FLORE
Repository istituzionale dell'Università degli Studi di Firenze

A Cross-media Model for Automatic Image Annotation

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

A Cross-media Model for Automatic Image Annotation / Lamberto Ballan, Tiberio Uricchio, Lorenzo Seidenari, Alberto Del Bimbo. - ELETTRONICO. - (2014), pp. ---. (ACM International Conference on Multimedia Retrieval (ICMR) Glasgow, UK April 1-4).

Availability:

The webpage https://hdl.handle.net/2158/855512 of the repository was last updated on 2019-07-03T00:11:47Z

Publisher:

ACM

Terms of use:

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

Publisher copyright claim:

La data sopra indicata si riferisce all'ultimo aggiornamento della scheda del Repository FloRe - The above-mentioned date refers to the last update of the record in the Institutional Repository FloRe

(Article begins on next page)

17 November 2026

A Cross-media Model for Automatic Image Annotation

Lamberto Ballan*, Tiberio Uricchio*, Lorenzo Seidenari, and Alberto Del Bimbo
Media Integration and Communication Center (MICC), Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy

ABSTRACT

Automatic image annotation is still an important open problem in multimedia and computer vision. The success of media sharing websites has led to the availability of large collections of images tagged with human-provided labels. Many approaches previously proposed in the literature do not accurately capture the intricate dependencies between image content and annotations. We propose a learning procedure based on Kernel Canonical Correlation Analysis which finds a mapping between visual and textual words by projecting them into a latent meaning space. The learned mapping is then used to annotate new images using advanced nearest-neighbor voting methods. We evaluate our approach on three popular datasets, and show clear improvements over several approaches relying on more standard representations.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Experimentation, Measurement

Keywords

Image annotation, Multi-label, Cross-media analysis, Tag relevance, Nearest-neighbor voting

1. INTRODUCTION

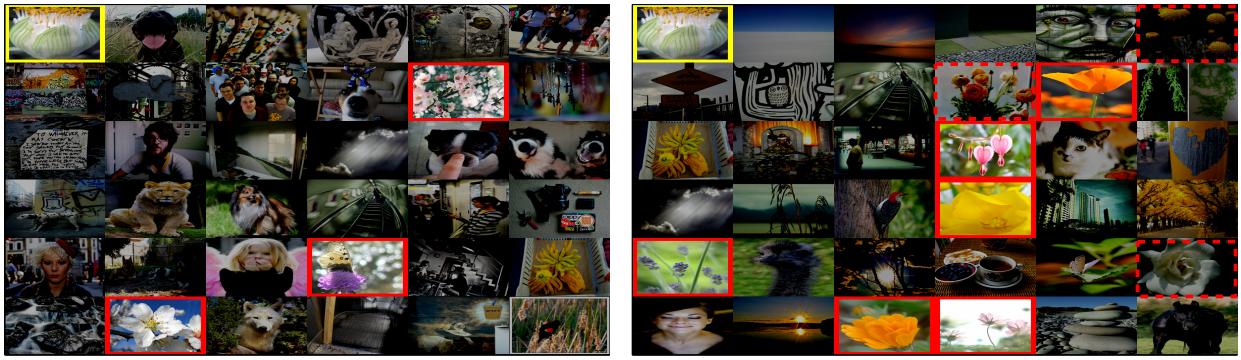
The exponential growth of media sharing websites, such as Flickr or Picasa, and social networks such as Facebook, has led to the availability of large collections of images tagged with human-provided labels. These tags reflect the image content and can thus be exploited as a loose form of labels and context. Several researchers have explored ways to use images with associated labels as a source to build classifiers or to transfer their tags to similar images [3, 18, 9, 16, 15, 29]. Automatic image annotation is therefore a very active subject of research [19, 27, 2, 17, 28, 25] since we can clearly increase performance of search and indexing over image collections that are machine enriched with a set of meaningful labels. In this work we tackle the problem of assigning a finite number of relevant labels (or tags) to an image, given the image appearance and some prior knowledge on the joint distribution of visual features and tags based on some weakly and noisy annotated data.

The main shortcomings of previous works in the field are twofold. The first is the well-known *semantic gap* problem, which points to the fact that it is hard to extract semantically meaningful entities using just low level visual features. The second shortcoming arises from the fact that many parametric models, previously presented in the literature, are not rich enough to accurately capture the intricate dependencies between image content and annotations. Recently, nearest-neighbor based methods have attracted much attention since they have been found to be quite successful for tag prediction [18, 9, 16, 23, 29]. This is mainly due to their flexibility and capacity to adapt to the patterns in the data as more training data is available. The base ingredient for a vote based tagging algorithm is of course the source of votes: the set of K nearest neighbors. In challenging real world data it is often the case that the vote casting neighbors do not contain enough statistics to obtain reliable predictions. This is mainly due to the fact that certain tags are much more frequent than others and can cancel out less frequent but relevant tags [9, 16]. It is obvious that all voting schemes can benefit from a better set of neighbors. We believe that the main bottleneck in obtaining such ideal neighbors set is the semantic gap. We address this problem using a cross-modal approach to learn a representation that maximizes the correlation between visual features and tags in a common semantic subspace.

In Figure 1 we show our intuition with an example provided by real data. We compare for the same query, a flower close-up, the first thirty-five most similar examples provided by the visual features and by our representation. The first thing to notice is the large visual and semantic difference between the sets of retrieved neighbors by the two approaches. Note also that some flower pictures, which we highlight with a dashed red rectangle, were not tagged as such. Second, note how the result presented in Figure 1(b) have more and better ranked *flower* images than the one in Figure 1(a).

*indicates equal contributions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR'14, April 1–4, 2014, Glasgow, United Kingdom.
Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00.



(a) Baseline (b) Our Method

Figure 1: Nearest neighbors found with baseline representation (a) and with our proposed method (b) for a flower image (first highlighted in yellow in both figures) from the MIRFlickr-25K dataset. Training images with ground truth tag *flower* are highlighted with a red border. Nearest neighbors are sorted by decreasing similarity and arranged in a matrix using a row-major convention. Dashed red lines indicate flower pictures not tagged as such.

Indeed with the result set in Figure 1(a) it is not possible to obtain a sufficient amount of meaningful neighbors and the correct tag *flower* is canceled by others such as *dog* or *people*.

In this paper we present a cross-media approach that relies on Kernel Canonical Correlation Analysis (KCCA) [10, 11] to connect visual and textual modalities through a common latent meaning space (called *semantic space*). Visual features and labels are mapped to this space using feature similarities that are observable inside the respective domains. If mappings are close in this semantic space, the images are likely to be instances of the same underlying semantic concept. The learned mapping is then used to annotate new images using a nearest-neighbor voting approach. We present several experiments using different voting schemes. First, a simple NN voting similar to the seminal work of Makadia *et al.* [18], and second three advanced NN models such as TagRelevance [16], TagProp [9] and 2PKNN [25].

1.1 Contribution

Other existing approaches learn from both words and images, including previous uses of CCA [10, 21, 13, 6]. In contrast, we are the first to propose an approach that combines an effective cross-modal representation with advanced nearest-neighbor models for the specific task of automatic image annotation.

In the following we show that, if combined with advanced NN schemes able to deal with the class-imbalance (i.e. large variations in the frequency of different labels), our cross-media model achieves high performance without requiring heavy computation such as in the case of metric learning frameworks with many parameters (as in [9, 25]).

We present experimental results for two standard datasets, Corel5K [3] and IAPR-TC12 [8], obtaining highly competitive results. We report also experiments on a challenging dataset collected from Flickr, i.e. the MIRFlickr-25K dataset [12], and our results show that the performance of the proposed method is boosted even further in a realistic and more interesting scenario such as the one provided by weakly-labeled images.

2. RELATED WORK

In the multimedia and computer vision communities, jointly modeling images and text has been an active research area in the recent years. A first group of methods uses mixture models to define a joint distribution over image features and labels. The training images are used by these models as components to define a mixture model over visual features and tags [14, 4, 2]. They can be interpreted as non-parametric density estimators over the co-occurrence of images and labels. In another group of methods based on topic models (such as LDA and pLSA), each topic represents a distribution over image features and labels [1, 20]. These kind of generative models may be criticized because they maximize the generative data likelihood, which is not optimal for predictive performance. Another main criticism of these models is their need for simplifying assumptions in order to do tractable learning and inference.

Discriminative models such as support vector machines have also been proposed [7, 26]. These methods learn a classifier for each label, and use them to predict whether a test image belongs to the class of images that are annotated with a particular label. A main criticism of these works resides in the necessity to define in advance the number of labels and to train individual classifiers for each of them. This is not feasible in a realistic scenario like the one of web images. Despite their simplicity, nearest-neighbor based methods for image annotation have been found to give state-of-the-art results [18, 9, 25]. The intuition is that similar images share common labels. The common procedure of the existing nearest-neighbor methods is to search for a set of visually similar images and then to select a set of relevant associated tags based on a tag transfer procedure [18, 16, 9]. In all these previous approaches, this similarity is determined only using image visual features.

3. APPROACH

The proposed method is based on KCCA which provides a common representation for the visual and tag features. We refer to this common representation as *semantic space*. Similarly to [10, 13] we use KCCA to connect visual and textual modalities, but our method is designed to effectively tackle the particular problem of image auto-annotation. In Section 3.1 we present our visual and text features with their respective kernels; next we briefly describe KCCA (Section 3.2) and the different NN schemes (Section 3.3). In Fig-

ure 2 we show an embedding computed with ISOMAP [22] of the visual data and its semantic projection. We randomly pick three tags to show how the semantic projection that we learn with KCCA better suits the actual distribution of tags with respect to the visual representation. The semantic projection improves the separation of the classes, allowing a better manifold reconstruction and, as our experiments will confirm, an improvement on precision and recall on different datasets.

3.1 Visual and Tags Views

3.1.1 Visual Feature Representation and Kernels

We directly use the 15 features provided by the authors of [9, 24][1]. These are different types of global and local features commonly used for image retrieval and categorization. In particular we use two types of global descriptors: Gist and color histograms with 16 bins in each channel for RGB, LAB, HSV color spaces. Local features include SIFT and robust hue descriptors, both extracted densely on a multi-scale grid or for Harris-Laplacian interest points. The local feature descriptors are quantized using k-means and then all the images are represented as bag-of-(visual)words histograms. The histograms are also computed in a spatial arrangement over three horizontal regions of the image, and then concatenated to form a new global descriptor that encodes some information of the global spatial layout.

In this work we use χ^2 exponential kernels for all visual features $f \in \mathcal{F}$:

$$K_{\chi^2}(h_i^f, h_j^f) = exp\left(-\frac{1}{2A}\sum_{k=1}^{d}\frac{(h_i^f(k) - h_j^f(k))^2}{(h_i^f(k) + h_j^f(k))}\right), \quad (1)$$

where A is the mean of the χ^2 distances among all the training examples, d is the dimensionality of a particular feature descriptor f and h_i^f is its respective histogram representation. It has to be noticed that all the feature descriptors are L1-normalized. Finally, all the different visual kernels are averaged to obtain the final visual representation. We obtain the kernel between two images I_i, I_j via kernel averaging:

$$K_v(I_i, I_j) = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} K_{\chi^2}(h_i^f, h_j^f). \quad (2)$$

3.1.2 Tag Feature Representation and Kernel

We use as tag features the traditional bag-of-words which records which labels are named in the image, and how many times. Supposing V is our vocabulary size, i.e. the total possible words used for annotation, each tag-list is mapped to an V-dimensional feature vector $h = [w_1, \cdots, w_V]$, where w_i counts the number of times the i-th word is mentioned in the tag list. In our case this representation is highly sparse and often counts are simply 0 or 1 values. We use these features to compute a linear kernel that corresponds to counting the number of tags in common between two images:

$$K_t(h_i, h_j) =< h_i, h_j >= \sum_{k}^{V} h_i(k)h_j(k). \quad (3)$$

[1]These features are available at: `http://lear.inrialpes.fr/people/guillaumin/data.php`.

3.2 Kernel Canonical Correlation Analysis

Given two views of the data, such as the ones provided by visual and textual modalities, we can construct a common representation. Canonical Correlation Analysis (CCA) seeks to utilize data consisting of paired views to simultaneously find projections from each feature space such that the correlation between the projected representations is maximized. In the literature, the CCA method has often been used in cross-language information retrieval, where one queries a document in a particular language to retrieve relevant documents in another language. In our case, the algorithm learns two semantic projection bases, one per each modality (i.e. the v view is the visual cue while the t view is the tag-list cue).

More formally, given N samples from a paired dataset $\{(v_1, t_1), \ldots, (v_N, t_N)\}$, where $v_i \in \mathbb{R}^n$ and $t_i \in \mathbb{R}^m$ are the two views of the data, the goal is to simultaneously find directions w_v^* and w_t^* that maximize the correlation of the projections of v onto w_v and t onto w_t. This is expressed as:

$$w_v^*, w_t^* = \arg\max_{w_v, w_t} \frac{\hat{E}[\langle v, w_v\rangle\langle t, w_t\rangle]}{\sqrt{\hat{E}[\langle v, w_v\rangle^2]\hat{E}[\langle t, w_t\rangle^2]}} = \arg\max_{w_v, w_t} \frac{w_v^T C_{vt} w_t}{\sqrt{w_v^T C_{vv} w_v w_t^T C_{tt} w_t}}, \quad (4)$$

where $\hat{E}$ denotes the empirical expectation, C_{vv} and C_{tt} respectively denote the auto-covariance matrices for v and t data, and C_{vt} denotes the between-sets covariance matrix. The solution can be found via a generalized eigenvalue problem [11].

The common CCA algorithm can only recover linear relationships, it is therefore useful to kernelize it by projecting the data into a higher-dimensional feature space by using the kernel trick. Kernel Canonical Correlation Analysis (KCCA) is the kernelized version of CCA. To this end, we define kernel functions over v and t as $K_v(v_i, v_j) = \phi_v(v_i)^T\phi_v(v_j)$ and $K_t(t_i, t_j) = \phi_t(t_i)^T\phi_t(t_j)$. Here, the idea is to search for solutions of w_v,w_t that lie in the span of the N training instances $\phi_v(v_i)$ and $\phi_t(t_i)$:

$$w_v = \sum_i \alpha_i\phi_v(v_i),$$
$$w_t = \sum_i \beta_i\phi_t(t_i), \quad (5)$$

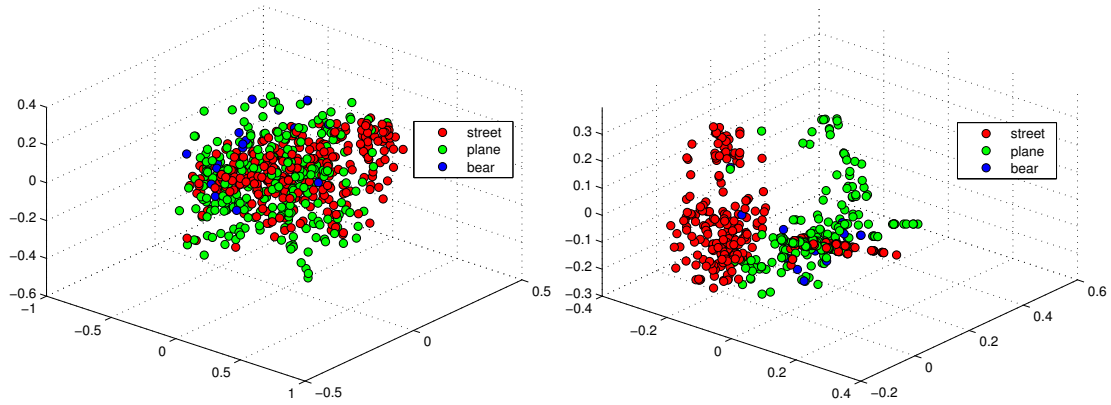
where $i \in \{1, \cdots, N\}$. The objective of KCCA is thus to identify the weights $\alpha, \beta \in \mathbb{R}^N$ that maximize:

$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta} \frac{\alpha^T K_v K_t \beta}{\sqrt{\alpha^T K_v^2 \alpha \beta^T K_t^2 \beta}}, \quad (6)$$

where K_v and K_t denote the $N \times N$ kernel matrices over a sample of N pairs. As shown by Hardoon [11], learning may need to be regularized in order to avoid trivial solutions. Hence, we penalize the norms of the projection vectors and obtain the standard eigenvalue problem:

$$(K_v + \kappa I)^{-1} K_t (K_t + \kappa I)^{-1} K_v \alpha = \lambda^2 \alpha. \quad (7)$$

The top D eigenvectors of this problem yield basis $A = \left[\alpha^{(1)} \ldots \alpha^{(D)}\right]$ and $B = \left[\beta^{(1)} \ldots \beta^{(D)}\right]$ that we use to compute the semantic projections of any vector v_i, t_i.



(a) Visual Space (b) Semantic Space

Figure 2: Visualization of three labels (Corel5K): (a) distribution of image features in the visual space (b) distribution of the same images after projecting into the semantic space learned using KCCA. Note the clearer distinction of the clusters in the semantic space.

3.2.1 Implementation Details

In order to avoid degeneracy with non-invertible Gram matrices and to increase computational efficiency we approximate the Gram matrices using the Partial Gram-Schmidt Orthogonalization (PGSO) algorithm provided by Hardoon *et al.* [11].As suggested in [11] the regularization parameter κ is found by maximizing the difference between projections obtained by correctly and randomly paired views of the data on the training set. In the experiments we have optimized both the parameters of the PGSO algorithm (i.e. κ and T); however, we found as a good starting configuration the setting $T = 30$ and $\kappa = 0.1$. We also found important swapping the use of visual and textual spaces as Hardoon [11] fixes A to be unit vectors while computing B on the basis of the two kernels.

3.3 Automatic Image Annotation Using Nearest-Neighbor Models in the Semantic Space

The intuition underlying the use of nearest-neighbor methods for automatic image annotation is that similar images share common labels. Following this key idea, we have investigated and applied several NN schemes to our semantic space in order to automatically annotate images. We briefly describe these models below.

For all baseline methods the K neighbours of a test image I_i are selected as the training images I_j for which our averaged test kernel value $K_v(I_i, I_j)$, defined in Eq. 2, scores higher. In case the semantic space projection is used, the K neighbors are computed using:

$$d(\psi(I_i), \psi(I_j)) = 1 - \frac{\psi(I_i)^T \cdot \psi(I_j)}{\|\psi(I_i)\|_2 \cdot \|\psi(I_j)\|_2}, \qquad (8)$$

where $\psi(I_i)$ is the semantic projection of a test image I_i. The projection of I_i is defined as $\psi(I_i) = K_v(I_i, \cdot)^T A$, where $K_v(I_i, \cdot)$ is the vector of kernel values of a sample I_i and all the training samples. Note that we only use the *visual* view of our data both for training and test samples.

3.3.1 Nearest-Neighbor Voting

Given a test image, we project onto the semantic space and identify its K Nearest-Neighbors. Then we merge their labels to create a tag-list by counting all tag occurrences on the K retrieved images, and finally we re-order the tags by their frequency. If we fix K to a very small number (e.g. $K = 2$) this approach is similar to the ad-hoc nearest-neighbor tag transfer mechanism proposed by Makadia *et al.* [18].

3.3.2 Tag Relevance

Li *et al.* [16] proposed a tag relevance measure based on the consideration that if different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content. Following this idea it can be assumed that, given a query image, the more frequently the tag occurs in the neighbor set, the more relevant it might be. However, some frequently occurring tags are unlikely to be relevant to the majority of images. To account for this fact the proposed tag relevance measurement takes into account both the distribution of a tag t in the neighbor set for an image I and in the entire collection:

$$tagRelevance(l, I, K) := n_t[N(I, K)] - Prior(t), \qquad (9)$$

where n_t is an operator counting the occurrences of t in the neighborhood $N(I, K)$ of K similar images, and $Prior(t)$ is the occurrence frequency of t in the entire collection.

3.3.3 TagProp

Guillaumin *et al.* [9] proposed an image annotation algorithm in which the main idea is to learn a weighted nearest neighbor model, to automatically find the optimal combination of multiple feature distances. Using $y_{it} \in \{-1, +1\}$ to represent if tag t is relevant or not for the test image I_i, the probability of being relevant given a neighborhood of K images $I_j \in N(I_i, K) = \{I_1, I_2, \ldots, I_K\}$ is:

$$p(y_{it} = +1) = \sum_{I_j \in N(I_i,K)} \pi_{ij}\ p(y_{it} = +1 | N(I_i, K)), \qquad (10)$$

$$p(y_{it} = +1 | N(I_i, K)) = \begin{cases} 1 - \epsilon & \text{for } y_{it} = +1, \\ \epsilon & \text{otherwise} \end{cases} \qquad (11)$$

$$\pi_{ij} \geq 0, \quad \sum_{I_j \in N(I_i,K)} \pi_{ij} = 1, \qquad (12)$$

where π_{ij} is the weight of a training image I_j of the neighborhood $N(I, K)$ and $p(y_{it} = +1|N(I_i, K))$ is the prediction of tag t according to each neighbor in the weighted sum.

The model can be used with rank-based (RK) or distance-based weighting; the latter can be learnt by using a single distance (referred to as the SD variant) or using metric learning (ML) over multiple distances. Furthermore, to compensate for varying frequencies of tags, a tag-specific sigmoid is used to scale the predictions, to boost the probability for rare tags and decrease that of frequent ones. Sigmoids and metric parameters can be learned by maximizing the log-likelihood $\sum_{I_i,t} \ln p(y_{it})$.

3.3.4 2PKNN

Verma and Jawahar [25] proposed a two phase method: a first pass is employed to address the class-imbalance by constructing a balanced neighborhood for each test image and then a second pass, where the actual tag importance is assigned based on image similarity.

The problem of image annotation is formulated similarly as in Guillaumin *et al.* [9], by finding the posterior probabilities:

$$P(y_{it}|I_i) = \frac{P(I_i|y_{it})P(y_{it})}{P(I_i)}. \quad (13)$$

Given a test image I_i, and a vocabulary $Y = \{t_1, t_2, \ldots, t_M\}$, the first phase collects a set neighborhoods T_{it} for each tag $t \in Y$ by selecting at least the nearest M training images annotated with t. The neighborhood of image I_i is then given by $N(I_i) = \bigcup_{t \in Y} T_{it}$. It should be noticed that a tag can have less than M training image and therefore $N(I_i)$, may still be a lightly unbalanced set of tags.

On the second phase of 2PKNN, given a tag $t \in Y$, the probability $P(I_i|t)$ is estimated by the neighborhood defined in phase one for image I:

$$P(I_i|t) = \sum_{I_j \in N(I_i)} \exp(-D(I_i, I_j)) p(y_{it} = +1|N(I_i)), \quad (14)$$

where $p(y_{it} = +1|N(I_i))$ is the presence of tag t for image I_i as in Guillaumin *et al.* [9] and $D(I_i, I_j)$ is the distance between image I_i and I_j.

In the simplest version of this algorithm $D(I_i, I_j)$ is just a scaled version of the distance $wD(I_i, I_j)$, where w is a scalar. Authors in [25] also propose a more complex version where $D(I_i, I_j)$ can be parameterized as a Mahalanobis distance where the weight matrix can be learned in a way that the resulting metric will pull the neighbors from the T_t belonging to ground-truth tags closer and push far the remaining ones.

4. EXPERIMENTS

We evaluate the performance of our cross-media model for automatic image annotation on three popular datasets and we compare it to closely related work.

4.1 Datasets

Corel5K. The Corel5K dataset [3] has been the standard evaluation benchmark in the image annotation community for around a decade. It contains 5,000 images which are annotated with 260 labels and each image has up to 5 different labels (3.4 on average). This dataset is divided into 4,500 images for training and 500 images for testing.

IAPR-TC12. This dataset was introduced in [8] for cross-language information retrieval and it consists of 17,665 training images and 1,962 testing images. Each image is annotated with an average of 5.7 labels out of 291 candidate.

(a) **Corel5K**

	NN-voting		TagRel [16]		TagProp [9]		2PKNN [25]	
	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA
P	26	**37**	25	**36**	29	**35**	36	**42**
R	30	**36**	35	**37**	35	**40**	38	**46**
N+	135	**139**	**151**	144	144	**149**	169	**179**

(b) **IAPR-TC12**

	NN-voting		TagRel [16]		TagProp [9]		2PKNN [25]	
	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA
P	32	**56**	27	**57**	37	**58**	46	**59**
R	21	**25**	26	**28**	22	**26**	29	**30**
N+	**235**	213	**258**	246	225	**235**	**272**	259

(c) **MIRFlickr-25K**

	NN-voting		TagRel [16]		TagProp [9]		2PKNN [25]	
	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA	Baseline	KCCA
P	34	**51**	38	**50**	37	**55**	16	**56**
R	26	**35**	22	**37**	26	**36**	6	**25**
N+	17	**18**	18	18	18	18	16	**18**

Table 1: This table shows the results of several configurations of our method based on KCCA and baselines on the Corel5K , IAPR-TC12 and MIRFlickr-25K datasets.

MIRFlickr-25K. The MIRFlickr-25K dataset has been recently introduced to evaluate keyword-based image retrieval methods. The set contains 25,000 images that were downloaded from Flickr and for each one of these images the tags originally assigned by the users are available (as well as EXIF information fields and other metadata such as GPS). It is a very challenging dataset since the tags are weak labels and not all of them are actually relevant to the image content. There are also many meaningless words. Therefore a pre-processing step was performed to filter out these tags. To this end we matched each tag with entries in Wordnet and only those tags with a corresponding item in Wordnet were retained. Moreover, we removed the less frequent tags, whose occurrence numbers are below 50. The result of this process is a vocabulary of 219 tags. The images are also manually annotated for 18 concepts (i.e. labels) that are used to evaluate the automatic annotation performances. As in [24], the dataset is divided into 12,500 images for training and 12,500 images for testing.

4.2 Evaluation Measures

We evaluate our models with standard performance measures, used in previous work on image annotation. The standard protocol in the field is to report Precision and Recall for fixed annotation length [3]. Thus each image is annotated with the n most relevant labels (usually, as in this paper, the results are obtained using $n = 5$). Then, the results are reported as mean precision **P** and mean recall **R** over the

	Previously reported results													ML		
	CRM [14]	InfNet [19]	NPDE [27]	MBRM [4]	SML [2]	TGLM [17]	GS [28]	JEC-15 [9]	TagProp σRK [9]	TagProp σSD [9]	RF-opt [5]	KSVM-VT [26]	2PKNN [25]	TagProp σML [9]	2PKNN ML [25]	**Our best result**
P	16	17	18	24	23	25	30	28	26	28	29	32	**39**	33	**44**	**42**
R	19	24	21	25	29	29	33	33	34	35	40	**42**	40	42	**46**	**46**
N+	107	112	114	122	137	131	146	140	143	145	157	**179**	177	160	**191**	**179**

Table 2: This table shows the results of our method and related work on the Corel5K dataset (as reported in the literature). JEC-15 refers to the JEC [18] implementation of [9] that uses our 15 visual features.

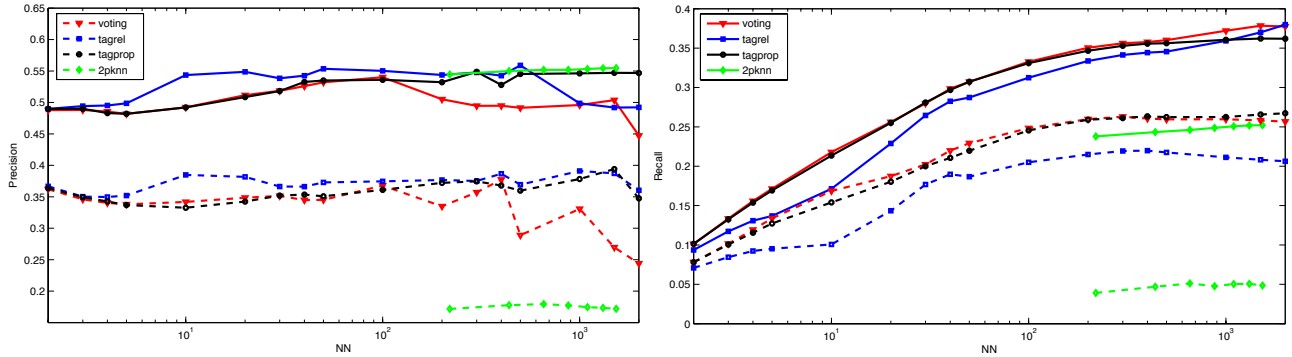


Figure 3: Precision and recall of all the methods on MIRFlickr-25k varying the number of nearest neighbors. Dashed lines represent baseline methods. Note that 2PKNN implicitly define the size of the neighborhood based only on the number of images per labels.

ground-truth labels; **N+** is often used to denote the number of labels with non-zero recall value. Note that each image is forced to be annotated with n labels, even if the image has fewer or more labels in the ground truth. Therefore we will not measure perfect precision and recall figures.

4.3 Results

As a first experiment we compare our method with the corresponding nearest neighbor voting schemes. It can be seen from Table 1 that our approach improves over baseline methods in every setting on all datasets. Precision is boosted notably, confirming the better separation of the classes in the semantic space (as previously discussed in Section 3). Also recall is improved by a large margin on Corel5K and MIRFlickr-25k. On IAPR-TC12 recall improvement is less pronounced. We believe this is due the different amount of textual annotation: IAPR-TC12 has an average of 5.7 tags per image (TPI) and up to 23 TPI while on Corel5K and MIRFlickr-25k the average TPI is respectively 3.4 and 4.7 with a maximum of 5 and 17 TPI respectively. Recalling that we are predicting $n = 5$ tags per image, recall is harder to improve on this dataset.

We conduct an evaluation of how the amount of neighbours affect the performance for both our method and the baseline on the challenging MIRFlickr-25k dataset. As can be seen from Figure 3 the KCCA variants (solid lines) of the four considered voting schemes systematically improve both precision and recall for any amount of nearest neighbors used. Note that in both cases, a similar pattern emerges due the natural instability of NN methods.

It is interesting to note that while recall gets better as the neighborhood gets bigger, saturating at near $2,000$ neighbours, precision depends on the algorithm chosen. Basic voting and Tag Relevance show an improvement until 200 neighbors and then begin decreasing; TagProp improves until saturates at around 900.

2PKNN misses a direct parameter to choose the dimension of the neighborhood, but it implicitly defines it by choosing at most M images per label. However, while it has a clear advantage on Corel5K and IAPR-TC12, both as a baseline and after the projection, it fails to achieve comparable performance on MIRFlickr-25K. We believe that this is due to the noisy and missing tags of MIRFlickr-25K, a notable difference on this more realistic and challenging dataset.

Comparing with the state of the art, on Tables 2 and 3, our method achieves better performance than all previous works while it is comparable with the state of the art method 2PKNN [25] on Corel5K. Our method performs slightly worse than 2PKNN in metric learning configuration. However, metric learning involves a learning procedure with many parameters that rise the complexity of optimization and undermines scalability.

Our method, once learned the semantic space, continues to work in what we call an open world setting. In this setting that is indeed more realistic, the amount of tags per image evolves over time. That is the case of big data from social media and, more in general, from the web.

We also report in Table 4 a comparison with the methods presented in [9, 24] using per-image average precision (iAP). This measure indicates how well a method identifies

	Previously reported results							ML		
	MBRM [4]	GS [28]	JEC-15 [9]	TagProp σSD [9]	RF-opt [5]	KSVM-VT [26]	2PKNN [25]	TagProp σML [9]	2PKNN ML [25]	**Our best result**
P	24	32	29	41	44	47	**49**	46	**54**	**59**
R	23	29	19	30	31	29	**32**	35	**37**	**30**
N+	223	252	211	259	253	268	**274**	266	**278**	**259**

Table 3: This table shows the results of our method and related work on the IAPR-TC12 dataset (as reported in the literature).

	Previously reported results					ML	
	random	SVM v	SVM t	SVM v+t	TagProp RK	TagProp ML	**Our best result**
iAP	5.6	44.2	32	45	46.3	47.3	**50.8**

Table 4: This table shows the results of our method and related work [24] on the MIRFlickr-25k dataset.

relevant concepts for a given image. Our method combining the 2PKNN voting scheme, without metric learning, with the semantic projection outperforms all the other methods.

4.3.1 Qualitative Analysis

In Figure 4 we present some anecdotal evidence for our method (from the MIRFlickr-25k dataset). It can be seen that TagProp and TagRel perform better in general for the baseline representation and our proposed KCCA variant. It has to be noted that for challenging images where visual features can be deceiving our cross-modal approach allows to retrieve more tags. As an example see the first two rows: a close-up of a flower and a cloudy sunset with a road. For the first one it is not surprising that visual features do not provide enough good neighbors to retrieve the *flower* tag. For the second one none of the baseline method can retrieve the *sunset* and *cloud* tags; we believe that this is due to the lack of color features. In this two cases it is clear that semantically induced neighbors in the common space can boost the accuracy.

Another challenging example is shown at row five: a *girl* is depicted behind an object that hides a part of the face. This image component do not have enough visual neighbors to retrieve its tags. With our representation we are able to retrieve *girl* and *portrait* in the first three voting schemes and also *people* in the TagProp voting scheme, though *face* and *woman* may be considered correct even if not present in the ground truth tags.

5. CONCLUSIONS

We presented a cross-media model based on KCCA to automatically annotate images. We learn semantic projections for both textual and visual data. This representation is able to provide better neighbors for voting algorithms. The experimental results show that our method makes consistent improvements over standard approaches based on a single-view visual representation as well as other previous work that also exploited tags. We report also experiments on a challenging dataset collected from Flickr and our results show that the performance of the proposed method is boosted even further in a realistic scenario such as the one provided by weakly-labelled images. Possible extensions of this work include the exploration of how richer textual and semantic cues from natural language annotations might also improve our model.

Acknowledgments. L. Ballan was supported by a grant from the Tuscany Region, Italy, for the AQUIS-CH project (POR CRO FSE 2007-2013).

6. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE TPAMI*, 29(3):394–410, 2007.

[3] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, 2002.

[4] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proc. of CVPR*, 2004.

[5] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *Proc. of ECCV*, 2012.

[6] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for internet images, tags, and their semantics. *IJCV*, in press, 2013.

[7] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE TPAMI*, 30(8):1371–1384, 2008.

[8] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proc. of LRECW*, 2006.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. of ICCV*, 2009.

[10] D. R. Hardoon and J. Shawe-Taylor. KCCA for different level precision in content-based image retrieval. In *Proc. of IEEE CBMI*, 2003.

[11] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[12] M. J. Huiskes and M. S. Lew. The MIR flickr retrieval evaluation. In *Proc. of ACM MIR*, 2008.

[13] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153, 2012.

[14] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proc. of NIPS*, 2003.

[15] L.-J. Li and L. Fei-Fei. OPTIMOL: Automatic online picture collection via incremental model learning.

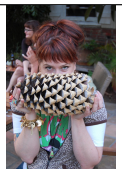
	Baselines				KCCA models			
	NN-voting	TagRel	TagProp	2PKNN	NN-voting	TagRel	TagProp	2PKNN
	dog graffiti people black art	dog graffiti animal people house	graffiti dog people face art	graffiti dog people face art	**flower** flowers pink green spring	**flower** flowers pink green red	**flower** flowers green pink white	graffiti dog people face art
	sky clouds water landscape trees	clouds **sky** landscape water trees	clouds **sky** water landscape trees	clouds **sky** water landscape trees	clouds **sky** landscape **sunset** blue	clouds **sky** **sunset** landscape **cloud**	clouds **sky** landscape **sunset** beach	clouds **sky** water landscape trees
	japan art water dog trees	japan zoo dog trees art	japan water dog park art	japan water dog park art	**portrait** **girl** **tree** street green	**portrait** **girl** woman **tree** trees	**portrait** **girl** green **tree** trees	japan water dog park art
	pink flower japan baby portrait	pink baby japan cake crochet	pink japan flower japanese vintage	pink japan flower japanese vintage	**food** chocolate cake fruit red	**food** cake chocolate dog crochet	**food** chocolate cake red fruit	pink japan flower japanese vintage
	japan **people** man street bicycle	japan man **people** bicycle animal	japan **people** animal kid eye	japan **people** animal kid eye	**portrait** **girl** girls hair face	**portrait** **girl** face woman hair	**portrait** **girl** face **people** woman	japan **people** animal kid eye
	street architecture beach white snow	street snow architecture beach home	beach street people portrait landscape	beach street people portrait landscape	beach **sea** clouds **sky** **water**	beach **sea** sunset ocean clouds	beach **sea** clouds ocean **water**	beach street people portrait landscape
	green garden people flower spring	green waterfall garden bird colours	green grass garden feet water	green grass garden feet water	dog **animal** zoo green dogs	dog **animal** animals puppy dogs	dog **animal** zoo dogs green	green grass garden feet water

Figure 4: Anecdotal results of the baseline methods and our proposed representation for a set of challenging images (MIRFlickr-25K dataset). The tags are ordered by their relevance scores.

IJCV, 88(2):147–168, 2010.

[16] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE TMM*, 11(7):1310–1322, 2009.

[17] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proc. of ECCV*, 2008.

[19] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proc. of ACM CIVR*, 2004.

[20] F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: Constraining the latent space. In *Proc. of ACM Multimedia*, 2004.

[21] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proc. of ACM Multimedia*, 2010.

[22] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[23] T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo. An evaluation of nearest-neighbor methods for tag refinement. In *Proc. of IEEE ICME*, 2013.

[24] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the mirflickr set. In *Proc. of ACM MIR*, 2010.

[25] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *Proc. of ECCV*, 2012.

[26] Y. Verma and C. V. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *Proc. of BMVC*, 2013.

[27] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proc. of ACM CIVR*, 2005.

[28] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metexas. Automatic image annotation using group sparsity. In *Proc. of CVPR*, 2010.

[29] A. Znaidia, , H. Le Borgne, and C. Hudelot. Tag completion based on belief theory and neighbor voting. In *Proc. of ACM ICMR*, 2013.