The Tangent Kernel Approach to Illumination-Robust Texture Classification

S. Verzakov, P. Paclík, and R.P.W. Duin

Information and Communication Theory Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
s.verzakov, p.paclik, r.p.w.duin@ewi.tudelft.nl

Abstract. Co-occurrence matrices are proved to be useful tool for the purpose of texture recognition. However, they are sensitive to the change of the illumination conditions. There are standard preprocessing approaches to this problem. However, they are lacking certain qualities. We studied the tangent kernel SVM approach as an alternative way of building illumination-robust texture classifier. Testing on the standard texture data has shown promising results.

1 Introduction

Often, it is impractical to keep experimental conditions strictly constant or redo full sensor recalibration. Imprecisions in calibration causes poorer generalization of the recognition system. Such effects can be compensated by increasing of learning sets sizes or by special data treatment: preprocessing or (which is somewhat similar to that) building robust recognition systems.

In this work we focus on building a texture classification system robust to the variability in illumination conditions. The common procedure to deal with this problem is to perform the full equalization of image histogram or mere the contrast stretching. Applying these techniques, one must decide what is the standard histogram form. It is not easy to figure out if the chosen one suits well for the purpose of the discrimination between different types of textures. Also, these methods may still leave some amount of illumination fluctuations because of the variability in data.

We propose an alternative approach which consists in a modification of a similarity measure between textures which is robust to the changes in the illumination. As a texture description we use co-occurrence matrices [1, 2] and employ tangent kernel SVM [3, 4] in order to built a robust classifier.

Our approach can be applied to any histogram-like type of data: biomedical data (histograms of DNA content), and normalized spectra. Basically, it may be used with any data represented by non-negative features with fixed sum of elements and suffering from the imprecise (linear) calibration.

The rest of the paper is structured as follows. The next section contains short review of tangent kernel SVM. Then, in section 3 it is shown how this approach

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 1009–1016, 2005.

can be applied to the illumination-robust texture recognition. Section 4 contains the description of data set and discussion of the results of numerical experiments. In section 5 we conclude our work.

2 Tangent Kernel SVM Technique

For reader convenience we provide a short account of ideas from [3, 4]. Suppose that two-class classification problem has to be solved: having an object $\mathbf{x} \in \mathbb{R}^d$, a label $y \in \{-1, +1\}$ should be assigned, defining, to which one of the two classes it belongs. In other words, the task consists of building a classification function

$$y = f(\mathbf{x}): \ \mathbb{R}^d \longrightarrow \{-1, +1\}$$

Here we assume that $f(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$, and $g : \mathbb{R}^d \to \mathbb{R}$ is the smooth discriminant function.

Suppose also that we have a prior knowledge that a one-parametric transformation $\mathcal{L}_t : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, $t \in \mathbb{R}$ does not change the class membership of the object. To simplify the task one can demand more stronger condition: an invariance of discriminant function g

$$g(\mathcal{L}_t\mathbf{x}) - g(\mathbf{x}) = 0$$

Assuming that $\mathcal{L}_t$ satisfies

$$\begin{aligned} \mathcal{L}_0\mathbf{x} &= \mathbf{x} \\ \mathcal{L}_{t_1}\mathcal{L}_{t_2} &= \mathcal{L}_{t_1+t_2} \end{aligned} \tag{1}$$

invariance property can be reformulated in a differential form

$$\left.\frac{\partial g(\mathcal{L}_t\mathbf{x})}{\partial t}\right|_{t=0} = 0$$

which transforms after some algebra into

$$\begin{aligned} \partial_{\mathbf{x}}^{\mathrm{T}} g(\mathbf{x})\ \mathcal{M}\mathbf{x} &= 0 \\ \mathcal{M} &\equiv \left.\frac{\partial \mathcal{L}_t}{\partial t}\right|_{t=0} \\ \partial_{\mathbf{x}} &\equiv \left(\frac{\partial}{\partial x^{(1)}}, \ldots, \frac{\partial}{\partial x^{(d)}}\right)^{\mathrm{T}} \end{aligned} \tag{2}$$

The condition (2) is supposed to be valid for all $\mathbf{x}$ from the data domain. Thus, it puts constraint on the possible choice of g.

Another approximate approach of taking into account Eq. (2) consists in the adding a penalty term

$$\begin{aligned} r(g; \mathbf{X}) &\equiv \frac{1}{2}\sum_i \left[\partial_{\mathbf{x}_i}^{\mathrm{T}} g(\mathbf{x}_i)\mathcal{M}\mathbf{x}_i\right]^2 = \sum_i \partial_{\mathbf{x}_i}^{\mathrm{T}} g(\mathbf{x}_i)\mathbf{C}_i\partial_{\mathbf{x}_i} g(\mathbf{x}_i) \\ \mathbf{C}_i &\equiv (\mathcal{M}\mathbf{x}_i)(\mathcal{M}\mathbf{x}_i)^{\mathrm{T}} \end{aligned}$$

to the original learning criterion by which minimization g meant to be found (e.g. inverse margin, noise to signal ratio, etc.). This transforms original minimization task into

$$\begin{aligned} g^* &= \arg\min_g R_\gamma(g; \mathbf{X}, \mathbf{y}) \\ R_\gamma(g; \mathbf{X}, \mathbf{y}) &= (1-\gamma) R(g; \mathbf{X}, \mathbf{y}) + \gamma r(g; \mathbf{X}) \\ \gamma &\in [0, 1) \end{aligned} \tag{3}$$

where R is the original learning criterion, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ is the training data set, $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ contains labels of training objects and parameter γ defines how is the penalty term important with regard to the R.

If we decide that g has linear $g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$ and choose as a learning algorithm the SVM [5] with $R = \frac{1}{2}\|\mathbf{w}\|^2$, then the modified minimization criterion takes a form

$$R_\gamma = \frac{1-\gamma}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\mathbf{w}^{\mathrm{T}}\mathbf{C}\mathbf{w}$$

$$\mathbf{C} = \sum_i \mathbf{C}_i$$

By substitution

$$\begin{aligned} \tilde{\mathbf{w}} &= \left[(1-\gamma)\mathbf{I} + \gamma\mathbf{C}\right]^{1/2} \mathbf{w} \\ \tilde{\mathbf{x}} &= \left[(1-\gamma)\mathbf{I} + \gamma\mathbf{C}\right]^{-1/2} \mathbf{x} \end{aligned}$$

we recover the original form of the SVM criterion: $R_\gamma = \frac{1}{2}\|\tilde{\mathbf{w}}\|^2$. So, to implement this technique we need only to redefine the matrix of inner products (kernel matrix)

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = \tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{y}} = \mathbf{x}^{\mathrm{T}} \left[(1-\gamma)\mathbf{I} + \gamma\mathbf{C}\right]^{-1} \mathbf{y} \tag{4}$$

If there is overlap between classes, then a soft-margin version of SVM algorithm should be used. The modifications are straightforward and do not change Eq. (4). For the sake of brevity we will use LTK-SVM (Linear Tangent Kernel SVM) abbreviation to name the soft margin SVM algorithm with kernel defined by Eq. (4). The extension of this approach to the non-linear discriminant functions g is possible [3] but it is beyond of the scope of this paper.

3 Derivation of the Tangent Kernel for n-Dimensional Histograms

Suppose that the input of the recognition system is a continues distribution density function of an n-dimensional random vector $\boldsymbol{\eta}$ such that

$$\boldsymbol{\eta} = e^t \boldsymbol{\xi}(\boldsymbol{\theta})$$

where $\boldsymbol{\xi}$ is a measurement made on a perfectly calibrated device, t is a parameter responsible for the imprecisions in sensor calibration and $\boldsymbol{\theta}$ is a set of parameters which defines intra- and inter-class variability.

The distribution of the "observed" $\boldsymbol{\eta}$ can be expressed in terms of the distribution of "ideal" $\boldsymbol{\xi}$ as

$$\rho_{\boldsymbol{\eta}}(z_1, \ldots, z_n) = e^{-nt}\rho_{\boldsymbol{\xi}}(e^{-t}z_1, \ldots, e^{-t}z_n)$$

To achieve better classification rates we need to built classifier robust to the data transformations caused by the operator

$$\mathcal{L}_t^{sc}\rho(z_1, \ldots, z_n) = e^{-nt}\rho(e^{-t}z_1, \ldots, e^{-t}z_n)$$

This operator satisfies conditions Eq. (1) and thus, we can employ tangent kernel SVM. By taking the first derivative of $\mathcal{L}_t^{sc}\rho$ at $t = 0$, we find that

$$\mathcal{M}^{sc}\rho(z_1, \ldots, z_n) = -n\rho(z_1, \ldots, z_n) - \sum_{j=1}^{n} z_j \partial_{z_j}\rho(z_1, \ldots, z_n)$$

In practice one deals with histograms (e.g. co-occurrence matrices) not with distribution densities. Assuming that $\mathbf{P} = (P_{k_1,\ldots,k_n})$ such a multi-dimensional histogram (properly normalized to be an estimation of a density function) we redefine the $\mathcal{M}^{sc}$ operator as

$$(\mathcal{M}^{sc}\mathbf{P})_{k_1,\ldots,k_n} = -nP_{k_1,\ldots,k_n} - \sum_{j=1}^{n} z_j^{(k_j)}(\Delta_j\mathbf{P})_{k_1,\ldots,k_n}$$

Here, $z_j^{(k_j)}$ are the centers of histogram bins in the j-th direction and Δ_j is the operator taking the (smoothed) finite differences of the histogram $\mathbf{P}$ in the the same direction j. Assuming that $\mathbf{u}(\mathbf{P})$ is unfolding of an n-dimensional array into a column vector, it is possible to write down the penalty term

$$r = \mathbf{w}^{\mathrm{T}}\mathbf{C}^{sc}\mathbf{w}$$
$$\mathbf{C}^{sc} = \sum_i \mathbf{u}(\mathcal{M}^{sc}\mathbf{P}_i)\mathbf{u}(\mathcal{M}^{sc}\mathbf{P}_i)^{\mathrm{T}}$$

Thus, the new similarity measure reads as

$$\tilde{k}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}}\{(1-\gamma)\mathbf{I} + \gamma\mathbf{C}^{sc}\}^{-1}\mathbf{y}$$

4 Numerical Experiments

As data for our experiments we used Brodatz textures 1.3.04 and 1.3.05 from the [6]. Each image of 1024-by-1024 size and 8-bit depth was split into 64 128-by-128 non-overlapping patches. To imitate the change in the illumination conditions each patch as a whole was multiplied by the randomly generated number uniformly distributed in the region $[1-\alpha; 1]$. The α values in the range from 0 to 0.5

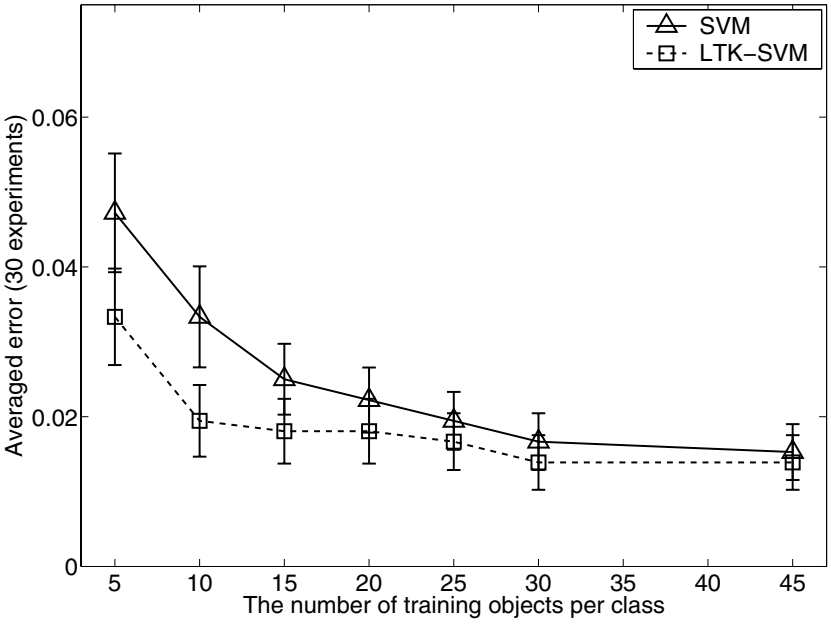

Fig. 1. Learning curves for 16-by-16 co-occurrence matrices data set. $\alpha = 0$

were used to create a number data sets. Afterwards, for each such data set we computed 128 (by the number of patches) 2-dimensional co-occurrence matrices of 16-by-16 and 32-by-32 sizes which served as an input of classifiers. Neither contrast stretching nor histogram equalization was applied to data in any way in all experiments presented in this paper.

We studied the difference in the performance of the conventional linear SVM classifier and its tangent kernel version. To see how useful the usage of prior knowledge can be for different sizes of training set the learning curves were computed. The results were obtained by averaging over 30 hold-out experiments: for each experiment we randomly took out 20% of all objects, the rest 80% objects were used as the training pool. The learning curves were obtained for each hold-out experiment by training classifiers on the sequence of the nested training sets generated from the training pool.

In all experiments the ν regularization parameter [7] of SVM/LTK-SVM classifier was optimized by grid search over the set of predefined values. For each candidate value the preliminary classifier was trained on 75% data randomly selected for the training procedure. The other 25% were used to measure the performance and select the actual ν value. Using this value, the final classifier was trained on the whole currently available training data set. Linear tangent kernel SVM was trained at a number of γ values (no internal optimization). For the smoothing of finite differences we used Savitsky-Golay filter of the first order. The size of the filter window w was always set to 3.

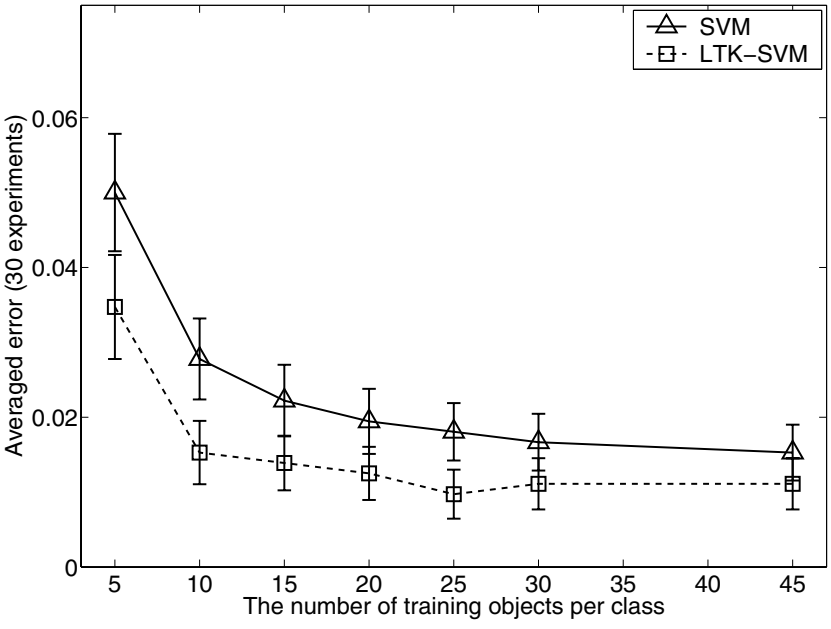

Fig. 2. Learning curves for 16-by-16 co-occurrence matrices data set. $\alpha = 0.1$

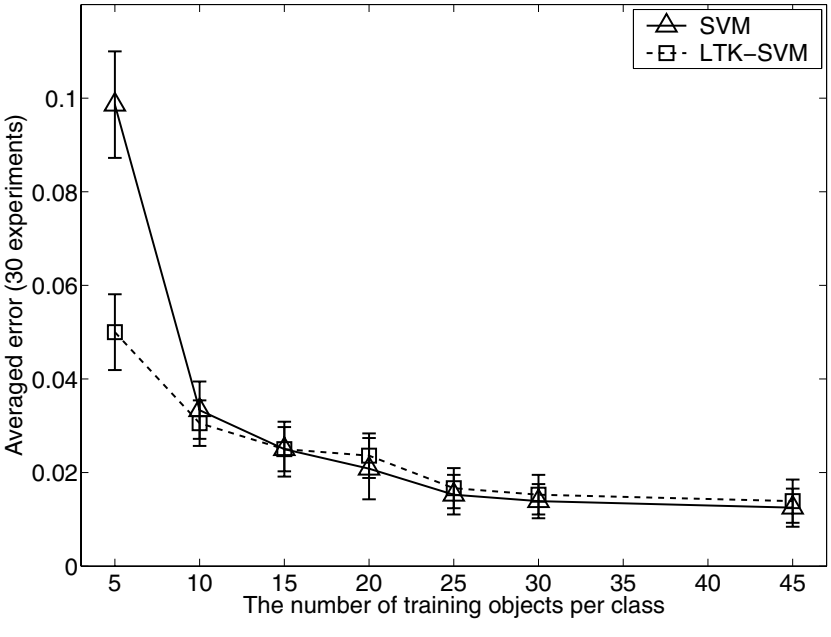

Fig. 3. Learning curves for 16-by-16 co-occurrence matrices data set. $\alpha = 0.5$

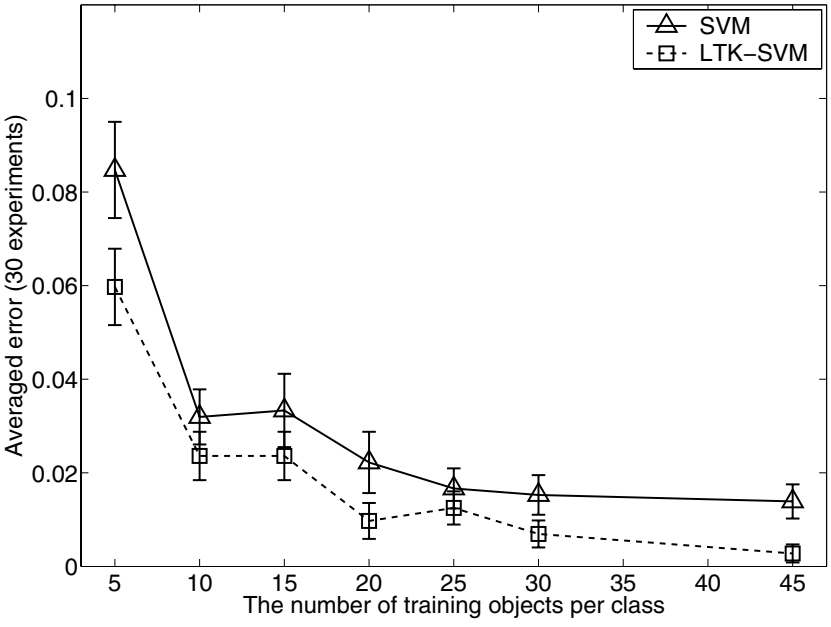

Fig. 4. Learning curves for 32-by-32 co-occurrence matrices data set. $\alpha = 0.5$

Figures 1, 2 and 3 show the learning curves of the conventional linear SVM and LTK-SVM ($\gamma = 0.95$) on the 16-by-16 co-occurrence matrices. It can be seen that even at $\alpha = 0$ (original illumination of images) the applying of modified similarity leads to the better classification rates. The effect is more recognizable at $\alpha = 0.1$. However, larger variability in the illumination ($\alpha = 0.5$) cannot be compensated by the use of LTK-SVM.

On the other hand, LTK-SVM can cope with the such an amount of illumination variability being applied to the 32-by-32 co-occurrence matrices (Fig. 4). Probably, this effect can be explained by the fact that the similarity measure being derived for continues distributions gives better results for histograms with finer bins.

5 Conclusions

We studied the possibility of application of the tangent kernel approach to the stabilization of illumination conditions for better texture classification. The tangent kernel approach shows significant advantages at smaller sample sizes comparing to the conventional SVM. Unlike the preprocessing methods, proposed technique can be applied directly to the co-occurrence matrices even when raw images are unavailable. The use of the tangent kernel approach does not require

to select invariant features or select standard form of image histogram. All this makes it to be a promising tool in many practical situations. However, there are open questions: e.g. what is the optimal strategy to select the tradeoff parameter γ or how necessary and convenient may be the exploitation of nonlinear kernels. Definitely, more study is needed including testing on the real-world data.

Acknowledgments

This research was supported by the Technology Foundation STW, applied science division of NWO and the technology program of the Ministry of Economic Affairs.

References

1. R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. 3(6):610–621, November 1973.
2. R.M. Haralick. Statistical and Structural Approaches to Texture. *Proceedings of the IEEE*, 67:786–804, 1979.
3. B. Schölkopf. *Support Vector Learning*. PhD thesis, Munich, 1997.
4. B. Schölkopf, P. Y. Simard, A. J. Smola, and V. N. Vapnik. Prior knowledge in support vector kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 640–646, Cambridge, MA, 1998. MIT Press.
5. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, USA, 2000.
6. USC-SIPI Image Database.
7. B. Schölkopf, A. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.