On the Regularization of Image Semantics by Modal Expansion

Jose Costa Pereira Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
josecp@ucsd.edu, nuno@ece.ucsd.edu

Abstract

Recent research efforts in semantic representations and context modeling are based on the principle of task expansion: that vision problems such as object recognition, scene classification, or retrieval (RCR) cannot be solved in isolation. The extended principle of modality expansion (that RCR problems cannot be solved from visual information alone) is investigated in this work. A semantic image labeling system is augmented with text. Pairs of images and text are mapped to a semantic space, and the text features used to regularize their image counterparts. This is done with a new cross-modal regularizer, which learns the mapping of the image features that maximizes their average similarity to those derived from text. The proposed regularizer is class-sensitive, combining a set of class-specific denoising transformations and nearest neighbor interpolation of text-based class assignments. Regularization of a state-of-the-art approach to image retrieval is then shown to produce substantial gains in retrieval accuracy, outperforming recent image retrieval approaches.

1. Introduction

Object recognition, scene classification, or image retrieval are challenging problems for computer vision. In the last decades, they have been solved with recourse to statistical decision theory and machine learning. These solutions have two major components: an image representation, obtained by projecting images into some *feature space*, and a *classification architecture*, which maps that representation into a recognition, classification, or retrieval (RCR) function. Over the last decade there have been significant advances in both areas, *e.g*. the SIFT [15] or HoG [4] features and many new classification architectures [8, 13, 24, 30]. While advances continue to be made in these areas, there is a sense that performance will asymptote and these solutions are not sufficient to solve all RCR problems.

This has spurred two recent research trends, which expand the RCR problem along the directions of *semantic abstraction* and *contextual modeling*. These two extensions are, in fact, applications of the *same principle* to the design of a feature space and a classification architecture. The shared principle is that this design *cannot account only for the class of interest, but has to leverage the detection of many other classes*. We refer to it as the *task expansion* principle: an RCR task, *e.g*. dog recognition, cannot be solved without solving many other RCR tasks, *e.g*. the detection of 1) concepts that provide context for dogs, or 2) semantic attributes that make up a dog. Semantic abstraction applies the principle to the design of feature spaces, context modeling to the design of classifiers.

The benefits of task expansion are now well established. Many contextual representations have proven useful for object recognition [2, 11, 27]. These methods have shown that exploiting the presence, elsewhere in an image, of contextually related objects (*e.g*. "dog-house", "bone", "backyard", "ball", etc.) improves the detection of an object of interest (*e.g*. "dog"). This has motivated *context-based architectures* of ever increasing complexity [19, 25]. The intuition behind semantic abstraction is that a space of low-level features, such as SIFT, is too far removed from the RCR goal. After all, people do not describe dogs as bags of edges and textures, but as conceptual entities with certain properties, *e.g*. "has legs", "is hairy", "chews bones", "lives on the backyard", "chases cats", etc. RCR performance should thus improve by designing *semantic feature spaces*, where features are themselves image classification scores for many such properties. This entails defining a vocabulary of *semantic concepts*, building the associated detectors, and using the vector of classification scores for semantic image representation. Such representations are widely used in image retrieval [26, 29] and, more recently, in object detection [6, 7] and scene classification [14, 23, 28].

From a statistical point of view, task expansion is a form of *regularization*. A natural extension of task expansion is the *modality expansion principle*. This states that the design of RCR architectures *cannot account only for visual information*. It is also inspired by perception, where RCR problems are always solved in the context of strong *cognitive priors* (*e.g*. concept taxonomies) *not necessarily*

learned from vision. For example, much of our understanding of contextual relationships is acquired by reading books, speaking with others, touching objects, etc. As in task expansion, these priors are regularizers, which can be implemented indirectly, in a data-driven manner, by *using data from non-visual modalities to constrain the learning of visual models*. Modality expansion achieves this working on a semantic space, where features are not tied to visual representation. In general, it is not harder to learn a classifier from text or speech than from images. On the contrary, it is usually easier, because the semantics of text are more *explicit* than those of images.

In this work, we introduce a solution for the image retrieval problem based on modality expansion. As is usual for the design of semantic feature spaces, a vocabulary of semantic concepts is first defined and a set of training examples collected per concept. The only difference is that these examples are image-text pairs, instead of images alone. Since these sets are usually collected on the web, where most images have associated text, this is quite simple. Pairs of semantic classifiers are then learned for images and text, and training examples from the two modalities mapped to the semantic space. This usually leads to a noisy set of features for images and a much cleaner set of features for text. The latter are then used to regularize the former. This consists of learning the mapping of the image-based semantic features that maximizes their average similarity to the text-based semantic features. This regularizer is finally used to build an image retrieval system. Images in a retrieval database are projected onto the semantic space, the resulting semantic feature vectors regularized and used as image representation in a query-by-example retrieval system. Experimental results show that the proposed *regularized image semantics* (RIS) retrieval method substantially outperforms both a state-of-the-art semantic image retrieval system and a retrieval system that combines images and text.

2. Semantic representation

In this section, we briefly review the semantic space mappings used in our work, and explain how they can be exploited for the proposed text-based regularization. Throughout the text the terms "semantic class" and "semantic concept" are used interchangeably to refer to the category to which the image or text belongs to.

2.1. Semantic space

Let $\mathcal{G} = \{\mathcal{I}_1, \ldots, \mathcal{I}_G\}$ be a set of images, where each entry $\mathcal{I}_i$ is represented in a low-level feature space $\mathcal{X}$, *e.g.* an histogram of visual-words, sampled from a random variable $\mathbf{X}$. This set is augmented with a concept vocabulary $\mathcal{L} = \{z_1, z_2, ..., z_L\}$, sampled from a random variable Z. Each concept induces a probability density, $P_{\mathbf{X}|Z}(\mathbf{x}|z)$, on $\mathcal{X}$. An image $\mathcal{I}_i$ is labeled by computing its posterior probability under each of the concept classes,

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

Given a set of manually labeled training examples per concept, this can be done by: 1) learning the concept distributions $P_{\mathbf{X}|Z}(\mathbf{x}|z), \forall z$ and applying Bayes rule, or 2) learning a discriminant mapping. Image $\mathcal{I}_i$ is finally represented by the probability vector π_i of its assignment to all concepts. The simplex of all such probability vectors is denoted the *semantic space* $\mathcal{S}$. An example of the projection of images in such space is given in Figure 1-(a). The data is a subsample from three classes of the Wikipedia dataset [21], *viz.* "History", "Royalty" and "Warfare".

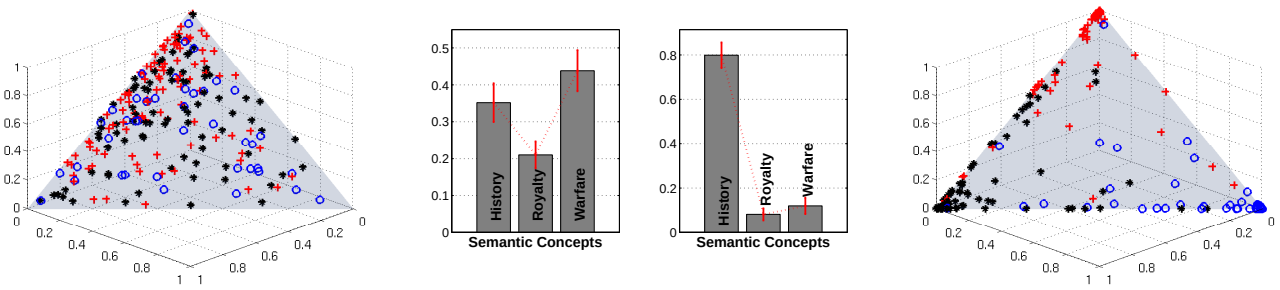
2.2. Regularization

The representation of images in terms of a collection of visual concepts has two main advantages. The first is that it is very robust to a number of confounding factors that frequently plague RCR problems, *e.g.* that sky can be blue on sunny days, grey on cloudy days, or orange during a sunset. This has been exploited to substantially improve the performance of image retrieval systems in the past [22, 26]. The second is that it maps images into an abstract space, where they can be easily combined with other sources of data. This follows from the fact that all the steps above could be equally applied to a dataset $\mathcal{G}$ from any modality other than images.

In this work, we exploit this fact, to design better image retrieval systems. It is assumed that the training set for the design of semantic labeling systems includes both text and images, *i.e.* $\mathcal{G} = (\mathcal{I}_1, \mathcal{T}_1), \ldots, (\mathcal{I}_G, \mathcal{T}_G\}$. The procedure of the previous section is then applied to the text documents, to learn a mapping from text documents to the semantic space. The goal is to leverage the fact that, due to the reduced ambiguity of text classification, the semantic space representation of text is usually much cleaner than that of images. This is illustrated in Figure 1. Note how the semantic feature vectors derived from text in Figure 1-(c) have much smaller variance than those derived from images in Figure 1-(a). Also note how the distributions of the different classes in $\mathcal{S}$ have much smaller overlap. This can also be observed in Figure 1-(b), which shows the average vectors in the simplex for all entries of class "History". The distribution derived from text assigns much higher probability to the "History" concept than that derived from images.

3. Regularization of the semantic space

In this section we introduce a method that relies on the semantic feature vectors derived from text to *regularize* those derived from images. Building on the terminology of [21], this is denoted *cross-modal regularization*.



(a) visual data (b) vectors of average probabilities (c) textual data

Figure 1. Semantic space created from three Wikipedia classes (*viz.* "History", "Royalty" and "Warfare"). Projections of images and text onto this space are shown in (a) and (c), respectively. The two average probability vectors (and respective error bars) for the "History" class are shown in (b) – images on the left and text on the right.

3.1. Cross-modal regularization on the probability simplex

Cross-modal regularization addresses the problem of using an *auxiliary* source of information $\mathcal{A}$ to regularize the space where the information from a *data* source $\mathcal{D}$ is to be represented. In this work we consider the case where both the auxiliary and the data sources are represented in a probability simplex. Let $\mathcal{A} = \{a_1, a_2, \dots a_N\}$ and $\mathcal{D} = \{d_1, d_2, \dots d_N\}$ be two samples from auxiliary and data source, respectively. Points in $\mathcal{A}$ and $\mathcal{D}$ are L-dimensional probability vectors x. These are vectors on the $(L-1)$-*simplex* $\mathcal{S}$, *i.e.* have non-negative components, $x^{(k)} \geq 0$, that add to one, $\sum_{k=1}^{L} x^{(k)} = 1$. It is assumed that there is a one-to-one correspondence between the points in $\mathcal{A}$ and $\mathcal{D}$, namely that each vector d_i in $\mathcal{D}$ is a noisy estimate of a corresponding vector a_i in $\mathcal{A}$. The goal is to find the transformation

$$\begin{aligned} H : \mathcal{S} &\rightarrow \mathcal{S} \\ d &\rightarrow a \end{aligned}$$

that makes the noisy data observations as "similar as possible" to the cleaner observations from the auxiliary source. In this section, we consider the case where H is a *linear transformation*:

$$A = DH \tag{2}$$

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

$$\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{pmatrix} = \begin{pmatrix} d_1^T \\ d_2^T \\ \vdots \\ d_N^T \end{pmatrix} \begin{pmatrix} h_1 & h_2 & \cdots & h_L \end{pmatrix} \tag{3}$$

and h_i are the columns of H. Since this has no solution, in general, we seek the best H in the least squares sense, under the constraint that the transformed vector has to lie in $\mathcal{S}$, *i.e.*

$$d_i^T h_k \geq 0, \quad \forall i = 1 \dots N, \forall k = 1 \dots L \tag{4}$$

and

$$d_i^T H \mathbf{1} = 1, \quad \forall i = 1 \dots N \tag{5}$$

The problem can be transformed to the canonical form

$$b = Mx, \tag{6}$$

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

$$\underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} d_1^T & 0 & \cdots & 0 \\ 0 & d_1^T & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & d_1^T \\ d_2^T & 0 & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & 0 & d_N^T \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_L \end{pmatrix}}_{x}. \tag{7}$$

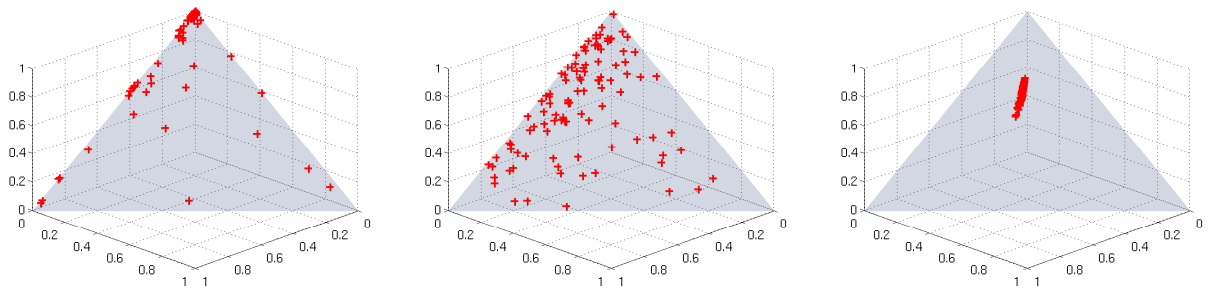
Further introducing the $N \times L^2$ matrix

$$S = \begin{pmatrix} d_1^T & d_1^T & \cdots & d_1^T \\ d_2^T & d_2^T & \cdots & d_2^T \\ \vdots & \vdots & & \vdots \\ d_N^T & \cdots & d_N^T & d_N^T \end{pmatrix} \tag{8}$$

the least squares solution of (2) under the constraints of (4) and (5) is given by the optimization

$$\begin{aligned} x^* &= \arg\min_x \| Mx - b \|_2^2 \\ \text{subject to:} \quad & Mx \succeq \mathbf{0} \\ & Sx = \mathbf{1} \end{aligned} \tag{9}$$

Since the constraints are affine, the feasible set is convex and the optimization problem is convex whenever $M^T M$ is positive definite. Note that $M^T M$ is known, directly obtained from the data $\mathcal{D}$. The learning procedure is summarized in Algorithm 1.



(a) (b) (c)

Figure 2. Cross-modal regularization in the probability simplex. The figure shows the probability vectors derived from the (a) auxiliary, and (b) data sources, and (c) the regularized data distribution for one class. The data was created from the "Wikipedia" dataset, using text as auxiliary and images as data sources.

Algorithm 1 compute regularization operators (9)

input: training set of images and auxiliary data

$\forall$ classes $i = 1, 2, \ldots, L$

$\mathcal{D}_i = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$

$\mathcal{A}_i = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$

1 compute vectors of posterior probabilities

$d_k \leftarrow \Psi(\mathcal{I}_k)$

$a_k \leftarrow \Theta(\mathcal{T}_k)$

2 for each concept: $i = 1, \ldots, L$

solve: $x^* = \arg\min_x \parallel Mx - b \parallel_2^2$

s.t. $Mx \succeq \mathbf{0}$

$Sx = \mathbf{1}$

where M, b are defined in (7) and S in (8).

output: set of regularization operators $\mathcal{H}$

$\mathcal{H} = \{H_1, H_2, \ldots, H_L\}$

In our implementation, the quadratic programming problem of (9) is solved by the method of [9, 10]. In all experiments, the matrix $M^T M$ was found to be positive definite, making the solution found by this procedure a global minimum. From (7), the regularization matrix H can be assembled by sequential extraction of the columns h_i from x^*. Given an example d from the data source, the regularization consists of the transformation

$$d' = H^T d \tag{10}$$

Figure 2 illustrates the regularization procedure for data from one of the three Wikipedia classes of Figure 1. The auxiliary source is the text, and the data source the image corpus. The probability vectors derived from text in Figure 2-(a) cluster tightly in the upper corner of the space, but those derived from images in Figure 2-(b) are much noisier. After regularization they cluster much more tightly, in the neighborhood of the upper corner of the simplex. This is the least squares compromise between the distribution expected from text, and the noisy distribution observed from the images.

3.2. Class-sensitive regularization

In general, a linear regularizer is not rich enough for problems involving real datasets. Better performance can usually be obtained with a non-linear regularizer $H(d)$. One possibility would be to kernelize the problem of (9). This is usually possible for quadratic problems with affine constraints. An alternative route, that we pursue in this work, is to make the regularization class-sensitive. This is frequently better for supervised learning problems, where training data is available per class. In these problems a non-linear regularizer can be learned by combining a set of linear operators with a non-linear weighting function.

Let A_i, D_i be the matrices of examples collected from the auxiliary and data sources for concept $Z = i$. A linear regularizer H_i is learned per concept, using the procedure of the previous section, and the non-linear regularizer defined as

$$\Phi(d) = \sum_i w_i(d) H_i^T d \tag{11}$$

The *weighting functions* $w_i(d)$ are non-negative and sum to one $w_i(d) \geq 0, \sum_i w_i(d) = 1, \forall d \in \mathcal{S}$, defining a soft partition of the simplex $\mathcal{S}$. Note that, since $\mathcal{S}$ is the range space of the class-specific regularizers H_i and (11) is a convex combination of their outputs, then $\Phi(d) \in \mathcal{S}$.

The weighting functions $w_i(d)$ are of the form

$$w_i(d) = f(d; (d_j, a_j) \in \mathcal{L}_i) \tag{12}$$

where $\mathcal{L}_i$ is the training set used to learn H_i, *i.e.* the semantic feature vectors from images and text of concept $Z = i$. The method we adopt, is to 1) learn class-weighting functions $w_i'(a)$ from the auxiliary source, using standard machine learning methods, and 2) *transfer* their scores to the data source, *i.e.* use

$$w_i(d) = f(d; d_j, w_i'(a_j), (d_j, a_j) \in \mathcal{L}_i). \tag{13}$$

When the auxiliary data is text, a_j is already a good estimate of the posterior distribution of assignment of example j to the semantic classes, and it suffices to use $w'_i(a_j) = a_{ji}$. To transfer these weighting functions, we rely on a simple *nearest-neighbor interpolation*

$$w_i(d) = w'_i(a_{j^*}) = a_{j^*i}, \quad (d_j, a_j) \in \mathcal{L}_i \quad (14)$$

$$j^* = \arg\max_j S(d, d_j) \quad (15)$$

where $S(\cdot,\cdot)$ is a similarity function between probability vectors. This is the probability of assignment to semantic class i of the auxiliary example j^* corresponding to the data source example d_{j^*} most similar to d. The image regularization procedure is summarized by Algorithm 2.

Algorithm 2 cross-modal regularization (11)

input: set of training images and auxiliary data $\mathcal{P} = \{(d_1, a_1), (d_2, a_2), \ldots, (d_N, a_N)\}$ and d image to regularize

1 find j^* according to (15) (d_k, a_k)

2 $w(d) \leftarrow a_{j^*}$

$\Phi(d) \leftarrow \sum_i w_i(d) H_i^T d$

output: regularized image $\Phi(d)$

3.3. Image retrieval

In image retrieval, the goal is to find, from an image database $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_N\}$, the image that most resembles a query image $\mathcal{Q}$. A popular solution to this problem is the *query by semantic example* (QBSE) method of [22], which poses retrieval as a nearest neighbor operation in the semantic space $\mathcal{S}$. This consists of returning the database image $\mathcal{F}_{j^*}$ such that

$$j^* = \arg\max_j S(q, f_j) \quad (16)$$

where q and f_j are the semantic feature vectors associated with $\mathcal{Q}$ and $\mathcal{F}_j$ respectively.

We introduce a new method, denoted *regularized image semantics* (RIS). This consists of applying (11) to all images. The regularization is done off-line for the database images in $\mathcal{F}$. However, for the query $\mathcal{Q}$, the search of (15) must be performed at retrieval time. This is a nearest neighbor operation over all semantic feature vectors used to learn the class regularizers H_i, *i.e.* all training examples from all semantic classes. Since this set can be large, the computational cost can be substantial. However, we have noted that, once the databases features f_j are regularized, the regularization of the query does not produce substantial additional gains. Hence, in our implementation, queries are not regularized and the regularization has *no computational cost* for the retrieval operation. This is quite remarkable since, as will be shown in the following section, the gains in retrieval accuracy can be substantial.

4. Evaluation

Several experiments were performed to evaluate the performance of RIS.

Representation: In all experiments, images were represented with the *bag-of-words* (BOW) model of [3], using SIFT descriptors quantized with a $1,024$ visual word codebook. The text representation was based on *latent Dirichlet allocation* (LDA) [1]. An LDA model is learned from a text corpus, and used to compute the probability of each text under 100 hidden topics. The probability vectors are used for text representation. For semantic classification, both visual word histograms and hidden topic probabilities were fed to a multi-class logistic regression [5].

Datasets: Three datasets were used. "TVGraz" [12] contains $2,058$ image/text pairs of 10 semantic categories, "Wikipedia" [21] $2,866$ pairs from 10 categories, and "Pascal sentences" [20] $1,000$ pairs from 20 categories. These datasets have different properties. Pascal is a challenging visual dataset. The added text features create a context for each picture, but they are not as semantically rich as a full text article. Both image and text classification is low. On Wikipedia, classes are broad ("History", "Art", "Literature", etc.), but contain both high quality images and text. Both relate to the category they belong to but, for images, the intra-class variability is quite large. On this dataset, image classification has low accuracy, but text classification is accurate. TVGraz classes report to narrow object classes ("Caltech-like"). The text, although often less stylistic than those of other datasets, does relate to the class. This leads to the largest semantic classification accuracies for both images and text. All datasets were split into train and test sets (in the range of 70-80% and 30-20% respectively). Table 1 summarizes this information.

Table 1. Test set size and uni-modal classification accuracy, for both images and text, on all datasets.

Classifier	TVGraz	Wikipedia	Pascal
Image	59%	30%	25%
Text	91%	84%	65%
size	500	693	300

Experiments: Image retrieval experiments were conducted comparing our method, RIS, to a purely visual method, QBSE [22, 23], and a more recent retrieval method that combines information from images and text, *Text-to-Image Translator* (TTI) [17, 18]. In the latter, text and image co-

occurrences are used to learn a transformation that transfers information from text to images. For best performance, feature transformations should be class specific, and produce a measure of confidence that an image-text pair relates to a concept. Repeating this process across transformations produces a vector of confidence measures for all concepts. TTI was implemented with code provided by its authors. The centered normalized correlation was used as the similarity function

$$S(p,q) = \frac{(p-\mu_p)^T (q-\mu_q)}{||p-\mu_p||\,||q-\mu_q||}.$$

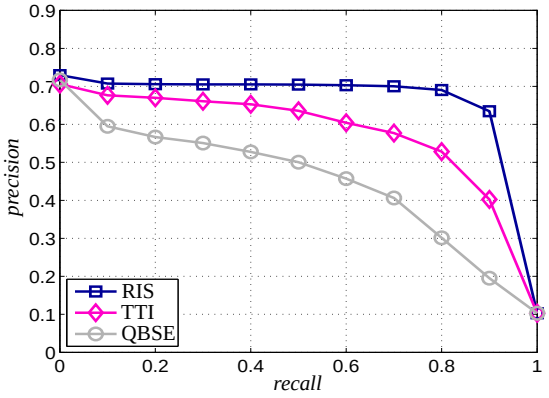
Retrieval performance was evaluated with standard metrics [16]: 1) *precision-recall* (PR) curves, and 2) *mean average precision* (mAP), *i.e*. average precision at the ranks where recall changes.

Results: Table 2 summarizes the mAP results for all datasets. In all cases, the gains of semantic space regularization are quite large. The mAP gain of RIS over QBSE ranges from 44% to 119%. This is strong evidence in support of the multimodal expansion principle. Gains of this magnitude are virtually impossible to obtain from better machine learning, or better image features. Figure 3-(a) presents the average PR curves obtained on TVGraz (similar curves were obtained for the other datasets and are omitted for brevity). Note that the average precision in RIS is quite high and approximately constant as a function of recall. This indicates that RIS has much better generalization than QBSE. This is not surprising, since good generalization is a trademark of effective regularization.

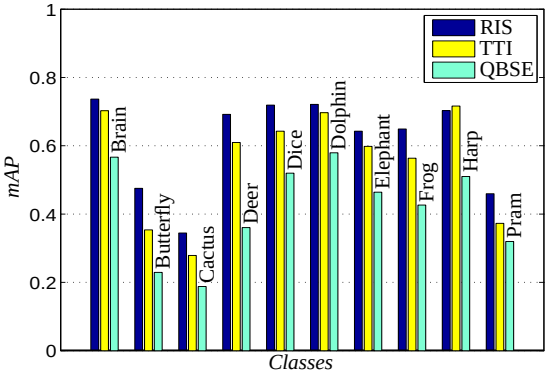
Table 2. Summary of mAP scores. These mAP scores are *per-query*, *i.e*. mean average precision is averaged over all queries. Gains in mAP scores towards our proposed retrieval method (RIS) are shown in (%) .

Method	TVGraz		Wikipedia		Pascal	
	mAP	%	mAP	%	mAP	%
RIS	**0.62**	-	**0.35**	-	**0.21**	-
TTI [17]	0.56	*11*	0.34	*3*	0.19	*11*
QBSE [22]	0.43	*44*	0.16	*119*	0.13	*62*
Random	0.1	*520*	0.1	*250*	0.05	*320*

When compared to TTI, RIS achieved superior performance on all datasets. The mAP gains ranged from 3% to 11%, which are non-trivial improvements. Figure 3-(b) shows the mAP scores obtained per-class on TVGraz (again the results were similar on other datasets), showing superior RIS performance in almost all classes. The average PR curve for RIS on Figure 3-(a) is also higher than that of TTI at all levels of recall, and more constant. Again, this indicates stronger regularization and better generalization. Overall, RIS was clearly superior to TTI.



(a) average PR curves



(b) *per-class* mAP scores

Figure 3. Retrieval evaluation on TVGraz: (a) average PR curves and (b) *per-class* mean average precision.

Figure 4 shows a retrieval example under each method. The top four retrieval results for a butterfly query image are shown.

We have empirically shown the usefulness of accurate auxiliary information in the regularization of images on a probability simplex. Furthermore, we note that our method needs this auxiliary data only to learn the regularizers. Other competitive methods such as [17] require an image/text association in order to produce a confidence that the pair belongs to a certain class. A demo of our method is available in the following URL: `http://www.svcl.ucsd.edu/~josecp/ris/`.

Although it was not explicitly tested, it is a straightforward extension of this work to add more sources of information to learn the regularization operators (*e.g*. audio, or video features where available). This flexibility results from the abstract space where the regularization operators have their domain and co-domain.

Acknowledgments

This work was funded by FCT graduate Fellowship SFRH/BD/40963/2007 and NSF grant CCF-0830535.

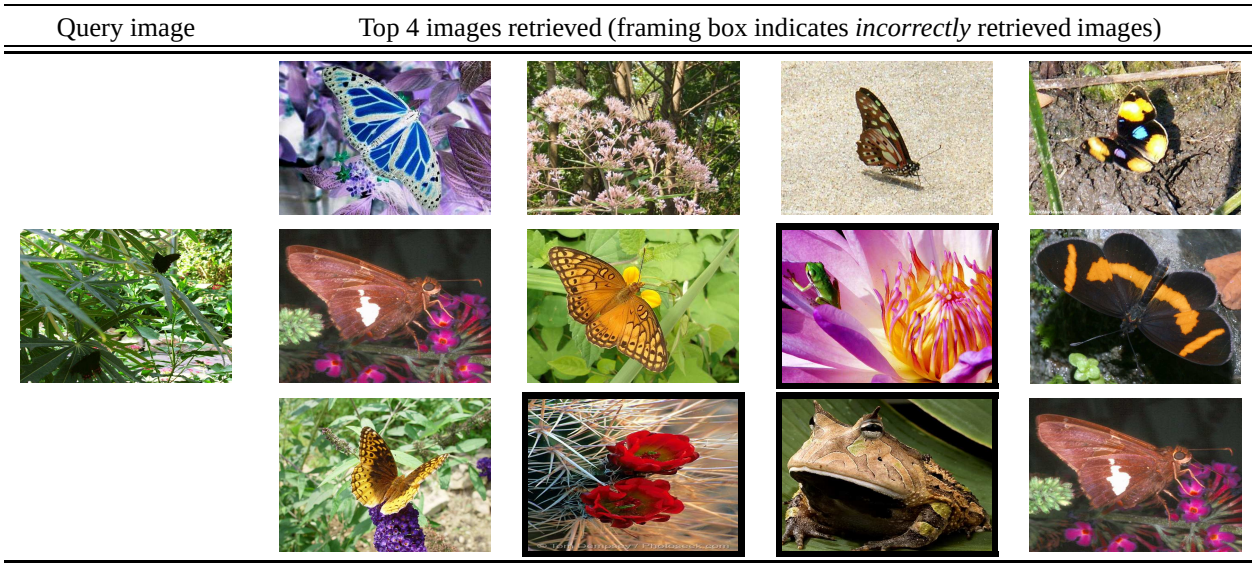


Figure 4. Query image of a *butterfly* (left) with top four retrieval results (right). Top row shows results with our proposed method, RIS, middle row uses TTI [17] and bottom row uses QBSE [22].

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR, MIT Press*, 3:993–1022, 2003. 5

[2] P. Carbonetto, N. Freitas, and K. Barnard. A statistical model for general contextual object recog. *ECCV*, pages 350–362, 2004. 1

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop Stat. Learn. Comp. Vision in ECCV*, volume 1, pages 1–22, 2004. 5

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR, IEEE*, volume 1, pages 886–893, 2005. 1

[5] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *JMLR, MIT Press*, 9:1871–1874, 2008. 5

[6] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR, IEEE*, pages 2352–2359, 2010. 1

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR, IEEE*, pages 1778–1785, 2009. 1

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007. 1

[9] P. Gill, W. Murray, M. Saunders, and M. Wright. Procedures for optimization problems with a mixture of bounds and general linear constraints. *ACM Trans. Math. Software*, 10(3):282–298, 1984. 4

[10] P. Gill, W. Murray, and M. Wright. *Numerical Linear Algebra and Optimization*, volume 1. Addison-Wesley, 1991. 4

[11] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008. 1

[12] I. Khan, A. Saffari, and H. Bischof. TVGraz: Multi-Modal Learning of Object Categories by Combining Textual and Visual Features. In *AAPR Workshop*, pages 213–224, 2009. 5

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR, IEEE*, volume 2, pages 2169–2178, 2006. 1

[14] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. *NIPS*, 2010. 1

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[16] C. Manning, P. Raghavan, and H. Schütze. *An introduction to information retrieval*. Cambridge Univ. Press, 2008. 6

[17] G. Qi, C. Aggarwal, and T. Huang. Towards semantic knowledge propagation from text corpus to web images. In *ACM Int. Conf. WWW*, pages 297–306, 2011. 5, 6, 7

[18] G. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang. Towards Cross-Category Knowledge Propagation for Learning Visual Concepts. In *CVPR, IEEE*, pages 897–904, 2011. 5

[19] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV, IEEE*, pages 1–8, 2007. 1

[20] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Workshop on Creating Speech and Language Data with AMT*, pages 139–147. NAACL HLT, 2010. 5

[21] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. In *ACM Int. Conf. Multimedia*, pages 251–260, 2010. 2, 5

[22] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Trans. Multimedia*, 9(5):923–938, 2007. 2, 5, 6, 7

[23] N. Rasiwasia and N. Vasconcelos. A study of query by semantic example. In *CVPR, IEEE*, pages 1–8, 2008. 1, 5

[24] M. Riesenhuber and T. Poggio. Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2:1019–1025, 1999. 1

[25] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, pages 1–15, 2006. 1

[26] J. Smith and S. Chang. VisualSEEk: a fully automated content-based image query system. In *ACM Int. Conf. Multimedia*, pages 87–98, 1997. 1, 2

[27] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *ICCV, IEEE*, pages 273–280, 2008. 1

[28] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recog. using classemes. *ECCV*, pages 776–789, 2010. 1

[29] N. Vasconcelos and A. Lippman. Library-based Coding: a Representation for Efficient Video Compression and Retrieval. In *DCC, IEEE*, pages 121–130, 1997. 1

[30] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 1