Countering the false positive projection effect in nonlinear asymmetric classification

Serhiy Kosinov, *Student Member, IEEE* Stéphane Marchand-Maillet, *Member, IEEE*
and Thierry Pun, *Member, IEEE*

Computer Vision and Multimedia Lab
University of Geneva, 1211 Geneva, Switzerland
Phone: +41 (22) 379 1099, email: kosinov@cui.unige.ch
Phone: +41 (22) 379 7631, email: marchand@cui.unige.ch
Phone: +41 (22) 379 7627, email: pun@cui.unige.ch

Countering the false positive projection effect in nonlinear asymmetric classification

Serhiy Kosinov, *Student Member, IEEE* Stéphane Marchand-Maillet, *Member, IEEE* and Thierry Pun, *Member, IEEE*

***Abstract*—This work concerns the problem of asymmetric classification and provides the following contributions. First, it introduces the method of KDDA - a kernelized extension of the distance-based discriminant analysis technique that treats data asymmetrically and naturally accommodates indefinite kernels. Second, it demonstrates that KDDA and other asymmetric nonlinear projective approaches, such as BiasMap and KFD are often prone to an adverse condition referred to as the *false positive projection effect*. Empirical evaluation on both synthetic and real-world data sets is carried out to assess the degree of performance degradation due to false positive projection effect, determine the viability of some schemes for its elimination, and compare the introduced KDDA method with state-of-the-art alternatives, achieving encouraging results.**

***Index Terms*—kernel methods, discriminant analysis**

I. Introduction

ASYMMETRIC classification considers a learning problem where a given target class must be distinguished from all of the other classes. In this scenario, the samples from the target class, usually referred to as positive class, and the rest of the data, referenced as the negative class, are not treated equally due to substantial differences in prior probabilities, misclassification costs, etc. Such distinction, when modeled explicitly, has been previously shown to improve classification accuracy for undersampled and unbalanced data sets [1], [2], [3].

Situated in the context of asymmetric classification, this paper provides the following contributions. First, it introduces the method of KDDA - a kernelized extension of the distance-based discriminant analysis technique with asymmetric data treatment that admits indefinite kernels. Second, it demonstrates that KDDA and other nonlinear projective approaches, such as BiasMap [3] and Kernel Fisher Discriminant, KFD [4], are prone to an adverse condition referred to as the *false positive projection effect.* An undesirable consequence of the said condition results in sizeable areas of the input space being erroneously associated with the target class. Using an illustrative geometric interpretation, we study the circumstances that lead to false positive projection effect, and consider possible strategies for its mitigation.

Extensive experiments are carried out on both synthetic and real-world data sets to help assess the degree of performance degradation due to false positive projection effect, determine the viability of different schemes for its elimination, and compare the proposed KDDA method with state-of-the-art alternatives.

II. Kernel distance-based discriminant analysis

A. DDA overview

The approach of distance-based discriminant analysis (DDA) [5] has been shown to improve classification performance as a result of the chosen non-parametric formulation focused on pairwise inter-observation distances, robustification of some of the interpoint distances, and an inherent feature selection component. Relying on an asymmetric formulation, the method derives a linear transformation T by iteratively optimizing an approximation of the logarithm of the following criterion:

$$J(T) = \frac{\left(\prod_{i=1}^{N_X}\prod_{j=i+1}^{N_X} \Psi\left(d_{ij}^W(T)\right)\right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X}\prod_{j=1}^{N_Y} d_{ij}^B(T)\right)^{\frac{1}{N_X N_Y}}}, \qquad (1)$$

where N_X, N_Y are the numbers of observations in data sets X and Y that represent positive and negative classes, respectively, the numerator and denominator of (1) characterize the geometric means of the within- and between-class Euclidean distances d_{ij}^W, d_{ij}^B, parametrized by T and defined as $d_{ij}^W = \sqrt{(x_i - x_j)^T TT^T (x_i - x_j)}$ and $d_{ij}^B = \sqrt{(x_i - y_j)^T TT^T (x_i - y_j)}$, respectively, for $\{x_i\}_{i=1}^{N_X} \in \mathbb{R}^n$, $\{y_j\}_{j=1}^{N_Y} \in \mathbb{R}^n$, and $\Psi(\cdot)$ denotes a modified Huber robust estimation function [6] that mitigates the influence of outliers and helps avoid numerical difficulties due to zero length transformed distances. The theoretical underpinnings motivating the above formulation become clear when a logarithm of (1) is considered. Indeed, $\log J(T)$, being a weighted sum of log-barrier functions, may be viewed as an extended formulation of analytic center machine (ACM) method that finds a separating hyperplane as an analytic center of the classifier version space [7].

In the below sections, we address the limitations of the DDA linked to its reliance on Euclidean distances and linearity of the sought transformation by introducing KDDA, a kernel-based extension of the technique.

B. Kernelized Distance-based Discriminant Analysis

Let us suppose that there is a space $\mathscr{F}$ where samples of training data can be mapped via $\Phi : \mathbb{R}^n \to \mathscr{F}$, such that there exists a kernel function $k(x,y) = (\Phi(x))^T \Phi(y)$,

Computer Vision and Multimedia Lab
University of Geneva, 1211 Geneva, Switzerland
Phone: +41 (22) 379 1099, email: kosinov@cui.unige.ch
Phone: +41 (22) 379 7631, email: marchand@cui.unige.ch
Phone: +41 (22) 379 7627, email: pun@cui.unige.ch

where $x, y \in \mathbb{R}^n$ and $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. We will also assume that the discriminative transformation is sought in $\mathscr{F}$ as a projection matrix ω of size $[\mathscr{N}_{\mathscr{F}} \times d]$, where $\mathscr{N}_{\mathscr{F}}$ is the dimensionality of $\mathscr{F}$, and d is the dimension of the derived discriminative projection subspace, such that the columns of ω lie in the span of all training samples mapped in $\mathscr{F}$:

$$\omega = \left[\sum_i^N \alpha_i^{(1)} \Phi(z_i) \;\; \sum_i^N \alpha_i^{(2)} \Phi(z_i) \;\; \ldots \;\; \sum_i^N \alpha_i^{(d)} \Phi(z_i) \right], \tag{2}$$

where z_i is one of the $N_X + N_Y$ samples from the training data compound matrix $Z = \left[\begin{array}{cc} X & Y \end{array} \right]$. The distances between images of samples x and y projected from $\mathscr{F}$ by solution ω are thus expressed as:

$$\begin{aligned} \mathscr{D}_{xy}^2(\omega) &= \left(\Phi(x) - \Phi(y)\right)^T \omega\omega^T \left(\Phi(x) - \Phi(y)\right) \\ &= \sum_j^d \left(\sum_i^N \alpha_i^{(j)} (k(z_i, x) - k(z_i, y)) \right)^2 . \end{aligned} \tag{3}$$

In matrix notation (3) can be simplified as:

$$\mathscr{D}_{xy}^2(\omega) \equiv \mathscr{D}_{xy}^2(P) = \mathbf{tr}\left(P^T H_{xy} P\right) \tag{4}$$

where $P \in \mathbb{R}^{N \times d}$ is the sought nonlinear transformation represented as a matrix collecting all of the $\alpha_i^{(j)}$ coefficients, $H_{xy} = (K_x - K_y)(K_x - K_y)^T$, and $K_s = [k(z_1, s), k(z_2, s), \ldots, k(z_N, s)]^T$ denotes a vector of kernel evaluations for sample s over all of the training data.

In view of (4), the logarithm of the DDA optimization criterion (1) can now be expressed in terms of distances projected from a richer, possibly infinite-dimensional feature space $\mathscr{F}$:

$$\begin{aligned} \log J(P) = & \frac{2}{N_X(N_X - 1)} \sum_{i=1}^{N_X} \sum_{j=i+1}^{N_X} \log \Psi\left(\mathscr{D}_{ij}^W(P)\right) \\ & - \frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \log \mathscr{D}_{ij}^B(P) \end{aligned} \tag{5}$$

The treatment of the obtained criterion differs only slightly compared to the linear case. Similarly to the way it is done in the DDA [5], we express convex parts of the criterion by their respective piece-wise linear approximations majorized by quadratics [8], while the concave parts are linearized. The former simple algebraic manipulation relies on the Cauchy-Schwarz inequality, while the latter is a direct consequence of the concavity of the log-function[1], whose combined application leads to the following approximation:

$$\begin{aligned} \mu_{\log J}(P, \bar{P}) = & \frac{1}{N_X(N_X - 1)} \mathbf{tr}\left(P^T \mathbb{K}_X B(\bar{P}) \mathbb{K}_X^T P\right) \\ & + \frac{1}{2 N_X N_Y} \mathbf{tr}\left(P^T \mathbb{K}_{XY} C \mathbb{K}_{XY}^T P\right) \\ & + \frac{2}{N_X N_Y} \mathbf{tr}\left(P^T \mathbb{K}_{XY} G(\bar{P}) \mathbb{K}_{XY}^T \bar{P}\right) \\ & + const, \end{aligned} \tag{6}$$

where $\bar{P}$ is the current solution, $\mathbb{K}_X$, $\mathbb{K}_{XY}$ are Gram matrices of kernel inner products evaluated over X and all data, respectively, and B, C, G are positive semi-definite design matrices independent of P. Elements b_{ij} of B are defined as:

$$b_{ij} = \begin{cases} -\dfrac{\bar{w}_{ij}}{\Psi\left(\mathscr{D}_{ij}^W(\bar{P})\right)} & \text{if} \quad i \neq j; \\ -\displaystyle\sum_{k=1, k \neq i}^{N_X} b_{ik} & \text{if} \quad i = j; \end{cases} \tag{7}$$

where $\bar{w}_{ij}$ is a weight of the Huber function majorizer, that in this case is equal to 1 if $\Psi(\mathscr{D}_{ij}^W(\bar{P}))$ is less than the robustness threshold c, or $c / \Psi(\mathscr{D}_{ij}^W(\bar{P}))$ otherwise. For matrices C and G, their non-zero elements m_{ij} are defined as:

$$m_{ij} = \begin{cases} r_{ij} & \text{for } i \in [1; N_X] \\ & \text{and } j \in [N_X + 1; N], \\ r_{ij} & \text{for } i \in [N_X + 1; N] \\ & \text{and } j \in [1; N_X], \\ -\displaystyle\sum_{k=1, k \neq i}^{N_X + N_Y} m_{ik} & \text{for } i = j, \end{cases} \tag{8}$$

where r_{ij} is equal to -1 and $-1 / \left(\mathscr{D}_{ij}^B(\bar{P})\right)^2$ for C and G, respectively.

The approximations used to derive $\mu_{\log J}(P, \bar{P})$ are chosen so as to ensure that the resulting expression's value is never less than the objective to be minimized, and thus provdies an upper bound of the criterion (5). By optimizing (6) iteratively, every subsequent iteration achieves a goal function value that is better or at least as good as the one from the previous iteration, which leads to covergence under the practically reasonable objective boundedness assumption. This iterative process has been previously shown to attain more robust as well as better quality local minima, compared to the standard optimization techniques, such as gradient descent and SQP with trust region approximations. More formally, such an iterative scheme that constitutes the core of the KDDA, the kernelized extension of the distance-based discriminant analysis method, can be written as the following algorithm:

Algorithm 1. *KDDA*

```
1. Assign a starting point $\bar{P} = \bar{P}_0 \in \mathbb{R}^{N \times d}$,
   set convergence tolerance $\epsilon$;
2. Find a successor point $P_s$ :
   $P_s = \arg\min_P \mu_{\log J}(P, \bar{P})$,
   subject to a regularization constraint;
3. If $\log J(\bar{P}) - \log J(P_s) < \epsilon$, then stop;
4. Set $\bar{P} = P_s$, go to 2.
```

III. False positive projection effect

A. Geometric illustration

We now turn to the discussion of the *false positive projection* effect, a condition that often arises when a projective

[1] For any $\bar{x} > 0$ we have: $\log(x) \leq \bar{x}^{-1} x + \log(\bar{x}) - 1$.

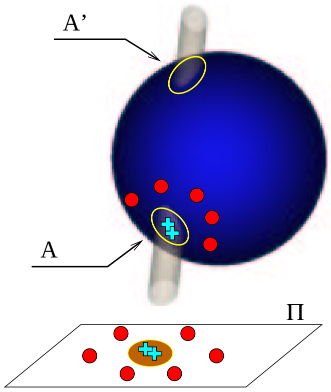


Fig. 1. A sketch of the false positive projection of region A'.

nonlinear classifier utilizing a Gaussian kernel learns a decision function that erroneously associates regions of the input space with the positive class.

A sketch of this adverse effect is shown in Figure 1. Here, an asymmetric projective nonlinear classifier, such as KDDA or BiasMap, learns a discriminative projection Π that ensures maximum compactness of the positive class observations, depicted as crosses, relative to the scatter of the negative class observations, depicted as circles. One may notice, however, that the obtained decision region in the projection plane Π corresponds to two distinct parts of the spherical mapping manifold in feature space $\mathscr{F}$: A and A'. In this setup, all test data mapped into A' are classified as positive, assuming such mapping is possible [9], even though the region contains none of the training samples of the positive class to support such a decision, and likely corresponds to the input space areas where the negative class is far more probable. Thus, a false positive projection of A' takes place. Some examples of the occurrence of this adverse condition with KDDA, BiasMap and KFD classifiers are demonstrated on simple 2D data sets in Figure 2. The data samples belonging to the positive and negative classes are shown as crosses and circles, respectively, while yellow-colored areas highlight the regions of input space classified as positive.

Alternatively, the false positive projection occurrence in the KDDA approach may be thought of as caused by a considerable multiplicity of solutions x^* of $\Phi(x^*)^T\omega = u$, for $u \in U$, where U is a region of projection of positive examples in Π. While this conjecture certainly merits a separate investigation into preventive modifications such as multiplicity-reducing signed distance inequalities, its universal applicability is yet to be established. Therefore, in the following discussion we will focus only on the method-independent post-processing strategies, i.e. the techniques that do not alter the method in question, but are applied once the learning process has been completed.

B. *Line tracing elimination strategy*

In order to summarize the description provided in the previous section and be able to formulate a simple post-processing strategy for elimination of the false positive projection effect, we make an observation analogous to that used in the cluster assignment rule of the support vector clustering method, SVC [10]: a data sample is subject to false positive projection if it is classified as positive, but lies across the decision boundary with respect to all of the positive class training samples. This prompts a straightforward strategy based on sampling or "tracing" classification decisions along the simplest possible linear paths between a test sample and positive class training data, leading to the following algorithm for detecting and rejecting the predictions on the test samples erroneously classified as positive.

Algorithm 2. *FPP elimination by line tracing*

```
1. Obtain a candidate test sample $t$
   classified as positive;
2. Select sets $\mathcal{L}_i = \{\lambda t + (1-\lambda)x_i : \lambda \in [0,1]\}$,
   $\forall x_i \in X$, $i = 1 \ldots N_X$;
3. If each of $\mathcal{L}_i$ has a sample classified as
   negative, declare false positive projection
   and reject positive classification decision on $t$.
```

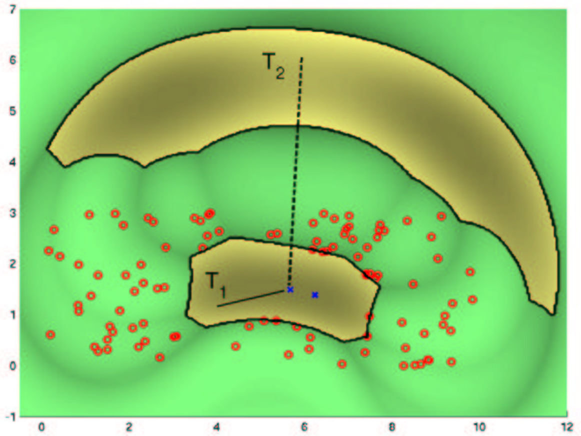
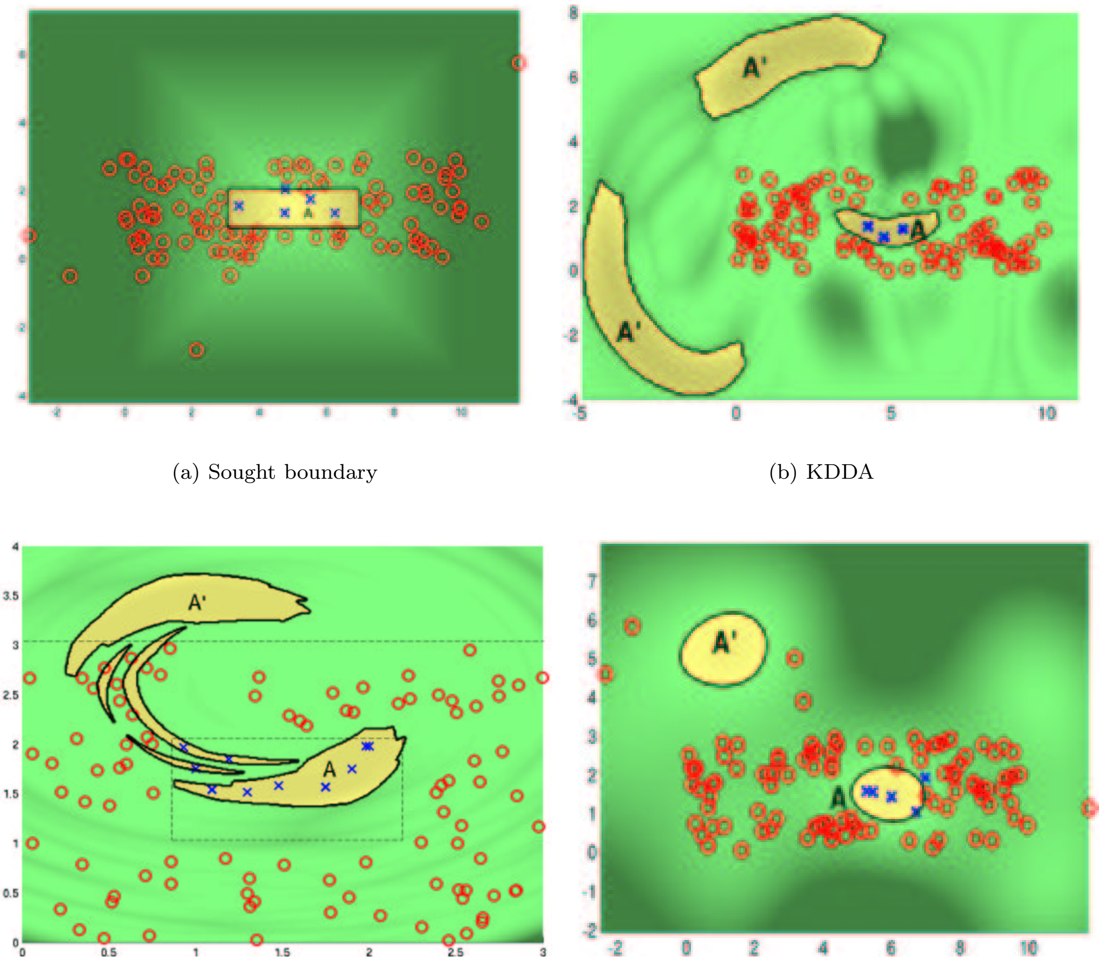


Fig. 3. Applying line-tracing strategy for the false positive projection (FPP) effect elimination: positive classification decision is accepted for T_1, but rejected for T_2

An illustration of the above algorithm applied to the KDDA method is shown in Figure 3. Here, two sample straight lines are traced in the input space from candiadate test points T_1 and T_2. While a positive classification decision is retained on T_1, it is rejected on T_2, since on every straight line connecting it to the positive samples of the training data there exist points classified as negative. The latter fact is detected by verifying the classification decisions in the learned nonlinear projection, on points sampled from sets $\mathcal{L}_i$ using a simple uniform sampling technique as adoped in the SVC method, and switching to a Newton-Raphson root-finding routine when necessary.

C. *Filter classifier elimination strategy*

The simplicity of the above described elimination strategy comes at a price of having to impose crude linear constraints on the obtained decision boundary, which may negate the benefits of learning a complex nonlinear classi-



(a) Sought boundary

(b) KDDA

(c) BiasMap

(d) Kernel Fisher Discriminant

Fig. 2. An illustration of an ideal class separation boundary (2(a)) and examples of the false positive projection (FPP) effect occurrence in various nonlinear projective methods, (2(b)-2(d)). Input space regions subject to FPP are denoted as A'. All methods use Gaussian kernels with $\sigma = 2$.

fier. For that reason, we also consider a filter classifier elimination strategy, implemented by introducing a high-recall add-on classifier that limits the input space domain admissible for positive classification. That is, a given test sample must be predicted as positive by both classifier in question and the filter. Practically, the filter is implemented as a multiple-hyperplane classifier [11].

IV. Experimental results

First, we conducted a series of experiments on the synthetic nested cuboid data sets, an 2D example of which was earlier shown in Figure 2(a). The positive and negative class observations were sampled inside and outside of randomly generated cuboids with the imbalance ratio of 100, and submitted to classification by KDDA, K-BiasMap and KFD using a Gaussian kernel with $\sigma = 2$. Due to substantial class imbalance, the classification performance is separately calculated over the positive and negative class instances. The true positive rate a^+, or sensitivity, is the fraction of the positive class samples predicted correctly. Similarly, the true negative rate a^-, or specificity, is the fraction of the negative class samples predicted correctly. The overall performance is thus assessed by evaluating geometric mean accuracy $\mathbf{GM} = \sqrt{a^+ \times a^-}$ that takes into account prediction accuracy on both classes [12], and specificity $\mathbf{SP} = a^-$ designed to measure the effect of false positives on classification performance. The results achieved by the three methods alone (denoted *none*, meaning no FPP elimination strategy is used) as well as their performance enhanced by the FPP elimination techniques (denoted *tracing* and *filter*, respectively) are listed in Table I. The reported figures demonstrate a statistically significant improvement in specificity for KDDA and KFD methods leading to an overall geometric mean accuracy increase, while at the same time pointing out the overly conservative nature of the BiasMap method where the changes are not significant.

For our content-based multimedia retrieval experiments we chose ETHZ80 collection [13], containing 3280 high-resolution color images whose visual information was rep-

TABLE I
GM accuracy and specificity (in %%) for nested cuboid synthetic data set

Method	None		Line tracing		Filter	
	GM	**SP**	**GM**	**SP**	**GM**	**SP**
KDDA	70.0 (±2.4)	96.6 (±1.1)	70.9 (±2.7)	99.6 (±0.1)	75.4 (±2.7)	99.6 (±0.2)
BiasMap	55.3 (±3.0)	99.1 (±0.5)	54.7 (±3.2)	99.5 (±0.2)	55.3 (±3.3)	99.6 (±0.1)
KFD	65.1 (±3.5)	73.3 (±5.9)	76.5 (±2.6)	99.6 (±0.2)	75.3 (±3.2)	99.7 (±0.1)

TABLE II
GM accuracy and specificity (in %%) for ETHZ80 image collection

	KFD			BiasMap			KDDA		
	none	tracing	filter	none	tracing	filter	none	tracing	filter
GM	82.7	82.7	82.7	58.0	59.2	71.4	76.8	77.2	82.2
SP	94.5	94.6	94.7	50.0	54.0	74.4	79.8	80.7	83.6

resented by 286-dimensional feature vector containing 166 global color histogram and 120 Gabor filter texture descriptors. extracted by the *Viper* system [14]. Kernel parameters were determined by cross-validation so as to maximize performance of KFD, and fixed afterwards. The obtained results for each method in terms of averaged GM accuracy and specificity in the "one-against-all" classification scenario are given in Table II. The reported figures generally confirm the hypothesis that false positive projection elimination strategies increase specificity leading to a better GM accuracy. These findings also demonstrate that even simple post-processing methods, such as line tracing, may sometimes be sufficient to enhance classification performance, while further benefits may be extracted from more sophisticated techniques, such as the filter method of an add-on high recall classifier.

V. Conclusion

We have presented the method of KDDA - a kernelized extension of the distance-based discriminant analysis approach. The proposed technique has been shown to be prone to a typical failure scenario referred to as the *false positive projection* effect, also present in a number of other nonlinear projective techniques. We have also suggested and empirically evaluated strategies for avoiding the above mentioned adverse effect on a number of synthetic data sets and on the task of content-based image retrieval, achieving encouraging results.

References

[1] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz, "Applying support vector machines to imbalanced datasets.," in *Proceedings of the 15th European Conference on Machine Learning (ECML'04)*, 2004, pp. 39–50.

[2] K. Veropoulos, N. Cristianini, and C. Campbell, "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI99)*, Stockholm, Sweden, 1999, pp. 55–60.

[3] X. Zhou, A. Garg, and T. Huang, "A discussion of nonlinear variants of biased discriminants for interactive image retrieval," in *Proceedings of CIVR'04*, Dublin, Ireland, 2004, pp. 353–364.

[4] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. 1999, pp. 41–48, IEEE.

[5] Serhiy Kosinov, Stéphane Marchand-Maillet, and Thierry Pun, "Iterative majorization approach to the distance-based discriminant analysis," in *Proceedings of the 28th Annual Conference of the GfKl 2004*, Dortmund, Germany, March 9–11 2004.

[6] P. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[7] T. B. Trafalis and A. M. Malyscheff, "An analytic center machine," *Machine Learning*, vol. 46, pp. 203–223, 2002.

[8] W. Heiser, "Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis," *Recent advances in descriptive multivariate analysis*, pp. 157–189, 1995.

[9] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[10] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and V. N. Vapnik, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[11] Serhiy Kosinov and Ivan Titov, "Large margin multiple hyperplane classification for content-based multimedia retrieval," "Machine Learning Techniques for Processing Multimedia Content", ICML Workshop on Machine Learning Techniques for Processing Multimedia Content, Bonn, Germany (accepted), August 11 2005.

[12] Miroslav Kubat and Stan Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proc. 14th International Conference on Machine Learning*, 1997, pp. 179–186.

[13] Bastian Leibe and Bernt Schiele, "Analyzing appearance and contour based methods for object categorization," in *International Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, Wisconsin, June 2003, pp. 409–415.

[14] David McG. Squire, Wolfgang Müller, Henning Müller, and Jilali Raki, "Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback," in *The 11th Scandinavian Conference on Image Analysis*, Kangerlussuaq, Greenland, june 1999, pp. 143–149.