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. We consider the problem of multi-modal pattern recognition under the assumption that a 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 sensors is employed as an appropriate joint scale corresponding to the idea of combining modalities at the sensor level. This contrasts with the commonly adopted method of combining classifiers inferred from each specific modality. However, a significant risk in combining linear spaces is that of overfitting. To address this, we set out a stochastic method for encompassing modal-selectivity that is intrinsic to (that is to say, theoretically contiguous with) the selected kernel-based pattern-recognition approach.

The principle of kernel selectivity supervision is then applied to the problem of signature verification by fusing several on-line and off-line kernels into a complete training and verification technique.

1 Introduction

It is often appropriate to treat observed phenomena via several distinct feature modalities (frequently with differing measurement scales) for the purposes of pattern recognition [1,2]. Such feature scales $x_i \in \mathbb{X}_i$ may be such that it is convenient, or even necessary, to treat real-world objects $\omega \in \Omega$ via a pair-wise similarity measure over these features $\big(x_i(\omega'), x_i(\omega'')\big)$. It is therefore assumed that mode-specific functions $K_i(x_i', x_i'')$ can be delimited over the output scales of the sensors in question $\mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$. The various $K(x', x'')$ functions constitute a *kernel* if they embed the sensor output $\mathbb{X}_i$ into a linear space via analogy with the inner-product. This condition is satisfied if the Kernel function defines a semidefinite matrix over any finite set of measured objects. The embedding may be of a significantly (even infinitely) different dimensionality to that of the original sensor scale, depending on the kernel characteristics.

J.A. Benediktsson, J. Kittler, and F. Roli (Eds.): MCS 2009, LNCS 5519, pp. 324–334, 2009.

Kernel-based multi-modal pattern recognition presents a number of difficulties and advantages over classical pattern-recognition in consequence of its pairwise nature. 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.

The class of discriminative classifiers known as Support Vector Machines (SVMs) may thus be employed for two-class pattern recognition within $\mathbb{R}^n$, once modalities are combined via the joint kernel $K(\mathbf{x}', \mathbf{x}'') = \sum_{i=1}^{n} x_i' x_i''$ (this approach can also be used for highly-complex kernel-represented modalities [3,4,5]).

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 only after selection has taken place. Feature selection (FS) techniques are of two broad types: *filters* and *wrappers* [6].

Filters are applied to the feature set irrespective of classification methodology, in contrast to wrappers. In this case, selection is either continuous (via weighting of the features) or else carried-out through absolute inclusion/exclusion of features from the total set. Wrappers, while considering feature selection in conjunction with classification, do not, in general, seek to do so via a single algorithmic approach (ie one in which FS is implicit in the process of classification itself - an exception being [7]). This is perhaps because of the danger of sample variability; if classification and FS progress interdependently, outliers can potentially affect the process disproportionately in the earlier stages. If, on the other hand, there exists a method of assigning selectivity *a priori*, this danger is mitigated to a large extent. Ideally, we require a range of behaviours, from the complete absence of selection, to the selection of only singular features.

In the following paper, we show, following [9] and [10], how selectivity may be incorporated into the Relevance Kernel Machine (RKM) [4,5], a continuous wrapper FS method previously described by the authors. The desired selectivity is achieved through a meta-parameter that controls the tendency of the RKM to generate zero components in the orientation of the decision plane (and hence the degree of elimination of constituent kernels). Thus, the selectivity parameter corresponds directly to model complexity, with the appropriate level of selectivity determined by cross validation or (in future work) via information-theoretic considerations.

[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.

The Relevance Kernel Machine with supervised selectivity is then applied 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

Let objects $\omega \in \Omega$, measured by n features with modality-specific scales $x_i(\omega) \in \mathbb{X}_i$, be allocated to one of two classes $y(\omega) \in \mathbb{Y} = \{-1, 1\}$. For convenience, we assume an underlying distribution in the set of observable feature values and associated class indices; $\big(x_1(\omega), ..., x_n(\omega), y(\omega)\big) \in \mathbb{X}_1 \times ... \times \mathbb{X}_n \times \mathbb{Y}$. 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 i.i.d. The kernel approach demands only that a real value similarity function exists - it thus obviates the distinction between different kinds of feature scales, so that we can assume that all the modality-specific features $x_i(\omega) \in \mathbb{X}_i$ are real-valued: $\mathbb{X}_i = \mathbb{R}$.

Functions $\varphi_1\,(x_1, ..., x_n \,|\, a_1, ..., a_n\,, b, y)$ with $y = \pm 1$ are thus two parametric families of probability densities in the composite feature space $\mathbb{X}_1 \times ... \times \mathbb{X}_n$. We assume marginally overlapping concentrations, such that the two together can be associated with a discriminant hyperplane $\sum_{i=1}^{n} a_i x_i + b \gtrless 0$. We further associate improper (ie non-unity integral) densities with the distributions:

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

The constant h then represents the extent to which the classes are equivalent to a uniform distribution over their respective half-spaces. The parameter c determines the extent to which the classes overlap.

The direction vector $(a_1, ..., a_n)$ of the discriminant hyperplane $\sum_{i=1}^{n} a_i x_i + b \gtrless 0$ will, in the absence of a training mechanism, be considered a random vector distributed in accordance with some specific prior density $\Psi(a_1, ..., a_n \mid \mu)$ parametrized by μ. No such constraint is assumed in b, hence, $\Psi(a_1, ..., a_n, b \mid \mu) \propto \Psi(a_1, ..., a_n | \mu)$.

With respect to the training set, the *a posteriori* joint distribution density of the parameters of the discriminant hyperplane is consequently proportional to the product $P(a_1, ..., a_n, b \,|\, X, Y, \mu) \propto \Psi(a_1, ..., a_n \mid \mu) \times \Phi(X \mid Y, a_1, .., a_n, b)$.

The objective of training is thus to maximise the *a posteriori* density:

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

This correlates to the training criterion:

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

Note that if we set $C = 2rc$, with r the common variance of the independent constituent variables (having zero mean), and omit the parameter μ (such that $\Psi(a_1, ..., a_n \mid \mu) = \Psi(a_1, ..., a_n)$ is the joint normal distribution), we obtain the classical SVM over the real-valued features $x_{ij} \in \mathbb{X}_i = \mathbb{R}$ with the direction vector elements $a_i \in \mathbb{X}_i = \mathbb{R}$ constituting a discriminant hyperplane in $\mathbb{X}_1 \times ... \times \mathbb{X}_n = \mathbb{R}^n$ such that:

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

Specifically, if the kernels $K_i(x'_i, x''_i) : \mathbb{X}_i \times \mathbb{X}_i \rightarrow \mathbb{R}$ defined for the sensor features $x_i \in \mathbb{X}_i$ are inserted into (2), we obtain the optimization:

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

It is important to note that, in general, the elements a_i of the hyperplane direction vector exist in the embedding space $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$, rather than the original feature space $\mathbb{X}_i$.

A central advantage of SVMs, in terms of their capacity for overfitting, is that at the minimum of the training criterion (such that $a_i = \sum_{j:\ \lambda_j > 0} \lambda_j y_j x_{ij} \in \tilde{\mathbb{X}}_i$), 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 determined only by those Lagrange multipliers with $\lambda_j \geq 0$ in the dual form of (3), ie the *support objects*. The dual problem, which can be solved by quadratic-programming is thus :

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

The following section will consider a distinct form of the *a priori* distribution $\Psi(a_1, ..., a_n \mid \mu)$, that gives rise to a feature- and kernel-selective SVM, such that the parameter μ controls the desired selectivity level.

3 The Continuous Training Technique with Supervised Selectivity

We first assume a conditional normal distribution for the direction elements a_i in relation to independent random variances given by r_i:

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

There is hence a hyper-ellipsoidal relationship between the direction elements a_i.

We further assume that the reciprocated variances are gamma distributed (a reasonable, maximum-entropy-based assumption for positive-constrained scale variables), ie: $\gamma\big((1/r_i) \,|\, \alpha, \beta\big) \propto (1/r_i)^{\alpha-1} \exp\left(-\beta\,(1/r_i)\right)$ (with means $E(1/r_i) = \alpha/\beta$ and variances $E\left((1/r_i)^2\right) = \alpha/\beta^2$). We then set the following parameter relations to enable convenient characterisation of the distribution; $\alpha = (1+\mu)^2\big/2\mu$, $\beta = 1/2\mu$.

There is hence now a parametrically-defined set of distributions in the direction elements a_i, dependant only on $\mu : \mu \geq 0$ (where $E(1/r_i) = (1+\mu)^2$ and $E\left((1/r_i)^2\right) = 2\mu(1+\mu)^2$).

In behavioral terms it should be noted that, as $\mu \to 0$, we find that $1/r_i \cong ... \cong 1/r_n \cong 1$. However, as μ increases, this identity constraint is progressively relaxed.

Proceeding with the derivation, we now eliminate the inverse variances as follows. Firstly, we note that the joint distribution of independent inverse variances with respect to μ is proportional to the product:

$$G(r_1, ..., r_n \,|\, \mu) \propto \left(\prod_{i=1}^n (1/r_i)\right)^{(1+\mu)^2/2\mu - 1} \exp\left(-1/2\mu \sum_{i=1}^n (1/r_i)\right).$$

The maximum of the joint *a posteriori* density function $P(a_1, ..., a_n, b, r_1, ..., r_n \,|X, Y, \mu)$ then gives us the required training criterion: we see that it is proportional to the product: $\Psi(a_1, ..., a_n \,|\, r_1, ..., r_n)\ G(r_1, ..., r_n \,|\, \mu)\ \Phi(X \,|\, Y, a_1, .., a_n, b)$.

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

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

$\varepsilon > 0$ is the inclusion criterion for features: it is thus a sufficiently small positive real number. In general, a smaller r_i will imply a smaller a_i. As $r_i \to \varepsilon$, the ith feature will affect the discriminant hyperplane $\sum_{i=1}^n a_i x_i + b \gtrless 0$ increasingly weakly.

Again, we obtain the kernel-based training criterion by substituting into (6) $K_i(a_i, a_i)$ for a_i^2 and replacing $a_i x_{ij}$ by $K_i(a_i, x_{ij})$ to give:

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

As with SVMs, there is no explicit need to evaluate either the $a_i \in \mathbb{R}$ in (6) or the $a_i \in \tilde{\mathbb{X}}_i$ in (7); it is sufficient merely to establish the non-zero Lagrange multipliers $\lambda_j \geq 0$ in the dual representation $a_i = r_i \sum_{j:\,\lambda_j>0} y_j \lambda_j x_{ij}$. We do this via quadratic-programming using a modification of (5):

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

This gives the Kernelised decision hyperplane:

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

In distinction to the discriminant hyperplanes for standard SVMs (4), features are effectively assigned weights r_i, so that as $r_i \to 0$, the influence of the respective features diminishes. However, as it stands, the weights are unknown in (7).

Solving this optimization problem for fixed μ, involves the application of the Gauss-Seidel iteration to the variable sets $(a_1, ..., a_n, b, \delta_1, \ldots, \delta_N)$ and $(r_1, ..., r_n)$, with initiation values of $(r_i^0 = 1,\ i = 1, ..., n)$. Once the solution $\lambda_1^k, \ldots, \lambda_N^k$, i.e. $(a_1^k, ..., a_n^k)$, is found at the k th iteration with the current approximations $(r_1^k, \ldots, r_n^k)$, the revised values of the variances $(r_1^{k+1}, ..., r_n^{k+1})$ are defined as

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

Convergence of the procedure occurs in $\approx 10 - 15$ steps for typical problems, suppressing redundant features through the allocating of very small (but always non-zero weights) r_i defining the discriminant hyperplane (9).

In summary, the training criterion for Relevance Kernel Machine (RKM) [4,5] is set out in (6). The feature selectivity of this SVM generalisation is parametrically determined by $\mu : 0 \leq \mu < \infty$. As $\mu \to 0$, variances tend toward unity (10), and the RKM degenerates to the classical SVM (2). Contrarily, when $\mu \to \infty$, we have from (6) that $\sum_{i=1}^{n}\big[(1/r_i)a_i^2 + (1+\mu)\ln r_i\big] + C\sum_{j=1}^{N}\delta_j \to \min$; actually a significantly more selective training criterion than the original RKM (without supervised selectivity): $\sum_{i=1}^{n}\big[(1/r_i)a_i^2\ +\ \ln r_i\big]\ +\ C\sum_{j=1}^{N}\delta_j\ \to\ \min$ [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

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

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.

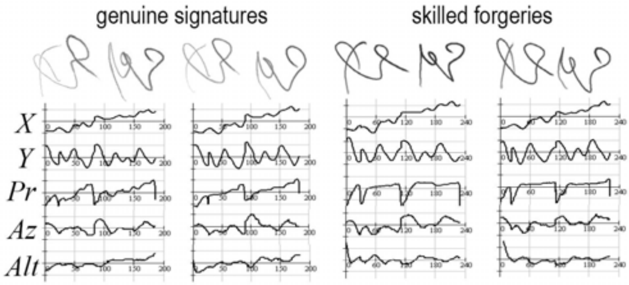


Fig. 1. Off-line (images) and on-line (signals) representation of signatures

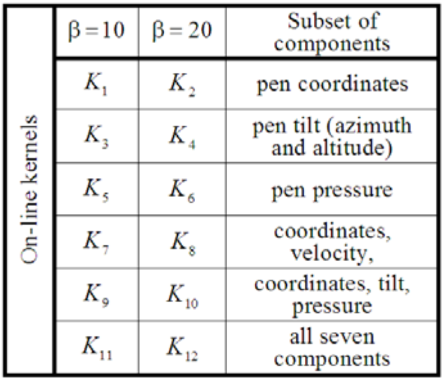
In this work, we tested 12 different metrics in the set of on-line signatures and 4 metrics computed from the pictorial off-line representation. So, all in all, we combined 16 different on-line and off-line kernels listed in Table 1.

4.2 Metrics in the Set of On-Line Signatures

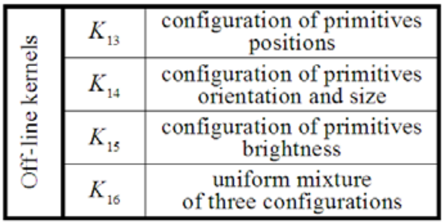
Each on-line signature is represented by a multi-component vector signal which initially includes five components $\mathbf{x}_t = (x_t^1 \cdots x_t^n)$: two pen tip coordinates (X, Y), pen tilt azimuth (Az) and altitude (Alt), and pen pressure (Pr) (Fig. 1). We supplement the signals with two additional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

Table 1. The kernels studied in the experiments



	$\beta=10$	$\beta=20$	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

$w(\omega',\omega'')$ 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 [11].

For the given pattern P, the recursive scheme described in [11] 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 [11], we define the distinction measure (metric) of the trees R' and R'' as follows:

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

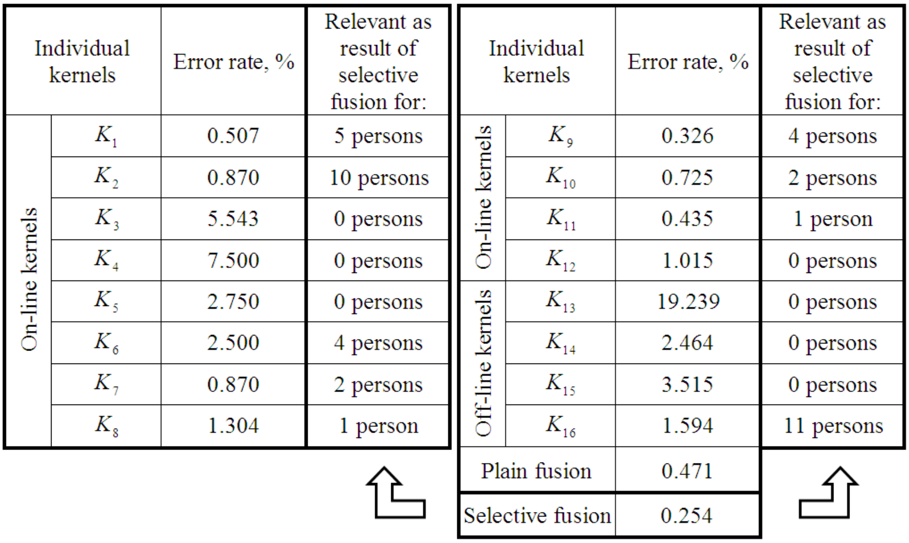
where the sum is taken over all pairs $(Q'_n, Q''_n) \in R' \bigcap R''$. We competitively applied three basic distinction measures of the form (13) $\rho_1(R', R'') = \rho(R', R'' \mid 1, 0, 0)$, $\rho_2(R', R'') = \rho(R', R'' \mid 0, 1, 0)$, $\rho_3(R', R'') = \rho(R', R'' \mid 0, 0, 1)$, and the 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 [12] 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

Table 2. Error rates for single kernels versus kernel fusion



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

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 $\{K_1(\omega',\omega''),\dots,K_{16}(\omega',\omega'')\}$, 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 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.

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

2. Jannin, P., Fleig, O.J., Seigneuret, E., Grova, C., Morandi, X., Scarabin, J.M.: A data fusion environment for multimodal and multi-informational neuronavigation. Computer Aided Surgery 5(1), 1–10 (2000)
3. Sonnenburg, S., Rätsch, G., Schäfer, C.: A general and efficient multiple kernel learning algorithm. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, December 5-8 (2005)
4. Sulimova, V., Mottl, V., Tatarchuk, A.: Multi-kernel approach to on-line signature verification. In: Proceedings of the 8th IASTED International Conference on Signal and Image Processing, Honolulu, Hawaii, USA, August 14-16 (2006)
5. Mottl, V., Tatarchuk, A., Sulimova, V., Krasotkina, O., Seredin, O.: Combining pattern recognition modalities at the sensor level via kernel fusion. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 1–12. Springer, Heidelberg (2007)
6. Guyon, I.M., Gunn, S.R., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction, Foundations and Applications. Springer, Heidelberg (2006)
7. Li, J., Zha, H.: Simultaneous classification and feature clustering using discriminant vector quantization with applications to microarray data analysis. In: Proceedings of the IEEE Computer Society Bioinformatics Conference, Palo Alto, CA, August 14-16, pp. 246–255 (2002)
8. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on Pattern Recognition and Machine Intelligence 22(1), 63–84 (2000)
9. Tatarchuk, A., Mottl, V., Eliseyev, A., Windridge, D.: Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective Support Vector Machines. In: Proceedings of the 19th International Conference on Pattern Recognition, Tampa, USA, December 8-11 (2008)
10. Mottl, V., Lange, M., Sulimova, V., Ermakov, A.: Signature verification based on fusion of on-line and off-line kernels. In: Proceedings of the 19th International Conference on Pattern Recognition, Tampa, USA, December 8-11 (2008)
11. Lange, M., Ganebnykh, S., Lange, A.: Moment-based pattern representation using shape and grayscale features. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4477, pp. 523–530. Springer, Heidelberg (2007)
12. SVC 2004: First International Signature Verification Competition, `http://www.cs.ust.hk/svc2004/index.html`