MULTI-KERNEL APPROACH TO ON-LINE SIGNATURE VERIFICATION

Valentina Sulimova
Tula State University, Tula, Russia
vsulimova@yandex.ru

Vadim Mottl, Alexander Tatarchuk
Computing Center of the Russian Academy of Sciences,
Moscow, Russia
vmottl@yandex.ru

ABSTRACT

The problem of on-line signature verification is considered within the bounds of the kernel-based methodology of pattern recognition and, more specifically, SVM principle of machine learning. In accordance with this methodology, any set of on-line signatures as vector signals of individual length is represented by a two-argument function which measures the pairwise similarity between respective signals and possesses the properties of a kernel, i.e. inner product in a hypothetical linear space. Since the SVM principle completely predefines the algorithms of both training and recognition, it remains only to choose a kernel produced by an appropriate metric in the set of signatures, so that the genuine signatures of the same person would be much closer to each other than those of different persons. However, different viewpoints of signature similarity lead, a priori, to different kernels. We propose a principle of fusing several kernels into an entire training and verification technique. Experiments with the data base of the World Signature Verification Competition 2004 have shown that multi-kernel verification essentially decreases the error rate in comparison with decision rules based on single kernels.

KEY WORDS

Signal recognition, on-line signature verification, time warping, kernel function, kernel fusion.

1. Introduction

On-line signatures are vector signals of individual length which carry information on both visual signature patterns and dynamics of writing. In contrast to the off-line approach to signature verification which deals only with the final image scanned from a document, the on-line approach requires the presence of the signer at the time of capture, but, in return, makes use of much greater amount of information [1,2].

Signature verification is a two-class signal recognition problem, in which it is required to check the hypothesis that the given signature belongs to the claimed person. During the more than 20-years long history of studying the problem of on-line signature verification [3], a plenty of ideas have been proposed and tested, practically all of which fall into two groups – feature-based [4] and function-based [5,6] methods. Any method of on-line signature verification, yields, finally, a metric in the set of signature signals.

It is just this aspect of the problem of signature verification which is addressed in this paper. The mathematically most complete methodology of utilizing dissimilarity or, in the inversed form, similarity between entities for allocating them over some classes is the kernel-based approach to pattern recognition [7]. The main notion of this approach is that of the kernel function in the respective set of entities, which is defined as a symmetric two-argument function possessing the property to form positive definite matrix for any finite collection of entities. Such a function defines a metric in the set of entities like inner product in a linear space leads to the respective Euclidean metric, thus, it allows for mentally embedding the original set of entities into a linear space with inner product.

This property of kernel function allows for reformulating practically any of existing methods of on-line signature verification in the kernel-based terms. Such a reformulation suggests a natural way of easily combining several different methods into an entire verification technique which is expected to outperform each of the original ones.

In this paper, we apply the general principles of fusing several kernels worked up in [8,9] to the problem of on-line signature verification. The advantages of the multi-kernel approach to on-line signature verification are demonstrated by experiments with the data base of the World Signature Verification Competition 2004.

2. Metrics and kernels in the set of signals produced by on-line signatures

2.1 Metric in the set of on-line signatures

The standard structure of the vector signal representing an on-line signature $\omega=(\mathbf{x}_s, s=1,...,N)$ includes five components [10]: pen tip coordinates (x_s^{hor}, x_s^{ver}), pen tilt azimuth and altitude (x_s^{az}, x_s^{alt}), and pen pressure x_s^{pr}. The signal is also supplied by time stamp t_s and button status (pen-down, pen-up).

In addition, we supplement the vector signal by two variables computed from the coordinates – pen's velocity x_s^{vel} and acceleration x_s^{acc}. So, we consider on-line signature signals as consisting of seven components.

For comparing pairs of signals $\left[\omega'=(\mathbf{x}'_s, s=1,...,N'),\ \omega''=(\mathbf{x}''_s, s=1,...,N'')\right]$ of different lengths, we use the principle of dynamic time warping [6] with the purpose of aligning the sequences, i.e. bringing them to a common length. Each version of alignment $w(\omega',\omega'')$ is equivalent to a renumbering the elements in both sequences $\omega'_w=(\mathbf{x}'_{w,s'_k}, k=1,...,N_w)$, $\omega''_w=(\mathbf{x}''_{w,s''_k}, k=1,...,N_w)$, $N_w \geq N'$, $N_w \geq N''$.

Let W be the set of all alignments of two signals ω' and ω''. The best alignment $\hat{w}(\omega',\omega'')$ is defined by the condition

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

$$\hat{w}(\omega',\omega'') = \arg\min_{w\in W}\left\{\sum_{k=1}^{N_w} \| \mathbf{x}'_{w,s'_k} - \mathbf{x}''_{w,s''_k} \|^2 + \beta\sum_{k=2}^{N_w}\left(I[s'_{k+1}=s'_k]+I[s''_{k+1}=s''_k]\right)\right\}, \quad (1)$$

where $I[\cdots]$ is indicator function of the respective event, which equals 1 if the condition in brackets is met and 0 otherwise. The first sum in the criterion (1) is the average squared Euclidean distance between corresponding elements of the vector signals, and the second sum imposes penalty β upon each repetition of elements.

Before time warping, each component of each vector signal has to be normalized, for instance, by the mean value and standard deviation:

$\bar{x} = (1/N)\sum_{s=1}^{N} x_s = 0$, $\sigma = (1/N)\sum_{s=1}^{N}(x_s - \bar{x})^2 = 1$.

Once the best alignment (1) is found, it appears natural to consider the value

$$\rho(\omega',\omega'') = \sum_{k=1}^{N_{\hat{w}}} \| \mathbf{x}'_{\hat{w},s'_k} - \mathbf{x}''_{\hat{w},s''_k} \|^2$$

as the degree of dissimilarity between the signals ω' and ω''. It is easy to prove that the two-argument function $\rho(\omega',\omega'')$ is a metric in the set of on-line signatures:

$$\rho(\omega,\omega)=0,\ \rho(\omega',\omega'')\ge 0 \text{ if } \omega'\ne\omega'',\quad \rho(\omega',\omega''')\le\rho(\omega',\omega'')+\rho(\omega'',\omega'''). \quad (2)$$

2.2 Transformation of a metric into radial kernel

Whereas metric (2) evaluates dissimilarity of on-line signatures, function

$$K(\omega',\omega'') = \exp\left[-\alpha\rho^2(\omega',\omega'')\right] \quad (3)$$

will have the sense of their pair-wise similarity. If coefficient α is large enough, this function will form positive semidefinite matrix $\left[K(\omega_i,\omega_j);\ i,j=1,...,N\right]$ for a finite collection of signals, for instance, the training set.

A two-argument function $K(\omega',\omega'')$ defined in a set of real-world entities of arbitrary kind $\omega\in\Omega$ is said to be kernel function in Ω, if it forms positive semidefinite matrices $\left[K(\omega_i,\omega_j); i,j=1,...,m\right]$ for all finite subsets of this set [7]. Any kernel $K(\omega',\omega'')$ embeds the set of entities into a real linear space with inner product $\Omega\subseteq\tilde{\Omega}$, in which the null element $\phi\in\Omega$ and linear operations $\omega'+\omega'': \tilde{\Omega}\times\tilde{\Omega}\to\tilde{\Omega}$ and $c\omega: R\times\tilde{\Omega}\to\tilde{\Omega}$ are defined in a special way, whereas the role of inner product is played by the kernel function itself [11].

This circumstance makes it possible to develop very simple algorithms of pattern recognition for, generally speaking, arbitrary real-world entities by exploiting practically all known methods which had been worked up for linear spaces.

Function (3) is usually called radial kernel function produced by a metric.

2.3 The finite set of kernels studied in experiments

The original metric (2) and kernel (3) produced by it are determined by several characteristics of the signal processing procedure, thus, we have, actually, a family of kernels. We don't pursue here the aim to choose the "most appropriate" metric, which would lead to the kernel providing the best accuracy of signature verification. Our aim is to show advantages of the approach utilizing several metrics (kernels) at once, as against that based on a single predefined metric (kernel).

We study the set of kernels formed by two parameters – first, the subset $X\subseteq X^* = \left\{x_s^{hor}, x_s^{ver}, x_s^{az}, x_s^{alt}, x_s^{pr} x_s^{vel}, x_s^{acc}\right\}$ in the full set of seven components constituting the vector signals of on-line signatures $\mathbf{x}_s = (x_s^i, i\in X)$, and, second, the value of positive penalty β in the time warping criterion (1). Despite the important role played by positive coefficient α in the transformation (3) of a metric into kernel, we use, in this work, the fixed value $\alpha = 0.25$ chosen experimentally and do not vary it.

Six subsets of signal components are studied, each with two values of warping penalty $\beta=10$ and $\beta=20$. So, the full number of kernels amounts to twelve (Table 1).

Table 1. The kernels studied in the experiments.

Kernel		Subset of signal components
$K_1(\omega',\omega'')$ $\beta=10$	$K_2(\omega',\omega'')$ $\beta=20$	pen coordinates $X^*=\left\{x_s^{hor}, x_s^{ver}\right\}$
$K_3(\omega',\omega'')$ $\beta=10$	$K_4(\omega',\omega'')$ $\beta=20$	pen tilt $X^*=\left\{x_s^{az}, x_s^{alt}\right\}$
$K_5(\omega',\omega'')$ $\beta=10$	$K_6(\omega',\omega'')$ $\beta=20$	pen pressure $X^*=\left\{x_s^{pr}\right\}$
$K_7(\omega',\omega'')$ $\beta=10$	$K_8(\omega',\omega'')$ $\beta=20$	coordinates, velocity, acceleration $X^*=\left\{x_s^{hor}, x_s^{ver}, x_s^{vel}, x_s^{acc}\right\}$
$K_9(\omega',\omega'')$ $\beta=10$	$K_{10}(\omega',\omega'')$ $\beta=20$	coordinates, tilt, pressure $X^*=\left\{x_s^{hor}, x_s^{ver}, x_s^{az}, x_s^{alt}, x_s^{pr}\right\}$
$K_{11}(\omega',\omega'')$ $\beta=10$	$K_{12}(\omega',\omega'')$ $\beta=20$	all components $X^*=\left\{x_s^{hor}, x_s^{ver}, x_s^{az}, x_s^{alt}, x_s^{pr} x_s^{vel}, x_s^{acc}\right\}$

3. The support-vector method of training in the linear space of on-line signatures produced by a single kernel

A commonly adopted kernel-based approach to the two-class pattern recognition problem is widely known under the name of Support Vector Machines (SVM) [7].

Let a kernel $K(\omega',\omega'')$ be defined in the set of all signals $\omega\in\Omega$ produced by on-line signatures. Since any kernel embeds the original set of entities into a linear space $\Omega\subset\tilde{\Omega}$ supplied with inner product $K(\omega',\omega'')$ produced by a continuation of the given kernel function, all real-world entities $\omega\in\Omega$ along with hypothetical results of linear operations $\omega'+\omega''$ and $c\omega$ can be considered as vectors in that linear space.

The class of linear functions in $\tilde{\Omega}$, just as in any linear space, is defined by two parameters $\vartheta\in\tilde{\Omega}$ and $b\in\mathbb{R}$

$$y(\omega) = K(\vartheta,\omega)+b,\ \omega\in\tilde{\Omega},\ \text{in particular},\ \omega\in\Omega. \quad (4)$$

We shall call parameter ϑ the direction element of the linear function. The parameters $\vartheta\in\tilde{\Omega}$ and $b\in\mathbb{R}$ determine a classification of the set of signatures into two classes:

$$y(\omega)=K(\vartheta,\omega)+b>0\to\text{class(1)},\ y(\omega)\le 0\to\text{class(-1)}. \quad (5)$$

The only reasonable choice of ϑ will be a linear combination of really existing objects $\hat{\vartheta}=\sum_{j=1}^{N} a_j\omega_j$ in accordance with the linear operations induced in the extended set $\tilde{\Omega}$ by the kernel function $K(\omega',\omega'')$. As being inner product in $\tilde{\Omega}$, the kernel function is linear with respect to its arguments, hence, the linear function resulting from training will include the values of the kernel function only for objects existing in reality $\hat{y}(\omega)=\sum_{j=1}^{N} a_j K(\omega_j,\omega)$.

Let $\Omega^*=\{(\omega_j, g_j);\ j=1,...,N\}$ be a training set of genuine signatures of a client $g_j=1$ and signatures of other people including, maybe, skilled forgeries $g_j=-1$. The SVM principle of training is aimed at finding the optimal discriminant hyperplane (5) which classifies the entities of the training set as precisely as possible by solving the quadratic programming problem [7]

$$\begin{cases} K(\vartheta,\vartheta)+C\sum_{j=1}^{N}\delta_j \to \min(\vartheta,b,\delta_j,\ j=1,...N), \\ g_j y(\omega_j;\vartheta,b)=g_j\left[K(\vartheta,\omega_j)+b\right]\ge 1-\delta_j,\ \delta_j\ge 0,\ j=1,...N. \end{cases} \quad (6)$$

Here $C>0$ is a sufficiently large coefficient providing a trade-off between the intent to have the greatest margin of separating the two classes and the number of errors in the training set.

Such a formulation of the training problem immediately leads to the optimal direction element being linear combination of elements of the training set

$$\hat{\vartheta}=\sum_{j:\lambda_j>0} g_j\lambda_j\omega_j\ , \quad (7)$$

where $\lambda_j\ge 0$ are Lagrange multipliers at inequality constraints in (6), which are solutions of the dual quadratic programming problem

$$\begin{cases} \sum_{j=1}^{N}\lambda_j-(1/2)\sum_{j=1}^{N}\sum_{l=1}^{N}\left[g_j g_l K(\omega_j,\omega_l)\right]\lambda_j\lambda_l \to \max, \\ \sum_{j=1}^{N} g_j\lambda_j=0,\ \ 0\le\lambda_j\le C/2,\ j=1,...,N. \end{cases} \quad (8)$$

As a rule, only a small number of Lagrange multipliers differ from zero $\lambda_j>0$. The respective entities of the training set are called support entities (support vectors, since they are treated as elements of the linear space $\tilde{\Omega}$), because only these elements affect the direction element of the optimal discriminant hyperplane (7). It is easy to see that the constant b is defined by the formula

$$b=\frac{\sum_{j:0<\lambda_j<C/2}\sum_{k=1}^{N} g_k K(\omega_j,\omega_k)\lambda_j\lambda_k+\frac{C}{2}\sum_{j:\lambda_j=C/2} g_j}{\sum_{j:0<\lambda_j<C/2}\lambda_j}. \quad (9)$$

The result of training is completely represented by the finite set of support entities (signatures) being subset of the training set $\tilde{\omega}\in\Omega^{\sup}\subset\Omega^*=\{(\omega_j,g_j);\ j=1,...,N\}$ and Lagrange multipliers at them $\lambda(\tilde{\omega})$. New entities $\omega\in\Omega$ which did not participate in the training set are to be compared with the support entities and classified by the rule

$$y(\omega)=\sum_{\tilde{\omega}\in\Omega^{\sup}} g(\tilde{\omega})\lambda(\tilde{\omega})K(\tilde{\omega},\omega)+b>0 \text{ or } <0. \quad (10)$$

4. Principles of combining several kernels

The idea of combining several kinds of object representation in pattern recognition systems is considered in the literature primarily in context of multimodal establishing the identity of a person. Face, voice, fingerprint, off-line and on-line signatures are examples of different modalities, fusion of which promises to essentially improve the reliability of identification [12].

Different modalities imply capturing different kinds of signals, and, from this point of view, an on-line signature verification system is a unimodal one. However, there is a deep analogy between different physical modalities and different metrics or kernels defined in the same set of signals. It appears reasonable to consider the problem of combining several kernels within the bounds of the same physical modality as combining different informational modalities.

The way of fusing several modalities by combining classifiers independently built for each modality has received primary attention of the pattern recognition community [13]. A technique of fusing classifiers as a way of multi-kernel on-line signature verification is considered below in Section5.

Techniques of fusing several kernels have shown high efficiency in combining several sources of miscellaneous sources of information, but lead to the challenging computational problem of quadratically-constrained quadratic optimization [14,15]. Below, in Section 6. we propose an alternative technique of kernel fusion and apply it to the problem of multi-kernel on-line signature verification.

5. Sum rule of fusing kernel-based classifiers

The schemes of classifier fusion that have been discussed in the literature are very different [13]. A common theoretical framework for combining classifiers developed in [16] is based on the assumption that the output of each particular i th classifier has the form of the a posteriori class-membership probabilities

$$\{p_i(1\,|\,\mathbf{x}_i),...,p_i(m\,|\,\mathbf{x}_i)\},\ \sum_{k=1}^{m} p_i(k\,|\,\mathbf{x}_i)=1,$$

with respect to the entity represented at the input by the feature vector $\mathbf{x}_i$. Under the assumption that each class $k=1,...,m$ is modeled by independent probability distributions in the classifier-specific feature spaces $\mathbf{x}_i$, $i=1,...,n$, the authors investigated various strategies of final decision making, derived the respective rules of combining the outputs of classifiers, and experimentally compared all the rules on several data sets. As a result, it turned out that the so-called sum rule of classifier fusion

$$\hat{k}=\arg\max_{k=1,...,m}\left[(1-n)p(k)+\sum_{i=1}^{n} p_i(k\,|\,\mathbf{x}_i)\right],$$

where $\{p(1),...,p(m)\}$ are a priori probabilities of classes, essentially outperforms all other schemes.

In the two-class situation $k\in\{1,-1\}$, it will be enough to use only one a posteriori probability at the output of each classifier, for instance, $p_i(1\,|\,\mathbf{x}_i)$, and only one a priori probability $p(1)$:

$$y(\mathbf{x}_1,...,\mathbf{x}_n)=\sum_{i=1}^{n} y_i(\mathbf{x}_i)+d\begin{cases} >0\to\hat{k}=1, \\ <0\to\hat{k}=-1, \end{cases} \quad (11)$$

where $y_i(\mathbf{x}_i)=p_i(k\,|\,\mathbf{x}_i)-p(1)$, $d=p(1)-1/2$. If there are no a priori preferences between the classes, we have $p(1)=1/2$ and $d=0$.

It is just this method which we use as the prototype for fusing several kernel-based decision rules of on-line signature verification. Each particular classifier based on the specific kernel $K_i(\omega',\omega'')$ is determined by the kernel-specific set of support signatures along with their Lagrange multipliers $\{\lambda(\tilde{\omega}),\ \tilde{\omega}\in\Omega_i^{\sup}\}$ and produces the score function (10) to be applied to new signatures:

$$y_i(\omega)=\sum_{\tilde{\omega}\in\Omega_i^{\sup}} g(\tilde{\omega})\lambda(\tilde{\omega})K(\tilde{\omega},\omega)+b_i\,. \quad (12)$$

However, this score has no explicit probabilistic meaning. But what allows for immediate summation of the classifier scores in (11) is the equal scale of all the outputs rather than their probabilistic nature itself. In accordance with the SVM principle of training (6), the outputs of the kernel based classifiers are equally scaled, too, in this case, to the interval $(-1,1)$:

$$y_i(\omega)\begin{cases}\geq 1 \rightarrow \hat{g}=1 \text{ with full confidence,}\\ =0 \rightarrow \text{neutral decision } \hat{g}=?,\\ \leq -1 \rightarrow \hat{g}=-1 \text{ with full confidence.}\end{cases}$$

For this reason, to construct a multi-kernel on-line signature verification decision rule on the basis of the classifier fusion principle, we fuse the kernel-specific classifiers by simply summing their scores (12):

$$y(\omega)=\frac{1}{n}\sum_{i=1}^{n} y_i(\omega)\begin{cases}\geq 1 \rightarrow \hat{g}=1, \text{ full–confidence acceptance,}\\ >0 \rightarrow \hat{g}=1, \text{ acceptance,}\\ <0 \rightarrow \hat{g}=1, \text{ rejection,}\\ \leq -1 \rightarrow \hat{g}=-1, \text{ full–confidence rejection.}\end{cases} \quad (13)$$

6. Subset of relevance kernels resulting from kernel fusion

Several recent papers considered the problem of kernel fusion. In this paper, we apply the method proposed in [8] to signals produced by on-line signatures.

This method is essentially underlain by the idea originally proposed in [17] as a means of constructing Relevance Vector Machines (RVM). In accordance with the RVM methodology, in contrast to the principle of Support Vector Machines (SVM) considered in Section 3. the discriminant hyperplane is to be defined not by the entities of the training set occurring in the active inequality constraints in (6) and called support entities, but by a linear combination of, generally speaking, all training set elements, in which, however, the majority of coefficients are close to zero. The remaining entities immediately forming the discriminant hyperplane are called in [17] relevance entities (vectors).

In this paper, following [8], the mathematical approach developed in [17] is used in a completely "transversal" way – not for choosing most appropriate entities in the training set, but for choosing most appropriate kernels. By analogy with [17], we call the kernels selected as result of training the relevance kernels.

In the case of the on-line signature verification problem, the subset of relevance kernels resulting from training will be individual for each person and emphasize the specificity of his/her signatures.

Let $K_i(\omega',\omega'')$, $i=1,...,n$, be the kernel functions defined on the same set of on-line signatures $\omega\in\Omega$. These kernels embed the set Ω into different linear spaces $\Omega\subset\tilde{\Omega}_i$, $i=1,...,n$, with different inner products and, respectively, different linear operations. It is convenient to treat the n linear spaces jointly as Cartesian product

$$\tilde{\tilde{\Omega}}=\tilde{\Omega}_1\times\ldots\times\tilde{\Omega}_n=\left\{\overline{\omega}=<\omega_1,...,\omega_n>:\ \omega_i\in\tilde{\Omega}_i\right\} \quad (14)$$

formed by ordered n-tuples of elements from $\tilde{\Omega}_1,...,\tilde{\Omega}_n$. The kernel function (i.e. inner product) in this linear space can be defined as the sum of the kernel functions (inner products) of the corresponding components in any two n-tuples $\overline{\omega}'=<\omega'_1,...,\omega'_n>$ and $\overline{\omega}''=<\omega''_1,...,\omega''_n>$:

$$K(\overline{\omega}',\overline{\omega}'')=\sum_{i=1}^{n}K_i(\omega'_i,\omega''_i),\ \ \overline{\omega}',\overline{\omega}''\in\tilde{\tilde{\Omega}}. \quad (15)$$

An actual signature signal $\omega\in\Omega$ will be represented by its n-fold repetition $\overline{\omega}=<\omega,...,\omega>\in\tilde{\tilde{\Omega}}$. Then, any real-valued linear function $\Omega\rightarrow\mathbb{R}$ specified by the choice of parameters $\overline{\vartheta}\in\tilde{\tilde{\Omega}}$ and $b\in\mathbb{R}$ will define a discriminant hyperplane in the combined space $\tilde{\tilde{\Omega}}$

$$y(\omega)=\sum_{i=1}^{n}K_i(\vartheta_i,\omega)+b>0\rightarrow\text{acceptance},<0\rightarrow\text{rejection}\ ,(16)$$

where $<\vartheta_1,...,\vartheta_n>\in\tilde{\tilde{\Omega}}$ is a combination of hypothetical elements of particular linear spaces produced by particular kernel functions $K_i(\omega',\omega'')$.

Thus, to define a discriminant function in the set of signatures by combining several kernel functions $K_i(\omega',\omega'')$, we have, first of all, to choose, as parameters, one element in each of linear spaces $\vartheta_i\in\tilde{\Omega}_i$ into which the kernel functions embed the original set $\Omega\subseteq\tilde{\Omega}_i$. It should be marked that the smaller the norm of the ith parameter in its linear space $\|\vartheta_i\|^2=K_i(\vartheta_i,\vartheta_i)$, the lesser the influence of the respective summand on the value of the function (16). If $K(\vartheta_i,\vartheta_i)\rightarrow 0$, i.e. $\vartheta_i\cong\phi_i\in\tilde{\Omega}_i$, the ith kernel function will practically not affect the function.

This means that the parametric family of discriminant functions (16) implies also an instrument of emphasizing "adequate" kernel functions with respect to the available training set and suppressing "inadequate" ones. Which kernel functions should be considered as adequate is the key question for providing a good generalization performance of the decision rule when it is applied to signatures not represented in the training set.

Since the class of discriminant functions incorporating all the kernels is chosen (16), it is possible to apply to the training set of genuine signatures of a client $g_j=1$ and forgeries $g_j=-1$ the same SVM criterion of training (6), but in the combined space $\tilde{\tilde{\Omega}}$, for this time. Such a criterion implies minimization of the squared norm of the combined direction element $\|\overline{\vartheta}\|^2\rightarrow\min$, $\overline{\vartheta}=(\vartheta_1,...,\vartheta_n)\in\tilde{\tilde{\Omega}}$. However, the norm in $\tilde{\tilde{\Omega}}$ may be measured in several ways. In particular, any linear combination of kernel functions with nonnegative coefficients also possesses all the properties of norm $\|\overline{\vartheta}\|^2=\sum_{i=1}^{n}(1/r_i)K_i(\vartheta_i,\vartheta_i)$. In this case, the criterion $\sum_{i=1}^{n}(1/r_i)K_i(\vartheta_i,\vartheta_i)\rightarrow\min$ will try to avoid kernels with small r_i. If $r_i=0$, and the respective kernel will not participate in forming the discriminant function. Thus, the values r_i play the role of weights with which the kernels participate in forming the decision rule.

The idea of adaptive training consists in jointly inferring the direction elements ϑ_i and the weights r_i from the training set by additionally penalizing large weights:

$$\begin{cases}\sum_{i=1}^{n}\left[(1/r_i)K_i(\vartheta_i,\vartheta_i)+\log r_i\right]+\\ \quad C\sum_{j=1}^{N}\delta_j\rightarrow\min(\vartheta_1,...,\vartheta_n,r_1,...,r_n,b,\delta_j,\ j=1,...N),\ (17)\\ g_j\left[\sum_{i=1}^{n}K_i(\vartheta_i,\omega_j)+b\right]\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,...N.\end{cases}$$

This adaptive training criterion displays a pronounced tendency to emphasize the kernels which are "adequate" to the training data and to suppress up to negligibly small values the weights r_i at "redundant" ones. We call relevance kernels those of the initial set of kernel which obtain essentially nonzero values in the result of minimizing the criterion (17).

The reasoning for the adaptive training criterion (17) is a paraphrase, in slightly different terms, of the reasoning for the RVM principle of training [17]. It is based on treating the unknown direction elements $\vartheta_i \in \tilde{\Omega}_i$ in each of the linear spaces $\tilde{\Omega}_i$ as hidden independent random variables whose mathematical expectations coincide with the respective null elements $\text{M}(\vartheta_i) = \phi_i \in \tilde{\Omega}_i$. The parameter r_i has the sense of the unknown mean-square distance of the random direction element from the null element. Then (17) is equivalent to finding the joint maximum-likelihood estimate of the variables $\vartheta_1,...,\vartheta_n$ and their variances $r_1,...,r_n$ under the additional assumption that each direction element ϑ_i is a priori normally distributed in the respective linear space $\tilde{\Omega}_i$.

It can be shown [8] that the following iterative procedure solves the problem (17):

$$\vartheta_i^k = r_i^{k-1} \sum_{j:\lambda_j^k > 0} g_j \lambda_j^k \omega_j , \quad (18)$$

$$r_i^k = (r_i^{k-1})^2 \sum_{j:\lambda_j^k > 0} \sum_{l:\lambda_l^k > 0} K_i(\omega_j, \omega_l) \lambda_j^k \lambda_l^k . \quad (19)$$

At each iteration k, the Lagrange multipliers $\lambda_1^k \geq 0,...,\lambda_N^k \geq 0$ are to be found as the solutions of the dual quadratic programming problem having the same structure as (8):

$$\begin{cases} \sum_{j=1}^N \lambda_j - \frac{1}{2} \sum_{j=1}^N \sum_{l=1}^N \left[g_j g_l \sum_{i=1}^n r_i^k K(\omega_j, \omega_l) \right] \lambda_j \lambda_l \to \max, \\ \sum_{j=1}^N g_j \lambda_j = 0, \ 0 \leq \lambda_j \leq C/2, \ j = 1,...,N. \end{cases}$$

Updating the constant b^k by analogy with (9) does not offer any difficulty. As a rule, the process converges in 10-15 steps.

The abstract variables $\vartheta_i^{(k)} \in \tilde{\Omega}_i$ (18) are linear combinations of the training-set signatures in the sense of linear operations induced by the kernel functions as inner products in the respective linear spaces. Substitution of (18) and (19) into (16) eliminates $\vartheta_i^{(k)}$ and gives the completely constructive estimate of the discriminant function (16) after each iteration:

$$\hat{d}^k(\omega) = \sum_{i=1}^n r_i^{(k-1)} \sum_{j:\lambda_j^{(k)} > 0} g_j \lambda_j^{(k)} K_i(\omega_j, \omega) + b^k > 0 \text{ or } < 0 .$$

7. Structure of experiments

In the experiment, we used the database of the Signature Verification Competition 2004 [18] 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.

Six different metrics were simultaneously measured for each pair of signature signals in accordance with the time warping (Section 2.1 , and, respectively, twelve different kernels were computed which are specified in Table 1 in Section 2.3 .

8. Experimental results

We tested 14 ways of training, namely, based on each of the initial kernels $K_1(\omega', \omega''),..., K_{12}(\omega', \omega'')$ separately (Section 3.), fusion of classifiers resulted from each kernel (Section 5.), and fusion of all kernels (Section 6.). The error rates in the total test set of 2360 signatures are shown in Table 2.

It is well seen that the combined kernel obtained by kernel fusion essentially outperforms each of the single ones and the technique based on the sum rule of classifier fusion.

But as to the sum rule of classifier fusion, this technique ranks essentially below not only the kernel fusion but also some of single kernels. This result is attributable to the very idea of combining classifiers, because this technique essentially rests on the assumption that signal properties perceived by single classifiers are independent of each other. This assumption is quite plausible if completely different modalities are fused, like signatures, face images and voice, but if the difference between the kernels to be combined is caused only by different methods of measuring the pair-wise similarity in the same set of signals, the respective informational modalities cannot be considered as independent.

For each of 40 persons whose signatures made the data set, the procedure of kernel fusion 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 $r_i \leq 10^{-5}$. Table 3 shows the chosen relevance kernels in accordance with the numbering in Table 1.

9. Conclusions

Establishing the identity of a person by his/her on-line signature is inevitably concerned with the necessity of measuring similarity between vector signals of different length produced by signatures. For comparing pairs of on-line signatures, we use the principle of dynamic time warping with the purpose of aligning the signals. The numerical result of signal comparison is represented in the form of a kernel function, what allows us to apply mathematically most advanced methods of pattern recognition to the problem of on-line signature verification.

However, a kernel in the set of vector signals of different length can be generated in different ways, and it is impossible to choose the most appropriate kernel a priori. To overcome this difficulty, we have developed a method of fusing several kernels into an entire on-line signature verification technique.

The experiments on a large data base of on-line signatures has shown that the proposed technique essentially outperforms as classifiers based on single kernels as well as the principle of combining several single-kernel classifiers. More over, the new technique has shown its ability to choose the most relevant kernel that emphasizes the individual specificity of the signature of each person.

Table 2. Error rates in the test set for the single initial kernels versus classifier fusion and kernel fusion.

Single kernels												Classif. fusion	Ker-nel fusion
1	2	3	4	5	6	7	8	9	10	11	12		
Number and percentage of errors in the total amount of 2360 test signatures													
12 0.51 %	16 0.68 %	562 23.81 %	485 20.55 %	65 2.75 %	81 3.43 %	42 17.80 %	18 0.76 %	12 0.51 %	14 0.59 %	19 0.81 %	16 0.68 %	42 1.78 %	9 0.38 %

Table 3. Relevance kernels chosen by the kernel fusion algorithm for each of 40 probationers.

Probationers																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevance kernels (in accordance with numbering in Table 1)																			
12	1	5	5	5	1	1	2	1	2	10	1	10	1	8	1	5	5	11	1
Probationers																			
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Relevance kernels (in accordance with numbering in Table 1)																			
9	1	1	2	5	9	5	1	10	10	1	1	5	2	1	1	2	5	1	1

References

[1] R. Plamondon, G. Lorette. Automatic signature verification and writer identification – the state of the art. *Pattern Recognition, 22*, 1989, 107-131.

[2] R. Plamondon, S. N. Srihari On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. on Pattern Recognition and Machine Intelligence, 22*(1), 2000, 63-84.

[3] H.D. Crane, J.S. Ostrem. Automatic signature verification using a three-axis force-sensitive pen. *IEEE Trans.on Systems, Man and Cybernetics, SMC-13*(3), 1983, 329-337.

[4] J. Richiardi, H. Ketabdar, A. Drygajlo. Local and global feature selection for on-line signature verification. *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, Seoul, South Korea, 2005, 625-629.

[5] J.J. Igarza, I. Goirizelaia, K. Espinosa, I. Hernáez, R. Méndez, J. Sánchez. On-line handwritten signature verification using Hidden Markov Models. *Progress in Pattern Recognition, Speech and Image Analysis*. LNCS Vol. 2905, November 2003, Springer, 391-399.

[6] R.Martens, L. Claesen. Dynamic programming optimisation for on-line signature verification. *Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97)*, Ulm, Germany, 1997, 2, 653-656.

[7] V. Vapnik. *Statistical Learning Theory* (John-Wiley & Sons, Inc., 1998),

[8] V. Mottl, O. Seredin, O. Krasotkina, I. Muchnik. Fusion of potential functions in reconstructing dependences from empirical data. *Doklady Mathematics, 71*(2), 2005, 315-319.

[9] V. Mottl, O. Krasotkina, O.Seredin, I. Muchnik. Principles of multi-kernel data mining. *Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition*, Leipzig, 2005, 52-61.

[10] D. Sakamoto, H. Morita, T. Ohishi. Y. Komiya T. Matsumoto. On-line signature verification algorithm incorporating pen position, pen pressure and pen inclination trajectories. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, 2001, Vol. 2, 993-996.

[11] V. Mottl. Metric spaces admitting linear operations and inner product. *Doklady Mathematics 67*(1), 2003, 140–143.

[12] A. Ross, A.K. Jain. Multimodal biometrics: An overview. *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2004, 1221-1224.

[13] *Multiple Classifier Systems (Roli F., Kittler J., Eds). Proceedings of the Third International Workshop (MCS)*, Calgary, Italy, 2002.

[14] G.R.G. Lanckriet, Cristianini N., Ghaoui L.E., Bartlett P., Jordan M.I. Learning the kernel matrix with semidefinite programming. *J. Machine Learning Research*, 5, 2004, 27–72.

[15] F.R. Bach, G.R.G. Lankriet, M.I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. *Proceedings of the 21th International Conference on Machine Learning*, Banff, Canada, 2004.

[16] J. Kittler, M. Hatef, R. Duin, J. Matas. On combining classifiers. *IEEE Trans. on PAMI, 20*(3), 1998, 226-239.

[17] C.M. Bishop, M.E. Tipping. Variational relevance vector machines. In: *C. Boutilier and M. Goldszmidt (Eds.), Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, 2000, 46–53.

[18] SVC 2004: First International Signature Verification Competition. http://www.cs.ust.hk/svc2004/index.html