 1,...,\ell$ with corresponding label set Y : $y_i \in \{\pm 1\}$ $i = 1,...,\ell$.

Definition 2.1 (feature combination set) Functions $\phi(x) : R^n \to R$ could be regard as one feature combination of sample x. A set of such function $\phi_i(x)$, $i = 1,..., p$ could form the feature combination set $\{\phi_1(x),...,\phi_p(x)\}$, and map the original training set $X \subset R^n$ to $X' \subset R^p$ in a high-dimensional space F. Furthermore, a feature combination set is extended, when the original feature set is a subset of the combination set.

Definition 2.2 (*r*th-combination set) If each function ϕ_i only depends on $r\ (r \le n)$ features at most, then the extended combination set is a *r*th-combination set

Definition 2.3 (*r*th-poly-combination) When ϕ is a *r*th-degree polynomial function, the corresponding *r*th-combination is *r*th-poly-combination. For example, the complete *2*nd-poly-combination set of sample x_i, could be listed as follows,

$$x_{i,1}, x_{i,2},..., x_{i,n}, x_{i,1}^2,..., x_{i,n}^2, x_{i,1}x_{i,2},..., x_{i,n-1}x_{i,n} \quad (1)$$

where $x_{i,j}$ is the *j*th-feature of sample x_i.

Theorem 2.4 (Cover Theorem on Linear Separability [10]) A classification problem is more likely to be linearly separable than in original low dimension space R^n, when it is cast into a high dimension space *F* with over-complete non-linear mapping. In the case of *r*th-combination, we called the dataset to be ϕ-separable, when

$$y[(\sum_{i=1}^{p} w_i\phi_i(x)) - b] > 0 \text{, where}$$
$$x \in X,\ \ y \in Y, w_i \in R, b \notin R \qquad (2)$$

In the special case of poly combinations, probability that particular dichotomy picked at random is $\varphi - separable$, and it will be

$$P(l, p) = (\tfrac{1}{2})^{l-1} \sum_{i=0}^{p-1} \binom{l-1}{i} \qquad (3)$$

Suppose the equation 2 is a specific definition of a non-linear decision function for separating the dataset, where ϕ_i is redefined as a non-linear function parameterized by *i*th feature. A similarly conclusion can be drawn from Cover's theorem.

Corollary 2.5 (Non-linear Separability) A pattern classification casting in linear *r*th-combination set is more likely to be non-linearly separable by equation 2 than in the original space R^n.

Corollary 2.6 Linear *r*th-combination set will have the same linear separability as it in the original space R^n.

Proof Suppose the Linear transformation matrix for *r*th-combination set is $A \in R^{p\times n}$, each sample $x \in R^n$ will be mirrored to $x' = Ax$ in the space F. According to the definition of rth-combination, $Rank(A) = n$.

Suppose we have the linear decision function in original space, which is $y_i(w^T x_i - b) > 0$, according to the existence of solutions theorem to systems of linear equations, we can always find w', which satisfied the equation $w'^T A = w^T$. Therefore the dataset is linear separable in space F by function $y_i(w'^T x_i' - b) > 0$.

Suppose we have the linear decision function in space F by function $y_i(w'^T x_i' - b) > 0$. Then obviously, according to equation $x' = Ax$, the data set is also linear separable in the original space.

Consequently, the equivalent is proved.

To conclude, in most cases, the more feature combinations are constructed the more discriminating ability will be provided. But with a finite training set, a high-dimensional feature space is almost empty and many separators may perform well on the training data [7]. Moreover, as $n << p$ in the most time, the feature combination space F is high redundant. Some feature selection algorithm should be employed to enhance the performance of successive learning procedure. Hereby a new framework of 3-step feature combination and selection is proposed.

2.2 Optimal Feature Combinations Generating

The goal of our 3-step feature combination and selection is to find the optimal subset of feature combinations to enhance the efficiency of successive learning procedure. As it is shown in fig. 1, the proposed system is constructed by 3 key steps.

The first one is irrelevance filter. Since the new features (combinations) generated in the following step is very large, it is necessary to remove initial irrelevant features in the earlier steps. Relief algorithm [5] is a prominent method in this area. It assigns a relevance value to each feature according to the feature difference between the sample and the nearest "hit" (another sample of the same class) and nearest "miss" (another sample of the same class). We therefore could set a threshold for the relevance values to divide the feature set into relevant and irrelevant features. If the training set is already highly redundant, K-means algorithm can be used to further reduce the feature set.

The second one is feature combination generator. It tries to construct more relevant features by generating an over-complete set of redundant features. The resulting set is always very large, for example, a full *2^nd^*-poly-combination set in R^n will has the dimension of $n(n+3)/2$. Therefore, high order feature combinations are seldom applicable without prior-knowledge.

The last one is optimal combinations selector. The selection algorithm is quite specific here, due to the following characteristics: the scale is very large in both features and samples, the feature set is highly redundant and continuous valued. Therefore some heuristic method should be used in this procedure, for example, RFE [3], boosting, SFFS (Sequential Floating Forward Selection) and SFBS (Sequential Floating Backward Selection [8].

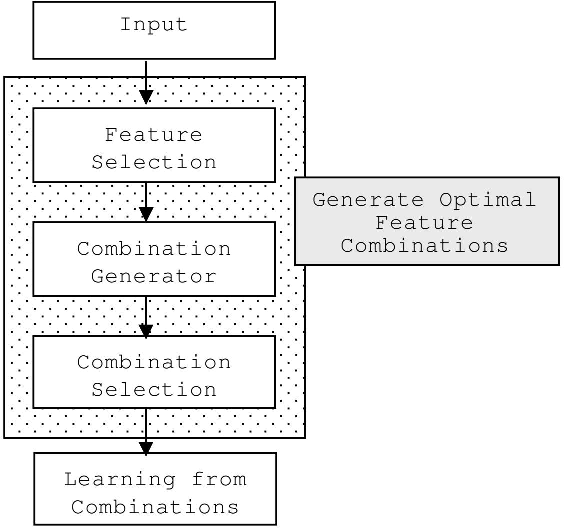

Figure 1. 3-Step Feature Combination and Selection

The optimal feature combination set is always relative to the successive learning procedure. Based on the L/N (linear/non-linear) properties of feature combinations and learning procedure, systems based on our framework can be divided into 4 categories, N+L, N+N, L+N, and L+L. As the last one is useless according to the theorem of separability, we test the remaining solutions in next section.

3. EXPERIMENT RESULTS

3.1 Non-Linear combinations for Linear SVM Classification

With the using of kernel tricks, Vapnik extended SVM to non-linear classification and made SVM more applicable to most pattern recognition problems. The cost for this extension, however, is that the final decision function is only obtained as kernel expansions which are parameterized by SVs (Support Vectors). Therefore, the computation complexity increases dramatically, when the SV set is very large. A few algorithms have been proposed to solve this problem, for example [1] introduced reduced set method to compact the size of SV set. Here, based on the idea of feature combinations, another solution is given.

For a SVM classifier with 2^{nd}-degree polynomial kernel $K(x_i, x_j) = (1 + x_i \cdot x_j)^2$, the feature space F with dimension $p = n(n+3)/2$ is given by the full 2^{nd}-poly-combination set,

$$x^* = (x_1, x_2, ..., x_n, x_1^2, x_2^2, ..., x_n^2, x_1x_2, x_1x_3, ..., x_{n-1}x_n) \quad (4)$$

An optimal hyper-plane could be found in the feature space F to separate training data (x_i, y_i) , where $x_i \in R^n$ and $y_i \in \{-1,1\}$. The decision function $f(x)$ should be,

$$f(x) = w \cdot x^* + b = \sum_{x_i \in SV} a_i y_i K(x_i, x) + b$$

where $w \in F$. (5)

However the dimensionality of feature space *F* used to be very high or infinite. Evaluating equation (5) will be time consuming. In this paper, feature selection algorithms it adopted for polynomial kernel. By this way the penalty of kernel expansions is avoided.

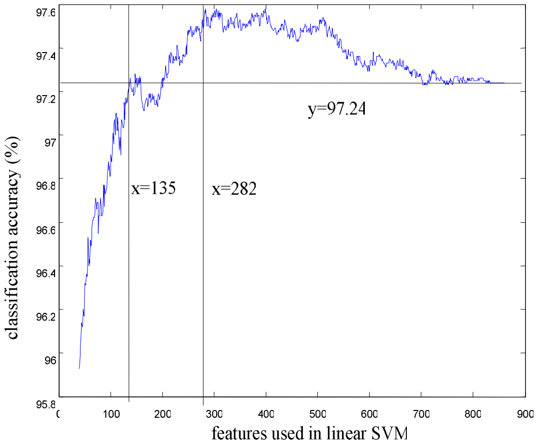

Figure 2. Optimal feature combination set selection. The x axis is the feature number we used in each step of RFE feature selection, and the y axis is the linear SVM classification accuracy.

To evaluate our method, about 6000 grayscale face examples are collected from various sources, covering the out-of-plane rotation in the range of $[-20^0, +20^0]$. They are roughly aligned by eyes and mouth. For each aligned face example, a synthesized face example is generated by a random in-plane-rotation in the range of $[-15^0, +15^0]$.

This creates a training set of 12,000 (8000 for training, 4000 for testing) face examples, which is cropped and re-scaled to the size of 20x20. Also the same number non-face examples for training set and testing set are collected from 100,000 images containing no faces. An over-complete set of Haar-like features was extracted on this training set. We first use Relief and a variant of K-means algorithm [4] to reduce the original feature set from 45891 features to 40 features, and then expand the feature set to the full *2^nd^*-poly-combination set with dimensionality 860. After that, RFE was used to reduce the feature set gradually.

Fig. 2 shows the performance of linear SVM classifier on each step of feature selection. We can find that most features x_k^* are irrelevant or redundant. Actually when we removed feature gradually, the classification performance keeps being improved until 2/3 features was removed, and only 15.6% features is required for the linear SVM classifier with the same accuracy as the 2^{nd}-polynomial SVM classifier.

3.2 Non-Linear Combinations for Boosting Classification

Based on above training set, the approach of N+N was evaluated. First we use adaboost selection 40 features from the 45000 Haar-like features [9]. Then we using 2^{nd}-polynomial feature combinations to expand the feature set to 860 features. At last we train another adaboost classifier on the new training set with 40 features.

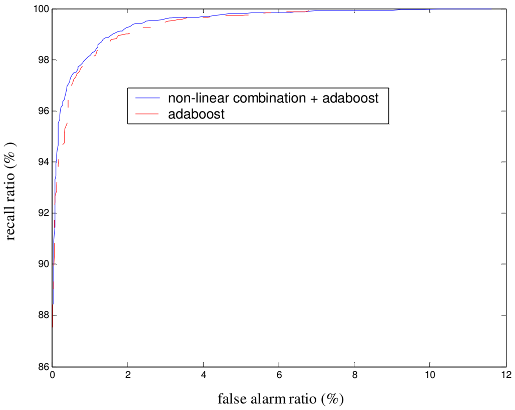

Figure 3. ROC curves for different approaches

In fig. 3, we compare the ROC curves of these two boosting model. It shows that the performance of the non-linear combination is almost the same as the performance of the pure boosting model.

3.3 Linear Combinations for Image orientation detection

Automatic detection of image orientation is a very difficult but important task in a digital image management system. Humans identify the correct orientation of an image through the contextual information or object recognition. Unfortunately, present works of computer vision still cannot establish a sufficient and effective way to interpret high level structure of objects in real world images. Therefore, the only way to deal with this problem is to exploit the the low-level features from the images rather than the content of the images.

Based on the low-level features, we represent the image orientation detection problem as a four-class classification problem, i.e. given an image from a scanner or a digital camera, determine its correct orientation from among the four possible ones: ω_1 ⇔0°, ω_2 ⇔90°, ω_3 ⇔180°, and ω_4 ⇔270°.

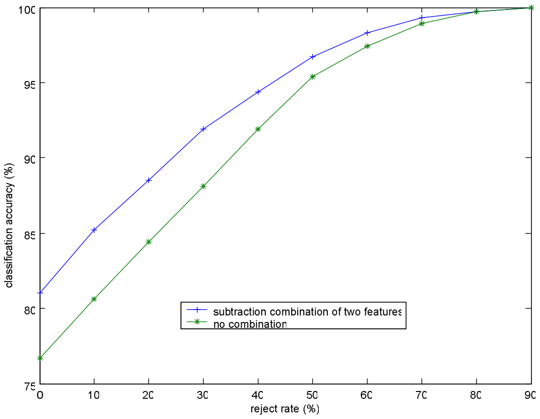

Figure 4. Comparison of feature combination with 2000 features in terms of classification accuracy vs. reject rate.

For each image, as described in [6], we extract 475 low level features, which consists 150 color moment (CM) features and 325 edge directional histogram (EDH) features. However, based only on these features, the performance of AdaBoost algorithm is not satisfactory in our experiment.

Motivated by the fact that the positive and negative samples are extracted from images in different orientations, e.g. 0° is for positive samples, 90°, 180° and 270° are for negative samples, we use the subtraction operation to combine two features together and form a new feature. Thus, if an image is rotated, the value of combined feature must change. In order to avoid the combination explosion, we only combine two features which have the same meaning. The total number of combined features is 6175.

We use the training data from the Corel photo gallery and the number of training examples is 5,416, and the test set size is 5,422.

From fig. 4, we can see clearly that by combining features with subtraction operation, the accuracy with rejection rate 0% increases from 76.7% to 81.0%. It shows that the feature combination is very effective. Note that the classification accuracy of SVM on this training set and testing set is 78.4%.

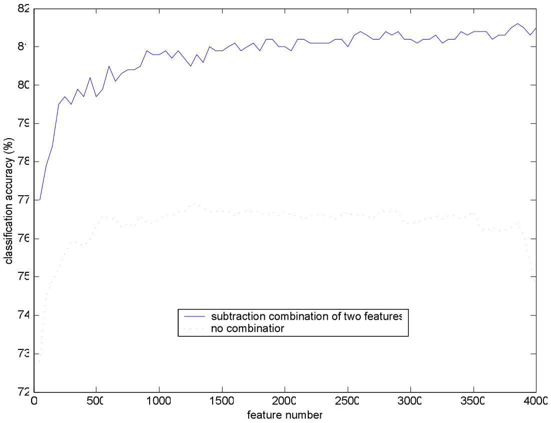

Figure 5. Comparison of feature combination in terms of classification accuracy vs. feature number when reject rate is 0%.

From fig. 5, we can see that feature combination plays an import role in boosting the performance. By feature combination, the more features selected by AdaBoost, the better the classification accuracy we can obtain. If we do not adopt any feature combination, i.e. training on the original 475 dimensional features, the accuracy is lower and the over fitting phenomenon appears when the feature number is more than 3500. Again it shows the effectiveness of the feature combination.

4. SUMMARY AND CONCLUSIONS

The feature combination framework presented in this paper targets effective pattern recognition on images. It is constructed by 3 steps. The first step is based on Relief and K-means algorithm, which filters out redundant and irrelevant features. The second step is a feature combination generator, which provides more relevant feature for further feature selection. The last step selects the final optimal subset of feature combinations. We recommend boosting and RFE algorithms, depending on the situation. The linear feature combination is best for boosting classifier. And RFE algorithm is more suitable for non-linear feature combination.

Combined with linear SVM and boosting algorithm, our system achieves pretty good performance in both accuracy and speed. Further improvement will be focused on the feature selection algorithms in step 1 and 3.

5. REFERENCES

[1] C. J. C. Burges. Simplified support vector decision rules. In L. Saitta, editor, Proc. 13th International Conference on Machine Learning, pp. 71-77, San Mateo, CA, 1996. Morgan Kaufmann.

[2] C. Papageorgiou, and T. Poggio, “Trainable Pedestrian Detection”, Proc. of International Conference on Image Processing, Kobe, Japan, Oct. 1999

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. BIOWulf Technical Report, 2000.

[4] J. Bins, "Feature Selection of Huge Feature Sets in the Context of Computer Vision", Ph. D. Dissertation, Computer Science Department, Colorado State University, 2000.

[5] I. Kononenko. “Estimation attributes: Analysis and extensions of relief”. In Proc. of the European Conference on Machine Learning, Catana, Italy, pp. 171--182. Springer Verlag, 1994.

[6] L. Zhang, M. J. Li, H. J. Zhang, Boosting image orientation detection with indoor vs. outdoor classification, Proc. IEEE Workshop on Applications of Computer Vision (WACV), 2002

[7] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, Machine Learning Proceedings of the Fifteenth International Conference(ICML '98), pp. 82--90, San Francisco, California, 1998. Morgan Kaufmann.

[8] P. Pudil, J. Novovicova and J. Kittler, “Floating Search Methods in Feature Selection”, Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994

[9] P. Viola and M.J. Jones, "Robust real-time object detection", ICCV Workshop on Statistical and Computation Theories of Vision, 2001

[10] Simon Haykin "Neural Networks: A Comprehensive Foundation", Prentice-Hall, pp. 257-260, 1999.

[11] Turk, M., Pentland, A. Face Recognition Using Eigenfaces. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1992.