We introduce in this paper a novel image annotation approach based on
maximum margin classification and a new class of kernels. The method
goes beyond the naive use of existing kernels and their restricted com-
bined in order to design "model-free" transductive kernels applicable to
interconnected image databases.

Let \( S = \{x_1, \ldots, x_n, \ldots, x_m\} \) denote an image database described in an \( n \)-dimensional input space. We assume that only the first \( l \) (\( l \ll m \)) vectors of \( S \) are labeled (a.k.a. annotated), i.e., \( \{x_1, \ldots, y_i\} \) are known; here \( y_i \in \{-1, +1\} \) and \( r \) is the number of possible labels used for annotation.

Our approach considers image annotation as a multi-label classification
problem in which a sample \( x_i \) may have more than one label, i.e.,
\( r > 1 \), with \( y_{ik} = +1 \) iff \( x_i \) has the \( k^{th} \) label and \( y_{ik} = -1 \) otherwise. Our objective is to build an optimal kernel map and a decision criterion \( f \) in order to infer the unknown label vectors \( \{y_1, \ldots, y_m\} \).

We adopt the max-margin classification [4] approach in order to learn a classifier \( f(x) = W^T \Phi(x) \) that balances training error and model complex-
ity. This classifier corresponds to
\[
\underset{f}{\text{argmin}} \ R(f) + \sum_{i=1}^{l} \ell(f(x_i), y_i),
\]
where \( R \) is a regularizer, \( \ell(f(x_i), y_i) \) is the loss associated with a pre-
diction \( f(x_i) \) when the true output is \( y_i \) and \( y_i > 0 \) balances these two terms. For nonlinear classification, \( \Phi \) maps the input data (in \( S \)) into a high dimensional space \( \mathcal{H} \) such that \( W \) can separate labeled data \( \{x_1, \ldots, x_l\} \).

Following the kernel trick [3], the function \( f \) may also be expressed as a linear combination of symmetric, continuous and positive (semi) def-
inite kernel functions. A kernel (denoted \( K \)) is defined on two samples \( x_i, x_j \) as \( K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \). The closed form of \( K(\cdot, \cdot) \) may also be defined as a collection of existing kernels including linear, poly-
nomial and histogram intersection; but the underlying mapping \( \Phi(\cdot) \in \mathcal{H} \) is usually implicit, i.e., it does exist but it is not necessarily known and may be infinite dimensional.

Our proposed method, in contrast to usual kernel methods, finds an ex-
plicit and finite dimensional kernel map. According to Vapnik’s VC-
theory [4], a finite dimensional kernel map, with a bounded related VC-
dimension, avoids loose generalization bounds and may guarantee better performance.

Our goal is to find hyperplane parameters \( W \) as well as a Gram (kernel)
matrix \( K = \Phi^T \Phi \) where each column \( \Phi_i \) corresponds to an explicit
mapping of \( x_i \) into a high dimensional space (i.e., \( \Phi(x_i) = \Phi_i \)). The learned mapping \( \Phi \) must i) guarantee linear separability of data in \( S \), ii) ensure good generalization performance by maximizing the margin, iii) approximate the input data, and also iv) ensure positive definiteness of \( K \) by construction, i.e., without adding further constraints. Consider the constraint
\[
\underset{B \in \mathbb{R}^{p \times m}}{\text{min}} \frac{1}{2} \|B\|_F^2 + \frac{1}{2} \|W\|_F^2 + \frac{1}{2} \|X - BY - [B_{0 \times p} W'] [F \ C]\|_F^2
\]
subject to \( \|B\|_2 = 1, \forall i = 1, \ldots, p \).

Here \( C \in \mathbb{R}^{m \times m} \) is a diagonal matrix with \( C_{ii} = 1 \) for labeled data, \( B_{0 \times p} \) and \( W' \) are \( n \times p \) and \( r \times p \) zeros matrices respectively, \( X \approx B \Phi \) is factorized using an overcomplete basis \( B \in \mathbb{R}^{n \times p} \) (i.e., \( p > n \)) and a new kernel map \( \Phi \in \mathbb{R}^{p \times m} \).

According to [4], the VC-dimension (related to a family of classi-
ifiers) depends also on the dimension of the learned kernel map and this
may affect generalization, especially if this dimension is very high. Since
the actual (intrinsic) dimension of the learned kernel map \( \Phi \) is unknown,
we choose the number of basis \( p \) to be sufficiently large such that the
factorization term (in right-hand side of Eq. 2) tends to zero for an
infinite number of solutions. Then, the actual (intrinsic) dimension is found
by regularizing Eq. 2 using the Frobenius norm \( \|B\|_F^2 \) which has similar
effect as the nuclear norm where \( \mu \geq 0 \) controls the rank of \( \Phi \).

For a better conditioning of Eq. 2, we adopt transductive inference [1, 5]
which assumes that close data in a high-density area of the input space
should have similar labels [2]. This assumption, therefore, enables label
diffusion from training to the test data (see toy example in Fig. 1).

By representing \( x_i \)'s as vertices \( v_j \) and their pairwise similarities as edges \( e_{ij} \), the smoothness assumption between \( v_i \) and \( v_j \) is modeled by the differences between \( f(x_i) \) and \( f(x_j) \), i.e.,
\[
\frac{1}{4} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\Phi x_i - \Phi x_j\|_2^2 \leq \frac{1}{2} \text{tr}(W^T \Phi \Phi^T W),
\]
where the graph Laplacian \( L = D - A \) is defined by the affinity matrix
\( A \) whose elements \( A_{ij} = 1_{\{v_i \in N(v_i)\}} \cdot s(x_i, x_j) \) and \( D = \text{diag}(A1) \) with \( I \)
being the all-one vector of length \( m \). Here \( s(\cdot, \cdot) \) is a visual similarity and \( N_i(v) \) is the set of the \( k \)-nearest neighbors of \( v \).

Now, we obtain the complete form of our transductive learning prob-
lem as
\[
\underset{B \in \mathbb{R}^{p \times m}}{\text{min}} \frac{1}{2} \|B\|_F^2 + \frac{1}{2} \text{tr}(W^T [L_\Phi + \gamma C] W) +
\frac{1}{2} \text{tr}(X^T [B_{0 \times p} W'] [F \ C]\|_F^2,
\]
subject to \( \|B\|_2 = 1, \forall i = 1, \ldots, p \).

This minimization problem makes it possible to learn both a de-
cision criterion and a kernel map that guarantee linear separability in a
high dimensional space and good generalization performance (see Fig. 1).

Experiments conducted on image annotation, show that, indeed, our
obtained kernel achieves at least comparable results with related state of
the art methods on the MSRC and the Corel5K databases.

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A
geometric framework for learning from labeled and unlabeled ex-


[3] B. Schölkopf and A. Smola. Learning with Kernels: Support Vec-
tor Machines, Regularization, Optimization, and Beyond. The MIT


risk in a problem of pattern recognition. Automation and Remote