

Selectivity Supervision in Combining Pattern-Recognition Modalities by Feature- and Kernel-Selective Support Vector Machines

A. Tatarchuk, V. Mottl
Computing Center of the Russian
Academy of Sciences, Moscow, Russia
aitech@yandex.ru, vmottl@yandex.ru

A. Eliseyev
Moscow Institute of Physics and
Technology, Moscow, Russia
andreyel@gmail.com

D. Windridge
Centre for Vision, Speech and Signal
Processing, University of Surrey,
Guildford, UK
D.Windridge@surrey.ac.uk

Abstract

Multi-modal pattern recognition must frequently truncate the set of initially available modalities. When a kernel-based approach is adopted within each modality, the problem of modality selection becomes mathematically analogous to that of wrapper-based feature selection. In this paper, we revise two implicitly wrapper-based methods of SVM-embedded selective kernel combination, the Relevance and Support Kernel Machines, so as to equip them with the ability to preset the desired level of feature-selectivity. Hence, a continuous axis of nested feature selection models is obtained, ranging from the absence of selectivity to the selection of single features. We thus unite the distinct processes of selection and classification within the two techniques in manner suitable for general application within Kernel-based multi-modal pattern recognition.

1. Introduction

Multimodal pattern recognition systems utilize several distinct feature modalities, often with different scales, to represent specific phenomena [1,2]. Feature scales $x_i \in \mathbb{X}_i$ may be quite complicated, so that frequently the only way of treating real-world objects $\omega \in \Omega$ is via pair-wise comparison of their features $(x_i(\omega), x_i(\omega'))$ using modality-specific functions $K_i(x'_i, x''_i)$ defined in the output scales of the sensors $\mathbb{X}_i \times \mathbb{X}_i \rightarrow \mathbb{R}$. A function $K(x', x'')$ is a *kernel* if it forms a semidefinite matrix for any finite collection of objects. Hence, a kernel embeds the scale of the respective feature \mathbb{X}_i into a hypothetical linear space in which it plays the role of inner product. In particular, when $x_i(\omega) \in \mathbb{X}_i = \mathbb{R}$, the natural kernel will be the product $K_i(x'_i, x''_i) = x'_i x''_i$. Support Vector Machines (SVMs), originally designed for two-class pattern recognition learning in \mathbb{R}^n , can thus be used to combine modalities by employing a joint kernel $K(\mathbf{x}', \mathbf{x}'') = \sum_{i=1}^n x'_i x''_i$. This analogy is exploited by multi-kernel SVMs when more sophisticated kernel-represented modalities are to be combined [3,4,5].

In general, the danger of over-fitting makes it necessary to combine modality-specific features on a selective basis. Feature selection (FS) techniques are classed in the literature as *filters* and *wrappers* [6].

Filters, as distinct from wrappers, are applied to the feature set independently of classification technique. Selection can take the form of assigning continuous weights to the features or, more commonly, binary inclusion/exclusion decisions. Less often considered are composite mechanisms for classification/selection, such that FS is implicit in the process of classification *itself* (although see [7]) because of the danger of increased sample variance. However, if there exists a method of assigning the desired level of selectivity *a priori*, ranging from the full waiver of selection to the adoption of only single features, we potentially gain a tool for optimizing generalization performance training without attendant instability.

In this paper, we incorporate selectivity into the Relevance Kernel Machine (RKM) [4,5] and Support Kernel Machine (SKM) [3,5], representing archetypal examples of, respectively, continuous and binary wrapper FS methods. The RKM and SKM are represented as making the same Bayesian decision on the discriminant hyperplane inferred from the training set with differing *a priori* orientation distributions. To achieve the desired selectivity, it hence suffices to substitute the fixed distributions by a respective distribution family, so that a meta-parameter controls the tendency to generate zero components of orientation and thus the rate of suppression of elements in the respective feature/kernel. Increasing the selectivity parameter hence corresponds to decreasing the model complexity. The appropriate selectivity level is to be determined by, for instance, cross validation.

Experimental results with simulated data demonstrate the utility of this approach.

2. The statistical approach to constructing SVMs

Suppose the objects $\omega \in \Omega$ are partitioned into two classes $y(\omega) \in \mathbb{Y} = \{-1, 1\}$, and measured by n features with modality-specific scales $x_i(\omega) \in \mathbb{X}_i$. We also assume a probability distribution in the set of observable feature values and hidden class indices $(x_1(\omega), \dots, x_n(\omega), y(\omega)) \in \mathbb{X}_1 \times \dots \times \mathbb{X}_n \times \mathbb{Y}$, and that training set members $(X, Y) = \{x_{1j}, \dots, x_{nj}, y_j, j=1, \dots, N\}$, $x_{ij} = x_i(\omega_j)$, $y_j = y(\omega_j)$, are sampled independently. Since the kernel-based approach removes the mathematical distinction between different kinds of feature scales, we initially assume all the modality-specific features $x_i(\omega) \in \mathbb{X}_i$ to be real-valued $\mathbb{X}_i = \mathbb{R}$.

This work is supported by the Russian Foundation for Basic Research under grants 08-01-00695 and 06-01-08042, as well as by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 215078.

Let $\varphi(x_1, \dots, x_n | a_1, \dots, a_n, b, y)$ with $y = \pm 1$ be two parametric families of probability densities in the joint feature space $\mathbb{X}_1 \times \dots \times \mathbb{X}_n$ associated with a discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \geq 0$ and concentrated predominantly on opposite sides of it. We shall consider that the improper densities

$$\varphi(x_1, \dots, x_n | a_1, \dots, a_n, b, y) = \begin{cases} \text{const}, & y \left(\sum_{i=1}^n a_i x_i + b \right) > 1, \\ \exp \left[-c \left(1 - y \left(\sum_{i=1}^n a_i x_i + b \right) \right) \right], & y \left(\sum_{i=1}^n a_i x_i + b \right) < 1, \end{cases}$$

$\text{const}=1$, by convention, expresses the assumption that the random feature vectors of both classes of objects are uniformly distributed over their half-spaces, with parameter c controlling the probability of incorrect location.

Let, further, the direction vector (a_1, \dots, a_n) of the discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \geq 0$ be considered as a random vector distributed in accordance with *a priori* density $\Psi(a_1, \dots, a_n | \mu)$ parametrized by μ . No prior information is assumed concerning b , hence, $\Psi(a_1, \dots, a_n, b | \mu) \propto \Psi(a_1, \dots, a_n | \mu)$.

Consequently, the *a posteriori* joint distribution density of the parameters of the discriminant hyperplane w.r.t. the training set is proportional to the product $P(a_1, \dots, a_n, b | X, Y, \mu) \propto \Psi(a_1, \dots, a_n | \mu) \Phi(X | Y, a_1, \dots, a_n, b)$. It is natural to consider the maximum point of this *a posteriori* density as the object of training:

$$(\hat{a}_1, \dots, \hat{a}_n, \hat{b}) = \arg \max \left[\ln \Psi(a_1, \dots, a_n | \mu) + \ln \Phi(X | Y, a_1, \dots, a_n, b) \right].$$

It is easy to show that, under these assumptions, we obtain the training criterion:

$$\begin{cases} -\ln \Psi(a_1, \dots, a_n | \mu) + c \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (1)$$

In particular, if we assume $\Psi(a_1, \dots, a_n | \mu) = \Psi(a_1, \dots, a_n)$ to be the joint normal distribution of independent constituents with zero mathematical expectations and identical variance r , and set $C = 2rc$, we obtain the classical SVM with real-valued features $x_{ij} \in \mathbb{X}_i = \mathbb{R}$ and elements of the direction vector $a_i \in \mathbb{X}_i = \mathbb{R}$ forming a discriminant hyperplane in $\mathbb{X}_1 \times \dots \times \mathbb{X}_n = \mathbb{R}^n$:

$$\begin{cases} \sum_{i=1}^n a_i^2 + C \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (2)$$

In terms of the kernels $K_i(x'_i, x''_i): \mathbb{X}_i \times \mathbb{X}_i \rightarrow \mathbb{R}$ defined in the scales of arbitrary features $x_i \in \mathbb{X}_i$, the classical SVM (2) is formulated as the optimization problem

$$\begin{cases} \sum_{i=1}^n K_i(a_i, a_i) + C \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ y_j \left(\sum_{i=1}^n K_i(a_i, x_{ij}) + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (3)$$

Elements of the direction vector a_i do not exist in the

original feature scales \mathbb{X}_i , but rather in the hypothetical linear spaces $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$ into which the kernels embed them. This does not affect the SVM principle, since at the minimum point $a_i = \sum_{j: \lambda_j > 0} \lambda_j y_j x_{ij} \in \tilde{\mathbb{X}}_i$ the discriminant hyperplane

$$\sum_{j: \lambda_j > 0} \lambda_j y_j \sum_{i=1}^n K_i(x_{ij}, x_i) + b \geq 0 \quad (4)$$

is completely determined by Lagrange multipliers $\lambda_j \geq 0$ at the inequality constraints in (3), namely, by those of them which are positive and define the *support objects*.

In the following two Sections, we consider two versions of the *a priori* distribution $\Psi(a_1, \dots, a_n | \mu)$ resulting in two different feature-selective SVMs, in which the parameter μ will control the desired selectivity.

3. The RKM with supervised selectivity

The direction elements a_i are assumed to be conditionally normally distributed w.r.t. different random variances r_i :

$$\varphi(a_i | r_i) = (1/r_i^{1/2} (2\pi)^{1/2}) \exp(-1/2r_i a_i^2),$$

$$\Psi(a_1, \dots, a_n | r_1, \dots, r_n) \propto \left(\prod_{i=1}^n r_i \right)^{-1/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (1/r_i) a_i^2\right),$$

Let us then consider independent *a priori* gamma distributions of inverse variances $\gamma((1/r_i) | \alpha, \beta) \propto (1/r_i)^{\alpha-1} \exp(-\beta(1/r_i))$ with identical mathematical expectations $E(1/r_i) = \alpha/\beta$ and variances $E((1/r_i)^2) = \alpha/\beta^2$, and set $\alpha = (1+\mu)^2/2\mu$, $\beta = 1/2\mu$. We now have a parametric family of distributions defined only by $\mu \geq 0$, such that $E(1/r_i) = (1+\mu)^2$ and $E((1/r_i)^2) = 2\mu(1+\mu)^2$. If $\mu \rightarrow 0$, values $1/r_i$ approach identity $1/r_i \cong \dots \cong 1/r_n \cong 1$, however, if μ grows, the independent nonnegative values $1/r_i$ may differ arbitrarily. The joint *a priori* distribution of independent inverse variances will be proportional to the product

$$G(r_1, \dots, r_n | \mu) \propto \left(\prod_{i=1}^n 1/r_i \right)^{\alpha-1} \exp\left(-\beta \sum_{i=1}^n (1/r_i)\right).$$

The maximum point of the joint *a posteriori* density $P(a_1, \dots, a_n, b, r_1, \dots, r_n | X, Y, \mu)$, proportional to the product $\Psi(a_1, \dots, a_n | r_1, \dots, r_n) G(r_1, \dots, r_n | \mu) \Phi(X | Y, a_1, \dots, a_n, b)$, is considered as the object of training:

$$\begin{cases} \sum_{i=1}^n \left[(1/r_i) (a_i^2 + (1/\mu)) + ((1/\mu) + 1 + \mu) \ln r_i \right] + \\ \quad C \sum_{j=1}^N \delta_j \rightarrow \min(a_1, r_1, b, \delta_j), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \leq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N, r_i \geq \varepsilon, \end{cases} \quad (5)$$

where $\varepsilon > 0$ is a sufficiently small number. Smaller r_i implies smaller a_i , and the i th feature weakly affects the discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \geq 0$. This is especially apparent in the kernel form

$$\sum_{j: \lambda_j > 0} \lambda_j y_j \sum_{i=1}^n r_i K_i(x_{ij}, x_i) + b \geq 0, \quad (6)$$

obtained by replacing a_i^2 by $K_i(a_i, a_i)$ and $a_i x_{ij}$ by

$K_i(a_i, x_{ij})$. In contrast to the discriminant hyperplane in SVM (4), weights are now assigned to the features.

To solve the optimization problem (5) for a fixed μ , we apply the Gauss-Seidel iteration to the variable groups (a_1, \dots, a_n, b) and (r_1, \dots, r_n) , with initial values $(r_i^0=1, i=1, \dots, n)$. Each iteration, with the current approximations $(r_i^k=1, i=1, \dots, n)$, turns (5) into a slight modification of the usual SVM problem (2). Thus, once the solution (a_1^k, \dots, a_n^k) is found, the revised values of the variances $(r_1^{k+1}, \dots, r_n^{k+1})$ are defined as

$$r_i^{k+1} = \left((a_i^k)^2 + 1/\mu \right) / (1/\mu + 1 + \mu). \quad (7)$$

This procedure typically converges in 10-15 steps, and displays a pronounced tendency to suppress redundant features by allocating, maybe, very small but non-zero weight values r_i in the discriminant hyperplane (6).

The criterion (5) is thus the training principle for Relevance Kernel Machine (RKM) [4,5] with supervised selectivity parametrically determined by $0 \leq \mu < \infty$. If $\mu \rightarrow 0$ all the variances equal unity (7) and we obtain the usual SVM (2). If $\mu \rightarrow \infty$, we have $\sum_{i=1}^n [(1/r_i)(a_i^2 + 1/\mu) + (1/\mu + 1 + \mu) \ln r_i] \rightarrow \min$ in (5), which is a more selective training criterion than the original RKM $\sum_{i=1}^n [(1/r_i)a_i^2 + \ln r_i] + C \sum_{j=1}^N \delta_j \rightarrow \min$ [4].

4. The SKM with supervised selectivity

Now let the a priori density $\Psi(a_1, \dots, a_n | \mu)$ be defined by a convex function $q(a | \mu)$ as

$$\Psi(a_1, \dots, a_n | \mu) \propto \exp\left(-\sum_{i=1}^n q(a_i | \mu)\right).$$

Then, the training criterion (1) will turn into the form

$$\left\{ \begin{array}{l} \sum_{i=1}^n q(a_i | \mu) + c \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j=1, \dots, N. \end{array} \right. \quad (8)$$

For real-valued features $x_i \in \mathbb{R}$, we use the piecewisely linear and quadratic function of the reals $a_i \in \mathbb{R}$:

$$q(a_i | \mu) = 2\mu |a_i| \text{ if } |a_i| \leq \mu, \text{ or } = a_i^2 + \mu^2 \text{ if } |a_i| > \mu.$$

In the case of an arbitrary kernel-represented modality $x_i \in \mathbb{X}_i$, the equivalent function is defined in the linear closure of the respective feature scale $a_i \in \tilde{\mathbb{X}}_i$:

$$q(a_i | \mu) = \begin{cases} 2\mu \sqrt{K_i(a_i, a_i)} & \text{if } \sqrt{K_i(a_i, a_i)} \leq \mu, \\ \mu^2 + K_i(a_i, a_i) & \text{if } \sqrt{K_i(a_i, a_i)} > \mu. \end{cases}$$

The parameter $0 \leq \mu < \infty$ serves as the selectivity parameter of the feature/kernel combination technique. If $\mu=0$, the training problem (8) is the classical SVM $q(a_i | \mu) = \text{const} + a_i^2$ without feature selection ability, and if $\mu \rightarrow \infty$, it becomes the SKM $q(a_i | \mu) \propto \mu |a_i|$ with increasing selectivity as μ grows relative to c , potentially to the point of over-selectivity.

Let us now consider an index variable taking four values $p \in P = \{p^-, p, p^+, p^+\}$ and four respective intervals of the real axis $A_p, p \in P$, intersecting at the boundaries:

$$A_p = \{a: -\infty < a \leq -\mu\} \text{ if } p = p^-, A_p = \{a: -\mu \leq a \leq 0\} \text{ if } p = p^-, \\ A_p = \{a: 0 \leq a \leq \mu\} \text{ if } p = p^+, A_p = \{a: \mu \leq a < \infty\} \text{ if } p = p^+.$$

Vector $\mathbf{p} = (p_1, \dots, p_n)$ decomposes \mathbb{R}^n into 4^n areas $\mathbb{A}_{\mathbf{p}} = \{(a_1, \dots, a_n) \in \mathbb{R}^n: a_i \in A_{p_i}\}$, $\mathbf{p} \in \mathbf{P} = \{p^-, p^-, p^+, p^+\}^n$, some of which have common boundaries.

A preset vector index $\mathbf{p} \in \mathbf{P}$ turns (8) into a quadratic programming problem within the single area $\mathbb{A}_{\mathbf{p}}$

$$\left\{ \begin{array}{l} \sum_{i=1}^n q(a_i | \mu) + c \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j=1, \dots, N, (a_1, \dots, a_n) \in \mathbb{A}_{\mathbf{p}}. \end{array} \right.$$

Its solution $((a_{1,\mathbf{p}}, \dots, a_{n,\mathbf{p}}), b_{\mathbf{p}}, (\delta_{1,\mathbf{p}}, \dots, \delta_{N,\mathbf{p}}))$ can be straightforwardly found via standard computational means, e.g. via, Lagrange multipliers using the dual form. If $(a_{1,\mathbf{p}}, \dots, a_{n,\mathbf{p}})$ is an inner point of $\mathbb{A}_{\mathbf{p}}$, the found solution is that of the entire problem (8). If not, the combination of the boundaries on which it lies points at another area $\mathbb{A}_{\mathbf{p}'}$ with the lesser or, in any case, not greater achievable value if the criterion.

This is the idea of an iterative optimization procedure which provides finding the solution of the convex training problem (8) in a finite number of steps.

We call the training principle (8) the Support Kernel Machine with supervised selectivity. For the given training set, the particular value of $\mu \geq 0$ determines a subset

$\hat{I}_{\mu} \subseteq I = \{1, \dots, n\}$ of support features (kernels) with $a_i^2 > 0$ or, respectively, $K_i(a_i, a_i) > 0$. Only support kernels occur in the discriminant hyperplane

$$\sum_{j: \lambda_j > 0} \lambda_j y_j \sum_{i \in \hat{I}_{\mu}} K_i(x_{ij}, x_i) + b \geq 0,$$

in contrast to the Relevance Kernel Machine which assigns weights to the features (6) but retains all of them.

5. Adjusting the selectivity parameter

The selectivity parameter $0 \leq \mu < \infty$ determines a sequence of nested model classes of diminishing dimensionality applied to the training set, commencing with the usual SVM model. In theoretical terms, the decrease in dimensionality is implicit in the case of RKM but completely explicit for SKM. However, in both techniques the user is not directly or quantitatively aware of how the dimensionality depends on the value of the selectivity parameter.

The most effective method for choosing the value of the selectivity parameter that provides the best generalization performance of training measured is thus cross-validation. The following series of experiments will consequently employ ten-fold cross validation.

6. Results and conclusions

We simulated two entity classes within a hundred-dimensional feature space \mathbb{R}^n , $n=100$, as uniform distributions over two adjoining areas on opposite sides of a fixed hyperplane $\mathbf{a}^T \mathbf{x} \geq 0$ with the direction vector $\mathbf{a} = (a_1=1, a_2=0.8, \dots, a_5=0.2, a_6=0, \dots, a_{100}=0)$, only 5 elements of which differ from zero. Thus, there are five

features carrying diminishing amounts of information concerning the relation between class-membership and feature values, along with 95 confusion features.

The hyper-volumes of the two classes constitute a 100-dimensional cylinder oriented along the direction vector and transversally cut into two identical parts by the actual discriminant hyperplane. The Euclidean length of each of the cylinders and the distance between their common surface and the axis are equal to each other.

We generated a training set $N = N_1 = N_{-1} = 50 + 50 = 100$ and a test set $N_{test} = 50000 + 50000 = 100000$ as a result of independent uniform sampling within the two classes. For a series of increasing values of the selectivity parameter, we then inferred the discriminant hyperplane from the training set in accordance with the RKM (5) and SKM (8) training criteria with supervised selectivity. At each stage, we computed the error rate in the test set (which approximates the ground truth distribution given that N_{test} is large), along with a ten-fold cross-validated estimate of the error rate. The results of experiments are shown in Figure 1. Two indicative benchmarks for the error rate values can be given. First, there is the error rate 0.0045 computed in the problem domain for the discriminant hyperplane obtained by the classical SVM with the correct subset of 5 reasonable features (a_1, \dots, a_5) which is assumed to be known a priori. Secondly, there is the error-rate 0.0245 obtained for the single most informative feature a_1 taken in its own.

If $\mu = 0$, both techniques are seen to be equivalent to the usual SVM with all 100 features, consequently, the respective error rates in the experimental domain have the same value 0.0538. In both cases, the error rate demonstrates a minimum at the point at which feature-information and over-classification effects are balanced, however the morphology of these minima are different for the RKM and SKM models.

The minimum achievable error rate for the RKM 0.0052 is close to the lower benchmark value 0.0045, whereas for the SKM it equals 0.0124, i.e. is more than twice as great. This is presumably a consequence of the finer-balancing of the weighting of combined features in the RKM model in comparison to the hard selection of various training-set-specific feature subsets carried-out by the SKM.

When $\mu \rightarrow \infty$, the limit value of the error rate in the RKM 0.0137, is significantly smaller than the upper benchmark value 0.0245, because this technique only very rarely shrinks the feature subset to the single feature deemed to be individually the most informative for the given training set. At the same time, it is just this that the SKM does as μ becomes very large. Consequently, its asymptotic error-rate equals the upper benchmark value.

Finally, when μ exceeds some critical threshold, both RKM and SKM remove any still remaining features, and the error rate tends to 0.5.

We thus conclude that the error characteristics of the selective RKM and SKM models fall within typical

behavioural patterns, and are thus acceptable for use in general pattern recognition.

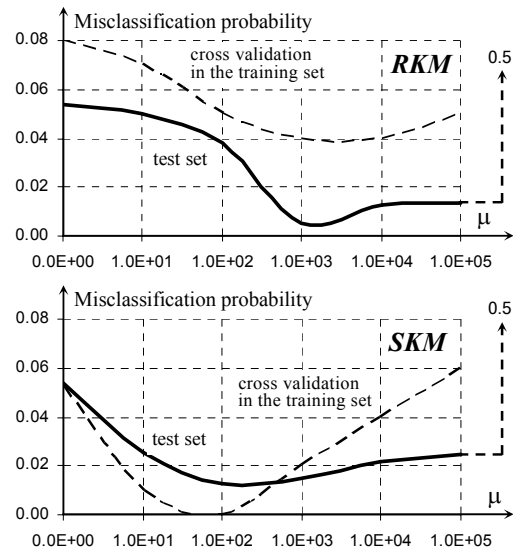


Figure 1. The test-set error rate of the discriminant hyperplane inferred from the training set and result of ten-fold cross validation for increasing values of the selectivity parameter $0 \rightarrow \mu \rightarrow \infty$.

References

- [1] Ross A., Jain A.K. Multimodal biometrics: An overview. *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, 2004. Vienna, Austria, pp. 1221-1224.
- [2] Jannin P, Fleig O.J, Seigneuret E, Grova C, Morandi X, Scarabin J.M. A data fusion environment for multimodal and multi-informational neuronavigation. *Computer Aided Surgery*, 2000, Vol. 5, No. 1, pp. 1-10.
- [3] Sonnenburg S, Rätsch G., Schäfer C. A general and efficient multiple kernel learning algorithm. *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, December 5-8, 2005.
- [4] Sulimova V., Mottl V., Tatarchuk A. Multi-kernel approach to on-line signature verification. *Proceedings of the 8th IASTED International Conference on Signal and Image Processing*. Honolulu, Hawaii, USA, August 14-16, 2006.
- [5] Mottl V., Tatarchuk A., Sulimova V., Krasotkina O., Seregin O. Combining pattern recognition modalities at the sensor level via kernel fusion. *Proceedings of the 7th International Workshop on Multiple Classifier Systems*. Czech Academy of Sciences, Prague, Czech Republic, May 23-25, 2007.
- [6] Guyon I. M., Gunn S. R., Nikravesh M., Zadeh L., Eds. *Feature Extraction, Foundations and Applications*. Springer, 2006.
- [7] Li J., Zha H. Simultaneous classification and feature clustering using discriminant vector quantization with applications to microarray data analysis. *Proceedings of the IEEE Computer Society Bioinformatics Conference*, Palo Alto, CA, August 14-16, 2002, pp. 246-255.