BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 2
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0030

New Proposed Fusion between DCT for Feature Extraction and NSVC for Face Classification

B. Nassih, M. Ngadi, A. Amine, A. El-Attar

Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University of Kenitra, Morocco
E-mails: Nassih.bouchra@univ-ibntofail.ac.ma Ngadi.mohammed@univ-ibntofail.ac.ma
amine_aouatif@univ-ibntofail.ac.ma adnane.elattar@usmba.ac.ma

***Abstract*: *Feature extraction is an interactive and iterative analysis process of a large dataset of raw data in order to extract meaningful knowledge. In this article, we present a strong descriptor based on the Discrete Cosine Transform* (*DCT*)*, we show that the new DCT-based Neighboring Support Vector Classifier* (*DCT-NSVC*) *provides a better results compared to other algorithms for supervised classification. Experiments on our real dataset named BOSS, show that the accuracy of classification has reached 99%. The application of DCT-NSVC on MIT-CBCL dataset confirms the performance of the proposed approach.*

***Keywords*: *Supervised learning, DCT, NSVC, shape recognition, SVM, feature extraction.*

1. Introduction

Today, the increasing of databases' size poses an unprecedented challenge for data mining. The researchers realized that the selection of variables consists in choosing from a set of large variables a subset of most interesting variables to perform supervised classification. The objective of supervised classification is to build, using a learning set, a classification model that allows to predict the belonging of a new example to a class. In other words, the objective is to identify the classes to which objects belong on the basis of their descriptive variables.

In recent years, some researchers have explored the possibility of extracting features in the frequency domain using the Discrete Cosine Transform (DCT). The results showed that this technique is promising and allows to have discriminant features in the frequency domain.

In this paper, we have applied the new DCT-based Neighbouring Support Vector Classifier (DCT-NSVC) [4] method to build decision rules in order to construct a classification system. This algorithm uses a set of vicinal kernel functions constructed based on supervised clustering in the kernel-induced function space. Comparison with supervised classification methods shows that the system has been able to provide better classification.

This paper is organized as follow: we begin in Section 2 with a brief review of DCT methods. Section 3 presents a mathematical study of our proposed method NSVC. The results of our experiments on reference datasets are presented in Section 4. Section 5 concludes the paper and gives a brief vision for future work.

2. Discrete cosine transform

In this section, we propose a description of feature extraction approach based on the DCT method [1, 5]. In the beginning, the DCT transform is applied to convert the image into the frequency domain and a first dimensionality reduction is performed by the rejection of the high-frequency components.

The DCT is widely used in signal and image processing, especially in compression, it has indeed an excellent property of “regrouping” of the energy: The information is mainly carried by the low frequency coefficients.

The application of the DCT causes the information of the image of the spatial domain to pass into an identical representation in the frequency domain. Why is this change of domain so interesting? Precisely because a conventional image admits a great continuity between the values of the pixels. Since high frequencies are reserved for rapid changes in the intensity of the pixel, they are generally minimal in an image. Thus, it is possible to represent all of the information of the image on very few coefficients, corresponding to rather low frequencies, the continuous component (average value of the image processed) having a great importance for the eye.

The DCT transformed matrix being orthogonal, it is accompanied by a method of inversion to be able to return in the spatial domain. Thus, after making modifications in the frequency domain, eliminating variations of the image that are almost invisible by the human eye, we return to a representation in the form of pixels.

This core performs the mathematics for the DCT of an $M\times M$ image $I(x, y)$ algorithm as defined by the equation below:

$$c(u,v)=\frac{2}{M}.\alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{M-1}I(x,y).\cos\left[\frac{(2x+1)u\pi}{2M}\right]\cos\left[\frac{(2x+1)v\pi}{2M}\right]$$

$$\text{for}\ \ u,v=0,1,\dots,M-1,$$

$$f(x)=\begin{cases}\sqrt{0.5}\ \text{for}\ u=0,\\ \ 1\ \text{otherwise}.\end{cases}$$

The local information of the image can be obtained using the blocks of the DCT. The principle is the following: the image is divided into blocks of pixel size. Each block is represented by the coefficients of the DCT. From these, only those at the top left of the block are the most relevant and useful. The information needed to achieve high classification accuracy is contained in the first DCT (low frequency) coefficients by zigzag scanning.

3. Neighboring support vector classifier (NSVC)

The use of the Support Vector Machine (SVM) was limited as the hypothesis that the training data are identically generated from unknown probability distributions which is not the case of real life applications and problems.

To overcome this problem, the proposed method, Neighbouring Support Vector Classification (NSVC) [6, 7], uses a set of vicinal kernel functions built based on supervised clustering in the feature space induced by the kernel. This algorithm is faster, simpler to implement and requires a small memory space.

The proposed approach includes two steps:

- Supervised clustering step based on SKDA Algorithm (for Supervised Kernel-based Deterministic Annealing, used to partition the training data in different vicinal areas).
- A training step where the SVM technique is used to minimize the Vicinal Risk function (VRM) under the constraints defined in clustering step based on SKDA.

Consider the following input output data together:

(1) $$(x_i, y_i\ )_{i=1}^{l},\quad x_i \in R^n,\ y_i\ \in \{-1, 1\},$$

where *l* is the number of input data points, and *n* is the dimension of the input space.

The vicinity functions $v(x_i)$ of the x_i data points are built if test data points satisfy two assumptions:

- The unknown density function is smooth in the neighbourhood of each point x_i.
- The function minimizing the functional risk is also smooth and symmetric in the neighbourhood of each point x_i.

The optimization problem based on the principle of VRM named vicinal linear SVM [8, 9], can then be formulated as

(2) $$\text{minimize } \Phi(w) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i,$$

$$\text{subject to } y_i \int_{v(x_i)} \left([< x, w > +b] p(x|v(x_i)\ \ )\right) dx\ \geq\ 1\ -\ \xi_i,$$

$$\xi_i\ \geq\ 0,\ \ i\ =\ 1, \ldots, l,$$

where *w* is a weight, *C* is a punishment constant for x_i, *b* is the offset, $v(x_i)$ is the vicinity associated with the test point x_i, and $p(x|v(x_i)\ \ )$ is the conditional probability of the respective vicinity in the input space.

The following theorem for the vicinal SVM solution is true (see [8] for a proof):

(3) $$f(x) =\ \sum_{i=1}^{l} y_i \beta_i\, L(x, x_i) + b,$$

where to define the coefficients β_i one has to maximize

(4) $$W(\beta) =\ \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j y_i y_j\, M(x_i, x_j)$$

$$\text{subject to } \sum_{i=1}^{l} \beta_i y_i,\ \ \beta_i \geq 0,$$

where $L(x, x_i)$ is called the mono-vicinal kernel and $M(x_i, x_j)$ is the bi-vicinal kernel of the vicinal SVM.

3.1. Supervised kernel-based deterministic annealing for NSVC

The clustering of training data in the feature space is a well-documented subject [10, 11]. It consists of non-linearly mapping the observed data of an input low-dimensional space to a high dimensional feature space using a kernel function, which facilitates the separation of linear data.

Denoting a non-linear transformation of the input space X to a high-dimensional space using a kernel function as:

$$\Phi : R^n \to F,$$
$$x_i \to \Phi(x_i), \quad j = 1, \dots, l,$$

where $\Phi(x_i)$ is the transformed point x_i.

All training data points are distributed in c vicinities/clusters in the feature space, where $\Phi_k(z)$ is the center of mass of the k-th vicinity residing in F. This is a similar representation to clustering based on the characteristic space of k-means:

$$\Phi_k = \sum_{i=1}^{l} \alpha_{ki} z_i , \quad k = 1, 2, \dots, c, \tag{5}$$

where c is the number of clusters, α_{ki} are the parameters to be defined by the clustering technique (SKDA) and $z_i = y_i \Phi(x_i)$ denotes the data points labeled in the feature space.

The classification problem is usually defined mathematically by a cost function to be minimized, for NSVC case, this function is the distortion function. Similar to the notation used in [12], we let $p(\Phi_k|z_i)$ denote the probability of association of points z_i mapped to the cluster center Φ_k. Using the square distance $D_k(z_i)$ between the center Φ_k and the training vector z_i, the distortion function in the function space becomes

$$J_\Phi = \sum_{i=1}^{l} \sum_{k=1}^{c} p(z_i) p(\Phi_k|z_i)\, D_k(z_i). \tag{6}$$

Since no a priori knowledge of the distribution of data is assumed, over all possible distributions which give a given value of J_Φ we choose the one that maximizes the conditional Shannon entropy in the characteristic space:

$$H_\Phi = -\sum_{i=1}^{l} \sum_{k=1}^{c} p(z_i)\, p(\Phi_k|z_i) \log p(\Phi_k|z_i). \tag{7}$$

The optimization problem can be reformulated as the minimization of the Lagrangian:

$$F_\Phi = J_\Phi - T H_\Phi, \tag{8}$$

where T is the Lagrange multiplier.

To determine the α_{ki} parameter, we minimize the free energy function F w.r.t the likelihood of association [12], which is related to the Gibbs distribution as

$$p(\Phi_k|z_i) = \frac{p(\Phi_k) e^{\frac{-D_k(z_i)}{T}}}{\sum_{m=1}^{c} p(\Phi_m) e^{\frac{-D_m(z_i)}{T}}}, \tag{9}$$

where $p(\Phi_k)$ is the mass probability for k-th cluster

$$p(\Phi_k) = \sum_{i=1}^{l} p(z_i)\, p(\Phi_k|z_i). \tag{10}$$

And so the energy function is

$$F_\Phi^* = \min_{p(\Phi_k|z_i)} (J_\Phi - T H_\Phi) = -T \sum_{i=1}^{l} p(z_i) \log \sum_{k=1}^{c} p(\Phi_k) e^{\frac{-D_k(z_i)}{T}}. \tag{11}$$

The partial derivative of F w.r.t Φ_k:

$$\frac{\partial(F_{\Phi}^{*})}{\partial(\Phi_k)} = 0. \tag{12}$$

Accordingly

$$\sum_{i=1}^{l} p(z_i)p(\Phi_k)e^{\frac{-D_k(z_i)}{\mathrm{T}}}[z_i - \Phi_k] = 0. \tag{13}$$

By dividing by the normalization factor

$$Z_{z_i} = \sum_{m=1}^{c} p(\Phi_m)e^{\frac{-D_m(z_i)}{T}}, \tag{14}$$

and so,

$$\sum_{i=1}^{l}\frac{p(z_i)p(\Phi_k)e^{\frac{-D_k(\mathrm{z_i})}{T}}}{Z_{z_i}} z_i = \sum_{i=1}^{l}\frac{p(z_i)p(\Phi_k)e^{\frac{-D_k(z_i)}{T}}}{Z_{z_i}} \Phi_k. \tag{15}$$

Using Equation (10) leads to:

$$\sum_{i=1}^{l} p(z_i)\, p(\Phi_k|z_i)z_i = \sum_{\mathrm{i=1}}^{\mathrm{l}} p(z_i)\, p(\Phi_k|z_i)\Phi_k, \tag{16}$$

$$\Phi_k = \sum_{i=1}^{l}\frac{p(z_i)p(\Phi_k|z_i)}{\sum_{i=1}^{l} p(z_i)p(\Phi_k|z_i)} z_i = \sum_{i=1}^{l} \alpha_{ki} z_i. \tag{17}$$

Finally, we obtain the expression of α_{ki} that will be used to construct the vicinal kernel for NSVC functions

$$\alpha_{ki} = \frac{p(z_i)p(\Phi_k|z_i)}{\sum_{j=1}^{l} p(z_j)p\left(\Phi_k|z_j\right)}. \tag{18}$$

3.2. NSVC with the feature space partitioning

The optimization problem based on feature space partitioning is formulated as follows [13]:

$$\text{minimize } \Phi(w) = \frac{1}{2}w^T w + C\sum_{k=1}^{K}\xi_k, \tag{19}$$

$$\text{subject to } \; y_k\int_{V(\Phi_k)} ([< z, w > +b]p(z|\Phi_k))dz \geq 1 - \xi_k,$$

$$\xi_k \geq 0, \; k = 1, \ldots, K,$$

where $V(\Phi_k)$ represents the k-th vicinity associated with the mass center Φ_k in the feature space, and $p(z|\Phi_k \;)$ is the conditional probability of respective vicinity in the feature space.

According to Bayes theorem

$$p(z_i|\Phi_k) = \frac{p(z_i)p(\Phi_k|z_i)}{p(\Phi_k)} = \frac{p(z_i)p(\Phi_k|z_i)}{\sum_{j=1}^{l} p(z_j)p\left(\Phi_k|z_j\right)}. \tag{20}$$

By comparing Equation (18) and Equation (21), we get:

$$\Phi_k = \sum_{i=1}^{l} p(z_i|\Phi_k)z_i, \tag{21}$$

and the optimization constraint becomes:

$$y_k\int_{V(\Phi_k)} ([\langle z, w\rangle + b]p(z|\Phi_k))dz= \tag{22}$$

$$= y_k[\langle\int_{V(\Phi_k)} p(z|\Phi_k)zdz\, , w\rangle + \int_{V(\Phi_k)} bp(z|\Phi_k)dz] =$$

$$= y_k[\langle\sum_{i=1}^{l} p(z_i|\Phi_k)z_i\, , w\rangle + \sum_{i=1}^{l} bp(z_i|\Phi_k)] =$$

$$= y_k[\langle\Phi_k, w\rangle + b].$$

Let define the mono- and bi-vicinal kernels as

(23) $$L_k(x) = \sum_{i=1}^{l} y_i \alpha_{ki} K(x, x_i) + b, \quad k = 1, 2, \dots, K,$$

(24) $$M_{km}(x) = \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_{ki} \alpha_{mj} K(x_i, y_j) + b, \quad k, m = 1,2, \dots, K,$$

where the α_{ki} parameters are obtained from the SKDA clustering step.

The decision boundary is

(25) $$f(x) = \sum_{k=1}^{c} \beta_k y_k L_k(x) + b,$$

where β_k is the coefficient that maximizes the dual function

(26) $$\text{maximize } W(\beta) = \sum_{k=1}^{c} \beta_k - \frac{1}{2} \sum_{k,m=1}^{c} \beta_k \beta_m y_k y_m M_{km}(x),$$

$$\text{subject to} \sum_{k=1}^{c} \beta_k y_k, \ \beta_k \ \geq \ 0.$$

In order to obtain a sparse solution at the cost of the extra clustering procedure, a good selection of the number of clusters is required.

4. Simulation and experimental results

4.1. BOSS dataset description

The purpose of the conducted experiments is to test the robustness and performance of the proposed approach for distinguishing face from non faces in different situations of facial expression, pose and luminance. In this context we used a set of real images established within our research team. This dataset is named BOSS, the selected images are almost frontal images with variations of poses, illumination and expressions. The normalization of all the images of the BOSS dataset is done by the cascade detector of Viola-Jones algorithm to remove unnecessary parts of the image. The retained size for the images is 19×19 pixels relative to the position of the eyes. The original dataset contains 2,000 images with 949 faces and 1051 non-faces.

4.2. MIT-CBCL dataset description

In addition to the previous dataset, we seek to demonstrate the performance of our method on a very well-known dataset used by the research community, namely the MIT-CBCL dataset.

The training set consists of 6,977 cropped images (2,429 faces and 4,548 non-faces). In our experiment, each image is normalized to 19×19 pixels. Fig. 1 shows some face images in the training and test sets.

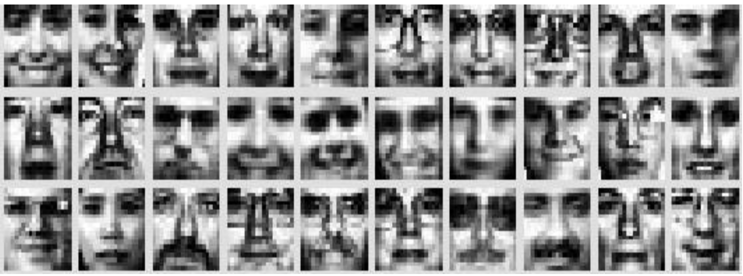

Fig. 1. A subset of MIT database used for classification

4.3. Experimental results

An accurate and robust face classification system was developed and tested. This system exploits the feature extraction capabilities of the DCT used by the NSVC classifier to increase the robustness to variations conditions.

After several experimental tests, the best accuracy was reached by the polynomial kernel function, which provided the most stable results after several tests. The classification accuracy on the BOSS datasets is given in Table 1.

Table 1. Classification results with NSVC.

Method	Polynomial kernel (q)	Accuracy (%)
DCT-NSVC	2	**99.63**
GABOR-NSVC	7	98.75
HOG-NSVC	2	97.15

The proposed technique performed much better compared to the other technique of feature extraction like HOG and Gabor combined NSVC.

Next, to demonstrate the effectiveness of our proposed algorithm, we have compared its performance to state-of-the art supervised classifiers. The obtained results are shown in Fig. 2.

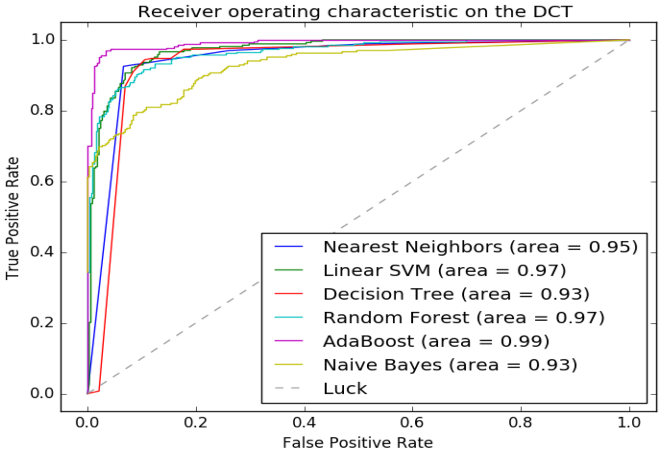

Fig. 2. Comparison results of different classifiers methods

We can conclude that almost all algorithms give good precision results: KNN (95%), SVM-linear (97%), Decision Tree (93%), Ranam Forest (97%), Adaboost (99%), Naive Bayes (93%).

As can be seen in Fig. 2, the DCT-Adaboost gives the best result of 99% accuracy. From Table 1 and Fig. 2, we can clearly observe that the NSVC outperforms the other classifiers.

4.4. Validation of experimental results

The classification results of our proposed method DCT-NSVC evaluated on the MIT-CBCL dataset are represented on the Table 2.

Table 2. Classification results for the MIT Data Set.

Kernel	Parameter	Accuracy (%)
Linear	–	99.00
Polynomial	q=2	**99.78**

The results show that the DCT with NSVC polynomial kernel function achieves the highest performance in terms of the accuracy (99.78%).

Finally, it's clear from the two evaluations that the proposed DCT-NSVC surpasses all the other sets classifiers on both Datasets. The main evaluation criteria used for comparison is the accuracy of the classification, without forgetting the adaptation to effectively manage practical applications where learning data can come from different environments.

5. Conclusion

In this paper, we presented a robust feature selection algorithm for face classification. The proposed approach consists first on extracting the image features in the frequency domain using the DCT transform. Then the classification used the NSVC method. The evaluation and comparison of the performance of the proposed approach was carried out using two public and real datasets, the best accuracy obtained exceeds 99% in both of them.

References

1. Amine, A., S. Ghouzali, M. Rziza et al. Investigation of Feature Dimension Reduction Based DCT/SVM for Face Recognition. – In: IEEE Symposium on Computers and Communications, 2008 (ISCC'2008), IEEE, 2008, pp. 188-193.
2. Chen, P.-Y., C.-C. Huang, C.-Y. Lien et al. An Efficient Hardware Implementation of HOG Feature Extraction for Human Detection. – IEEE Transactions on Intelligent Transportation Systems, Vol. **15**, 2014, No 2, pp. 656-662.
3. Teoh, S. S., T. Braunl. Performance Evaluation of HOG and Gabor Features for Vision-Based Vehicle Detection. – In: IEEE International Conference on Control System, Computing and Engineering (ICCSCE'15), IEEE 2015, pp. 66-71.
4. Ngadi, M., A. Amine, H. Hachimi, A. El-Attar. A New Optimal Approach for Breast Cancer Diagnosis Classification. – International Journal of Imaging and Robotics, Vol. **16**, 2016, Issue No 4, pp. 25-36.
5. Gupta, I., S. Kaur, P. Sahni et al. Novel Human Age Estimation System Based on DCT Features and Locality-Ordinal Information. – In: International Conference on. Inventive Computation Technologies (ICICT'16), 2016, IEEE, pp. 1-4.
6. Yang, X., A. Cao, Q. Song, G. Schaefer, Y. Su. Vicinal Support Vector Classifier Using Supervised Kernel-Based Clustering. – Artificial Intelligence in Medicine, 2014.
7. Cao, A., Q. Song, X. Yang, S. Liu, C. Guo. Mammographic Mass Detection by Vicinal Support Vector Machine. – In: IEEE International Joint Conference on Neural Networks, IEEE, Budapest, Hungary, Vol. **3**, 2004, pp. 1953-1958.
8. Vapnik, V. The Nature of Statistical Learning Theory. 2nd Edition. NY, USA, Springer Verlag, 2000.
9. Chapelle, O., J. Weston, L. Bottou, V. Vapnik. Vicinal Risk Minimization. – Advances in Neural Information Processing Systems, Vol. **13**, MIT Press, MA, USA, 2000, pp. 416-422.
10. Camastra, F., A. Verri. A Novel Kernel Method for Clustering. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **27**, 2005, No 5, pp. 801-805.

11. Leski, J. Fuzzy c-Varieties/Elliptotypes Clustering in Reproducing Kernel Hilbert Space. – Fuzzy Sets and Systems, Vol. **141**, 2004, No 2, pp. 259-280.

12. Rose, K. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. – Proceedings of the IEEE, 1998, pp. 2210-2239.

13. Zheng, B., S. W. Yoon, S. S. Lam. Breast Cancer Diagnosis Based on Feature Extraction Using a Hybrid of k-Means and Support Vector Machine Algorithms. – Expert Systems with Applications, Vol. **41**, 2014, No 4, pp. 1476-1482.

Received 15.08.2017; Second Version 08.03.2018; Accepted 02.04.2018