Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems

Jose Costa Pereira, Nuno Vasconcelos

Department of Electrical and Computer Engineering
University of California San Diego
Engineering Building 1, La Jolla, CA 92037, USA

Abstract

In query-by-semantic-example image retrieval, images are ranked by similarity of semantic descriptors. These descriptors are obtained by classifying each image with respect to a pre-defined vocabulary of semantic concepts. In this work, we consider the problem of improving the accuracy of semantic descriptors through cross-modal regularization, based on auxiliary text. A cross-modal regularizer, composed of three steps, is proposed. Training images and text are first mapped to a common semantic space. A regularization operator is then learned for each concept in the semantic vocabulary. This is an operator which maps the semantic descriptors of images labeled with that concept to the descriptors of the associated texts. A convex formulation of the learning problem is introduced, enabling the efficient computation of concept-specific regularization operators. The third step is the selection of the most suitable operator for the image to regularize. This is implemented through a quantization of the semantic space, where a regularization operator is associated with each quantization cell. Overall, the proposed regularizer is a non-linear mapping, implemented as a piecewise linear transformation of the semantic image descriptors to regularize. This transformation is a form of cross-modal domain adaptation. It is shown to achieve better performance than recent proposals in the domain adaptation literature, while requiring much simpler optimization.

Keywords: content-based image retrieval, query-by-example, domain adaptation, semantic representation, cross-modal regularization, class-specific regularization

1. Introduction

Image representation is a central component of computer vision problems such as image classification or content-based image retrieval (CBIR). In this context, the

Email addresses: josecp@ucsd.edu (Jose Costa Pereira), nuno@ece.ucsd.edu (Nuno Vasconcelos)

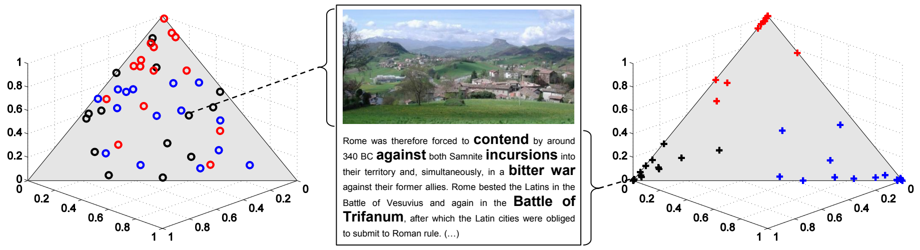


Figure 1: An excerpt from an article of the "Warfare" class from the Wikipedia dataset, with the corresponding image (middle). Left: representation of the image component of various articles from the dataset in a semantic space of three concepts ("History", "Royalty" and "Warfare"). Different colors correspond to different article classes (black for "Warfare", blue for "Royalty", and red for "History"). Right: similar representation for the text components. Note that the concept probability estimates are much noisier for images than text. In result, the image semantics are substantially more ambiguous than the text semantics.

design of visual features has been a subject of substantial interest. Early representations relied on explicit representation of low-level image properties such as color, texture, or shape, through color histograms [1], color moments [2, 3], Gabor wavelets [4], Fourier features [5], stochastic models [6], or shape contexts [7], among others. More recently, substantial effort has been devoted to the extension and robustification of these representations, through operations like normalization and spatial pooling, leading to modern descriptors such as SIFT [8], HoG [9], SURF [10], spatial pyramids [11], or Fisher vectors [12].

It was also realized, early on, that one of the limitations of these representations is a *semantic gap* [13] between strict visual similarity, i.e. similarity in terms of patterns color or texture, and human judgments of image similarity. This spurred significant interest in the development of representations that account for *semantic abstraction* [14, 15, 16, 17, 18, 19, 20, 21, 22]. In CBIR, such representations are designed by identifying a vocabulary of concepts of interest for the retrieval operation and learning classifiers for the detection of these concepts. Images are then classified and mapped to a space where each feature is a score for the detection of a concept. Several methods have been proposed to implement this approach, under different terminology. In this work, we adopt the framework of [18], which refers to the representation as a *semantic representation*, and relies on the vector of *posterior probabilities* of the image under the concepts in the vocabulary, as semantic feature vector. This feature vector is denoted a *semantic multinomial* (SMN). As illustrated in Figure 1, this representation maps each image into a point on the probability simplex. It should be noted that other implementations of semantic representation have been proposed in the literature, e.g. the query-by-example semantic retrieval method of [17], the *classeme* representation of [21], or the *object bank* of [22].

The representation of images in a semantic space has several advantages. First, the generalization from low-level features to semantic concepts enables similarity measures that correlate much better with the expectations of CBIR users [15, 18, 23]. Second, because semantic features are, by definition, discriminant for tasks like image categorization, the semantic representation enables the solution of these tasks with low-dimensional classifiers [24, 25]. Third, the semantic representation is naturally aligned with recent computer vision interest on *contextual modeling* [26, 27, 28, 29, 30]. This is of importance for tasks such as object recognition, where the detection of contextually related objects has been shown to improve the detection of certain objects of interest [31, 32, 33], or semantic segmentation, where the coherence of segment semantics can be exploited to achieve more robust segmentations [34, 35, 36]. Finally, due to their abstract nature, semantic spaces enable a unified representation for data from different content modalities, e.g. images, text, or audio. This opens up a new set of possibilities for multimedia processing, enabling operations such as *cross-modal* retrieval, where an image is used to search a database of texts and vice-versa [37], or where an audio clip is used to rank a set of images [38].

In this work, we exploit this support for cross-modal processing to design an improved image representation for CBIR. The basic idea is to leverage the fact that most images exist in a rich multimodal context, e.g. web-pages, which provide contextual information about the image content. In fact, some of this information may be much easier to model or classify than the image itself. For example, text classifiers tend to have higher accuracy than state-of-the-art image classifiers. Due to this, an SMN inferred from an image is likely to be more noisy than an SMN derived from an associated text document. This is illustrated in Figure 1, where SMNs derived from images scatter through the semantic space much more than those derived from text.

A question that arises naturally is whether it would be possible to exploit the presence of this text to denoise the semantic representation of the image. One possibility would be to simply replace the image SMN with the associated text SMN. This would reduce to the cross-modal retrieval scheme of [37], where a query image is matched to a database of texts. While effective, this solution is not fully general, since it assumes the availability of text for all images in the CBIR database. A more general solution is to collect a dataset of image-text pairs and *learn a transformation* that maps the ambiguous image semantics on the left of Figure 1 to the less ambiguous text semantics on the right. This transformation can then be applied to images that have no complementary text. Because this denoising operation is likely to enable better generalization for all retrieval operations we denote it as a *regularization* of the semantic image representation. Since text information is used to regularize visual information, the process is denoted *cross-modal regularization.* The denoised semantic representation is denoted as *regularized image semantics*.

We propose a cross-modal regularizer of image semantics (RIS) composed of three steps, illustrated in Figure 2. Training images and texts are first mapped to the semantic space. A regularization operator is then learned for each concept in the semantic vocabulary. This operator maps SMNs of images labeled with that concept to the SMNs of the associated texts. Because the transformation is linear on an affine space (probability simplex), and the objective function is to minimize the mean squared error of the mapping, the problem can be framed in a convex formulation, which lends itself to efficient optimization. The process results in a set of concept-specific regularization operators. The final step is a procedure for the selection of the most suitable regularization operator for the image to regularize. This can be seen as a quantization of the probability simplex, where each quantization cell is associated with a regularization

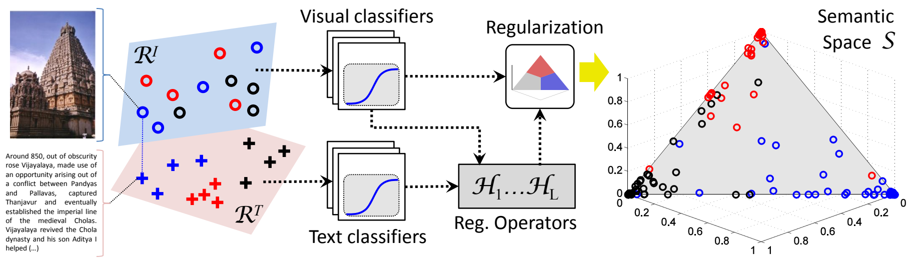


Figure 2: Proposed cross-modal regularizer of image semantics. Images and text are first mapped to the semantic space, using a set of classifiers. A regularization operator is learned per concept in the semantic vocabulary. The probability simplex is finally quantized and each of these operators assigned to a quantization cell. After regularization, images labeled with the same concept tend to cluster in a subspace of the simplex. This contains the vertices of the simplex associated with the concept and its contextually related concepts. As before, different colors correspond to images labeled with different concepts.

operator. Overall, the proposed regularizer is a non-linear mapping, implemented as a piecewise linear transformation of the image SMN to regularize. This is shown to enable better performance than other recent proposals in the domain adaptation literature [39, 40, 41, 42, 43], and requires a much simpler optimization.

The paper is organized as follows. Section 2 discusses previous related work. Section 3 reviews the fundamental concepts of semantic representation. The proposed operator is then introduced in Section 4. Section 5 presents an extensive experimental evaluation of the regularizer in the context of CBIR. Finally, some conclusions are presented in Section 6. A preliminary version of this work appeared in [44].

2. Related Work

CBIR has been a subject of research for many years. Popular retrieval systems such as QBIC [45] and Virage [46], sprung the first efforts for Internet-scale image search engines such as Visualseek [47] and Webseer [48]. These systems were based on similarity of low-level descriptors accounting for properties such as image color and texture. Semantic representations were first introduced in the video classification literature [14, 15, 16] and then extended to the CBIR literature. In this context, one of the first and most comprehensive efforts towards semantic representation was the ImageScape system [49]. [17] extended the popular *query-by-example* retrieval paradigm to the realm of semantic representations. Many other proposals have since been made in the CBIR, scene classification, object recognition, and video understanding literatures [18, 21, 22]. Some of these apply to special domains or specific sets of semantic concepts. For example, the space of attributes [19, 50, 20] is a mid-level semantic representation that has enjoyed substantial popularity in recent years [51, 52, 53, 28, 30].

Recently, deep convolutional neural networks have gained substantial popularity in the task of large scale visual recognition [54], including scenarios of joint image-text embeddings [55, 56].

The starting point for this work is the *query-by-semantic-example* retrieval paradigm of [18]. This introduced the SMN image representation and extended the minimum probability of error retrieval framework of [57] to the semantic domain. It consists of retrieving images by similarity of the associated SMNs. This was demonstrated to significantly improve the performance of the classical *query-by-visual-example,* where images are matched by similarity of visual descriptors [18]. It should be noted, however, that most of the ideas now proposed could be applied to most other semantic image representations in the literature. The question that we now investigate is whether it is possible to improve any such semantic representation by taking advantage of additional data modalities. In particular, whether given a training set of images and text, it is possible to learn a transformation that denoises the semantic representation of unseen images. This is expected to further improve QBSE performance.

Since it leverages text to improve image retrieval, cross-modal regularization is a form of *transfer learning.* This consists of transferring information from an *auxiliary* dataset to regularize a learning operation on a *target* dataset. Transfer learning is useful when learning is poorly constrained in the target domain, e.g. when too little training data is available. Several forms of transfer learning have been proposed. The most popular is probably *semi-supervised learning* [58], where a small set of labeled target data is augmented by a large auxiliary corpus of unlabeled data. These methods assume that the statistics of the target and auxiliary datasets are similar and are not directly applicable to cross-modal regularization. A second form is *multi-task* learning [59], where a common model and training data are shared for the solution of two or more learning tasks, e.g. the simultaneous classification of images and text. This is again unlike cross-modal regularization, where the goal is to learn improved image classifiers only. No text classification is performed after learning.

A third form of transfer learning is *model adaptation*, where auxiliary data is used to regularize the parameters of a target model, which can be either generative [60, 61, 62, 63, 64, 65] or discriminative [66, 67, 68, 69, 70]. Although this is sometimes denoted *domain adaptation,* the latter usually refers to methods that regularize the target feature space, rather than the models themselves. This is frequently implemented by learning a feature transformation that maximizes the similarity of feature vectors from target and auxiliary domains [71, 72, 73, 41, 74, 75]. Some methods have also been proposed to implement both domain and model adaptation [42]. The proposed approach to cross-modal regularization can be seen as a form of domain adaptation, although it has significant differences with respect to previous implementations of the former.

First, while domain adaptation assumes more auxiliary than target data, this is not the case for cross-modal regularization. Here, the problem is instead that data from the two modalities has different degrees of *semantic ambiguity:* cross-modal regularization is useful even if there is infinite image data. Second, most domain adaptation methods assume that auxiliary and target domains produce data of the same type, e.g. images taken under different views or from different datasets. This simplifies the problem in two ways. One, it enables simplifying assumptions, e.g. the existence of a smooth path through a sequence of subspaces between the auxiliary and target domains [73, 41], that does not hold for cross-modal regularization. Another, it implies the absence of a semantic gap between the two domains, leading to a simpler correspondence problem than that of cross-modal regularization. This assumption contradicts the essence of cross-modal regularization, where the goal is to leverage the smaller semantic ambiguity of text to regularize image classification.

Perhaps due to this, the notion of performing regularization in a semantic space has received little attention in the literature. Instead, domain adaptation is usually implemented through a global transformation between low-level features in the auxiliary and target domains. This is the case even for the few approaches previously proposed for cross-modal domain adaptation using images and text [43, 76]. These methods simply learn a feature transformation between the two spaces, denoted a *translator,* from co-occurrence counts of visual and text words. While global low-level transformations can be used for cross-modal regularization, our experiments show that they have weaker performance than the now proposed combination of semantic-specific regularization operators.

3. Semantic Image retrieval

In this section, we briefly review the semantic image retrieval framework used in this work.

3.1. Semantic space

A semantic representation consists of a mapping from a low-level feature space to a space where each feature has well defined semantics. Images are first represented in a low-level feature space $\mathcal{X}$, e.g. the space of SIFT descriptors sampled over a predefined image grid. Given a set of images $\mathcal{G} = \{\mathcal{I}_1, \dots, \mathcal{I}_G\}$, each image is represented as a bag of descriptors $\mathcal{I}_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,n}\}$. A vocabulary $\mathcal{L} = \{z_1, z_2, \dots, z_L\}$ of L semantic concepts is then defined. These can be broad classes, such as "indoors"', "sports", "forest", a finer grained set like object classes or object attributes, such as "tall" or "four-legged", or any other semantic classes of interest. This vocabulary is then used to design a classifier that assigns a score $\pi_{i,j}$ to each image $\mathcal{I}_i$ under each concept z_j. The vector of scores π_i are then regarded as the features of the image under the semantic representation. This can be seen as the projection of the image into a space $\mathcal{S}$ where each dimension corresponds to a concept in the vocabulary $\mathcal{L}$. The space $\mathcal{S}$ is usually denoted as the *semantic space*.

3.2. Semantic representations

Two main types of semantic representation have been investigated in the literature. In the first, the semantic space $\mathcal{S}$ consists of a set of mutually exclusive classes [18]. For example, the classes in a taxonomy used to organize an image database, where each image is placed in one and only one folder. In this case, the class label is a categorical random variable $Z \in \{1, \dots, L\}$ and the semantic representation of image $\mathcal{I}_i$ a vector of class probabilities $\pi_{i,j}$ that add up to one. In the second, $\mathcal{S}$ consists of a set of non-exclusive classes. For example, a set of binary attributes [19, 20] that can be simultaneously active for any $\mathcal{I}_i$. In this case, the class label is a multivariate Bernoulli random variable $Z \in \{0, 1\}^L$, i.e. a vector of independent binary random variables, and the entries of π_i do not add to one.

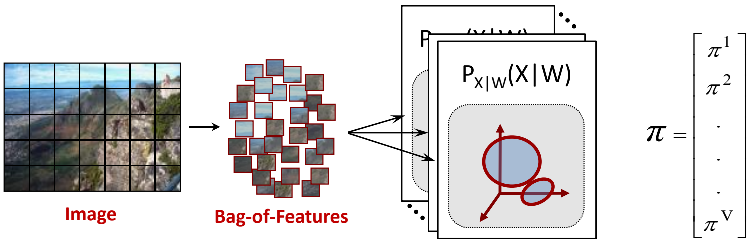


Figure 3: Example of the categorical semantic representation of an image. The image is decomposed into patches, whose probabilities are evaluated under several concept models. The vector π of posterior concept probabilities is the SMN image descriptor.

The distinction is somewhat artificial, since the first representation can be extended into a hierarchical taxonomy, where higher levels in the hierarchy are composed of broader images classes, containing images that belong to different classes in the subsequent levels. Attribute-based classes could be implemented at these higher levels [77, 78]. Similarly, a retrieval system that adopts the second representation must always have access to a disjoint set of classes, namely the classes used as groundtruth to optimize and evaluate the retrieval operation. This may only be used non-parametrically, e.g. retrieval may be based on a nearest-neighbor search, but must exist. Otherwise, no claims can be made about the optimality of the system, it is not clear what the system attempts to do, and no claims can be made that the system is preferable to any other system. The two representations can probably be best seen as alternative semantic views of an image database. One view based on generic semantics (attributes) that can be shared by all images, the other view based on categorical semantics that can be used to organize images into disjoint sets. The two views can also be combined, e.g. by expressing images as attribute vectors, mapping these vectors into a categorical variable (e.g. things that have "fur", and "ears" belong to the class "dog" if they also "eat meat" **or** to the class "cat" if they instead "eat fish"), and using the resulting probabilities as dimensions of $\mathcal{S}$.

3.3. Implementation

The regularization procedures proposed in this work can be applied to the two types of semantic representations. For simplicity of the presentation, we limit the discussion to the categorical view, and adopt the approach of [79]. The modifications needed to extend the regularization procedure to the multivariate Bernoulli representation are discussed in Appendix A.

Under the categorical representation of [79] image descriptors are considered samples from a random variable $\mathbf{X}$, concepts from a random variable $Z \in \{1, \ldots, L\}$, and each concept assumed to induce a probability density, $P_{\mathbf{X}|Z}(\mathbf{x}|z)$ on $\mathcal{X}$. Bayes rule then enables the representation of image $\mathcal{I}_i$ as a vector of posterior probability scores

$$\pi_{i,j} = P_{Z|\mathbf{X}}(j|\mathcal{I}_i). \tag{1}$$

An illustration is shown in Figure 3. In this way, the vector π_i defines a multinomial distribution, denoted as *semantic multinomial* (SMN) [18], and the semantic space $\mathcal{S}$ is a probability simplex, i.e. all dimensions of π_i are positive and add to one. Given a set of manually labeled training examples per concept, the posterior probabilities π_{i_j} can be learned in several manners. One possibility is to learn the concept distributions $P_{\mathbf{X}|Z}(\mathbf{x}|z), \forall z$ using the training set, and apply Bayes rule to compute the posteriors of (1). Another possibility is to learn a discriminative multi-class classifier, which produces estimates of the posterior probabilities directly. In this work, we adopt the latter strategy, which is implemented with the multi-class logistic regression package of [80].

In all our experiments, the classes used to define $\mathcal{S}$ are the groundtruth classes inherent to the optimality criterion used to 1) design the retrieval system and 2) measure its performance. For this reason, we will use the terms *semantics* or *classes* interchangeably in the remainder of this work. We note that this choice of semantics makes the mapping from $\mathcal{X}$ to $\mathcal{S}$ a discriminant feature transformation for the retrieval operation. Discriminant transformations, i.e. transformations informed by the groundtruth classes, are a commonly used feature extraction procedure in machine learning. In prior work, we have investigated alternative semantic configurations, e.g. using an expanded set of classes derived from various datasets, or using more abstract classes obtained through various combinations of the groundtruth classes [81]. These experiments have shown that the semantic representation is quite robust, as these variations only produced minor changes in retrieval accuracy. We do not repeat such studies here, partly because there is no reason to expect different results and partly because all currently available datasets that include both images and text only have categorical annotations.

3.4. Query by semantic example

Given a semantic space $\mathcal{S}$, image retrieval is implemented with the query-by-semantic example procedure of [18]. This consists of mapping all images $\mathcal{I}_i$ in a database into $\mathcal{S}$, by computing the associated SMNs π_i, and measuring image similarity with any measure of similarity between SMNs. Given a query image $\mathcal{I}_q$, and the associated SMN π_q, the database images are ranked by increasing values of $d(\pi_q, \pi_i)$ where $d(.,.)$ is a suitable measure of SMN distance. Several such measures can be used, in this work we adopt the Kullback-Leibler divergence

$$d(\pi_q, \pi_i) = \sum_{j=1}^{L} \pi_{q,j} \log \left(\frac{\pi_{q,j}}{\pi_{i,j}} \right). \tag{2}$$

3.5. Context and multi-modality

The semantic representation above has three properties of particular relevance for this work. First, it is a representation that encodes contextual dependencies between different concepts. For example, because most images of the "outdoors" class include "vegetation," the presence of the "vegetation" concept is a clue for image assignment to the "outdoors" class. The semantic representation encodes this contextual relationship by assigning image $\mathcal{I}_i$ to the two concepts with some probability. This enables image retrieval and classification systems to take contextual cues into account [18, 26].

Second, unlike $\mathcal{X}$, the semantic space $\mathcal{S}$ offers a unified representation for information from multiple modalities. For example, as illustrated in the right side of Figure 1, replacing the SIFT descriptors of $\mathcal{X}$ with descriptors extracted from text documents produces a semantic representation for text. This enables a broader representation of context than that possible from images alone: by augmenting the training set with text, it is possible to learn contextual dependencies from the latter. One immediate benefit is that, because text classification is less ambiguous than image classification, the probabilities of (1) tend to be much more accurate for the former. This is illustrated in Figure 1 and motivates cross-modal regularization, where a regularizer learned from a corpus of images and text is used to denoise the semantic representation of subsequent images.

Third, by projecting images and text in the same space, the semantic representation simplifies the regularization operation itself. Since semantic translation is an automatic side-effect of the semantic representation there is no need to learn a translator between the two modalities. This reduces the cross-modal regularization problem to one of domain adaptation between two homogeneous domains. In this way, domain adaptation is decoupled from semantic translation, and considerably simpler than in the low-level space $\mathcal{X}$, where a translator must always be learned [43, 76].

4. Cross-modal Regularization

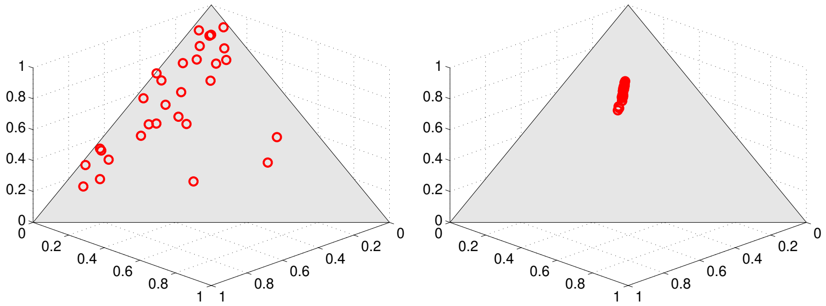
In this section we introduce the proposed cross-modal regularizer. In all equations d-dimensional vectors are represented as *column* ($d \times 1$) vectors and lowercase font, and matrices in uppercase.

4.1. Cross-modal regularization on the probability simplex

We consider the regularization problem where an auxiliary information source $\mathcal{A}$ is used to regularize the space where a *target data* source $\mathcal{T}$ is to be represented. It is assumed that a training sample $\{(a_1, t_1), \ldots, (a_N, t_N)\}$ of pairs of auxiliary and target examples is available. The regularizer is learned in two steps. First, both the auxiliary a_i and target t_i examples are mapped into a semantic space $\mathcal{S}$ associated with a vocabulary $\mathcal{L}$. This produces a sample of SMN pairs $(\pi_1^a, \pi_1^t), \ldots, (\pi_N^a, \pi_N^t)$, where π_i^a and π_i^t are L-dimensional probability vectors, i.e. vectors of non-negative components, $\pi_{i,k} \geq 0$, that add to one, $\sum_{k=1}^{L} \pi_{i,k} = 1$. It is assumed that the probabilities π_i^t associated with the target data are noisier than the probabilities π_i^a associated with the auxiliary source. This is usually the case when $\mathcal{T}$ is an image source and $\mathcal{A}$ a text source. The second step learns the transformation

$$\begin{array}{cccc} \Phi: & \mathcal{S} & \rightarrow & \mathcal{S} \\ & \pi^t & \rightarrow & \pi^a \end{array}$$

that makes the noisy target observations as "similar as possible" to the cleaner observations from the auxiliary source. This is implemented as a convex combination of class-specific linear regularizers. We start by discussing the learning of the linear regularizers and then discuss their combination in Section 4.5.



(a) Image SMNs (b) After regularization

Figure 4: Image SMNs before (a) and after (b) class-specific regularization.

4.2. *Linear regularizers*

In this section, we assume that all examples $(\pi_1^a, \pi_1^t), \ldots, (\pi_N^a, \pi_N^t)$, are extracted from text-image pairs of a single semantic class. To simplify the notation, we refer to π_i^a as a_i and π_i^t as t_i. A class-specific regularizer is then implemented through a linear transformation, H, such that

$$A = TH, \tag{3}$$

where A and T are the $N \times L$ matrices containing one example from $\mathcal{A}$ and $\mathcal{T}$, respectively, per row

$$\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{pmatrix} = \begin{pmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{pmatrix} \begin{pmatrix} h_1 & h_2 & \cdots & h_L \end{pmatrix} \tag{4}$$

and h_i are the columns of H. It is assumed that $N > L$ and (3) has no analytical solution. We seek the best H in the least squares sense, under the constraint that the transformed vector lies in $\mathcal{S}$, i.e.

$$t_i^T h_k \geq 0, \quad \forall i = 1 \ldots N, \forall k = 1 \ldots L \tag{5}$$

and

$$t_i^T H\mathbf{1} = 1, \quad \forall i = 1 \ldots N, \tag{6}$$

where $\mathbf{1}$ is the vector of all ones. This least squares problem can be written in the canonical form

$$\begin{aligned} x^* \quad &= \quad \arg\min_x \parallel Mx - b \parallel_2^2 \\ \text{subject to:} \quad & \quad Mx \succeq \mathbf{0} \\ & \quad Sx = \mathbf{1}. \end{aligned} \tag{7}$$

For this, it suffices to introduce the $N \times L^2$ matrix

$$S = \begin{pmatrix} t_1^T & t_1^T & \cdots & t_1^T \\ t_2^T & t_2^T & \cdots & t_2^T \\ \vdots & \vdots & & \vdots \\ t_N^T & \cdots & t_N^T & t_N^T \end{pmatrix} \tag{8}$$

and rewrite the transformation of (3) as

$$b = Mx, \tag{9}$$

where b and x are vectors of dimension NL and L^2, respectively, and M is a sparse matrix of dimensions $NL \times L^2$, as follows

$$\underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} t_1^T & 0 & \cdots & 0 \\ 0 & t_1^T & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & t_1^T \\ t_2^T & 0 & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & 0 & t_N^T \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_L \end{pmatrix}}_{x}. \tag{10}$$

Since the constraints are affine the feasible set is convex, and the optimization problem of (7) is convex whenever $M^T M$ is positive definite.

4.3. Positive definiteness of $M^T M$

To show that $M^T M$ is positive definite ($M^T M \succ 0$) it suffices to check that all its eigenvalues are positive. Since $M^T M$ is a block diagonal matrix of dimension $L^2 \times L^2$ with the structure

$$M^T M = \begin{pmatrix} B & 0 & \cdots & 0 \\ 0 & B & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & B \end{pmatrix}, \tag{11}$$

its eigenvalues are those of B, with multiplicity L. Furthermore, because the $L \times L$ matrix B is a sum of outer products of probability vectors

$$B = \sum_{i=1}^{N} (t_i t_i^T), \tag{12}$$

it has full-rank if there are at least L linearly independent t_i in this summation. In this case, $B \succ 0$, $M^T M \succ 0$, and the solution of (7) is a global minimum. Making

$N \gg L$ yields $rank(B) = L$ almost surely. In practice, the stochastic nature of t_i makes it sufficient to have $N = L$[1].

4.4. Learning

The optimization of (7) is a quadratic programming problem and can be solved by many standard optimization procedures. In our implementation, we use an active-set strategy (also known as a projection method) similar to that of [82, 83]. In all experiments, the matrix $M^T M$ was found to be positive definite, making the solution a global minimum. From (10), the regularization matrix H can be assembled by sequential extraction of the columns h_i from x^*. The procedure is summarized in Algorithm 1.

Algorithm 1 compute regularization operators (7)

input: train set of images and auxiliary data $\forall$ classes $i = 1, 2, \dots, L$
  $\mathcal{T}_i = \{\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_N\}$
  $\mathcal{A}_i = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N\}$
1 compute vectors of posterior probabilities
  $t_k \leftarrow \Psi(\mathcal{I}_k)$
  $a_k \leftarrow \Theta(\mathcal{X}_k)$
2 for each concept: $i = 1, \dots, L$
    solve: $x^* = \arg\min_x \| Mx - b \|_2^2$
      s.t. $Mx \succeq \mathbf{0}$
        $Sx = \mathbf{1}$
  where M, b are defined in (10) and S in (8).
output: set of regularization operators: $\mathcal{H} = \{H_1, H_2, \dots, H_L\}$

A conceptual illustration of the regularization is given in Figure 4. The figure shows the outcome of the regularization on a small sample of images from the "Warfare" class of the Wikipedia dataset, using a semantic space of three concepts ("Warfare", "History", and "Royalty"). The images are represented by their SMNs, shown in Figure 4a, which, due to the ambiguity of image classification, are scattered throughout the probability simplex. The auxiliary source is text. Figure 4b shows the result of the regularization of the image SMNs, t, with the transformation

$$\Phi(t) = H^T t. \tag{13}$$

The regularized SMNs cluster much more tightly in the neighborhood of the vertex of the simplex associated with the "Warfare" concept. This is the least squares compromise between the SMN distribution expected from the text, and the noisy distribution observed from the images.

[1] the number of training images per class (N) equal to the number of semantic concepts (L)

4.5. Class-adaptive regularization

So far, we have assumed that the class of the images to regularize is known. While this is usually the case during learning, it doesn't usually hold at run time, where the goal is to regularize SMNs of images outside the training set. In this case, it is necessary to select which of the regularization operators in the set $\mathcal{H} = \{H_1, H_2, \ldots, H_L\}$ is more suitable for a particular image t. This is a classification problem. Assuming the existence of auxiliary data a for image t, two strategies are possible.

(i) classify the auxiliary information, a, and apply to the image t the regularization operator corresponding to the resulting class. Only one operator is applied.

(ii) apply a convex combination of all regularization operators, where the combination coefficients are obtained from a regression or classification procedure over the auxiliary information a. Several regularization operators are combined.

The two procedures are summarized by Algorithm 2-(i) and Algorithm 2-(ii), respectively. When the auxiliary data is text, Algorithm 2-(i) applies a text classifier to text a, in order to determine its class j^*. The regularization operator learned from image-text pairs of this class is then applied to image t.

Algorithm 2-(i) classification-based regularization

input: set of regularization operators $\mathcal{H}$, and image-text pair (t, a), where t is the image to regularize and a its auxiliary information.

1 $j^* = \arg\max_j P(j|a), \quad \forall j = \{1, 2, \ldots, L\}$

2 $\Phi(t) \leftarrow H_{j^*}^T t$

output: regularized image $\Phi(t)$

On the other hand, Algorithm 2-(ii) computes a measure of the relevance $f_j(a)$ of class j for text a, which is then used to weight the contribution of operator H_j to the regularization of t. This allows the combination of all operators, according to their relative importance. Step 2 ensures that the weight vector, w, is a convex combination (i.e. adds up to one).

Algorithm 2-(ii) interpolation-based regularization

input: set of regularization operators $\mathcal{H}$, and image-text pair (t, a), where t is the image to regularize and a its auxiliary information.

1 $w_j(t) \leftarrow f_j(a), \quad \forall j = \{1, 2, \ldots, L\}$

$f_j()$ is a regression function for class j

2 $w \leftarrow \sigma(w)$

3 $\Phi(t) \leftarrow \sum_i w_i(t) H_i^T t$

output: regularized image $\Phi(t)$

Note that, in both cases, the overall regularizer is non-linear. Algorithm 2-(i) implements a piecewise linear regularization and Algorithm 2-(ii) a convex combination of linear regularizers (based on a non-linear weighting function). For simplicity, we denote Algorithm 2-(i) as *classification-based* regularizer and Algorithm 2-(ii) as *interpolation-based*.

4.6. Regularizing in the absence of auxiliar modality

In the previous section, we have assumed that auxiliary information a can be used to guide the choice of regularization operator for image t. This may not always be possible, since not all images possess auxiliary information. When this is the case, a possibility is to simply use the image t in place of a in line 1 of both classification and interpolation procedures. Another possibility is to use a *surrogate auxiliary datapoint*. This consists of finding, within the set of image/text pairs used to learn the regularization operators, the image t_{j^*} most similar to the image t being regularized. The text a_{j^*} associated with t_{j^*} is then used as a *surrogate* text for the regularization of t, using either Algorithm 2-(i) or (ii). This can be seen as a pre-processing procedure for images that lack text.

4.7. Classification and regression functions

There are many possibilities for implementing the classification and regression functions of Algorithms 2-(i) and (ii). Different methods frequently have different performance on different types of data. To evaluate the robustness of the proposed regularization to the choice of these functions, we consider three popular methods.

Logistic regression (LR) computes the posterior probability of a particular class by fitting the semantic features to a logistic function. Parameters are chosen to minimize the loss function,

$$\min_w \frac{1}{2} w^T w + C \sum_i \log(1 + \exp(-y_i w^T x_i)) \tag{14}$$

where y_i is the class label, x_i the input feature vector, and w a parameter vector. A multi-class LR returns a vector of posterior probabilities that can be used as weights in the interpolation scenario. For classification, we select the class of largest posterior probability. Our implementation of LR is based on the Liblinear package of [80].

Support vector machines (SVM) learn the separating hyperplane of largest margin between two classes, using

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_i \xi_i \tag{15}$$

$$\begin{aligned} s.t. \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \ \ \forall i \\ & \xi_i \geq 0 \end{aligned}$$

where w and b are the hyperplane parameters, y_i the class label, x_i input feature vectors, ξ_i slack variables that allow outliers, and $C > 0$ a penalty on the number of outliers. SVM classification can be used directly to select the regularization operator. For interpolation, the SVM scores $y_i w^T x_i$ can be converted into class probabilities through a calibration function. Our SVM implementation is based on the LibSVM [84] package.

Gaussian processes (GP) are a generalization of the Gaussian distribution. A GP defines a distribution over functions

$$f(x) \sim \mathcal{GP}(m(x), k(x, x^T)), \tag{16}$$

which is specified by a mean and covariance functions

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ k(x, x^T) &= \mathbb{E}[(f(x) - m(x))(f(x^T) - m(x^T))]. \end{aligned}$$

In this work, we adopt a squared-exponential covariance and affine mean, with a Gaussian likelihood function. This combination enables an exact inference procedure, which is implemented with the GPML [85] package.

5. Experiments

Several experiments were performed to evaluate the proposed *regularizer of image semantics*, denoted as "RIS". They are grouped in three sets. The first aimed to determine the best regularizer configuration, by comparing the performance of the classification and interpolation-based methods and different classification and regression functions. The second aimed to evaluate the robustness of the regularization to missing auxiliary information. Lastly, the third compared the proposed regularization procedure to a number of recently proposed domain adaptation methods.

5.1. Experimental set-up

All experiments are performed in the QBSE setting. In what follows, the terms retrieval set and database are used indistinguishably when referring to the repository of images being ranked. A query refers to the act of selecting one image from the database and using it to rank the remaining ones. Auxiliary information is only available for database images and always in the form of text modality. In some experiments, a percentage of the database images does not contain auxiliary information. Query images are never regularized.

Datasets: Three datasets are used in all experiments: "TVGraz" [86] contains $2,058$ image/text pairs of 10 semantic categories, "Wikipedia" [37] $2,866$ pairs from 10 categories, and "Pascal sentences" [87] $1,000$ pairs from 20 categories. These datasets have different characteristics. Pascal-sentences originates from a subset of Pascal VOC [88] images augmented with five sentences written by a human annotator [87]. The added text provides some context for each picture, but is not a semantically rich document. On both Wikipedia and TVGraz, the text is much more extensive and informative. On Wikipedia, classes are broad themes ("Media", "Music", "Biology", etc.), and intra-class image variability is quite large. On this dataset, image classification tends to have low accuracy. In fact, in the absence of additional information, many of the images are difficult to classify even for a human subject. On the other hand, text classes are fairly unambiguous. TVGraz contains narrow ("Caltech-like") object classes. The text, although less stylistic than that of Wikipedia, is informative of the class. This leads to fairly high classification accuracies for both images and text. Datasets were split into a training and test sets, in the range of 70-80% for the former and 30-20% for the latter, as detailed in Table 1. In each case, the training set is used to learn all semantic classifiers and regularization operators (both classification or interpolation functions and linear regularizers). The test set is then used in the retrieval experiments. These are implemented in a leave-one-out setting; repeating the retrieval operation with each image as a query and averaging results over all queries.

Table 1: Data split among training and test sets.

Dataset	train set	test set
TVGraz	1558	500
Wikipedia	2173	693
Pascal-senteces	700	300

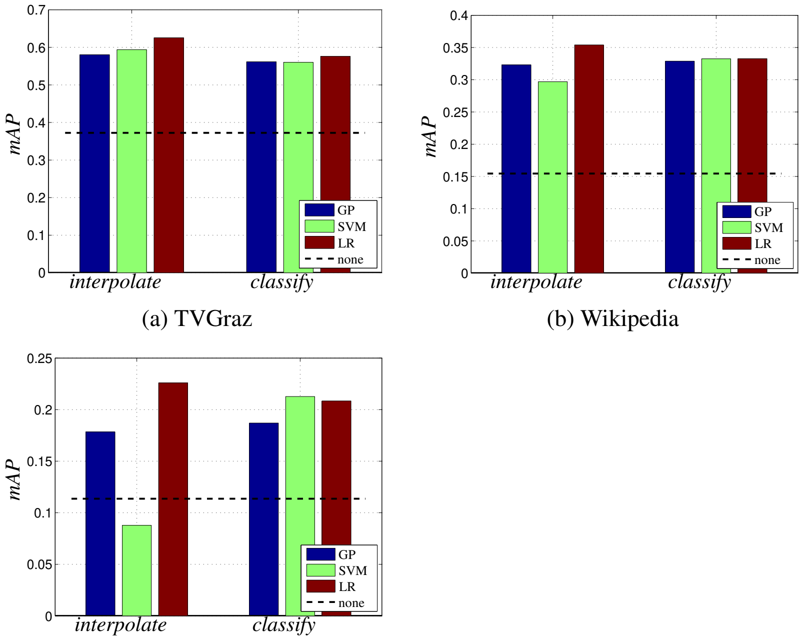
Representation: all images are represented as a *bag-of-words* (BOW) [89], using SIFT descriptors quantized with a $1,024$ visual word codebook. Text representation is based on *latent Dirichlet allocation* [90]. An LDA model is learned from all texts, and used to compute the probability of each text under 100 hidden topics. This probability vector is used for text representation. Both this and the image representation are mapped into a semantic space whose features are the classes that compose the dataset. This is implemented by designing a classifier Ψ of visual word histograms and a classifier Θ of hidden topic probabilities. In both cases, the classifier is a multi-class logistic regressor [80] and the semantic descriptor the vector of posterior probabilities of equation (1).

Evaluation metrics: retrieval performance is assessed with precision-recall curves. To facilitate comparisons of different methods, these are sometimes summarized by the mean average precision (mAP) or the R-Precision. The latter requires a set of known relevant documents (r) and the computation of the precision at recall r.

5.2. Regularization methods

A first set of experiments is designed to evaluate the effectiveness of various regularizer configurations. This includes classification vs. interpolation based regularization (Algorithm 2-(i) vs. 2-(ii)) and the choice of classification or interpolation function (GP, SVM or LR). In these experiments all database images have auxiliary text. Figure 5 compares the mAP of all regularization methods. In each graph, the dashed line labeled “none” represents QBSE without regularization. Since it makes use of no auxiliary information, this lower-bound can be seen as measure of the visual complexity of each dataset. It confirms that both Wikipedia and Pascal are significantly more challenging than TVGraz.

The figure shows that the benefits of regularization are substantial for all datasets. In some cases, the regularized mAP is more than double of that achieved without regularization. With the exception of SVM-based interpolation, all methods achieve significant gains in all datasets. In general, the relative gains over *vanilla* QBSE are largest



(a) TVGraz

(b) Wikipedia

(c) Pascal-sentences

Figure 5: Retrieval performance (mAP) of the various regularizer configurations on the three datasets. Each graph shows two groups of bars, referring to interpolation and classification methods implemented with GP, SVM, and LR. The dashed line denoted "none" indicates the mAP of QBSE without regularization.

for the more difficult datasets. Concerning the relative performances of the different regularizers, the two regularization strategies have similar performance, with a slight advantage for interpolation in TVGraz and Wikipedia and a slight advantage for classification in Pascal. With regards to the choice of regularization functions, SVMs tended to be weaker than GPs and LR for interpolation, but performed well under the classification strategy. Overall, the best performance was achieved by the LR implementation of interpolation-based regularization.

5.3. Coping with absent text

A second set of experiments is designed to evaluate the robustness of the regularization to missing auxiliary data. In these experiments only a percentage of the database images are complemented by text. These images are regularized with the interpolation-based regularizer as detailed in Algorithm 2-(ii), Section 4.5. For the regularization of the remaining images different weighting functions (w) are tested. Denoted:

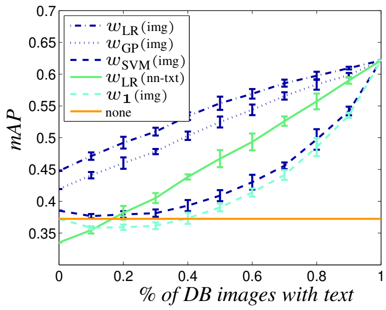
$$w_{\langle function \rangle}(\langle feature \rangle),$$

where the possible values for $\langle function \rangle$ and $\langle feature \rangle$ are listed in Table 2.

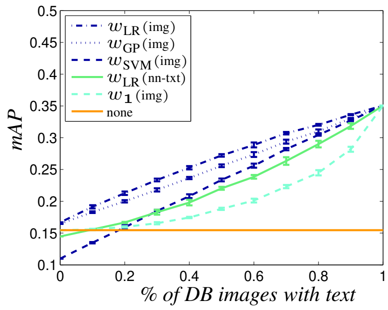
Table 2: Functions and features used to obtain the regularization weights for an image with no text.

⟨*function*⟩	⟨*feature*⟩
LR, GP, SVM or $\mathbf{1}$	image or *NN*-text

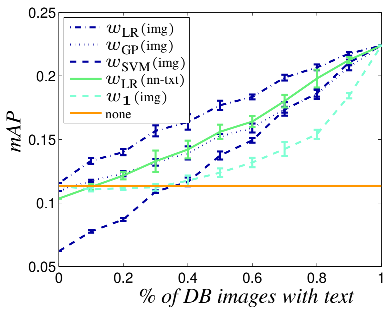
Each ⟨*function*⟩-⟨*feature*⟩ pair is an admissible combination to obtain regularization weights. Logistic regression (LR), support vector machines (SVM) and Gaussian processes (GP) are interpolation functions detailed in Section 4.7. Another possible function is the *identity* (denoted $\mathbf{1}$) that maps the ⟨*feature*⟩-vector directly to act as the weights. For image features all functions are tested: w_{LR}, w_{GP}, w_{SVM} and $w_{\mathbf{1}}$. Since the image has no text of its own, alternatively we can look for a *surrogate* text. Since the superiority of LR-based interpolation has already been established for text features in the previous section (Figure 5), when using these features (*NN*-text) we test only: w_{LR}. These experiments are repeated for various percentages of images with text. Each experiment is repeated five times, each using a different random set of such images.



(a) TVGraz



(b) Wikipedia



(c) Pascal-sentences

Figure 6: mAP of the different regularizers vs. the percentage of database images complemented by text, on the three datasets. Line labelled "none" corresponds to standard QBSE. Four functions are tested when using the image's own features: logistic regression, Gaussian process, SVM and the identity function, respectively "w_{LR}", "w_{GP}", "w_{SVM}" and "$w_{\mathbf{1}}$". When using a surrogate text for weight computation only LR-based regularization is tested and denoted "w_{LR}".

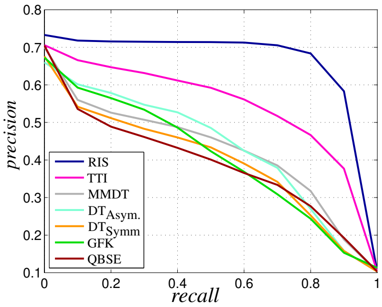
Figure 6 presents plots of mAP vs the % of images complemented by text. As before, we present the lower-bound of QBSE without regularization (labeled "none"). A second lower bound was computed by regularizing only the images that are complemented by text while applying the identity weights to the remaining images (labelled "$w_{\mathbf{1}}$(img)"). While superior to vanilla QBSE, this approach is not very robust. Its mAP degrades quickly as the percentage of text decreases. Better results are achieved by using a surrogate text to weigh the regularization operators applied to images without text. For clarity we only present the implementation of LR-based regularization for surrogate text (labelled "$w_{\mathbf{LR}}$(nn-txt)"). As mentioned, this method achieved superior performance when compared to GP and SVM. However, the surrogate text features underperform the image-driven selection of regularization operators. The remaining curves in each plot correspond to the implementation of this strategy with LR, GP, and SVM (labelled "$w_{\mathbf{LR}}$(img)", "$w_{\mathbf{GP}}$(img)" and "$w_{\mathbf{SVM}}$(img)" respectively). Among these, LR achieves the best results on all datasets.

Overall, the experiments of this and the previous section provide strong evidence for the benefits of regularization. Best results are obtained with an interpolation-based regularizer, using class-probabilities inferred with LR to weigh the class-specific regularization operators. This strategy proved quite robust to the absence of auxiliary text in the retrieval set. For images without text, good results are obtained by simply using the class probabilities derived from the image itself to weigh regularization operators. For example, on the harder Wikipedia and Pascal datasets, the mAP achieved with regularization was double that of baseline QBSE when only 60% of the images contained text. On the easier TVGraz dataset, where image-based estimates of class probability are more reliable, it improved on QBSE even when *no* images have auxiliary text. Interestingly, in all datasets, this regularization strategy also led to a nearly-linear increase in mAP with the percentage of database images complemented by text. For all these reasons, we only considered the LR implementation of interpolation-based regularization in the remaining experiments.

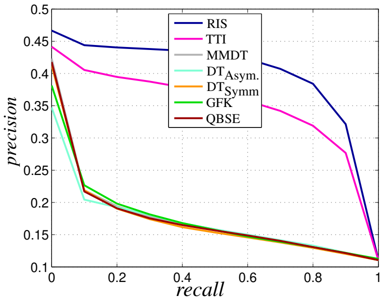
5.4. Comparison to alternative regularization methods

When compared to the previous literature, the proposed regularization in semantic space has the advantage of 1) not requiring a translation function, and 2) enabling the combination of class-specific regularizers. In this section, we report on experiments designed to evaluate the benefits of these properties. Since some of the competing methods assume image-text pairs for all examples, we only considered the scenario where all database images are complemented by text. For some methods (DT, GFK), the code provided by the authors produces matrices of similarity or distances between pairs of images. In these cases, retrieval was based on these distances. For methods that produced regularized image SMNs we used the set-up of the previous sections, i.e. QBSE with the KL divergence as similarity function. In all experiments, the proposed regularizer was implemented with the interpolation-based regularizer, using text features and logistic regression in the weighting function.

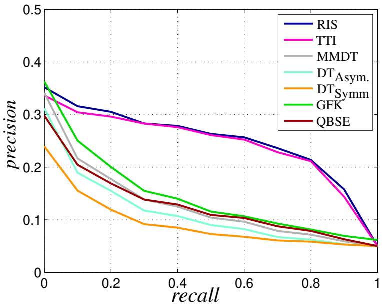
Previous approaches to cross-modal adaptation, e.g. [43, 76], represent images and text in low-level feature spaces and attempt to learn a translator function that maps text into the image domain. This is done by measuring co-occurrences of visual and text words on image-text pairs. To compare the proposed regularization approach with these



(a) TVGraz



(b) Wikipedia



(c) Pascal-sentences

Figure 7: Precision-recall curves of different regularizers on the three datasets. The proposed regularizer of image semantics is denoted as "RIS".

methods, we implemented an extension of the text-to-image translator (TTI) method of [43]. The implementation was based on code provided by the authors, which learns a translator function that assigns a confidence value to image/text pairs. This is a measure of how relevant the text is for the image. Preliminary experiments showed that best results were obtained by learning one translator per semantic class. In all experiments, each image/text pair in the retrieval set is represented by concatenation of the scores computed for all classes. Since queries have no text, query images were paired with the average text computed from the training set.

Previous approaches to both cross-modal and image-specific domain adaptation have proposed global transformations between the auxiliary and target domains. For example, the (DT) method of [39] learns the linear transformation, W, that minimizes the regularization cost $tr(W) - \log\det(W)$ subject to constraints that enforce (positive) similarity for a random sample of same-class object pairs. The choice of regularizer and constraints had been previously proposed in [91], where it is denoted information theoretic metric learning (ITML). Since the learned transformation is always symmetric positive definite, the method is denoted DT_{Symm}. A variant of this method, proposed in [40], uses a different objective function that does not enforce positive definitness. This is referred to as DT_{Asymm}. Max-Margin Domain Transforms

Table 3: Comparison of the mAP and R-precision scores of the proposed regularizer with those of previous approaches. Relative gains with respect to the latter are shown in (%). All methods were implemented with code provided by the authors.

Method	TVGraz		Wikipedia		Pascal		TVGraz	Wikipedia	Pascal
	mAP	%	mAP	%	mAP	%	R-Precision		
RIS	**0.622**	-	**0.356**	-	**0.224**	-	**0.554**	**0.272**	**0.182**
TTI [43]	0.531	*17*	0.323	*10*	0.220	*2*	0.476	0.259	0.168
MMDT [42]	0.405	*53*	0.155	*129*	0.115	*95*	0.400	0.158	0.114
GFK [41]	0.384	*62*	0.155	*129*	0.131	*71*	0.372	0.159	0.135
DT Symm. [39]	0.375	*65*	0.153	*133*	0.101	*122*	0.377	0.157	0.102
DT Asymm. [40]	0.425	*46*	0.152	*134*	0.118	*90*	0.396	0.148	0.120
QBSE [18]	0.372	*67*	0.155	*129*	0.114	*97*	0.368	0.156	0.107
Random	0.1	*522*	0.1	*256*	0.05	*348*	0.1	0.1	0.05

(MMDT) was later proposed in [42]. This is a combination of domain and model adaptation that optimizes an objective function of a discriminant classifier rather than the similarity measure used in [72]. Finally, we also consider the Geodesic Flow Kernel (GFK) method of [41] (GFK). This method models domain shift by integrating an infinite number of subspaces that establish a path between the auxiliary and target domains. It determines the optimal dimensionality of the subspaces in which to embed the two domains and constructs the geodesic curve connecting them through the Grassmann manifold. The geodesic distance is used to define a kernel that measures similarity between auxiliary and target data. For more details on these methods the reader is referred to the original publications. Preliminary experiments showed that they achieve best performance when applied in semantic space, i.e. using text SMNs as auxiliary and image SMNs as target data rather than their low-level representations. This is the configuration used in all experiments discussed below. Other than this, all methods were implemented with the code provided by the authors.

Figure 7 presents the 11-point interpolated precision-recall curves for all methods. Table 3 summarizes these results by presenting the mAP and R-precision computed over all queries[2]. These results support several conclusions. First, the class-specific transformation used by both the proposed regularizer and our extension of TTI achieves better regularization than the holistic transformation of the space used by the other methods. This seems to be particularly important on the datasets (Wikipedia and Pascal) where image classification is most ambiguous. Second, the simpler learning problem inherent to the representation in semantic space (no need to learn a translator function) enables further improvements. This is visible both by 1) the better performance of the proposed regularizer than TTI, and 2) the better performance of the global transform methods in the semantic space (observed in our preliminary experiments). Third,

[2]We note that some of the results reported in the table for TTI and QBSE are weaker than those reported in the earlier version of this work [44]. This is due to the fact that the similarity functions used for the image retrieval operation are different. The centered normalized correlation was used in [44], while we use the Kullback-Leibler divergence of (2) in this work. These functions yield slight variations in the mAP for certain dataset/method combinations. However, the differences are small and do not affect the conclusions of this work.

all methods outperformed QBSE in at least some datasets, with significant gains for the proposed regularizer.

Overall, these results confirm that the regularization of image semantics is beneficial (improvements over QBSE), and show that both the semantic representation (no need for translation) and the class-adaptive nature of the proposed regularizer are beneficial for image retrieval. Finally, these gains tend to be most significant when the ambiguity of image classification is largest, as in the Wikipedia and Pascal datasets. Figure 8 illustrates the robustness of the retrieval operation after semantic regularization, by presenting the top four matches for various query images from the three datasets. Each query is shown in a different row, displaying the query image on the left and the top matches on the right.

6. Conclusions

In this work, we have proposed a cross-modal domain adaptation method that exploits training text to learn a regularizer of image semantics. The resulting regularization was shown beneficial for image retrieval, where it led to significant performance improvements on various challenging datasets. While the largest gains (up to double mAP) were obtained for retrieval problems where all database images are complemented by text, the method was also shown successful when this is not the case. In fact, for some datasets, it enabled gains even when no text was available to the retrieval operation.

This robustness was justified by two properties of the proposed regularizer. The first is the semantic nature of the underlying image and text representation. This enables the modeling of contextual relationships between semantic concepts and establishes a unified space for image and text data. In result, the cross-modal regularization problem is reduced to one of adaptation between two homogeneous domains, i.e. there is no need to learn a translator between images and text. It was shown that, when compared to previous proposals to cross-modal regularization, this significantly simplifies the learning problem, enabling better generalization. The second is the implementation of the regularizer as a combination of class-specific regularizers. This leads to a piecewise-linear transformation of the image descriptors to regularize, which is highly non-linear but can be learned efficiently. When compared to previous approaches to domain adaptation in computer vision, the resulting regularizer is both more flexible and naturally aligned to the semantics of images and text. This was shown to enable significant gains in regularization performance.

Appendix A. Multivariate Bernoulli representation

In this appendix, we discuss the extension of the regularization procedure of Section 4.2 to the multivariate Bernoulli semantic representation. The only modification is to replace the constraint that regularized semantic descriptors must add up to one ($Sx = \mathbf{1}$) by a constraint that each concept probability must be less or equal to one

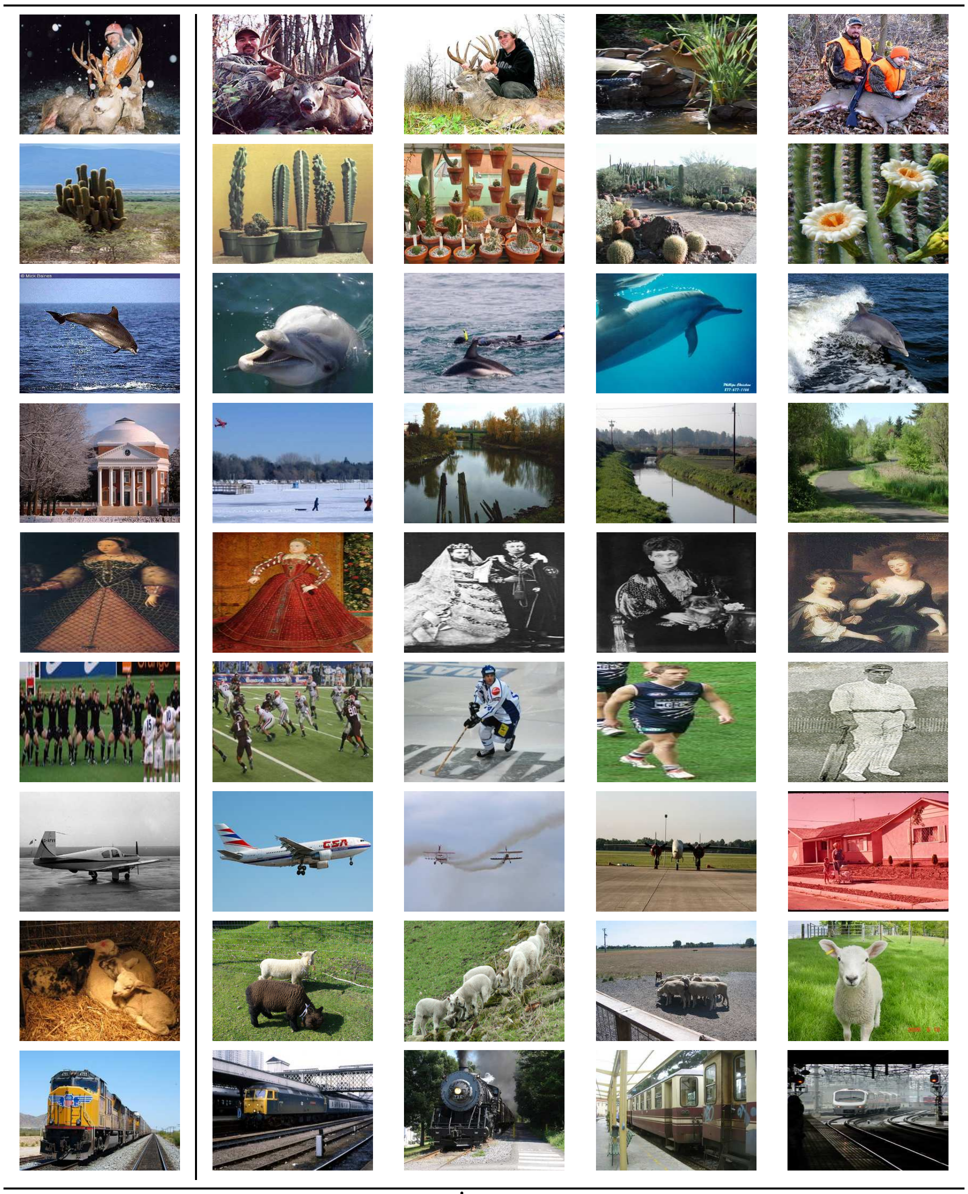


Figure 8: Retrieval examples (three queries of TVGraz (top), Wikipedia(middle) and Pascal-sentences (bottom)). In all cases the query image is shown on the leftmost column and top four database matches on the right.

($Mx \preceq \mathbf{1}$). The optimization problem of (7) is transformed into

$$\begin{aligned} x^* &= \arg\min_x \parallel Mx - b \parallel_2^2 \\ \text{subject to:} \quad & Mx \succeq \mathbf{0} \\ & Mx \preceq \mathbf{1} \end{aligned}$$

where M, S, and b are defined as before. The problem remains convex, and can be solved with the numeric procedures used in Section 4.2.

Acknowledgements

This work was funded by FCT graduate Fellowship SFRH/BD/40963/2007 from the Portuguese Ministry of Sciences and Education, and NSF grant CCF-0830535.

References

[1] M. Swain, D. Ballard, Color indexing, International Journal on Computer Vision, Springer, Springer 7 (1) (1991) 11–32. 2

[2] M. A. Stricker, M. Orengo, Similarity of color images, in: IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology, International Society for Optics and Photonics, 1995, pp. 381–392. 2

[3] M. A. Stricker, A. Dimai, Color indexing with weak spatial constraints, in: Electronic Imaging: Science & Technology, International Society for Optics and Photonics, 1996, pp. 29–40. 2

[4] B. S. Manjunath, W.-Y. Ma, Texture features for browsing and retrieval of image data, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 837–842. 2

[5] N. Vasconcelos, A. Lippman, A probabilistic architecture for content-based image retrieval, in: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, Vol. 1, IEEE, 2000, pp. 216–221. 2

[6] J. Mao, A. K. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, Pattern recognition 25 (2) (1992) 173–188. 2

[7] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522. 2

[8] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal on Computer Vision, Springer, Springer 60 (2) (2004) 91–110. 2

[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, Vol. 1, 2005, pp. 886–893. 2

[10] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Journal of Computer Vision and Image Understanding, Elsevier 110 (3) (2008) 346–359. 2

[11] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, Vol. 2, 2006, pp. 2169–2178. 2

[12] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2007, pp. 1–8. 2

[13] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380. 2

[14] N. Vasconcelos, A. Lippman, A bayesian video modeling framework for shot segmentation and content characterization, in: Content-Based Access of Image and Video Libraries, 1997. Proceedings. IEEE Workshop on, IEEE, 1997, pp. 59–66. 2, 4

[15] N. Vasconcelos, A. Lippman, Towards semantically meaningful feature spaces for the characterization of video content, in: Proc. IEEE International Conference on Image Processing, Vol. 1, 1997, pp. 25–28. 2, 4

[16] N. Vasconcelos, A. Lippman, Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing, in: Proc. IEEE International Conference on Image Processing, 1998, pp. 153–157. 2, 4

[17] J. Smith, M. Naphade, A. Natsev, Multimedia semantic indexing using model vectors, in: Proc. IEEE International Conference on Multimedia and Expo, Vol. II, 2003, pp. 445–448. 2, 4

[18] N. Rasiwasia, P. Moreno, N. Vasconcelos, Bridging the gap: Query by semantic example, IEEE Transactions on Multimedia 9 (5) (2007) 923–938. 2, 4, 5, 6, 8, 21

[19] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2009, pp. 1778–1785. 2, 4, 6

[20] A. Farhadi, I. Endres, D. Hoiem, Attribute-centric recognition for cross-category generalization, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2010, pp. 2352–2359. 2, 4, 6

[21] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recog. using classemes, European Conference on Computer Vision (2010) 776–789. 2, 4

[22] L. Li, H. Su, E. Xing, L. Fei-Fei, Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification, Advances in Neural Information Processing Systems. 2, 4

[23] N. Vasconcelos, From pixels to semantic spaces: Advances in content-based image retrieval, IEEE Transactions on Computers 40 (7) (2007) 20–26. 2

[24] N. Rasiwasia, N. Vasconcelos, Scene classification with low-dimensional semantic spaces and weak supervision, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2008, pp. 1–6. 3

[25] R. Kwitt, N. Vasconcelos, N. Rasiwasia, Scene recognition on the semantic manifold, in: Computer Vision–ECCV 2012, Springer Berlin Heidelberg, 2012, pp. 359–372. 3

[26] N. Rasiwasia, N. Vasconcelos, Holistic context models for visual recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 34 (2012) 902–917. 3, 8

[27] L. Lin, T. Wu, J. Porway, Z. Xu, A stochastic graph grammar for compositional object representation and recognition, Pattern Recognition 42 (7) (2009) 1297–1307. 3

[28] B. Siddiquie, R. S. Feris, L. S. Davis, Image ranking and retrieval based on multi-attribute queries, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2011, pp. 801–808. 3, 4

[29] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, S.-C. Zhu, I2t: Image parsing to text description, Proceedings of the IEEE 98 (8) (2010) 1485–1508. 3

[30] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, S.-F. Chang, Weak attributes for large-scale image retrieval, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2012, pp. 2949–2956. 3, 4

[31] D. Hoiem, A. Efros, M. Hebert, Putting objects in perspective, International Journal on Computer Vision, Springer 80 (1) (2008) 3–15. 3

[32] A. Torralba, K. Murphy, W. Freeman, M. Rubin, Context-based vision system for place and object recognition, in: Proc. IEEE International Conference on Computer Vision, 2008, pp. 273–280. 3

[33] P. Carbonetto, N. Freitas, K. Barnard, A statistical model for general contextual object recog., European Conference on Computer Vision (2004) 350–362. 3

[34] M. Vasconcelos, G. Carneiro, N. Vasconcelos, Weakly supervised top-down image segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006. 3

[35] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: Proc. IEEE International Conference on Computer Vision, 2007, pp. 1–8. 3

[36] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, European Conference on Computer Vision (2006) 1–15. 3

[37] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proc. ACM International Conference on Multimedia, 2010, pp. 251–260. 3, 15

[38] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, Ranking with local regression and global alignment for cross media retrieval, in: Proc. ACM International Conference on Multimedia, 2009, pp. 175–184. 3

[39] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: European Conference on Computer Vision, Springer, 2010, pp. 213–226. 4, 20, 21

[40] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Visual domain adaptation using regularized cross-domain transforms, Tech. rep., UCB/EECS-2010-106, EECS Department, University of California Berkeley (2010). 4, 20, 21

[41] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2012, pp. 2066–2073. 4, 5, 21

[42] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, T. Darrell, Efficient learning of domain-invariant image representations, in: International Conference on Learning Representations, 2013. 4, 5, 21

[43] G. Qi, C. Aggarwal, T. Huang, Towards semantic knowledge propagation from text corpus to web images, in: Proc. ACM International Conference on World Wide Web, 2011, pp. 297–306. 4, 6, 9, 19, 20, 21

[44] J. Costa Pereira, N. Vasconcelos, On the regularization of image semantics by modal expansion, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2012, pp. 3093–3099. 4, 21

[45] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, et al., Query by image and video content: the QBIC system, IEEE Transactions on Computers 28 (9) (1995) 23–32. 4

[46] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, C.-F. Shu, Virage image search engine: an open framework for image management, in: Electronic Imaging: Science & Technology, 1996, pp. 76–87. 4

[47] J. Smith, S. Chang, VisualSEEk: a fully automated content-based image query system, in: Proc. ACM International Conference on Multimedia, 1997, pp. 87–98. 4

[48] C. Frankel, M. Swain, V. Athitsos, Webseer: An image search engine for the world wide web, Tech. rep., University of Chicago, Computer Science Department (1996). 4

[49] M. S. Lew, Next-generation web searches for visual content, IEEE Transactions on Computers 33 (11) (2000) 46–53. 4

[50] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2009, pp. 951–958. 4

[51] X. Yu, Y. Aloimonos, Attribute-based transfer learning for object categorization with zero/one training example, in: European Conference on Computer Vision, Springer, 2010, pp. 127–140. 4

[52] T. L. Berg, A. C. Berg, J. Shih, Automatic attribute discovery and characterization from noisy web data, in: European Conference on Computer Vision, Springer, 2010, pp. 663–676. 4

[53] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2011, pp. 1641–1648. 4

[54] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks., in: Advances in Neural Information Processing Systems, Vol. 1, 2012, p. 4. 4

[55] J. Weston, S. Bengio, N. Usunier, Large scale image annotation: learning to rank with joint word-image embeddings, Machine learning 81 (1) (2010) 21–35. 4

[56] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., Devise: A deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems, 2013, pp. 2121–2129. 4

[57] N. Vasconcelos, Minimum probability of error image retrieval, IEEE Transactions on Signal Processing 52 (8) (2004) 2322–2336. 5

[58] X. Zhu, A. Goldberg, Introduction to semi-supervised learning, Morgan & Claypool, 2009. 5

[59] R. Caruana, Multitask learning: A knowledge-based source of inductive bias, Machine Learning 28 (1997) 41–75. 5

[60] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing 10 (1) (2000) 19–41. 5

[61] P. C. Woodland, Speaker adaptation for continuous density hmms: A review, in: ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition, 2001. 5

[62] M. Dixit, N. Rasiwasia, N. Vasconcelos, Adapted gaussian models for image classification, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2011, pp. 937–943. 5

[63] X. Zhou, N. Cui, Z. Li, F. Liang, T. S. Huang, Hierarchical gaussianization for image classification, in: Proc. IEEE International Conference on Computer Vision, IEEE, 2009, pp. 1971–1977. 5

[64] L. Fei-Fei, R. Fergus, P. Perona, A bayesian approach to unsupervised one-shot learning of object categories, in: Proc. IEEE International Conference on Computer Vision, 2003, pp. 1134–1141. 5

[65] R. Raina, A. Y. Ng, D. Koller, Constructing informative priors using transfer learning, in: Proc. ACM International Conference on Machine Learning, 2006, pp. 713–720. 5

[66] C. Do, A. Ng, Transfer learning for text classification, in: Advances in Neural Information Processing Systems, 2005. 5

[67] J. Yang, R. Yan, A. G. Hauptmann, Cross-domain video concept detection using adaptive svms, in: Proc. ACM International Conference on Multimedia, 2007, pp. 188–197. 5

[68] A. Bergamo, L. Torresani, Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach, in: Advances in Neural Information Processing Systems, 2010, pp. 181–189. 5

[69] W. Dai, Q. Yang, G.-R. Xue, Y. Yu, Boosting for transfer learning, in: Proc. ACM International Conference on Machine Learning, 2007, pp. 193–200. 5

[70] Y. Aytar, A. Zisserman, Tabula rasa: Model transfer for object category detection, in: Proc. IEEE International Conference on Computer Vision, 2011, pp. 2252–2259. 5

[71] H. Daumé III, Frustratingly easy domain adaptation, in: Association of Computational Linguistics, Vol. 1785, 2007, pp. 256–263. 5

[72] B. Kulis, K. Saenko, T. Darrell, What you saw is not what you get: Domain adaptation using asymmetric kernel transforms, in: Proc. IEEE International Conference on Computer Vision on Pattern Recognition, 2011, pp. 1785–1792. 5, 21

[73] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: An unsupervised approach, in: Proc. IEEE International Conference on Computer Vision, 2011, pp. 999–1006. 5

[74] L. Duan, D. Xu, I. Tsang, Learning with augmented features for heterogeneous domain adaptation, arXiv:1206.4660. 5

[75] F. Zhu, L. Shao, Enhancing action recognition by cross-domain dictionary learning, in: British Machine Vision Conference, 2013. 5

[76] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, Y. Yu, Translated learning: Transfer learning across different feature spaces, in: Advances in Neural Information Processing Systems, 2008, pp. 353–360. 6, 9, 19

[77] N. Vasconcelos, Image indexing with mixture hierarchies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001. 7

[78] N. Vasconcelos, Exploiting group structure to improve retrieval accuracy and speed in image databases, in: Proc. IEEE International Conference on Image Processing, Vol. 1, 2002, pp. I – 980–983. 7

[79] G. Carneiro, A. Chan, P. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410. 7

[80] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research, MIT Press 9 (2008) 1871–1874. 8, 14, 16

[81] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence to appear (3). 8

[82] P. Gill, W. Murray, M. Saunders, M. Wright, Procedures for optimization problems with a mixture of bounds and general linear constraints, ACM Transactions on Mathematical Software 10 (3) (1984) 282–298. 12

[83] P. Gill, W. Murray, M. Wright, Numerical Linear Algebra and Optimization, Vol. 1, Addison-Wesley, 1991. 12

[84] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 14

[85] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006. 15

[86] I. Khan, A. Saffari, H. Bischof, TVGraz: Multi-modal learning of object categories by combining textual and visual features, in: Workshop of the Austrian Association for Pattern Recognition, 2009. 15

[87] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, NAACL HLT, 2010, pp. 139–147. 15

[88] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 15

[89] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on Stat. Learn. in Comp. Vision, European Conference on Computer Vision, Vol. 1, 2004, pp. 1–22. 16

[90] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research, MIT Press 3 (2003) 993–1022. 16

[91] B. Kulis, P. Jain, K. Grauman, Fast similarity search for learned metrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12) (2009) 2143–2157. 20