Transductive Kernel Map Learning and its Application to Image Annotation

Dinh-Phong Vo
vo@enst.fr
Hichem Sahbi
sahbi@enst.fr

LTCI CNRS Telecom ParisTech
46 rue Barrault, 75013, Paris, France

Abstract

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 combinations in order to design "model-free" transductive kernels applicable to interconnected image databases. The main contribution of our method includes the minimization of an energy function mixing i) a reconstruction term that factorizes a matrix of interconnected image data as a product of a learned dictionary and a learned kernel map ii) a fidelity term that ensures consistent label predictions with those provided in a training set and iii) a smoothness term which guarantees similar labels for neighboring data and allows us to iteratively diffuse kernel maps and labels from labeled to unlabeled images. Solving this minimization problem makes it possible to learn both a decision criterion and a kernel map that guarantee linear separability in a high dimensional space and good generalization performance. Experiments conducted on image annotation, show that our obtained kernel achieves at least comparable results with related state of the art methods on the MSRC and the Corel5k databases.

1 Introduction

With the exponential growth of multimedia sharing spaces, such as social networks, visual contents are nowadays abundant. Searching these large collections requires a preliminary step of image annotation that translates visual contents into labels also known as keywords or concepts (see for instance [12]). Automatic image annotation is challenging due to the perplexity when assigning many possible labels to images and the difficulty to analyze rich and highly semantic contents. In annotation, image observations are first described using low-level features (color, texture, shape, etc.), and labels are then assigned to images using variety of inference techniques such as hidden Markov models [13], latent Dirichlet allocation [4], probabilistic latent semantic analysis [23], and support vector machines (SVMs) [9, 41, 45]. These inference techniques are used in order to model the correspondence between low level features and labels and allow us to predict keywords for unlabeled images.

Among existing image annotation approaches, machine learning ones are particularly successful and may be categorized into generative and discriminative. Generative methods model a priori knowledge and dependencies between image observations and their possible

http://dx.doi.org/10.5244/C.26.68

labels using for instance graphical models [15, 16, 19, 42]. In these models, the annotation process is based on maximizing a posterior probability using a variety of network inference techniques. This category of methods even though relatively successful suffers from complexity in modeling and inference especially when labels are taken from a large scale vocabulary. Alternative approaches are discriminative and consider image annotation as a classification problem [10, 11, 14, 40]. A vocabulary of labels is first defined, and a decision criterion is then learned for each label and used in order to identify images belonging to that label.

The two aforementioned categories of machine learning techniques are highly dependent on the learned concepts and may fail when the latter are highly semantic and difficult to model. In order to overcome these issues, recent discriminative approaches consider a priori knowledge and relationships between data and the learned concepts (context, shared features, etc.) [15, 17, 19, 25, 29, 31, 32, 37, 48]. The success of these image annotation methods, also depends on cardinality of the labeled data and the choice of the appropriate setting for learning. The inductive setting [10, 11, 14, 40] consists in building a decision function for each concept using labeled images, and uses that function in order to generalize across unlabeled images. In these methods, labeled data are usually scarce and expensive; only a very small fraction of training images is labeled and the unlabeled images may not follow the same distribution as the labeled ones, so learning using inductive techniques is clearly not appropriate. Alternatives [3, 46] may include the unlabeled data as a part of the learning process and this is known as transductive inference. The concept of transductive inference, or transduction, was pioneered by Vapnik (see for instance [46]). It relates to semi-supervised learning [6] and relies on the i) smoothness assumption which states that close data in a high-density area of the input space, should have similar labels [6] and ii) the cluster assumption which finds decision rules in low density areas of the input space [6]. Learning consists in building decision functions by optimizing the parameters of a learning model together with the labels of the unlabeled data (see for instance [3, 26, 27, 37]). When applied, these transductive methods turned out to be very useful in order to overcome the limited cardinality of the labeled images in image annotation [8, 18, 35, 50, 52].

Among popular learning techniques support vector machines [9, 41, 45] are well studied and proved to be performant in image annotation [20]; in SVMs, kernels are used in order to model visual similarity between images, and only images sharing the same concepts are expected to have high kernel values. The success of SVMs is therefore, highly dependent on the choice of kernels and usual ones, such the linear, the gaussian and the histogram intersection, may not be appropriate in order to capture the actual and the semantic similarity between images for some specific concepts. Better kernels based on tuning Mahalanobis distances were obtained by minimizing the ratio between intra and inter class distances [7, 24, 28] while others were designed using semidefinite programming [30]. In order to take extra advantage from different settings, multiple kernels (MKL) were also introduced [1, 2, 39, 43, 51] and consider convex (and possibly sparse) linear combinations of elementary kernels and proved to be more suitable [47]. With the current state of the art, MKL are considered as one of the most effective kernel design and combination techniques. Nevertheless, MKL based design hits at least two major limitations; On the one hand, and as mentioned earlier, these methods are limited by the cardinality of labeled data and they do not rely on any extra information in order to overcome that limitation, on the other hand they are mainly restricted to linear combinations of existing kernels only.

In this paper we introduce a novel transductive learning algorithm, for kernel learning and image classification and annotation. Our method is based on a constrained matrix fac-

torization which produces a kernel map that takes image data from the input space into a high dimensional space in order to guarantee their linear separability while maximizing their margin. This margin property, however, and as known [45], does not necessarily guarantee good generalization performance on the unlabeled set, if the latter is drawn from a different probability distribution compared to the labeled data. Therefore and beside maximizing the margin, our transductive approach includes a regularization term that enforces smoothness and low rankness in the resulting kernel map in order to correctly diffuse labels to the unlabeled data. Following our formulation, and in contrast to MKL, our learning model is not restricted to only convex linear combinations of existing kernels; indeed it is model-free. Furthermore, it also takes advantage from both labeled and unlabeled data and this results into better generalization performances as corroborated by our image annotation experiments.

The remainder of this paper is organized as follows. We introduce our transductive learning approach and kernel design in Section 2 and the implementation of our optimization procedure in Section 3. We illustrate in Section 4 the application of our method to image annotation using two datasets; MSRC and Corel5K. We conclude the paper in Section 5 while providing a possible extension for a future work.

2 Problem Formulation

Define $\mathcal{X} \subseteq \mathbb{R}^n$ as an input space corresponding to all the possible image features and let $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_m\}$ be a finite subset of $\mathcal{X}$ with an arbitrary order. This order is defined so only the first l label vectors of $\mathcal{S}$, denoted $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$ are given; here $\mathbf{y}_i \in \{-1, +1\}^r$ and r is the number of possible labels used for annotation. In many real-world applications only a few data is labeled (i.e., $l \ll m$) and its distribution may be different from the unlabeled data.

We can view $\mathcal{S}$ as a matrix $\mathbf{X}$ in which the i^{th} column $\mathbf{X}_i$ corresponds to $\mathbf{x}_i$ and $\mathbf{Y}$ is the label matrix of $\mathbf{X}$ where its i^{th} column $\mathbf{Y}_i$ corresponds to $\mathbf{y}_i$. Different from binary classification, in multi-label classification, a sample $\mathbf{X}_i$ may have more than one label, i.e., $r > 1$, with $\mathbf{Y}_{ik} = +1$ iff $\mathbf{X}_i$ has the k^{th} label and $\mathbf{Y}_{ik} = -1$ otherwise. Our objective is to build both a decision criterion and an optimal *kernel map* in order to infer the unknown label vectors $\{\mathbf{Y}_{l+1}, \ldots, \mathbf{Y}_m\}$.

2.1 Max Margin Inference for Multi-label Classification

The general classification problem aims to learn a classifier f, that minimizes training error and also generalize well on test data, as

$$\underset{f}{\operatorname{argmin}} \; \mathcal{R}(f) + \gamma_c \sum_{i=1}^{l} \ell\left(f(\mathbf{x}_i), \mathbf{y}_i\right), \tag{1}$$

$\mathcal{R}$ is a regularizer that controls model complexity, $\ell(f(\mathbf{x}_i), \mathbf{y}_i)$ is the loss associated with a prediction $f(\mathbf{x}_i)$ when the true output is $\mathbf{y}_i$ and $\gamma_c > 0$ balances these two terms. In the max-margin classification [45], $f(\mathbf{x}_i) = \mathbf{W}'\phi(\mathbf{x}_i)$ (for a well chosen $\mathbf{W}$) and ϕ is a mapping of the input data (in $\mathcal{X}$) into a high dimensional space $\mathcal{H}$. The dimension of $\mathcal{H}$ is usually sufficiently large (possibly infinite) in order to guarantee linear separability of data.
Assuming data linearly separable in $\mathcal{H}$, the max-margin inductive learning finds $\mathbf{W}$ (and

hence f) as

$$\underset{\mathbf{W}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \gamma_c \sum_{i=1}^{l} \ell\left(\mathbf{W}'\phi(\mathbf{x}_i), \mathbf{y}_i\right) \tag{2}$$

here $\mathbf{W}'$ is the transpose of $\mathbf{W}$ and $\|\mathbf{W}\|_F^2$ denotes the Frobenius norm. Following the kernel trick [41], one may show that the classification function f may also be expressed as a linear combination of symmetric, continuous and positive semi-definite functions (called kernels). A kernel (denoted κ) is defined on two samples $\mathbf{x}_i, \mathbf{x}_j$ as $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. The closed form of $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ may also be defined among a collection of existing kernels including linear, polynomial and histogram intersection; but the underlying mapping $\phi(x) \in \mathcal{H}$ is usually *implicit*, i.e., it does exist but it is not necessarily known and may be infinite dimensional.

We propose in the remainder of this section a new approach that builds *explicit* and finite dimensional kernel map. In contrast to usual kernels, such as the gaussian, the VC-dimension [45], related to a finite dimensional kernel map, is finite[1]. According to Vapnik's VC-theory [45], the finiteness of the VC-dimension avoids loose generalization bounds and may guarantee better performance.

2.2 Learning Low-Rank Kernel

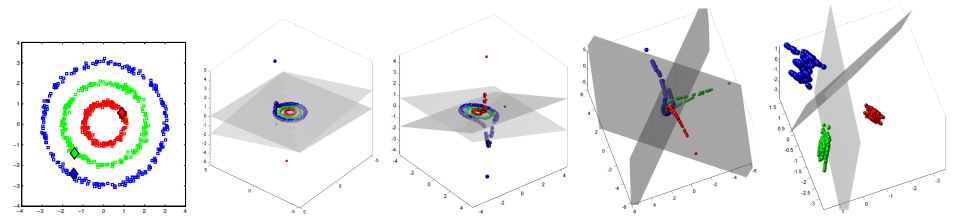
Now, we turn the problem into finding the hyperplane parameters $\mathbf{W}$ as well as a Gram (kernel) matrix $\mathbf{K} = \Phi'\Phi$ where each column Φ_i corresponds to an explicit mapping of $\mathbf{x}_i$ into a high dimensional space (i.e., $\phi(\mathbf{x}_i) = \Phi_i$). This mapping is designed in order to i) guarantee linear separability of data in $\mathcal{S}$, ii) to ensure good generalization performance by maximizing the margin, iii) to approximate the input data, and also iv) to ensure positive definiteness of $\mathbf{K}$ by construction, i.e., without adding further constraints. This results into the following constrained minimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{B},\Phi,\mathbf{W}} & \frac{\mu}{2}\|\Phi\|_F^2 + \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{\gamma_c}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{r\times p} & \mathbf{W}' \end{bmatrix} \begin{bmatrix} \Phi \\ \Phi\mathbf{C} \end{bmatrix} \right\|_F^2, \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{3}$$

here $\mathbf{C} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $\mathbf{C}_{ii} = 1_{\{1\leq i\leq l\}}$, $\mathbf{0}_{n\times p}$ and $\mathbf{0}_{r\times p}$ are $n\times p$ and $r\times p$ zeros matrices respectively, $\mathbf{X} \approx \mathbf{B}\Phi$ is factorized using an overcomplete basis $\mathbf{B} \in \mathbb{R}^{n\times p}$ (i.e., $p > n$) and a new kernel map $\Phi \in \mathbb{R}^{p\times m}$.

As discussed earlier, and according to [45], the VC-dimension (related to a family of classifiers) 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 Φ is unknown, we choose the number of basis p to be sufficiently large such that the factorization term (in the right-hand side term of Eq. 3) tends to zero for an infinite number of solutions. In practice, p is overestimated and set to $\max(l,n)+1$, and this guarantees that the above constrained minimization problem has a solution. Then, the actual (intrinsic) dimension is found by regularizing Eq. 3 by the Frobenius norm $\|\Phi\|_F^2$ which has similar effect as the nuclear norm where $\mu \geq 0$ controls the rank of $\mathbf{K}$. Indeed, the squared Frobenius norm is exactly the ℓ_2-norm on the eigenvalues of $\mathbf{K}$ and it is less likely to shrink these eigenvalues into zeros compared to the ℓ_1-norm (which is the nuclear norm). Nevertheless, it provides a closed form kernel solution and our experiments show that it indeed reduces the rank of the kernel map while allowing to learn effective classifiers.

[1] The VC-dimension is the maximum number of data samples, that can be shattered, whatever their labels.



(a) (b) $\gamma_s = 1, \gamma_c = 1$ (c) $\gamma_s = 10, \gamma_c = 1$ (d) $\gamma_s = 10^2, \gamma_c = 1$ (e) $\gamma_s = 10^3, \gamma_c = 10^2$

Figure 1: (a) This figure shows the input data where different colors stand for different classes; red-colored data are annotated with $(1 \quad -1)'$, blue-colored data with $(-1 \; 1)'$ and green-colored data with $(1 \; 1)'$. Note that just one training sample per class (diamond-shaped) is labeled while others are unlabeled. Our algorithm is initialized with $p = 10$ and after 5 iterations (before convergence) it reduces the ranks to 4. Figures (b,c,d) are the learned kernel maps (shown in 3d) and the obtained decision hyperplanes for different setting of the parameters γ_c and γ_s.

2.3 Transductive Setting

For a better conditioning of Eq. 3, we implement in this section the smoothness assumption discussed in Section 1. This makes it possible to design smooth kernel maps and to assign similar predictions to neighboring data (see toy example in Fig. 1).

We model the input data $\mathcal{S}$ using an adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V} = \{v_1, \ldots, v_m\}$ correspond to samples $\{x_i\}$ and edges $\mathcal{E} = \{e_{i,j}\}$ are the set of weighted links of $\mathcal{G}$. In the above definition, $\mathbf{x}_i \in \mathbb{R}^n$ is a feature vector (color, texture, etc.) while $e_{i,j} = (v_i, v_j, \mathbf{A}_{ij})$ defines a connection between v_i, v_j weighted by $\mathbf{A}_{ij}$. The latter is defined as $\mathbf{A}_{ij} = 1_{\{v_j \in \mathcal{N}_k(v_i)\}} \cdot s(\mathbf{x}_i, \mathbf{x}_j)$, here $s(\cdot,\cdot)$ is a visual similarity function and the neighborhood $\mathcal{N}_k(v_i)$ of a given node v_i, includes the set of the k-nearest neighbors of v_i. Notice that the neighborhood system is designed in order to guarantee that $\forall v_i, v_j \in \mathcal{V}$, $v_j \in \mathcal{N}_k(v_i)$ implies $v_i \in \mathcal{N}_k(v_j)$ and vice-versa. Considering $f(\mathbf{x}_i) = \mathbf{W}'\mathbf{\Phi}_i$ and $f(\mathbf{x}_j) = \mathbf{W}'\mathbf{\Phi}_j$, we define our regularizer as

$$\frac{\gamma_s}{4} \sum_{i=1,j=1}^{m} \|\mathbf{W}'\mathbf{\Phi}_i - \mathbf{W}'\mathbf{\Phi}_j\|^2 \mathbf{A}_{ij} \tag{4}$$

which can be rewritten as $\frac{\gamma_s}{2}\mathrm{tr}\,(\mathbf{W}'\mathbf{\Phi}\mathbf{L}\mathbf{\Phi}'\mathbf{W})$, here $\gamma_s \geq 0$ and $\mathbf{L}$ is the graph Laplacian defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})$ where $\mathbf{1}$ is the all-ones vector of length m. We obtain the complete form of our transductive learning problem as

$$\begin{array}{ll} \min\limits_{\mathbf{B},\mathbf{\Phi},\mathbf{W}} & \frac{\mu}{2}\|\mathbf{\Phi}\|_F^2 + \frac{1}{2}\mathrm{tr}\Big(\mathbf{W}'\left(\mathbf{I}_p + \gamma_s\mathbf{\Phi}\mathbf{L}\mathbf{\Phi}'\right)\mathbf{W}\Big) + \frac{\gamma_c}{2}\left\|\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} - \begin{bmatrix}\mathbf{B} & \mathbf{0}_{n\times p}\\ \mathbf{0}_{r\times p} & \mathbf{W}'\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}\\ \mathbf{\Phi C}\end{bmatrix}\right\|_F^2, \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{5}$$

with $\mathbf{I}_p$ the $p \times p$ identity matrix and again $\mathbf{C}$ is the diagonal $m \times m$ matrix for which the i^{th} diagonal element is fixed to 1 for a labeled sample, and 0 for an unlabeled one.

3 Optimization

It is clear that the minimization problem in Eq. 5 is not convex jointly w.r.t $\mathbf{B}, \Phi, \mathbf{W}$. We consider an alternative optimization procedure by solving three subproblems: we first optimize the matrix $\mathbf{W}$ and we update the basis $\mathbf{B}$, then we minimize the regularization criterion, the rank and the reconstruction error w.r.t Φ. This process is repeated until convergence; i.e., all the unknowns remain unchanged from one iteration to another. Different steps of the algorithm are shown in Algorithm 1; the superscript (t) is added to $\mathbf{W}$, $\mathbf{B}$ and Φ in order to show the evolution of their values through different iterations of the learning process.

Algorithm 1 Transductive Kernel Map Learning

Input: labeled $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{l}$ and unlabeled data $\{\mathbf{x}_i\}_{i=l+1}^{m}$
Initialization: set the adjacency matrix $\mathbf{A}$, $t \leftarrow 0$ and set $\Phi^{(0)}$ to a random full rank matrix.
Repeat steps (1+2) until convergence

1. Update $\mathbf{W}^{(t+1)}$ and $\mathbf{B}^{(t+1)}$ using Eq. 6,7 respectively.
2. Update $\Phi^{(t+1)}$ by taking the limit $\tilde{\Psi}$ of Eq. 8, with $\Psi^{(0)} = \Phi^{(t)}$.

Output: kernel maps $\{\tilde{\Phi}_i\}$ and labels $\{\mathbf{y}_i\}$ with $\mathbf{y}_i = \mathbf{W}'\tilde{\Phi}_i$.

Learning Basis and Classifier. Assuming fixed $\Phi^{(t)}$ (denoted simply as Φ) and enforcing the gradient of Eq. 5 to vanish (w.r.t $\mathbf{W}$) leads to

$$\mathbf{W}^{(t+1)} = \gamma_c \left(\mathbf{I}_p + \Phi \left(\gamma_c \mathbf{C} + \gamma_s \mathbf{L} \right) \Phi' \right)^{-1} \Phi \mathbf{C} \mathbf{Y}'. \tag{6}$$

Similarly, we find $\mathbf{B}^{(t+1)}$ by solving the dual problem

$$\underset{\Lambda}{\operatorname{argmax}} \left[\underset{\mathbf{B}}{\operatorname{argmin}} \left(\frac{1}{2} \|\mathbf{X} - \mathbf{B}\Phi\|_F^2 + \operatorname{tr}\left(\mathbf{B}'\Lambda\mathbf{B}\right) - \operatorname{tr}(\Lambda) \right) \right], \tag{7}$$

where Λ is the diagonal matrix whose entry Λ_{ii} is equal to the Lagrange multiplier λ_i associated with the i^{th} equality constraint in Eq. 5. After maximizing Eq. 7 w.r.t Λ, we obtain the optimal basis $\mathbf{B}^{(t+1)} = \mathbf{X}\Phi' \left(\Phi\Phi' + \Lambda^*\right)^{-1}$.

Learning Kernel Map. Considering fixed $\mathbf{B}^{(t+1)}$ and $\mathbf{W}^{(t+1)}$ (denoted simply as $\mathbf{B}$, $\mathbf{W}$ in the remainder of this section), and the previous kernel map solution $\Phi^{(t)}$, our goal is to find $\Phi^{(t+1)}$ by solving Eq. 5. The optimization problem in Eq. 5 admits a unique solution $\Phi^{(t+1)} = \tilde{\Psi}$ where $\tilde{\Psi} = \lim_{k \to \infty} \Psi^{(k)}$ and

$$\Psi_i^{(k)} = \left(\gamma_c \mathbf{B}'\mathbf{B} + (\gamma_s \mathbf{D}_{ii} + \gamma_c \mathbf{C}_{ii}) \mathbf{W}\mathbf{W}' + \mu \mathbf{I}_p \right)^{-1} \cdot \left[\gamma_c \mathbf{B}'\mathbf{X} + \gamma_c \mathbf{W}\mathbf{Y}\mathbf{C} + \gamma_s \mathbf{W}\mathbf{W}'\Psi^{(k-1)}\mathbf{A} \right]_i, \tag{8}$$

here $\Psi_i^{(k)}$ and $[.]_i$ stand for the i^{th} column of a matrix. Proof about this kernel map solution and its convergence to a fixed point are detailed in [19]. The process described in Eq. 8 allows us to recursively diffuse the kernel maps from the labeled to the unlabeled data, through the neighborhood system defined in the graph $\mathcal{G}$. The algorithm convergences when $\|\Psi^{(k)} - \Psi^{(k-1)}\| \leq \varepsilon$; (in practice, $\varepsilon = 10^{-2}$, and convergence usually happens in less than 100 iterations).

4 Experiments

In this section, we evaluate the performance of our transductive kernel map learning (denoted TKML) in image annotation and we compare it to closely related work. For that purpose, we use two standard datasets for experiments and comparison. The MSRC dataset includes 591 images from 23 categories excluding the category "horse" (as it has a very low cardinality). This dataset is divided into two subsets with equal cardinality; the first one is used for training and the other one for testing. Corel5K dataset contains 5000 images which are annotated with 260 keywords and each image has up to 5 keywords. This dataset is divided into 4500 images for training and 500 images for testing.

Features. Images in MSRC and Corel5K are processed using a rectangular grid in order to extract densely sampled SIFT features. These features are assigned to their nearest visual words using a trained codebook of size N ($N = 512$ in practice). Spatial information is also considered using a three-level pyramid including 1×1, 2×2, 1×3 cells, and one bag-of-word histogram is computed for each cell. Consequently, each image is encoded by a concatenated descriptor of length $(1 \times 1 + 2 \times 2 + 1 \times 3)N$. In order to capture various visual informations, we combine various local features [44] including SIFT, rgbSIFT, rgSIFT, hsvSIFT, cSIFT, opponentSIFT resulting into a final visual representation joining the six SIFT features.

Learning and Annotation. Given training and test images, we define our neighborhood system using an adjacency graph where each node corresponds to an image and an edge connects two images if they are visually similar. Using this neighborhood system, we run TKML in order to measure the membership of each keyword to different test images; these memberships correspond to the scores of the underlying classifiers. A keyword is then assigned to a test image iff the score associated to that keyword is among the 5 largest values. We also applied a variant of our TKML algorithm, that weights, for each keyword, the loss of positive and negative data differently especially when they are unbalanced. This variant is referred to as weighted TKML (wTKML). Our method uses following settings in all the experiments: i) the smoothness term is $\gamma_s = 1$, ii) the fidelity coefficient is $\gamma_c = 1$, iii) the low-rank coefficient is $\mu = 10^{-8}$, iv) the convergence threshold is $\varepsilon = 10^{-2}$, v) the max number of iterations is 5, and vi) the max number of iterations for diffusion is 100. Graph construction parameters, however, depend on data sets. For MSRC, the neighborhood size $k = 6$, euclidean metric $s(\cdot,\cdot)$; for Corel5K, the neighborhood size $k = 3$, histogram intersection metric $s(\cdot,\cdot)$. More details about the setting of these parameters are given in [49].

Evaluation Criteria. Different evaluation criteria are used in order to measure the quality of this annotation process including precision (denoted **P**), recall (denoted **R**) and positive recall (denoted **N**+); these criteria are defined as

$$\mathbf{P} = \mathbb{E}_{\omega}\Big(\frac{\text{number of images correctly annotated with a keyword } \omega}{\text{number of images annotated with } \omega}\Big)$$

$$\mathbf{R} = \mathbb{E}_{\omega}\Big(\frac{\text{number of images correctly annotated with a keyword } \omega}{\text{number of images annotated with } \omega \text{ in the ground truth}}\Big)$$

$$\mathbf{N}+ = \sum_{\omega} 1_{\{(\text{number of images correctly annotated with a keyword } \omega) \geq 1\}},$$

here the expectation $\mathbb{E}_{\omega}$ is with respect to all possible keywords $\{\omega\}$ in our dataset. We further benchmark the quality of keyword assignment using break-even point (denoted **BEP** [20]),

with

$$\textbf{BEP} = \mathbb{E}_{\omega}\Big(\frac{\text{number of images correctly annotated with a keyword } \omega \text{ in a sorted list of } N_{\omega} \text{ images}}{N_{\omega}}\Big)$$

here N_{ω} is the number of images annotated with ω in the ground truth and the list of N_{ω} images is sorted by decreasing classification scores. By varying the size of the sorted lists and taking the expectation of precision, with respect to this size, we obtain the mean average precision (denoted **mAP**).

4.1 Performance & Comparison

Inductive methods. We consider 3 state-of-the-art methods for comparison: (i) SVM classifiers [45]; (ii) MKL via SMO-MKL [48]; (iii) structured SVM for multi-label classification via M3L [22]. All of these methods are tested against four choices of kernels linear, RBF, χ^2, and histogram intersection. Parameters of each method are optimally tuned. For SMO-MKL, we use SMO solver with ℓ_2 regularization. Note that SMO-MKL has been extensively trained using 36 Gram matrices resulting from the combination of the 6 kernels (linear, χ^2, Histogram Intersection, and RBF with 3 bandwidth values) and the 6 descriptors mentioned earlier. We use libSVM [2] as the standard implementation for SVM while M3L and MKL implementations are taken from their original authors.

Transductive Methods. LapSVM [37] and TranSVM [26] are accounted for comparison. While they are based on large margin approach, LapSVM is more related to our method since both share smoothness regularization term. The two methods are also tested against four choices of kernels mentioned above. The implementation of LapSVM is taken from [37] and that of TranSVM from SVM^{light} [3].

Fig. 2 and Table 1 show results and comparison on the MSRC data set. A first conclusion indicates that methods relying on labeled training data as well as unlabeled test data provide better performance. The performance of our method, as shown in Fig. 2, is equivalent to the top one, i.e., structured SVMs (M3L) even though the latter takes also into account label interactions that help improving image annotation performance. Differences between methods become clearer on the Corel5K database (see Fig. 3). As expected inductive methods (including SVMs) perform worse than transductive ones and structured SVM (M3L) in terms of recall and mAP. Better recall, mAP and N+ performances are obtained by our method (wTKML) (see Fig. 3-right).

Finally, Table 2 shows results reported in the related work including the baseline JEC [36], the sparse coding method in [53], the graph-based method in [33, 50] and the TagProp [21]. Compared to these state-of-the-art methods, our method is very competitive in at least three out of the five evaluation criteria.

5 Conclusion

We introduced in this paper, a new transductive learning approach for kernel design and multi-label classification. Our contribution resides in the variational framework that allows us to explicitly design an "optimal" kernel map as a part of the learning process. When compared to baseline inductive methods as well as related transductive ones, our approach

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[3] http://svmlight.joachims.org/

shows competitive performance on the challenging image annotation task.
As a future work, we will investigate the application of our method to other tasks including interactive image retrieval.

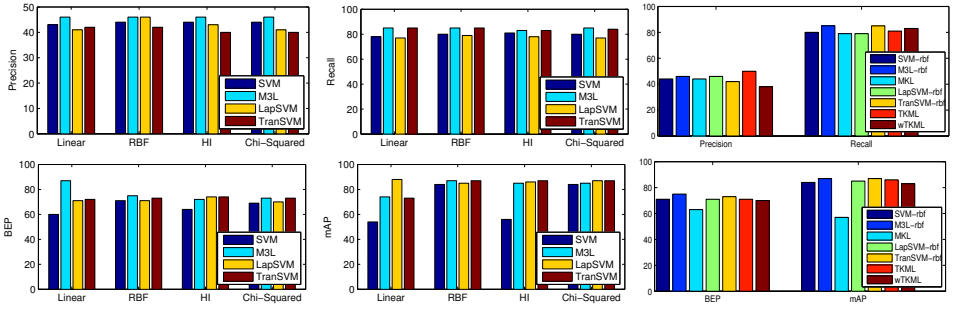


Figure 2: Image annotation performance w.r.t baseline methods on the MSRC database.

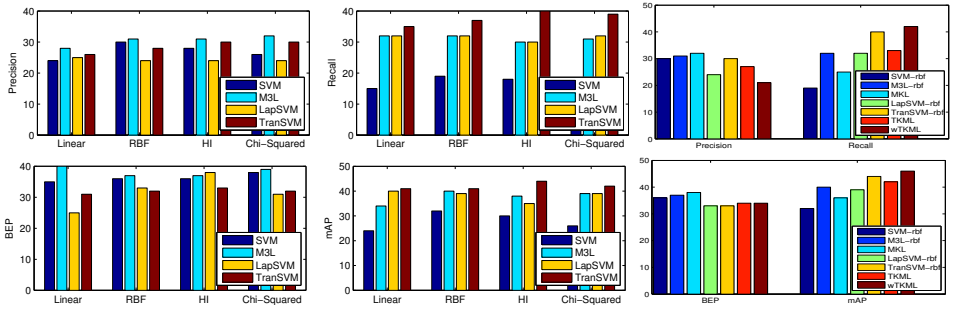


Figure 3: Image annotation performance w.r.t baseline methods on the Corel5K database.

	SVM				M3L				MKL	LapSVM				TranSVM				TKML	wTKML
	lin	rbf	hi	χ^2	lin	rbf	hi	χ^2		lin	rbf	hi	χ^2	lin	rbf	hi	χ^2		
N+	80	97	91	85	133	122	119	124	113	126	129	112	129	155	150	157	155	133	**173**

Table 1: This table shows the positive recall (N+) for different methods on the Corel5K database. As for MSRC, all the keywords are recalled.

	CRM [23]	InfNet [24]	NPDE [25]	MBRM [10]	SML [8]	TGLM [26]	MEG [27]	MSC [28]	JEC [29]	GS [7]	TagProp [30]	TKML	wTKML
P	16	17	18	24	23	25	25	25	27	30	**33**	27	21
R	19	24	21	25	29	29	31	32	32	33	**42**	33	**42**
N+	107	112	114	122	137	131	-	136	139	146	160	133	**173**

	PAMIR [31]	MSC [28]	TagProp [30]	TKML	wTKML
BEP	17	-	**36**	34	34
mAP	26	42	42	42	**46**

Table 2: This table shows the performances (with different evaluation criteria) of our proposed method and related work on the Corel5K dataset.

References

[1] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[2] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, December 2006.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29(3):394–410, March 2007.

[6] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[7] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. A new kernelization framework for Mahalanobis distance learning algorithms. *Neurocomputing*, 73(10-12):1570–1579, June 2010.

[8] Xiangyu Chen, Xiaotong Yuan, Shuicheng Yan, Jinhui Tang, Yong Rui, and Tat-Seng Chua. Towards multi-semantic image annotation with graph regularized exclusive group lasso. *ACM Multimedia*, page 263, 2011.

[9] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2010.

[10] Jia Deng and Alexander C Berg. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In *CVPR*, 2011.

[11] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-fei. What Does Classifying More Than 10 , 000 Image Categories Tell Us ? In *ECCV*, pages 71–84, 2010.

[12] Pinar Duygulu, Kobus Barnard, J.F.G de Freitas, and David A Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *ECCV*, 2002.

[13] Yariv Ephraim, Amir Dembo, and Lawrence R. Rabiner. A minimum discrimination information approach for hidden markov modeling. *IEEE Transactions on Information Theory*, 35(5):1001–1013, 1989.

[14] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009.

[15] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-Centric Recognition for Cross-category Generalization. In *CVPR*, 2010.

[16] Li Fei-fei and Li jia Li. What , Where and Who ? Telling the Story of an Image by Activity Classification , Scene Recognition and Object Categorization. In *Studies in Computational Intelligence- Computer Vision*. 2010.

[17] Shaolei Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[18] Rob Fergus, New York, and Antonio Torralba. Semi-supervised Learning in Gigantic Image Collections. In *NIPS*, pages 1–9, 2009.

[19] Vittorio Ferrari and Andrew Zisserman. Learning Visual Attributes. In *NIPS*, 2007.

[20] David Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, 2008.

[21] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009.

[22] Bharath Hariharan, Lihi Zelnik-Manor, S. V. N. Vishwanathan, and Manik Varma. Large Scale Max-Margin Multi-Label Classification with Priors. In *ICML*, 2010.

[23] Thomas Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

[24] Prateek Jain, Brian Kulis, Jason V Davis, and Inderjit S Dhillon. Metric and Kernel Learning using a Linear Transformation. *JMLR*, 2009.

[25] Jiwoon Jeon, Victor Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, pages 119–126, 2003.

[26] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[27] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer, 2002.

[28] Brian Kulis and U C Berkeley Eecs. Inductive Regularized Learning of Kernel Functions. In *NIPS*, number x, pages 1–9, 2010.

[29] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *IEEE CVPR*, pages 951–958, June 2009.

[30] Gert R G Lanckriet, Peter Bartlett, and Michael I Jordan. Learning the Kernel Matrix with Semidefinite Programming. *JMLR*, 5:27–72, 2004.

[31] Victor Lavrenko, R. Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[32] Zechao Li, Jing Liu, Xiaobin Zhu, Tinglin Liu, and Hanqing Lu. Image Annotation Using Multi-Correlation Probabilistic Matrix Factorization. In *ACM Multimedia*, pages 10–13, 2010.

[33] Dong Liu, Shuicheng Yan, Yong Rui, and Hong-Jiang Zhang. Unified Tag Analysis With Multi-Edge Graph. In *ACM Multimedia*, pages 25–34, 2010.

[34] Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[35] Zhigang Ma and Jasper Uijlings. Exploiting the Entire Feature Space with Sparsity for Automatic Image Annotation. In *ACM Multimedia*, pages 283–292, 2011.

[36] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *ECCV (3)*, pages 316–329, 2008.

[37] Stefano Melacci and Mikhail Belkin. Laplacian support vector machines trained in the primal. *J. Mach. Learn. Res.*, pages 1149–1184, July 2011.

[38] Donald Metzler and R. Manmatha. An inference network approach to image retrieval. In *CIVR*, pages 42–50, 2004.

[39] Alain Rakotomamonjy and Francis R Bach. SimpleMKL. *JMLR*, pages 1–34, 2008.

[40] Olga Russakovsky and Li Fei-fei. Attribute learning in large-scale datasets. In *ECCV*, 2010.

[41] B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, December 2001.

[42] R. Socher. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *2009 IEEE CVPR*, pages 2036–2043, June 2009.

[43] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

[44] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE PAMI*, 32(9):1582–1596, 2010.

[45] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[46] V. Vapnik and A. Sterin. On structural risk minimization or overall risk in a problem of pattern recognition. *Automation and Remote Control*, 10(3):1495–1503, 1977.

[47] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *ICML*, page 134, 2009.

[48] Bharath Hariharan S V N Vishwanathan. Efficient Max-Margin Multi-Label Classification with Applications to Zero-Shot Learning. Technical report, Microsoft Research Technical Report MSR-TR-2010-141, 2010.

[49] D-P. Vo and H. Sahbi. Transductive kernel learning. Technical report, Telecom ParisTech, 2012.

[50] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong jiang Zhang. Multi-label sparse coding for automatic image annotation. *2009 IEEE CVPR*, pages 1643–1650, June 2009.

[51] Mingrui Wu, Bernhard Schölkopf, and Gokhan Bakir. A Direct Method for Building Sparse Kernel Learning Algorithms. *Journal of Machine Learning Research*, 7:603–624, 2006.

[52] Ying Yuan, Fei Wu, Yueting Zhuang, and Jian Shao. Image Annotation by Composite Kernel Learning with Group Structure. In *ACM Multimedia*, pages 1497–1500, 2011.

[53] Shaoting Zhang, Junzhou Huang, Yuchi Huang, Yang Yu, Hongsheng Li, and Dimitris N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, pages 3312–3319, 2010.