Combining Pattern Recognition Modalities at the Sensor Level via Kernel Fusion

Vadim Mottl, Alexander Tatarchuk [1],
Valentina Sulimova, Olga Krasotkina, Oleg Seredin [2]

[1] Computing Center of the Russian Academy of Sciences
Vavilov St., 40, 117968 Moscow, Russia
vmottl@yandex.ru, aitech@yandex.ru

[2] Tula State University, Lenin Ave. 92, 300600 Tula, Russia
vsulimova@yandex.ru, krasotkina@uic.tula.ru, oseredin@yandex.ru

Abstract. The problem of multi-modal pattern recognition is considered under the assumption that the kernel-based approach is applicable within each particular modality. The Cartesian product of the linear spaces into which the respective kernels embed the output scales of single sensor is employed as an appropriate joint scale corresponding to the idea of combining modalities, actually, at the sensor level. From this point of view, the known kernel fusion techniques, including Relevance and Support Kernel Machines, offer a toolkit of combining pattern recognition modalities. We propose an SVM-based quasi-statistical approach to multi-modal pattern recognition which covers both of these modes of kernel fusion.

Keywords: Kernel-based pattern recognition; support vector machines, combining modalities, kernel fusion

1 Introduction

It is clear that no physical object can be immediately perceived by a computer. As an intermediary between real-world objects $\omega \in \Omega$ and the computer, always serves a formal variable $x(\omega): \Omega \to \mathbb{X}$ which plays the role of some computer-perceptible generalized feature of objects of a certain kind.

The space (scale) $\mathbb{X}$ of the generalized feature may have quite a complicated structure. For instance, the set of biometric traits used for establishing the identity of a person [1] includes face image, fingerprints, off-line and on-line signature, iris and retina images, ear form etc. In medical diagnosis, the typical kinds of information on a patient [2] are, in particular, numerical or nominal results of laboratory tests, X-ray, ultrasonic and MR images or tomograms, electro- and magneto-encephalograms. In public surveys, target properties of population representatives are measured in the form of answers to special-purpose questions each of which produces a specific set of possible response alternatives $\mathbb{X}$.

Any specific type of physical, biological, social or other phenomenon which is considered as characteristic for some real-world objects and expressed by a formal variable is called the specific modality of object representation in data analysis. In terms of the given modality, the original set of objects is substituted for their representations in the value space of an appropriate generalized feature $x(\omega) \in \mathbb{X}$, for instance, in the form of signals, images, questionnaire answers, or, in relatively rare simple situations, real-valued vectors.

The essence of the pattern recognition problem is extension of the information contained in the training set $(X,Y) = \{x(\omega_j), y(\omega_j), j=1,...,N\}$, $y(\omega_j) \in \mathbb{Y} = \{y^{(1)},...,y^{(m)}\}$, onto the entire scale of the respective feature $\hat{y}(x(\omega)): \mathbb{X} \to \mathbb{Y}$. In

many practical cases, no single modality is able to provide the acceptable reliability of recognition. The intent to increase the generalization performance of the resulting recognition rule has led to the concept of multimodal systems $\hat{y}\left(x_1(\omega),...,x_n(\omega)\right)$. In the comprehensive survey of multimodal biometrics [1], three levels of fusing several modalities are considered.

(a) Sensor or data level implies fusion of signals acquired immediately from sensors forming different initial object representations, and final decision making $\hat{y}\left(x(\omega)\right)$ on the basis of some resulting unified feature $x(\omega)=\varphi\left(x_i(\omega), i=1,...,n\right)$.

(b) Classifier score level presupposes fusion of scores of multiple classifiers produced by different modalities to be combined. The score of a single classifier usually has the meaning of posterior probability vector associated with class-membership hypotheses $\left[p^{(1)}\left(x_i(\omega)\right),..., p^{(m)}\left(x_i(\omega)\right); i=1,...,n\right]$. The final decision is to be made from the vector of classifier scores $\hat{y}\left(p_i^{(1)},..., p_i^{(m)}, i=1,...,n\right)$.

(c) Decision level implies fusing final decisions $\hat{y}\left(\hat{y}_i, i=1,...,n\right)$ made separately by single classifiers on the basis of each modality $\left[\hat{y}_i\left(x_i(\omega)\right), i=1,...,n\right]$.

Fusing modalities at the decision level (c) is considered in [1] to be rigid. As to the sensor (a) and classifier score level (b), the latter one is rated in [1] as the most preferable level of fusing several modalities, because the signals of initial sensors are of different physical nature and hardly lend themselves to combination. Therefore, over a long time, the researchers paid the main attention to classifier score fusion [3]. At the same time, it is noted in [1] that the sensor level of fusing modalities might yield better results, if only there were a chance to algorithmize it. The aim of this paper is studying the ways of such algorithmization under the assumption that the kernel-based methodology is applied as a means of inferring a recognition rule for each particular modality.

The essence of the kernel-based methodology [4,5] is expressed by the notion of a kernel function $K(x',x'')$ defined in the output scale of a particular sensor $\mathbb{X}\times\mathbb{X}\to\mathbb{R}$ and meant to be the only means of perceiving real-world objects $\omega\in\Omega$ by pair-wise comparing their generalized features $x'=x(\omega')$, $x''=x(\omega'')$. A two-argument function $K(x',x'')$ is said to be kernel if it forms a positive semidefinite matrix $\left[K\left(x(\omega_j),x(\omega_l)\right), j,l=1,...,N\right]$ for any finite collection of objects. In this case, it embeds the scale of the respective sensor $\mathbb{X}$ into a hypothetical linear space with inner product $\tilde{\mathbb{X}}\supseteq\mathbb{X}$ [6] in which the null element $\phi\in\tilde{\mathbb{X}}$ and linear operations $x'+x'': \tilde{\mathbb{X}}\times\tilde{\mathbb{X}}\to\tilde{\mathbb{X}}$ and $\alpha x: \mathbb{R}\times\tilde{\mathbb{X}}\to\tilde{\mathbb{X}}$ are defined in a special way. The role of the inner product is played by the kernel function $K(x',x'')$, which will be linear with respect to its arguments $K(\alpha'x'+\alpha''x'', x)=\alpha'K(x',x)+\alpha''K(x'',x)$ and produce the Euclidean metric along with Euclidean norm

$$\rho(x',x'')=\sqrt{K(x',x')+K(x'',x'')-2K(x',x'')},\ \|x\|=\rho(x,\phi)=\sqrt{K(x,x)}. \qquad (1)$$

If, at least, one kernel is defined in the output scale of each of several sensors $K_i(x_i',x_i'')$, $x_i',x_i''\in\mathbb{X}_i$, $i=1,...,n$, it appears natural to consider the Cartesian product of the respective linear spaces $\tilde{\mathbb{X}}=\tilde{\mathbb{X}}_1\times...\times\tilde{\mathbb{X}}_n$ and define an appropriate combined kernel (inner product) in it $K(\boldsymbol{x}',\boldsymbol{x}'')$, $\boldsymbol{x}=(x_1,...,x_n)\in\tilde{\mathbb{X}}$. The recent progress in the methodology of kernel fusion [7,8,9,10 has cleared the way for combining any modalities, actually, at the sensor level.

In terms of the kernel-based pattern recognition, the training set has the structure of n matrices of kernel values and N class-indices of objects:

$$(X,Y)=\{\mathbf{x}(\omega_j), y(\omega_j), j=1,...,N\}\Rightarrow\{\mathbf{K}_i, i=1,...,n, y(\omega_j), j=1,...,N\}. \quad (2)$$

In addition, it is required to hold the ability of computing the kernel values $K_i(x_i(\omega), x_i(\omega_j))$, $i=1,...,n$, for any new real-world object $\omega\in\Omega$ and all the objects ω_j represented in the training set. Thus, the burden of dealing with the initial representation of objects in terms of the given modalities completely falls on the kernel computing procedures which are not considered here.

In this paper, like all the above-cited papers, we restrict our consideration only to the two-class patter recognition problem $\mathbb{Y}=\{-1,1\}$. The common idea of all the known kernel fusion methods is the search for a combined kernel as linear combination of the particular ones $K(\boldsymbol{x}',\boldsymbol{x}'')=K(\boldsymbol{x}'',\boldsymbol{x}')=\sum_{i=1}^{n}\alpha_i K_i(x_i',x_i'')$, $\alpha_i\geq 0$. The particular kernel fusion methods differ from each other by the choice of the training criterion which essentially affects the coefficients forming the combined kernel.

The framework proposed in [7] is referred to as Support Kernel Machines (SKM) due to the fact that it produces coefficients $(\alpha_1,...,\alpha_n)$ the most part of which equals zero, so that only the remaining positive coefficients $\alpha_i>0$ indicate the active (support) kernels. However, this framework leads to the dual quadratically-constrained quadratic optimization problem (QCQP), which is essentially more challenging than the standard quadratic programming, or, in more detailed characterization, linearly constrained quadratic optimization problem (LCQP), which underlies the original SVM learning technique.

Another approach [8] leads to coefficients $(\alpha_1,...,\alpha_n)$ which tend to zeros at redundant kernels without complete nulling. We propose here to call kernel fusion techniques of such a kind Relevance Kernel Machines (RKM) due to their analogy with the idea of Relevance Vector Machines (RVM) proposed by Bishop and Tipping in [11] for another purpose, namely, for constructing single-kernel discriminant hyperplanes on a different basis than SVM. The technique proposed in [8] results in the so-called semi-infinite linear programming procedure (SILP), i.e. an algorithm of minimizing a linear function of a finite number of variables under a continuum set of inequality constraints.

The quasi-statistical approach to the problem of kernel fusion we develop in this paper covers both RKM and SKM modes. Like our earlier publication [9,10], the combined kernel is assumed to be simply the sum of the initial kernels $K(\boldsymbol{x}',\boldsymbol{x}'')=\sum_{i=1}^{n}K_i(x_i',x_i'')$. The discriminant hyperplane is sought immediately in the linear space of the combined generalized feature $\tilde{\mathbb{X}}=\tilde{\mathbb{X}}_1\times...\times\tilde{\mathbb{X}}_n$ in the form

$$f(\boldsymbol{x}(\omega))=f(x_1(\omega),...,x_n(\omega))=K(\boldsymbol{\vartheta},\boldsymbol{x}(\omega))+b=\sum_{i=1}^{n}K_i(\vartheta_i,x_i(\omega))+b\gtrless 0. \quad (3)$$

It is well seen that if the norm $\sqrt{K_i(\vartheta_i,\vartheta_i)}$ of a component of the direction vector $\boldsymbol{\vartheta}=(\vartheta_1,...,\vartheta_n)\in\tilde{\mathbb{X}}$ is small in its linear space $\vartheta_i\in\tilde{\mathbb{X}}_i$, the respective kernel $K_i(x_i',x_i'')$ will little affect the recognition rule.

We call the approach quasi-statistical, because the improper densities we use may have no finite integral over the respective space. Our approach leans upon a quasi-probabilistic assumption on the a priori distribution of independent random elements of the direction vector $\boldsymbol{\vartheta}=(\vartheta_1,...,\vartheta_n)\in\tilde{\mathbb{X}}$ in (3). Let $\psi(\vartheta)$ be a basic "spherical" density in some linear space, namely, the density with null mathematical expectation $\mathrm{E}(\vartheta)=\phi$

and mean-square distance from the null element $\mathrm{E}\left(\rho^2(\vartheta,\phi)\right)=1$. Then, if we assume m_i to be the dimensionality of a modality-specific linear space $\tilde{\mathbb{X}}_i$, the density

$$\psi_i(\vartheta_i \mid r_i)=\sqrt{1/r_i^{m_i}}\,\psi\left(\vartheta_i/\sqrt{r_i}\right) \tag{4}$$

will determine, in accordance with (1), the mean-square distance from the respective null element $\mathrm{E}\left(\rho_i^2(\vartheta_i,\phi_i)\right)=\mathrm{E}\left(K_i(\vartheta_i,\vartheta_i)\right)=r_i$.

We consider two kinds of basic a priori densities $\psi(\vartheta)$, namely, the quasi-normal and quasi-Laplace density. The kernel fusion problem is formulated as that of estimating the spatial variances $(r_1,...,r_n)$ along with the elementary direction vectors $(\vartheta_1,...,\vartheta_n)$ from the given training set (2). The normality assumption on the a priori distribution of the hidden direction elements ϑ_i leads to the RKM mode of kernel fusion, and the assumed Laplace distribution results in the SKM fusion mode.

We illustrate the proposed kernel fusion framework by its application to the problem of multi-kernel on-line signature verification.

2 The quasi-statistical approach to kernel fusion

Let $\{x_i(\omega)\in\mathbb{X}_i,\, i=1,...,n\}$ be the given set of generalized features $\boldsymbol{x}(\omega)=\left(x_1(\omega),...,x_n(\omega)\right)\in\mathbb{X}=\mathbb{X}_1\times...\times\mathbb{X}_n$ defined in a universe of real-world objects $\omega\in\Omega$. We shall consider the universe $\omega\in\Omega$ with its genuine partition into two subsets $y(\omega)\in\mathbb{Y}=\{-1,1\}$ as probability space producing a probability distribution in the set of pairs $\left(\boldsymbol{x}(\omega),\, y(\omega)\right)\in\mathbb{X}\times\mathbb{Y}$, and treat the training set $(X,Y)=\left\{\boldsymbol{x}(\omega_j),\, y(\omega_j),\, j=1,...,N\right\}$ as result of repeated independent sampling from this distribution.

It is assumed that some kernel $K_i(x_i',x_i'')$ is defined in each scale $\mathbb{X}_i$, and, so, all the scales as well as its Cartesian product are embedded into respective linear spaces $\tilde{\mathbb{X}}_i\supseteq\mathbb{X}_i$, $\tilde{\mathbb{X}}=\tilde{\mathbb{X}}_1\times...\times\tilde{\mathbb{X}}_n\supseteq\mathbb{X}=\mathbb{X}_1\times...\times\mathbb{X}_n$. So, any choice of a point $\boldsymbol{\vartheta}=(\vartheta_1,...,\vartheta_n)\in\tilde{\mathbb{X}}$ and a real number $b\in\mathbb{R}$ defines a discriminant hyperplane in $\tilde{\mathbb{X}}$ (3).

Let $\boldsymbol{\vartheta}\in\tilde{\mathbb{X}}$ be a fixed point in the Cartesian product of the linear spaces produced by the elementary kernels. If we assume this point as the direction vector of the discriminant hyperplane (3), it will remain to choose the threshold $b\in\mathbb{R}$. Let the accepted training strategy be expressed by the criterion

$$b(\boldsymbol{\vartheta})=\arg\min_{b\in\mathbb{R}} Q(b\mid\boldsymbol{\vartheta},X,Y)\,. \tag{5}$$

The main heuristic idea of our approach is treating the function

$$\Phi(\boldsymbol{\vartheta},b\mid X,Y)=\frac{\exp\left(-u\,Q(b|\boldsymbol{\vartheta},X,Y)\right)}{\int_{\tilde{\mathbb{X}}\times\mathbb{R}}\exp\left[-u\,Q(b'|\boldsymbol{\vartheta}',X,Y)\right]d\boldsymbol{\vartheta}'\,db'}=\frac{1}{D(X,Y)}\exp\left(-u\,Q(b|\boldsymbol{\vartheta},X,Y)\right),$$

where $u>0$ is free parameter, as a posteriori joint distribution density of the direction vector and threshold value under the assumption that no a priori information is available on the direction vector.

In addition, we assume that a priori information is available only on the direction vector $\boldsymbol{\vartheta}\in\tilde{\mathbb{X}}$, so, the respective density will be expressed in the improper form $\Psi(\boldsymbol{\vartheta},b)=\Psi(\boldsymbol{\vartheta})$ which is constant with respect to b. Thus, the a posteriori joint probability density of $\boldsymbol{\vartheta}$ and b will be the product $P(\boldsymbol{\vartheta},b\mid X,Y)=\Psi(\boldsymbol{\vartheta})\Phi(\boldsymbol{\vartheta},b\mid X,Y)=$

$(1/D(X,Y))\Psi(\boldsymbol{\vartheta})\exp\left(-u\,Q(b|\boldsymbol{\vartheta},X,Y)\right)$. Under these assumptions, we obtain the training criterion $(\hat{\boldsymbol{\vartheta}},\hat{b})=\arg\max_{\vartheta\in\tilde{\mathbb{X}},\,b\in\mathbb{R}} P(\boldsymbol{\vartheta},b|X,Y)$, or, in the equivalent form,

$$(\hat{\boldsymbol{\vartheta}},\hat{b})=\arg\min_{\vartheta\in\tilde{\mathbb{X}},\,b\in\mathbb{R}} J(\boldsymbol{\vartheta},b|X,Y),\quad J(\boldsymbol{\vartheta},b|X,Y)=-\ln\Psi(\boldsymbol{\vartheta})+u\,Q(b|\boldsymbol{\vartheta},X,Y). \quad (6)$$

3 The general SVM-based kernel fusion framework

Let the initial threshold-oriented training criterion (5) that occurs in the resulting criterion (6) be taken in the form

$$Q(b|\boldsymbol{\vartheta},X,Y)=\sum_{\substack{j:\,y(\omega_j)=1,\\ K(\boldsymbol{\vartheta},\boldsymbol{x}(\omega_j))+b<1}}\left\{1-\left[K(\boldsymbol{\vartheta},\boldsymbol{x}(\omega))+b\right]\right\}-\sum_{\substack{j:\,y(\omega_j)=-1,\\ K(\boldsymbol{\vartheta},\boldsymbol{x}(\omega_j))+b>-1}}\left\{1+\left[K(\boldsymbol{\vartheta},\boldsymbol{x}(\omega))+b\right]\right\}\to\min_b,$$

or, what is equivalent,

$$\sum\nolimits_j\delta_j\to\min\left(b\in\mathbb{R},\delta_j\in\mathbb{R}\right),\quad y_j\left[K\left(\boldsymbol{\vartheta},\boldsymbol{x}(\omega_j)\right)+b\right]\geq 1-\delta_j,\ \delta_j\geq 0.$$

In this case, in accordance with (6), we come to the following general training criterion:

$$\begin{cases}-\ln\Psi(\boldsymbol{\vartheta})+u\sum\nolimits_j\delta_j\to\min\left(\boldsymbol{\vartheta}\in\tilde{\mathbb{X}},\,b\in\mathbb{R},\,\delta_j\in\mathbb{R}\right),\\ y_j\left[K\left(\boldsymbol{\vartheta},\boldsymbol{x}(\omega_j)\right)+b\right]\geq 1-\delta_j,\ \delta_j\geq 0.\end{cases} \quad (7)$$

This training criterion differs from the usual kernel-based SVM [5] only in two aspects. First, the direction vector of the sought discriminant hyperplane $\boldsymbol{\vartheta}$ is interpreted as element of a hypothetical linear space $\tilde{\mathbb{X}}$ "spanned", by the accepted kernel $K(\boldsymbol{x}',\boldsymbol{x}'')$, over the Cartesian product of the particular scales $\mathbb{X}_i$ of single object representation modalities. Second, the deflection of this vector from the null is penalized by the a priori probability density $-\ln\Psi(\boldsymbol{\vartheta})$ assumed to have the maximum value if $\boldsymbol{\vartheta}=(\vartheta_1,...,\vartheta_n)=(\phi_1,...,\phi_n)$ instead of the usual squared norm $K(\boldsymbol{\vartheta},\boldsymbol{\vartheta})$ (1).

Kernel $K(\boldsymbol{x}',\boldsymbol{x}'')$ is the inner product in the combined linear space $\tilde{\mathbb{X}}=\tilde{\mathbb{X}}_1\times...\times\tilde{\mathbb{X}}_n$ of the object-representation modalities $\boldsymbol{x}(\omega)=\left(x_1(\omega),...,x_n(\omega)\right)$. For instance, any linear combination of the inner products in the particular modalities, i.e. particular kernels, with nonnegative coefficients will produce a combined kernel $K(\boldsymbol{x}',\boldsymbol{x}'')=\sum_{i=1}^n\alpha_i K_i(x_i',x_i'')$. However, in our case there is no need to consider more sophisticated combined kernels than the simple sum $K(\boldsymbol{x}',\boldsymbol{x}'')=\sum_{i=1}^n K_i(x_i',x_i'')$.

We restrict our consideration only to independent a priori distributions of the components in the combined direction vector $\boldsymbol{\vartheta}=(\vartheta_1,...,\vartheta_n)$, assuming that each of them $\psi_i(\vartheta_i\,|\,r_i)$ (4) depends on the specific unknown value of the spatial variance r_i. Then,

$$\Psi(\boldsymbol{\vartheta}\,|\,r_1,...,r_n)=\prod\nolimits_{i=1}^n\psi_i(\vartheta_i\,|\,r_i)=\sqrt{1\Big/\left(\prod\nolimits_{i=1}^n r_i^{m_i}\right)}\prod\nolimits_{i=1}^n\psi\left(\vartheta_i\Big/\sqrt{r_i}\right),$$

where the product in the denominator has the meaning of the concentration volume of this distribution in $\tilde{\mathbb{X}}$. Estimation of the concentration volume makes no sense in this problem, and we shall assume it to be preset, for instance, $\prod_{i=1}^n r_i^{m_i}=1$. So,

$$\Psi(\boldsymbol{\vartheta}\,|\,r_1,...,r_n)=\prod\nolimits_{i=1}^n\psi\left(\vartheta_i\Big/\sqrt{r_i}\right), \quad (8)$$

and we shall have in (7) $-\ln\Psi(\boldsymbol{\vartheta}) = -\sum_{i=1}^{n}\left[-\ln\psi\left(\vartheta_i/\sqrt{r_i}\right)\right]$ under the additional assumption $\sum_{i=1}^{n} m_i \ln r_i = 0$.

The unknown variances r_i of kernel-specific a priori distributions will be of crucial importance for the result of kernel fusion, therefore, we optimize (7) by $(r_1,...,r_n)$ along with other variables. Thus, we shall use the general kernel fusion criterion having the following structure:

$$\begin{cases} \sum_{i=1}^{n}\left[-\ln\psi\left(\vartheta_i/\sqrt{r_i}\right)\right] + u\sum_j \delta(\omega_j) \to \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i,\ r_i \in \mathbb{R}^+,\ b \in \mathbb{R},\ \delta(\omega_j) \in \mathbb{R}^+\right), \\ \sum_{i=1}^{n} m_i \ln r_i = 0, \quad y(\omega_j)\left[\sum_{i=1}^{n} K_i\left(\vartheta_i, x_i(\omega_j)\right) + b\right] \geq 1 - \delta(\omega_j). \end{cases} \tag{9}$$

Any particular kernel fusion technique will be specified only by the choice of the basic a priori density $\psi(\vartheta)$.

Generally speaking, it is problematic to evaluate the dimensionalities m_i of the kernel-specific linear spaces $\tilde{\mathbb{X}}_i$, which may be infinite. But for the given training set (2), the observed dimensionality cannot exceed the number of objects N, and the lower bound of m_i can be estimated from the respective positive semidefinite kernel matrix $\mathbf{K}_i = \left\{K_i\left(x_i(\omega_j), x_i(\omega_k)\right),\ j,k = 1,...,N\right\}$ as the number of its essentially positive eigenvalues.

4 Quasi-normal a priori distributions of modality-specific direction vectors: The Relevance Kernel Machine

For the standard normal density $\psi(\vartheta) = \left(1/\sqrt{2\pi}\right)\exp\left(-(1/2)\|\vartheta\|^2\right)$, we have $\ln\psi\left(\vartheta_i/\sqrt{r_i}\right) =$
$-\ln\sqrt{2\pi} - (1/2r_i)K_i(\vartheta_i,\vartheta_i)$, and the training criterion (9) gets the form with $C = 2u$:

$$\begin{cases} \sum_{i=1}^{n}(1/r_i)K_i(\vartheta_i,\vartheta_i) + C\sum_j \delta_j \to \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i,\ r_i \in \mathbb{R},\ b \in \mathbb{R},\ \delta_j \in \mathbb{R}\right), \\ \sum_{i=1}^{n} m_i \ln r_i = 0, \quad y_j\left[\sum_{i=1}^{n} K_i\left(\vartheta_i, x_i(\omega_j)\right) + b\right] \geq 1 - \delta_j,\ \ \delta_j \geq 0. \end{cases} \tag{10}$$

Theorem 1. *For fixed variances* $(r_i, i = 1,...,n)$, *the combined recognition rule following from the optimization problem (10) has the structure*

$$\hat{f}\left(\boldsymbol{x}(\omega)\right) = \hat{y}\left(x_1(\omega),...,x_n(\omega)\right) = \sum_{j:\hat{\lambda}_j > 0} y_j \hat{\lambda}_j \sum_{i=1}^{n} r_i K_i\left(x_i(\omega_j), x_i(\omega)\right) + \hat{b} \gtrless 0,$$

$$\hat{b} = -\frac{\sum_{j:0<\hat{\lambda}_j<C/2}\hat{\lambda}_j \sum_{l:\lambda_l>0} y_l \hat{\lambda}_l \sum_{i=1}^{n} r_i K_i\left(x_i(\omega_j), x_i(\omega_l)\right) + (C/2)\sum_{j:\hat{\lambda}_j = C/2} y_j}{\sum_{j:0<\hat{\lambda}_j<C/2}\hat{\lambda}_j}. \tag{11}$$

Here Lagrange multipliers $\hat{\lambda}_j$ *at the inequality constraints in (10) are solutions of the dual quadratic programming problem*

$$\begin{cases} \sum_{j=1}^{N}\lambda_j - (1/2)\sum_{j=1}^{N}\sum_{l=1}^{N}\left(y_j y_l \sum_{i=1}^{n} r_i K_i\left(x_i(\omega_j), x_i(\omega_l)\right)\right)\lambda_j\lambda_l \to \max, \\ \sum_{j=1}^{N} y_j\,\lambda_j = 0,\ \ 0 \leq \lambda_j \leq C/2,\ \ j = 1,...,N. \end{cases} \tag{12}$$

The training criterion (10) with fixed variances r_i is a generalization of the well-known SVM criterion [5]. The training-set objects ω_j whose Lagrange multipliers are

positive $\lambda_j > 0$ correspond to active constraints and are called support objects. It is easy to show that the hypothetical constituents $\hat{\vartheta}_i \in \tilde{\mathbb{X}}_i$ of the direction vectors of the optimal discriminant hyperplane in the combined linear space $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times ... \times \tilde{\mathbb{X}}_n$ are linear combinations of the support objects in the sense of specific linear operations produced by the respective kernels:

$$\hat{\vartheta}_i = r_i \sum_{j:\hat{\lambda}_j>0} y_j \hat{\lambda}_j x_i(\omega_j) \in \tilde{\mathbb{X}}_i . \tag{13}$$

However, there is no need to deal with these products of mathematical imagination, because they do not occur in the final formulas (11).

For finding the optimal values variances $(r_i, i = 1,...,n)$, we apply to (10) the Gauss-Seidel iteration with respect to both groups of variables $(\vartheta_i, i = 1,...,n, b)$ and $(r_i, i = 1,...,n)$:

$(r_i^{(k)}, i = 1,...,n) \to \left(\lambda_j^{(k+1)},\ j = 1,...,N\right)$ from (12),

$\left(\lambda_j^{(k+1)},\ j = 1,...,N\right) \to \vartheta_i^{(k+1)} = r_i^{(k)} \sum_{j:\lambda_j^{(k+1)}>0} y_j \lambda_j^{(k+1)} x_i(\omega_j)$ from (13),

$$\left(\vartheta_i^{(k+1)}, i = 1,...,n\right) \to \begin{cases} (r_i^{(k+1)}, i = 1,...,n) = \arg\min \sum_{i=1}^{n} (1/r_i) K_i(\vartheta_i, \vartheta_i), \\ \sum_{i=1}^{n} m_i \ln r_i = 0, \end{cases} \text{ from (10).} \tag{14}$$

It is easy to show that the solution of the problem (14) is expressed by the formulas

$$r_i^{(k+1)} = \frac{K_i(\vartheta_i^{(k+1)}, \vartheta_i^{(k+1)})}{m_i \exp\left\{\left(\sum_{l=1}^{n} m_l \ln\left[(1/m_l) K_l(\vartheta_l^{(k+1)}, \vartheta_l^{(k+1)})\right]\right) \Big/ \sum_{l=1}^{n} m_l\right\}}, \tag{15}$$

where, according to (13),

$$K_i(\vartheta_i^{(k+1)}, \vartheta_i^{(k+1)}) = (r_i^{(k)})^2 \sum_j \sum_l y_j y_l \lambda_j^{(k+1)} \lambda_l^{(k+1)} K_i\left(x_i(\omega_j), x_i(\omega_l)\right)$$

with summation only over the support objects $j : \lambda_j^{(k+1)} > 0$.

It is well seen from (11) that variances r_i occur in the recognition rule as weights at the respective kernels – the greater r_i, the greater the contribution of kernel $K_i(x_i', x_i')$ to the recognition rule. The iterative process of estimating the variances usually converges in 10-15 steps and displays a tendency to suppressing the weights at "redundant" kernels $r_i \to 0$ along with emphasizing $r_i \gg 0$ the kernels which are "adequate" to the trainer's data. We name this training mode the Relevance Kernel Machine, because it results in soft extraction of a relatively small number of most adequate kernels without full suppression of the others.

5 Quasi-Laplace a priori distributions of direction vectors: The Support Kernel Machine

Let us consider now the standard Laplace density $\psi(\vartheta) = (1/2)\exp\left(-(1/2)\| \vartheta \|\right)$. In this case, $\ln \psi\left(\vartheta / \sqrt{r_i}\right) = -\ln 2 - (1/2)\left(1/\sqrt{r_i}\right)\sqrt{K_i(\vartheta_i, \vartheta_i)}$, and we come to the criterion (9) in the following form with $C = 2u$:

$$\begin{cases} \sum_{i=1}^{n} \sqrt{(1/r_i) K_i(\vartheta_i, \vartheta_i)} + C \sum_j \delta_j \to \min\left(\vartheta_i \in \tilde{\mathbb{X}}_i,\ r_i \in \mathbb{R},\ b \in \mathbb{R},\ \delta_j \in \mathbb{R}\right), \\ \sum_{i=1}^{n} m_i \ln r_i = 0, \quad y_j\left[\sum_{i=1}^{n} K\left(\vartheta_i, x_i(\omega_j)\right) + b\right] \geq 1 - \delta_j,\ \delta_j \geq 0. \end{cases} \tag{16}$$

We have come to the formulation of the training problem which differs from the formulation proposed in [7] by the norm of the combined kernel, which is not squared in the criterion (16) in contrast to $\left(\sum_{i=1}^{n} d_i \sqrt{K_i(\vartheta_i,\vartheta_i)}\right)^2$ in the cited paper. We shall see that this distinction results in a considerable simplification of the optimization problem.

Theorem 2. *For fixed variances* $(r_1,...,r_n)$, *the combined recognition rule following from the optimization problem (16) has the structure*

$$\hat{f}\left(\mathbf{x}(\omega)\right) = \hat{y}\left(x_1(\omega),...,x_n(\omega)\right) = \sum_{j:\lambda_j>0} \hat{\lambda}_j y_j \sum_{i\in\hat{I}} \hat{\mu}_i K_i\left(x_i(\omega_j), x_i(\omega)\right) + \hat{b}. \qquad (17)$$

Here $\hat{\lambda}_j$ *are the solutions of the first dual optimization problem*

$$\begin{cases} \sum_{j=1}^{N} \lambda_j \to \max, \quad \sum_{j=1}^{N} \lambda_j y_j = 0, \ 0 \le \lambda_j \le C, \ j = 1,...,N, \\ \sum_{j=1}^{N} \sum_{l=1}^{N} y_j y_l K_i\left(x_i(\omega_j), x_i(\omega_l)\right) \lambda_j \lambda_l \le 1/r_i^2, \ i = 1,...,n, \end{cases} \qquad (18)$$

set $\hat{I} \subseteq \{1,...,n\}$ *is the set of active inequality constraints in (18) at the maximum point,* $\hat{\mu}_i$ *are the solutions of the system of linear equations*

$$\sum_{i\in\hat{I}} \left[\sum_{q:\hat{\lambda}_q>0} \ \sum_{l:0<\hat{\lambda}_l<C} g_l g_q \left(K_i(\omega_q,\omega_j) - K_i(\omega_q,\omega_l)\right) \hat{\lambda}_l \hat{\lambda}_q \right] \mu_i = g_j \sum_{q:0<\hat{\lambda}_q<C} g_q \hat{\lambda}_q - \sum_{q:0<\hat{\lambda}_q<C} \hat{\lambda}_q, \quad j: 0 < \hat{\lambda}_j < C, \qquad (19)$$

and $\hat{b}$ *is determined by the formula*

$$\hat{b} = \frac{\sum_{j:0<\hat{\lambda}_j<C} \hat{\lambda}_j - \sum_{i\in\hat{I}} \left(\sum_{j:\lambda_j>0} \sum_{l:0<\lambda_l<C} g_j g_l K_i(\omega_j,\omega_l) \hat{\lambda}_j \hat{\lambda}_l \right) \hat{\mu}_i}{\sum_{j:0<\hat{\lambda}_j<C} g_j \hat{\lambda}_j} \qquad (20)$$

In accordance with the terminology introduced in [7], the kernels $\left\{K_i(x_i',x_i'), i\in\hat{I}\right\}$ indicated by the subset of active constraints in (18) are the support kernels for the given training set, because only these kernels participate in the recognition rule (17). This is an alternative version of Support Kernel Machine first considered in [7].

The dual optimization problem (18) is that of maximizing a linear function under linear and quadratic constraints. The problems of this class, let us name it Quadratically Constraint Linear Programming (QCLP) is much simpler than those of Quadratically Constraint Quadratic Programming (QCQP) resulting from the framework studied in [7]. The QCLP problems lend themselves to an easy numerical solution by publicly available instruments.

The abstract constituents $\hat{\vartheta}_i \in \tilde{\mathbb{X}}_i$ of the direction vectors of the optimal discriminant hyperplane are linear combinations of the support objects:

$$\hat{\vartheta}_i = \hat{\mu}_i \sum_{j:\hat{\lambda}_j>0} \hat{\lambda}_j y_j x_i(\omega_j) \in \tilde{\mathbb{X}}_i. \qquad (21)$$

This fact is exploited in the Gauss-Seidel iterative procedure which is applied to the training criterion (16) for jointly optimizing it by $(r_i, i=1,...,n)$ and $(\vartheta_i, i=1,...,n, b)$:

$(r_i^{(k)}, i=1,...,n) \to \left(\lambda_j^{(k+1)}, \ j=1,...,N\right)$ and $\hat{I}^{(k+1)} \subseteq \{1,...,n\}$ from (18),

$\left(\lambda_j^{(k+1)}, \ j=1,...,N\right)$ and $\hat{I}^{(k+1)} \subseteq \{1,...,n\} \to \left(\hat{\mu}_i^{(k+1)}, i\in I^{(k+1)}\right)$ from (19),

$\left(\lambda_j^{(k+1)},\ j=1,...,N\right)$ and $\left(\hat{\mu}_i^{(k+1)}, i\in I^{(k+1)}\right)\rightarrow\left(\vartheta_i^{(k+1)},\ i=1,...,n\right)$ from (21),

$\left(\vartheta_i^{(k+1)}, i=1,...,n\right)\rightarrow\begin{cases}(r_i^{(k+1)}, i=1,...,n)=\arg\min\sum_{i=1}^{n}\sqrt{(1/r_i)K_i(\vartheta_i^{(k+1)},\vartheta_i^{(k+1)})},\\ \sum_{i=1}^{n} m_i\ln r_i=0,\end{cases}$ from (16).

The following formulas give the solution of the last optimization problem:

$$\sqrt{r_i^{(k+1)}}=\frac{\sqrt{K_i(\vartheta_i^{(k+1)},\vartheta_i^{(k+1)})}}{m_i\exp\left\{\left(\sum_{k=1}^{n} m_k\ln\left[(1/m_k)\sqrt{K_k(\vartheta_k^{(k+1)},\vartheta_k^{(k+1)})}\right]\right)\Big/\sum_{k=1}^{n} m_k\right\}},$$

where $K_i(\vartheta_i^{(k+1)},\vartheta_i^{(k+1)})=(\hat{\mu}_i^{(k+1)})^2\sum_j\sum_l y_j y_l\lambda_j^{(k+1)}\lambda_l^{(k+1)}K_i\left(x_i(\omega_j),x_i(\omega_l)\right)$ with summation over the support objects.

6 Experiments: Multi-kernel on-line signature verification

The problem of signature verification consists in testing the null hypothesis that the given signature belongs to the person who has claimed his/her identity against the alternative hypothesis that this is forgery. The approach to on-line signature verification presented in [12] is completely based on evaluating one or several kernels on the set of "all feasible" signals that may be produced by the pen's trajectory. Twelve different kernels were simultaneously computed for each pair of signature signals.

In the experiment, we used the database that contains signatures of 40 persons. For each person, the training set consists of 800 signatures, namely, 10 signatures of the respective person, 10 skilled forgeries (attempts to emulate the signature dynamics of this person), and 780 random forgeries formed by 390 original signatures of other 39 persons and 390 skilled forgeries for them. The test set for each person consists of 59 signatures, namely, 10 original signatures, 10 skilled forgeries, and 39 random forgeries. Thus, the total number of the test signatures for 40 persons amounts to 2360.

We tested 13 ways of training, namely, based on each of the initial kernels separately and RKM principle of fusing all the kernels (Section 4). The errors rates of single kernels in the total test set of 2360 signatures range from 0.51% to 23.81%. The RKM kernel fusion technique essentially outperforms each of the single kernels with 0.38% error rate.

For each of 40 persons whose signatures made the data set, the RKM procedure has selected only one relevance kernel which turned out to be most adequate to his/her handwriting. In each case, the relevance kernel obtained nonzero weight $r_i\geq 1.0$, whereas the weights at other kernels were assigned negligibly small values $r_i\leq 10^{-5}$.

7 Conclusions

The kernel-based view of the multi-modal pattern recognition problem stems from the assumption that, at least, one kernel is defined in the output scale of each of several sensors, and, so, each of the scales is embedded into a hypothetical kernel-specific linear space with inner product. The Cartesian product of these linear spaces appears to be just the expedient joint scale corresponding to the idea of combining modalities at the sensor level.

In these terms, the choice of a particular method of multi-modal pattern recognition boils down to the choice of an appropriate kernel in the resulting combined linear space. From this point of view, the known kernel fusion techniques, including Relevance and Support Kernel Machines, offer an appropriate toolkit of combining patter recognition modalities, actually, at the sensor level.

However, it remains open to question whether the sensor level of fusing modalities is really more preferable than the classifier combination level. In the companion paper [13], we set out to show that our approach to combining kernels leads, under some additional assumptions, to a new method of combining kernel-based classifiers, and offers a mathematical basis for comparison of two competing or, maybe, cooperating principles of kernel and classifier fusion.

Acknowledgments. This work is supported by the Russian Foundation for Basic Research, Grants 05-01-00679, 06-01-08042, 06-07-89249, and INTAS Grant 04-77-7347.

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. Comput Aided Surg., 2000, Vol. 5, No. 1, pp. 1-10.
3. Multiple Classifier Systems. Proceedings of the 1st - 6th International Workshops: Lecture Notes in Computer Science, Springer, 2001, 2002, 2003, 2004, 2005.
4. Aizerman M.A., Braverman E.M., Rozonoer L.I. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 1964, Vol. 25, pp. 821-837.
5. Vapnik V. Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.
6. Mottl V. Metric spaces admitting linear operations and inner product. Doklady Mathematics 67(1), 2003, 140–143.
7. Bach F.R., Lankriet G.R.G., Jordan M.I. Multiple kernel learning, conic duality, and the SMO algorithm. Proceedings of the 21th International Conference on Machine Learning, Banff, Canada, 2004.
8. 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.
9. Mottl V., Krasotkina O., Seredin O., Muchnik I. Principles of multi-kernel data mining. Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany, 2005. Lecture Notes in Computer Science, 3587 Springer, 2005, pp. 52-61.
10. Mottl V., Krasotkina O., Seredin O., Muchnik I. Kernel fusion and feature selection in machine learning. Proceedings of the 8th IASTED International Conference on Intelligent Systems and Control. Cambridge, USA, October 31 - November 2, 2005.
11. Bishop C.M., Tipping M.E. Variational relevance vector machines. In: *C. Boutilier and M. Goldszmidt (Eds.)*, Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 2000, pp. 46–53.
12. 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.
13. Windridge D., Mottl V., Tatarchuk A., Eliseyev A. The neutral point method for kernel-based combining disjoint training data in multi-modal pattern recognition. Proceedings of the 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, May 23-25, 2007.