Supervised Selective Combining Pattern Recognition Modalities and its Application to Signature Verification by Fusing On-Line and Off-Line Kernels

Alexander Tatarchuk[1], Valentina Sulimova[2], David Windridge[3], Vadim Mottl[1] and Mikhail Lange[1]

[1] Computing Center of the Russian Academy of Sciences,
Moscow, Russia
[2] Tula State University, Tula, Russia
[3] Centre for Vision, Speech and Signal Processing,
University of Surrey,Guildford, UK

Abstract. 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 pattern recognition. We consider the problem of multi-modal pattern recognition 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, in contrast to the commonly adopted level of combining classifiers inferred from each specific modality. In this paper, to avoid overfitting, we set out a family of related stochastic methods for encompassing modal-selectivity that are intrinsic to the chosen kernel-based pattern-recognition approach. The principle of kernel selectivity supervision is applied to the problem of signature verification by fusing several on-line and off-line kernels into an entire training and verification technique.

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 $\big(x_i(\omega'), x_i(\omega'')\big)$ using modality-specific functions $K_i(x'_i, x''_i)$ defined in the output scales of the sensors $\mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$. 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. Depending on the set of kernel values and the characteristics of the kernel, this embedding space can be of a significantly different dimensionality that of the original measurement domain.

In consequence of its pairwise nature, kernel-based multi-modal pattern recognition presents a number of difficulties over canonical pattern-recognition. In particular, the problem of the composition and selection of feature modalities becomes acute, since we cannot simply assume the Euclidean vectorisablity of composite data without explicit construction of a kernel in the composite space. This problem is further compounded by the potential presence of training data that is not equally represented within each modality - as sometimes occurs in census returns, or in independently-trained classification systems, for example, in multimodal biometrics[1].

However, when $x_i(\omega) \in \mathbb{X}_i = \mathbb{R}$, the kernel defined by the product $K_i(x_i', x_i'') = x_i' x_i''$ generates an appropriate and natural embedding of the multimodal data. 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 [3–5].

Despite the improved resilience of the SVM approach to over-fitting by virtue of its adjustment of capacity to the requirements of hyperplane description, it is often still 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 of training without attendant instability.

In this paper, we incorporate selectivity into the Relevance Kernel Machine (RKM) [4, 5] representing an archetypal example of a continuous wrapper FS method. The RKM is 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

[1] This missing data issue also occurs, albeit less acutely, in standard pattern recognition: the reason for its particularly problematic nature in kernel-based pattern-recognition is the inability to construct an embedding space when presented with an incomplete kernel Gram matrix w.r.t all of the measured objects.

complexity. The appropriate selectivity level is to be determined by, for instance, cross validation.

The Relevance Kernel Machine with supervised selectivity is applied here to the problem of signature verification which consists in testing the hypothesis that a given signature belongs to the person having claimed his/her identity. Depending on the initial data representation, it is adopted to distinguish between on-line and off-line signature verification [8]. Any method of signature verification is based, finally, on a metric or kernel in the set of signatures. The selective kernel fusion technique considered in this paper serves as a natural way of easily combining on-line and off-line methods into an entire signature verification procedure. Experiments with signature database SVC2004 have shown that the multi-kernel approach essentially decreases the error rate in comparison with verification based on single kernels.

2 A Bayesian strategy for determining the discriminant hyperplane

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 $\big(x_1(\omega), ..., x_n(\omega), y(\omega)\big) \in \mathbb{X}_1 \times ... \times \mathbb{X}_n \times \mathbb{Y}$, and that training set members $(X, Y) = \{x_{1j}, ..., x_{nj}, y_j,\ j = 1, ..., N\}$, $x_{ij} = x_i(\omega_j)$, $y_j = y(\omega_j)$, are sampled independently. Since the kernel-based approach removes the mathematical distinction between different kinds of feature scales, we assume all the modality-specific features $x_i(\omega) \in \mathbb{X}_i$ to be real-valued $\mathbb{X}_i = \mathbb{R}$.

Let $\varphi_1\,(x_1, ..., x_n \,|\, a_1, ..., a_n\,, b, y)$ with $y = \pm 1$ be two parametric families of probability densities in the joint feature space $\mathbb{X}_1 \times ... \times \mathbb{X}_n$ associated with a discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \gtrless 0$ and concentrated predominantly on opposite sides of it. We shall consider that the improper densities

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

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

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 $P(a_1, ..., a_n, b \,|\, X, Y, \mu) \propto \Psi(a_1, ..., a_n \,|\, \mu) \times \Phi(X \,|\, Y, a_1, .., a_n, b)$. It is natural to consider the maximum point of this *a posteriori* density as the object of training:

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

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

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

In particular, if we assume $\mathit{\Psi}(a_1, ..., a_n \,|\, \mu) = \mathit{\Psi}(a_1, ..., a_n)$ 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 $\mathbb{X}_1 \times ... \times \mathbb{X}_n = \mathbb{R}^n$:

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

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

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

In the general case, elements of the direction vector a_i do not exist in the original feature scales $\mathbb{X}_i$, but belong rather to the hypothetical linear closures $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$ 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 applicable to any new point $(x_i \in \mathbb{X}_i,\ i = 1, \ldots, n)$

$$\sum\nolimits_{j:\ \lambda_j > 0} \lambda_j y_j \sum\nolimits_{i=1}^{n} K_i(x_{ij}, x_i) + b \gtrless 0 \tag{4}$$

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*. To find the Lagrange multipliers, it is enough to solve the well-known dual quadratic-programming problem:

$$\begin{cases} \sum\limits_{j=1}^{N} \lambda_j - (1/2) \sum\limits_{j=1}^{N} \sum\limits_{l=1}^{N} y_j y_l \Big(\sum\limits_{i=1}^{n} K_i(x_{ij}, x_{il}) \Big) \lambda_j \lambda_l \to \max, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0,\ 0 \leq \lambda_j \leq C/2,\ j = 1, ..., N. \end{cases} \tag{5}$$

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

3 The continuous training technique with supervised selectivity

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

$$\begin{array}{l} \psi(a_i \,|\, r_i) = \left(1\Big/r_i^{1/2}(2\pi)^{1/2}\right) \exp\left(-(1/2r_i)a_i^2\right), \\ \Psi(a_1, ..., a_n \,|\, r_1, ..., r_n) \propto \\ \left(\prod_{i=1}^n r_i\right)^{-1/2} \exp\Big(-(1/2)\sum_{i=1}^n (1/r_i)a_i^2\Big). \end{array}$$

Let us then consider independent a priori gamma distributions of inverse variances $\gamma\big((1/r_i)\,|\,\alpha,\beta\big) \propto (1/r_i)^{\alpha-1}\exp\left(-\beta\,(1/r_i)\right)$ with identical mathematical expectations $E(1/r_i) = \alpha/\beta$ and variances $E\left((1/r_i)^2\right) = \alpha/\beta^2$, 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 $E\left((1/r_i)^2\right) = 2\mu(1+\mu)^2$. If $\mu \to 0$, values $1/r_i$ approach identity $1/r_i \cong ... \cong 1/r_n \cong 1$, 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

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

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

In the case of real-valued features $x_i \in \mathbb{R}$, the resulting training criterion has the form

$$\begin{cases} \sum_{i=1}^n \left[(1/r_i)\left(a_i^2 + (1/\mu)\right) + ((1/\mu)+1+\mu)\ln r_i\right] + \\ \qquad\qquad C\sum_{j=1}^N \delta_j \;\to \min\left(a_i \in \mathbb{R}, r_i, b, \delta_j\right), \\ y_j\left(\sum_{i=1}^n a_i x_{ij} + b\right) \geq 1 - \delta_j,\; \delta_j \geq 0,\; j = 1, ..., N, \\ \qquad\qquad\qquad\qquad r_i \geq \varepsilon, \end{cases} \tag{6}$$

where $\varepsilon > 0$ is a sufficiently small number. Smaller r_i implies smaller a_i, and the ith feature weakly affects the discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \gtrless 0$.

Replacing a_i^2 by $K_i(a_i, a_i)$ and $a_i x_{ij}$ by $K_i(a_i, x_{ij})$ in (6), yields the analogous training criterion for kernel-represented modalities $x_i \in \mathbb{X}_i$:

$$\begin{cases} \sum_{i=1}^n \left[(1/r_i)\big(K_i(a_i, a_i) + (1/\mu)\big) + \right. \\ \left.\big((1/\mu) + 1 + \mu\big)\ln r_i\right] + C\sum_{j=1}^N \delta_j \to \min\limits_{a_i \in \tilde{\mathbb{X}}_i, r_i, b, \delta_j}, \\ y_j\big(\sum_{i=1}^n K_i(a_i, x_{ij}) + b\big) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \ldots, N, r_i \geq \varepsilon. \end{cases} \tag{7}$$

If the weights r_i are fixed, there is no need to evaluate the real numbers $a_i \in \mathbb{R}$ in (6) or the abstract elements of linear closures $a_i \in \tilde{\mathbb{X}}_i$ in (7), it is enough to find the Lagrange multipliers $\lambda_j \geq 0$ in the representation $a_i = r_i \sum_{j:\, \lambda_j > 0} y_j \lambda_j x_{ij}$ by solving the dual quadratic-programming problem which is a slight modification of (5):

$$\begin{cases} \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l \big(\sum_{i=1}^{n} r_i K_i(x_{ij}, x_{il})\big)\lambda_j \lambda_l \to \max, \\ \sum_{j=1}^{N} y_j \lambda_j = 0,\ 0 \le \lambda_j \le C/2,\ j = 1, ..., N. \end{cases} \tag{8}$$

In the kernel form

$$\sum\nolimits_{j:\ \lambda_j > 0} y_j \lambda_j \sum\nolimits_{i=1}^{n} r_i K_i(x_{ij}, x_i)\ +\ b \gtrless 0 \tag{9}$$

of the discriminant hyperplane $\sum_{i=1}^{n} K_i(a_i, x_i) + b \gtrless 0$, in contrast to the discriminant hyperplane in SVM (4), weights are now assigned to the features, so that small r_i suppress the respective features.

However, the weights are unknown in (7). To solve this optimization problem for a fixed μ, we apply the Gauss-Seidel iteration to the variable groups $(a_1, ..., a_n, b, \delta_1, \dots, \delta_N)$ and $(r_1, ..., r_n)$, with initial values $(r_i^0 = 1,\ i = 1, ..., n)$. Once the solution $\lambda_1^k, \dots, \lambda_N^k$, i.e. $(a_1^k, ..., a_n^k)$, is found at the kth iteration with the current approximations $(r_1^k, \dots, r_n^k)$, the revised values of the variances $(r_1^{k+1}, ..., r_n^{k+1})$ are defined as

$$\begin{aligned} & r_i^{k+1} = \tilde{r}_i^{k+1} \ \text{ if } \ \tilde{r}_i^{k+1} \ge \varepsilon,\ r_i^{k+1} = \varepsilon \ \text{ otherwise}, \\ & \tilde{r}_i^{k+1} = \frac{(a_i^k)^2 + 1/\mu}{1/\mu + 1 + \mu} = \\ & \frac{\sum_{j:\lambda_j^k > 0} \sum_{l:\lambda_l^k > 0} y_j y_l \, (r_i^k)^2 K_i(x_{ij}, x_{il}) \lambda_j^k \lambda_l^k + 1/\mu}{1/\mu + 1 + \mu}. \end{aligned} \tag{10}$$

This procedure typically converges in 10-15 steps, and displays a pronounced tendency to suppress redundant features by allocating very small but non-zero weights r_i in the discriminant hyperplane (9).

The criterion (6) is thus the training principle for Relevance Kernel Machine (RKM) [4, 5] with supervised selectivity parametrically determined by $0 \le \mu < \infty$. If $\mu \to 0$ all the variances equal unity (10), and we obtain the usual SVM (2). If $\mu \to \infty$, we have $\sum_{i=1}^{n} \big[(1/r_i) a_i^2 + (1+\mu) \ln r_i\big] + C \sum_{j=1}^{N} \delta_j \to$ min in (6), which is a more selective training criterion than the original RKM $\sum_{i=1}^{n} \big[(1/r_i) a_i^2 +\ \ln r_i\big]\ +\ C \sum_{j=1}^{N} \delta_j\ \to$ min [4].

4 Signature verification via selective fusion of on-line and off-line kernels

4.1 Kernels produced by metrics

Let ω' and ω'' be two signatures represented by signals or images, and $\rho(\omega', \omega'')$ be a metric evaluating dissimilarity of signatures from a specific point of view. Then function

$$K(\omega', \omega'') = \exp\big[-\gamma\, \rho^2(\omega', \omega'')\big] \tag{11}$$

has the sense of their pair-wise similarity. If coefficient $\gamma > 0$ is large enough, this function will be a kernel in the set of signatures, usually called the radial kernel.

As a rule, it is impossible to know in advance which of possible metrics is more appropriate for a concrete person. The advantages of the multi-kernel approach to the problem of on-line signature verification were demonstrated in [4]. We extend here the kernel-based approach onto the problem of combining the on-line and off-line modalities (Figure 1) into an entire signature verification technique.

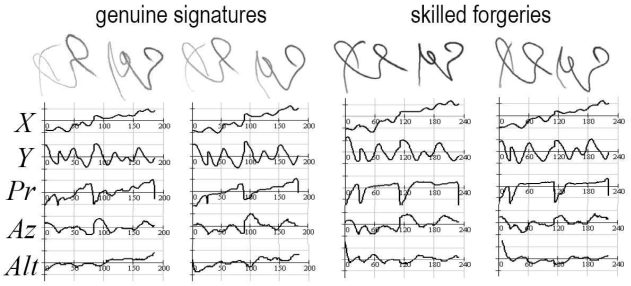


Fig. 1. Off-line (images) and on-line (signals) representation of signatures.

In this work, we tested 12 different metrics in the set of on-line signatures and 4 metrics computed from the pictorial off-line representation. So, all in all, we combined 16 different on-line and off-line kernels listed in Table 1.

Table 1. The kernels studied in the experiments

	β = 10	β = 20	Subset of components
On-line kernels	K_1	K_2	pen coordinates
	K_3	K_4	pen tilt (azimuth and altitude)
	K_5	K_6	pen pressure
	K_7	K_8	coordinates, velocity,
	K_9	K_{10}	coordinates, tilt, pressure
	K_{11}	K_{12}	all seven components

Off-line kernels	K_{13}	configuration of primitives positions
	K_{14}	configuration of primitives orientation and size
	K_{15}	configuration of primitives brightness
	K_{16}	uniform mixture of three configurations

4.2 Metrics in the set of on-line signatures.

Each on-line signature is represented by a multi-component vector signal which initially includes five components $\mathbf{x}_t = (x_t^1 \cdots x_t^n)$: two pen tip coordinates (X, Y), pen tilt azimuth (Az) and altitude (Alt), and pen pressure (Pr) (Fig. 1). We supplement the signals with two additional variables - pen's velocity and acceleration.

For comparing pairs of signals of different lengths $[\,\omega' = (\mathbf{x}'_s, s = 1, \ldots, N')$, $\omega'' = (\mathbf{x}''_s, s = 1, \ldots, N'')]$, we use the principle of dynamic time warping with the purpose of aligning the vector sequences [4]. Each version of alignment $w(\omega', \omega'')$ is equivalent to a renumbering of the elements in both sequences $\omega'_w = (\mathbf{x}'_{w,s'_k},\ k = 1, \ldots, N_w)$, $\omega''_w = (\mathbf{x}''_{w,s''_k},\ k = 1, \ldots, N_w)$, $N_w \geq N'$, $N_w \geq N''$. We tested 12 different metrics defined by 6 different subsets of signal components and 2 different values of the alignment rigidity parameter β [4] as shown in Table 1:

$$\rho(\omega', \omega''|\beta) = \min_w \sqrt{\sum\nolimits_{k=1}^{N_w} \|\mathbf{x}'_{w,s'_k} - \mathbf{x}''_{w,s''_k}\|^2}. \qquad (12)$$

4.3 Metrics in the set of off-line signatures

For comparing grayscale images (patterns) representing off-line signatures we apply the technique of tree-structured pattern representation proposed in [10].

For the given pattern P, the recursive scheme described in [10] produces a pattern representation R in the form of a complete binary tree of elliptic primitives (nodes) Q: $R = \{Q_n : 0 \leq n \leq n_{\max}\}$, where n is the node number of the level $l_n = \lfloor \log_2(n = 1) \rfloor$.

Let R' and R'' be a pair of tree-structured representations, and $R' \bigcap R''$ be their intersection formed by the pairs of nodes (Q'_n, Q''_n) having the same number n. For comparing any two corresponding nodes $Q'_n \in R'$ and $Q''_n \in R''$, a dissimilarity function $d(Q'_n, Q''_n) \geq 0$ can be easily defined through parameters of each primitive such as center vector, orientation vectors with their sizes (along two principal axes of the primitive), and the mean brightness value. Using these parameters, we define a loss function

$$D(Q'_n, Q''_n) = \begin{cases} d(Q'_n, Q''_n), \text{ if } Q'_n \text{ and/or } Q''_n \text{ are "end" nodes,} \\ 0, \text{ otherwise,} \end{cases}$$

where $d(Q'_n, Q''_n) = \alpha_1 d_1(Q'_n, Q''_n) + \alpha_2 d_2(Q'_n, Q''_n) + \alpha_3 d_3(Q'_n, Q''_n)$, $\alpha_1, \alpha_2, \alpha_3 \geq 0$, $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Here, $d_i(Q'_n, Q''_n)$ is a distinction function between the centers of the primitives, their orientation and size parameters, and the mean brightness values for $i = 1, 2, 3$, respectively.

Then, following [10], we define the distinction measure (metric) of the trees R' and R'' as follows:

$$\begin{aligned} &\rho(R', R'' \mid \alpha_1, \alpha_2, \alpha_3) = \textstyle\sum_{R' \bigcap R''} 2^{-l_n} D(Q'_n, Q''_n) = \\ &\alpha_1 \textstyle\sum_{R' \bigcap R''} 2^{-l_n} d_1(Q'_n, Q''_n) + \\ &\alpha_2 \textstyle\sum_{R' \bigcap R''} 2^{-l_n} d_2(Q'_n, Q''_n) + \\ &\alpha_3 \textstyle\sum_{R' \bigcap R''} 2^{-l_n} d_3(Q'_n, Q''_n), \end{aligned} \qquad (13)$$

where the sum is taken over all pairs $(Q'_n, Q''_n) \in R' \bigcap R''$.

We competitively applied three basic distinction measures of the form (13) $\rho_1(R', R'') = \rho(R', R'' \mid 1, 0, 0)$, $\rho_2(R', R'') = \rho(R', R'' \mid 0, 1, 0)$, $\rho_3(R', R'') = \rho(R', R'' \mid 0, 0, 1)$, and the uniform mixture $\rho_4(R', R'') = \rho(R', R'' \mid 1/3, 1/3, 1/3)$.

4.4 Signature database and results of experiments

In the experiment, we used the database of the Signature Verification Competition 2004 [11] that contains vector signals of 40 persons (Fig. 1). On the basis of these signals we generated grayscale images (256 × 256 pixels) with 256 levels of brightness corresponding to the levels of pen pressure in the original signals.

For each person, the training set consists of 400 signatures, namely, 5 signatures of the respective person, 5 skilled forgeries, and 390 random forgeries formed by 195 original signatures of other 39 persons and 195 skilled forgeries for them. The test set for each person consists of 69 signatures, namely, 15 genuine signatures, 15 skilled forgeries, and 39 random forgeries. Thus, the total number of the test signatures for 40 persons amounts to 2760.

For each pair of signature signals, 12 different on-line metrics and 4 off-line metrics were simultaneously computed and, respectively, 16 different kernels were evaluated (Table 1).

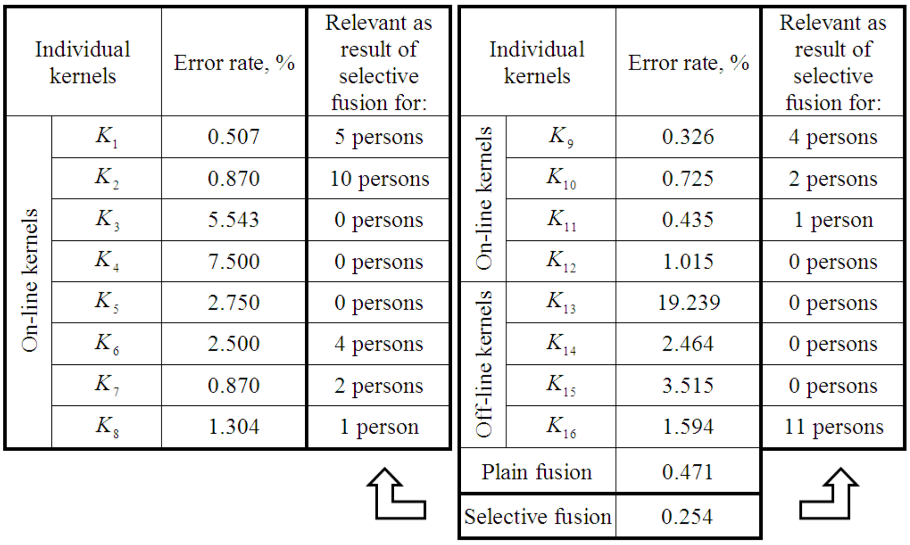
For each person, we tested 18 ways of training based, first, on each of the initial kernels separately $\{K_1(\omega', \omega''), \ldots, K_{16}(\omega', \omega'')\}$, second, on the plane fusion of all the individual kernels with equal weights $(1/16) \sum_{i=1}^{16} K_i(\omega', \omega'')$, and, third, on the selective fusion of all the 16 kernels using the continuous training technique (Section 3) with the selectivity level chosen via cross validation. The error rates in the total test set of 2760 signatures are shown in Table 2.

It is well seen that the combined kernel obtained by selective kernel fusion with individually chosen selectivity essentially outperforms each of the single ones. At the same time, for each of 40 persons whose signatures made the data set, the kernel fusion procedure has selected only one relevant kernel as the most adequate representation of his/her handwriting.

5 Conclusions

The kernel-based approach to signature verification enables harnessing the kernel-selective SVM as one of mathematically most advanced methods of pattern

Table 2. Error rates for single kernels versus kernel fusion



	Individual kernels	Error rate, %	Relevant as result of selective fusion for:
On-line kernels	K_1	0.507	5 persons
	K_2	0.870	10 persons
	K_3	5.543	0 persons
	K_4	7.500	0 persons
	K_5	2.750	0 persons
	K_6	2.500	4 persons
	K_7	0.870	2 persons
	K_8	1.304	1 person
On-line kernels	K_9	0.326	4 persons
	K_{10}	0.725	2 persons
	K_{11}	0.435	1 person
	K_{12}	1.015	0 persons
Off-line kernels	K_{13}	19.239	0 persons
	K_{14}	2.464	0 persons
	K_{15}	3.515	0 persons
	K_{16}	1.594	11 persons
	Plain fusion	0.471	
	Selective fusion	0.254	

recognition. This approach predefines the algorithms of both training and recognition, and it remains only to choose the kernel produced by an appropriate metric in the set of signatures, such that the genuine signatures of the same person would be much closer to each other than those of different persons. However, different understandings of signature similarity lead to different kernels.

The proposed kernel fusion technique automatically chooses the most appropriate subset of kernels for each person in the process of adaptive training. Experiments with signature data base SVC2004 demonstrate that verification results obtained by selective fusion of several on-line and off-line kernels in accordance with the proposed approach essentially outperforms the results based on both single kernels and their plane fusion.

6 Acknowledgements

This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013 under Grant agreement No. 215078, the Russian Foundation for Basic Research (Grants 05-01-00679, 06-01-08042, 08-01-00695, 09-01-00573), INTAS Grant YSF-06-1000014-6563.

References

1. A. Ross, A. K. Jain. Multimodal biometrics: An overview. *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, 2004, Vienna, Austria, pp. 1221-1224.
2. P. Jannin, O. J. Fleig , E. Seigneuret, C. Grova, X. Morandi, J. M. Scarabin. A data fusion environment for multimodal and multi-informational neuronavigation. *Computer Aided Surgery*, 2000, Vol. 5, No. 1, pp. 1-10.
3. S. Sonnenburg, G. Rä tsch, C. Schä fer. 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. V. Sulimova, V. Mottl, A. Tatarchuk. 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. V. Mottl , A. Tatarchuk ,V. Sulimova, O. Krasotkina , O. Seredin. 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. I.M. Guyon , S. R. Gunn , M. Nikravesh, L. Zadeh, Eds. *Feature Extraction, Foundations and Applications*. Springer, 2006.
7. J. Li, H. Zha. 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.
8. 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, pp. 63-84.
9. V. Mottl, M. Lange, V. Sulimova, A. Ermakov. Signature verification based on fusion of on-line and off-line kernels. *Proceedings of the 19th International Conference on Pattern Recognition*, Tampa, USA, December 8-11, 2008.
10. M. Lange, S. Ganebnykh, A. Lange. Moment-based pattern representation using shape and grayscale features. *Lecture Notes in Computer Science*, Vol. 4477, Springer, 2007, pp. 523-530.
11. *SVC 2004: First International Signature Verification Competition.* http://www.cs.ust.hk/svc2004/index.html