On the role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval

Jose Costa Pereira[1], Emanuele Coviello[1], Gabriel Doyle[2], Nikhil Rasiwasia[1,†], Gert R.G. Lanckriet[1], Roger Levy[2], and Nuno Vasconcelos[1]

[1]Dept. of Electrical and Computer Engineering
[2]Dept. of Linguistics
University of California, San Diego [†]currently with Yahoo! Labs, Bangalore
{josecp,ecoviell,gdoyle,rlevy}@ucsd.edu, nikhil.rasiwasia@gmail.com, {gert,nuno}@ece.ucsd.edu

***Abstract*—The problem of cross-modal retrieval from multimedia repositories is considered. This problem addresses the design of retrieval systems that support queries *across* content modalities, *e.g.*, using an image to search for texts. A mathematical formulation is proposed, equating the design of cross-modal retrieval systems to that of isomorphic feature spaces for different content modalities. Two hypotheses are then investigated, regarding the fundamental attributes of these spaces. The first is that low-level cross-modal correlations should be accounted for. The second is that the space should enable semantic abstraction. Three new solutions to the cross-modal retrieval problem are then derived from these hypotheses: correlation matching (CM), an unsupervised method which models cross-modal correlations, semantic matching (SM), a supervised technique that relies on semantic representation, and semantic correlation matching (SCM), which combines both. An extensive evaluation of retrieval performance is conducted to test the validity of the hypotheses. All approaches are shown successful for text retrieval in response to image queries and vice-versa. It is concluded that both hypotheses hold, in a complementary form, although the evidence in favor of the abstraction hypothesis is stronger than that for correlation.**

***Index Terms*—multimedia, content-based retrieval, multi-modal, cross-modal, image and text, retrieval model, semantic spaces, kernel correlation, logistic regression**

I. Introduction

Classical approaches to information retrieval are of a *uni-modal* nature [1–3]. Text repositories are searched with text queries, image databases with image queries, and so forth. This paradigm is of limited use in the modern information landscape, where multimedia content is ubiquitous. Due to this, *multi-modal* modeling, representation, and retrieval have been extensively studied in the multimedia literature [4–11]. In multi-modal retrieval systems, queries combining multiple content modalities (*e.g.* images and sound of a music video-clip) are used to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). These efforts have become increasingly widespread, due in part to large-scale research and evaluation efforts, such as TRECVID [12] and ImageCLEF [13], involving datasets that span multiple data modalities. However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario. Typically, the different modalities are fused into a representation that does not allow individual access to any of them, *e.g.* some form of dimensionality reduction of a large feature vector that concatenates measurements from images and text. Classical uni-modal techniques are then applied to the low-dimensional representation.

In this work, we consider a richer interaction paradigm, which is denoted *cross-modal* retrieval. The goal is to build content models that enable interactivity with content *across* modalities. Such models can then be used to design cross-modal retrieval systems, where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.* audio tracks). This form of retrieval can be seen as a generalization of current content labeling systems, where a primary modality is augmented with keywords, which can be subsequently searched. Examples include keyword-based image [14–16] and song [17–19] retrieval systems.

A defining property of cross-modal retrieval is the requirement that representations generalize across content modalities. This implies the ability to establish cross-modal links between the attributes (of different modalities) characteristic of each document, or document class. Detecting these links requires deeper content understanding than what is obtained by classical matching of uni-modal attributes. For example, while an image retrieval system can retrieve images of roses by matching red blobs, and a text retrieval system can retrieve texts about roses by matching the "rose" word, a cross-modal retrieval system must *understand* that the word "rose" matches the visual attribute "red blob". This is much closer to what humans do than simple color or word matching. Hence, cross-modal retrieval is a better context than uni-modal retrieval for the study of the fundamental hypotheses on multimedia modeling.

We exploit representations that generalize across content modalities to study two hypotheses on the joint modeling of images and text. The first, denoted the *correlation hypothesis*, is that explicit modeling of low-level correlations between the different modalities is important for the success of the joint models. The second, denoted the *abstraction hypothesis*, is that model benefits from semantic abstraction, *i.e.*, the representation of images and text in terms of semantic (rather than low-level) descriptors. These hypotheses are partly motivated by previous evidence that correlation, *e.g.*, correlation analysis on fMRI [20], and abstraction, *e.g.*, hierarchical topic models for text clustering [21] or hierarchical semantic representations for image retrieval [22], improve performance on uni-modal retrieval tasks. Three joint image-text models that exploit low-level correlation, denoted *correlation matching*, semantic abstraction, denoted *semantic matching*, and both, denoted *semantic correlation matching*, are introduced.

The correlation and abstraction hypotheses are then tested by measuring the retrieval performance of these models on two reciprocal cross-modal retrieval tasks: 1) the retrieval of text documents in response to a query image, and 2) the retrieval of images in response to a query text. These are basic cross-modal retrieval problems, central to many applications of practical interest, such as finding pictures that effectively illustrate a given text (*e.g.* illustrate a page of a story book), finding the texts that best match a given picture (*e.g.* a set of vacation accounts about a given landmark), or searching using a combination of text and images. Model performance on these tasks is evaluated with two datasets: TVGraz [23] and a novel dataset based on Wikipedia's featured articles. These experiments show that correlation modeling and abstraction yield independent benefits. In particular, the best results are obtained by a model that accounts for both low-level correlations — by performing a kernel canonical correlation analysis (KCCA) [24, 25] — and semantic abstraction — by projecting images and texts into a common semantic space [22] designed with logistic regression. This suggests that the hypotheses of abstraction and correlation are complementary, each improving the modeling in a different manner.

The paper is organized as follows. Section II discusses previous work in multi-modal and cross-modal multimedia modeling. Section III presents a mathematical formulation for cross-modal modeling and discusses the two fundamental hypotheses analyzed in this work. Section IV introduces the models underlying correlation, semantic, and semantic correlation matching. Section V summarizes an extensive experimental evaluation designed to test the hypotheses. Conclusions are presented in Section VI. A preliminary version of this work appeared in [26].

II. Previous Work

The problems of image and text retrieval have been the subject of extensive research in the fields of information retrieval, computer vision, and multimedia [2, 10, 12, 27, 28].

Unimodal retrieval: In all these areas, the emphasis has been on uni-modal approaches, where query and retrieved documents share a single modality [1, 2, 10, 29, 30]. For example, in [29] a query text, and in [30] a query image is used to retrieve similar text documents and images, based on low-level text (*e.g.* words) and image (*e.g.* DCTs) representations, respectively. However, this is not effective for all problems. The existence of a well known *semantic gap,* between current image representations and those adopted by humans, severely hampers the performance of uni-modal image retrieval systems [2].

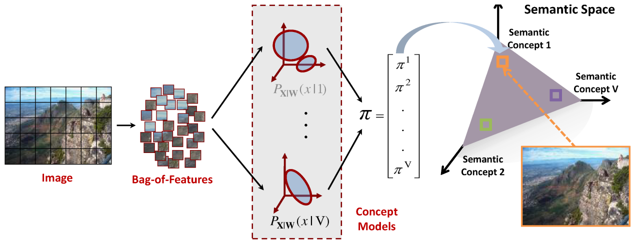


Fig. 1: Semantic space representation. An image is decomposed into a bag-of-features, and represented by the vector of its posterior probabilities with respect to the concepts in a semantic vocabulary $\mathcal{V}$.

Annotations: In general, successful retrieval from large-scale image collections requires that the latter be augmented with text metadata provided by human annotators. These manual annotations are typically in the form of a few keywords, a small caption, or a brief image description [12, 13, 27]. When this metadata is available, the retrieval operation tends to be uni-modal and ignore the images — the text metadata of the query image is simply matched to the text metadata available for images in the database. Because manual image labeling is labor-intensive, recent research has addressed the problem of automatic image labeling[1].

Labeling: A common assumption is that images can be segmented into regions, which can be described by a small word vocabulary. The focus is then on learning a probability model that relates image regions and words. This can be done by learning a joint probability distribution for words and visual features, e.g. using latent Dirichlet allocation models [14], probabilistic latent semantic analysis [31], histograming methods [32], or a combination of Bernoulli distributions for text and kernel-based models for visual features [33, 34]. Alternatively, it is possible to use categorized images to train a dictionary of concept models, e.g. Gaussian mixtures [16] or two-dimensional hidden Markov models [35], in a weakly supervised manner. The extent of association between images and concepts or words is measured by the likelihood of each image under these models. All these methods assume that each image or image region is associated with a single word.

Semantic space: An alternative representation, where images are modelled as weighted combinations of concepts in a pre-defined vocabulary, is proposed in [22]. Statistical models of the distribution of low-level image features are first learned for each concept. The posterior probability of the features extracted from each image, under each of the concept models, is then computed. The image is finally represented by the vector of these posterior concept probabilities. This can be interpreted as a vector of semantic features, establishing a semantic feature space where each dimension is associated with a vocabulary concept. Figure 1 illustrates how this descriptor, denoted a *semantic*

[1]Although not commonly perceived as being *cross-modal*, these systems support cross-modal retrieval, e.g., by returning images in response to explicit text queries.

multinomial (SMN), maps the image into the *semantic space*. All standard image analysis/classification tasks can then be conducted in the latter space, at a higher level of abstraction than that supported by low-level feature spaces. For example, image retrieval is formulated as retrieval by *semantic similarity,* by combining the semantic space with a suitable similarity function [22]. This allows assessments of image similarity in terms of weighted combinations of vocabulary words, and substantially extends the range of concepts that can effectively be retrieved. It also increases the subjective quality of the retrieval results, even when the retrieval system makes mistakes, since images are retrieved by similarity of their content semantics rather than plain visual similarity [36].

Multi-modal retrieval: In parallel with these developments, advances have been reported in multi-modal retrieval systems [8–13, 27]. These are extensions of the classic uni-modal systems, where a common retrieval system integrates information from various modalities. This can be done by fusing features from different modalities into a single vector [37–39], or by learning different models for different modalities and fusing their predictions [40, 41]. One popular approach is to concatenate features from different modalities and rely on unsupervised structure discovery algorithms, such as latent semantic analysis (LSA), to find multi-modal statistical regularities. A good overview of these methods is given in [39], which also discusses the combination of uni-modal and multi-modal retrieval systems. Multi-modal integration has also been applied to retrieval tasks including audio-visual content [42, 43]. In general, the inability to access each data modality individually (after the fusion of modalities) prevents the use of these systems for cross-modal retrieval.

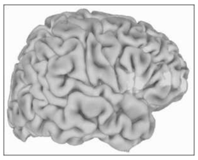
Cross-modal retrieval: Recently, there has been progress towards cross-modal systems. This includes retrieval methods for corpora of images and text [8, 44], images and audio [45, 46], text and audio [47], images, text, and audio [46, 48–51], or even other sources of data like EEG and fMRI [52]. One popular approach is to rely on manifold learning techniques [46, 48–52]. These methods learn a manifold from a matrix of distances between multi-modal objects. The multi-modal distances are formulated as a function of the distances between individual modalities, which allows to single out particular modalities or ignore missing ones. Retrieval then consists of finding the nearest document, on the manifold, to a multimedia query (which can be composed of any subset of modalities). The main limitation of these methods is the lack of out-of-sample generalization. Since there is no computationally efficient way to project the query into the manifold, queries are restricted to the training set used to learn the latter. Hence, all unseen queries must be mapped to their nearest neighbors in this training set, defeating the purpose of manifold learning.

An alternative is to learn correlations between modalities [45, 53]. For example, [45] compares canonical correlation analysis (CCA) and cross-modal factor analysis (CFA) in the context of audio-image retrieval. Both CCA and CFA perform a joint dimensionality reduction that extracts highly

(a)

Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in ”Time” magazine as saying, ”The new administration should have been given a chance to confer with the various groups interested in change.” Black hotel owner A. G. Gaston stated, ”I regret the absence of continued communication between white and Negro leadership in our city.” A white Jesuit priest assisting in desegregation negotiations attested, ”These demonstrations are poorly timed and misdirected.” Protest organizers knew they would meet with violence from the Birmingham Police Department but chose a confrontational approach to get the attention of the federal government. (...)

Home - Courses - Brain and Cognitive Sciences - A Clinical Approach to the Human Brain 9.22J / HST.422J A Clinical Approach to the Human Brain Fall 2006 Activity in the highlighted areas in the prefrontal cortex may affect the level of dopamine in the mid-brain, in a finding that has implications for schizophrenia. (Image courtesy of the National Institutes of Mental Health.) Course Highlights This course features summaries of each class in the lecture notes section, as well as an extensive set of readings. Course Description This course is designed to provide an understanding of how the human brain works in health and disease, and is intended for both the Brain and Cognitive Sciences major and the non-Brain and Cognitive Sciences major. (...)



(b)

Fig. 2: Two examples of image-text pairs: (a) section from the Wikipedia article on the Birmingham campaign (“History” category), (b) part of a Cognitive Science class syllabus from the TVGraz dataset (“Brain” category).

correlated features in the two data modalities. A kernelized version of CCA was also proposed in [53] to extract translation invariant semantics of text documents written in multiple languages. It was later used to model correlations between web images and corresponding captions, in [20]. Another approach is *re-ranking*: uni-modal retrieval is first performed using the query modality, and a second modality is used to re-rank the results [54, 55].

Rich annotation: Despite all these advances, current approaches tend to rely on a limited textual representation, in the form of keywords, captions, or small text snippets. We refer to these as forms of lighter annotation. This is at odds with the ongoing explosion of multimedia content on the web, where it is now possible to collect large sets of extensively annotated data. Examples include news archives, blog posts, or Wikipedia pages, where pictures are related to *complete* text articles, not just a few keywords. We refer to these datasets as *richly annotated*. While potentially more informative, rich annotation establishes a much more nuanced connection between images and text than light annotation. While keywords tend to be explicit image labels, many of the words in a rich text can be unrelated to the image used to illustrate it. For example, Figure 2a shows a section of the Wikipedia article on the “Birmingham campaign”, along with the associated image. Notice that, although related to the text, the image is clearly not representative of all the words in the article. The same is true for the web-page in Figure 2b, from the TVGraz dataset [23]. This is a course syllabus that, beyond the pictured brain, includes course information and other unrelated matters. A major long-term goal of modeling richly annotated data is to recover this *latent* relationship between the text and image components of a document, and exploit it in benefit of practical applications.

III. Fundamental Hypotheses

In this section, we present a novel multi-modal content modeling framework, which is flexible and applicable to rich content modalities. Although the fundamental ideas are

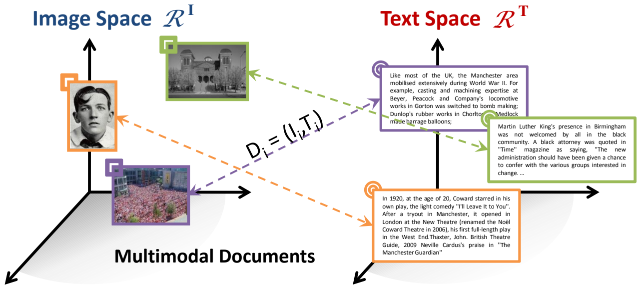


Fig. 3: A document (D_i) is a pair of an image (I_i) and a text (T_i) represented as vectors in feature spaces $\Re^I$ and $\Re^T$, respectively. Documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$.

applicable to any combination of modalities we restrict the discussion to documents containing images and text.

A. *The problem*

We consider the problem of information retrieval from a database $\mathcal{D} = \{D_1, \ldots, D_{|D|}\}$ of *documents* comprising *image* and *text* components. Such documents can be quite diverse: from a single text complemented by one or more images (*e.g.* a newspaper article) to documents containing multiple pictures and text sections (*e.g.* a Wikipedia page). For simplicity, we consider the case where each document consists of a single image and its accompanying text, *i.e.*, $D_i = (I_i, T_i)$. Images and text are represented as vectors in feature spaces $\Re^I$ and $\Re^T$, respectively, as illustrated in Figure 3. In this way, documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$. Given a text (image) query $T_q \in \Re^T$ ($I_q \in \Re^I$), the goal of *cross-modal retrieval* is to return the closest match in the image (text) space $\Re^I$ ($\Re^T$).

B. *Multi-modal modeling*

Whenever the image and text spaces have a natural correspondence, cross-modal retrieval reduces to a classical retrieval problem. Let

$$\mathcal{M} : \Re^T \rightarrow \Re^I$$

be an invertible mapping between the two spaces. Given a query T_q in $\Re^T$, it suffices to find the nearest neighbor to $\mathcal{M}(T_q)$ in $\Re^I$. Similarly, given a query I_q in $\Re^I$, it suffices to find the nearest neighbor to $\mathcal{M}^{-1}(I_q)$ in $\Re^T$. In this case, the design of a cross-modal retrieval system reduces to the design of an effective similarity function for determining the nearest neighbors.

In general, however, different representations are adopted for images and text, and there is no natural correspondence between $\Re^I$ and $\Re^T$. In this case, the mapping $\mathcal{M}$ has to be learned from examples. In this work, we map the two representations into intermediate spaces, $\mathcal{V}^I$ and $\mathcal{V}^T$, that have a natural correspondence. This consists of learning two mappings

$$\mathcal{M}_I : \Re^I \rightarrow \mathcal{V}^I \qquad \mathcal{M}_T : \Re^T \rightarrow \mathcal{V}^T$$

from each of the image and text spaces to two *isomorphic* spaces $\mathcal{V}^I$ and $\mathcal{V}^T$, connected by an invertible mapping

$$\mathcal{M} : \mathcal{V}^T \rightarrow \mathcal{V}^I.$$

Given a text query T_q in $\Re^T$, cross-modal retrieval reduces to finding the image I_r such that $\mathcal{M}_I(I_r)$ is the nearest neighbor of

$$\mathcal{M} \circ \mathcal{M}_T(T_q)$$

in $\mathcal{V}^I$. Similarly, given an image query I_q in $\Re^I$, the goal is to find text T_r such that $\mathcal{M}_T(T_r)$ is the nearest neighbor of

$$\mathcal{M}^{-1} \circ \mathcal{M}_I(I_q)$$

in $\mathcal{V}^T$. Under this formulation, the main problem in the design of a cross-modal retrieval system is the design of the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ (and the corresponding mappings $\mathcal{M}_I$ and $\mathcal{M}_T$).

C. *The fundamental hypotheses*

Since the goal is to design representations that generalize across content modalities, the solution of this problem requires some ability to derive a more *abstract* representation than the sum of the parts (low-level features) extracted from each content modality. Given that such abstraction is the hallmark of true image or text understanding, this problem enables the exploration of some central questions in multimedia modeling. Consider, for example, a query for a “swan”. While 1) a uni-modal image retrieval system can successfully retrieve images of “swans” in that they are the only white objects in a database, 2) a text retrieval system can successfully retrieve documents about “swans” because they are the only documents containing the word “swan”, and 3) a multi-modal retrieval system can simply match “white” to “white” and “swan” to “swan”, a cross-modal retrieval system cannot solve the task without understanding that “white is a visual attribute of swan”. Hence, cross-modal retrieval is a more effective paradigm for testing fundamental hypotheses in multimedia representation than uni-modal or multi-modal retrieval.

In this work, we exploit the cross-modal retrieval problem to test two such hypotheses regarding the joint modeling of images and text.

- $\mathcal{H}_1$ (**correlation** hypothesis): low-level cross-modal correlations are important for joint image-text modeling.
- $\mathcal{H}_2$ (**abstraction** hypothesis): semantic abstraction is important for joint image-text modeling.

The hypotheses are tested by comparing three possibilities for the design of the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ of cross-modal retrieval. In the first case, two feature transformations map $\Re^I$ and $\Re^T$ onto *correlated* d-dimensional *subspaces* denoted as $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively, which act as $\mathcal{V}^I$ and $\mathcal{V}^T$. This maintains the level of semantic abstraction of the representation while maximizing the correlation between the two spaces. We refer to this matching technique as *correlation matching* (CM). In the second case, a pair of transformations are used to map the image

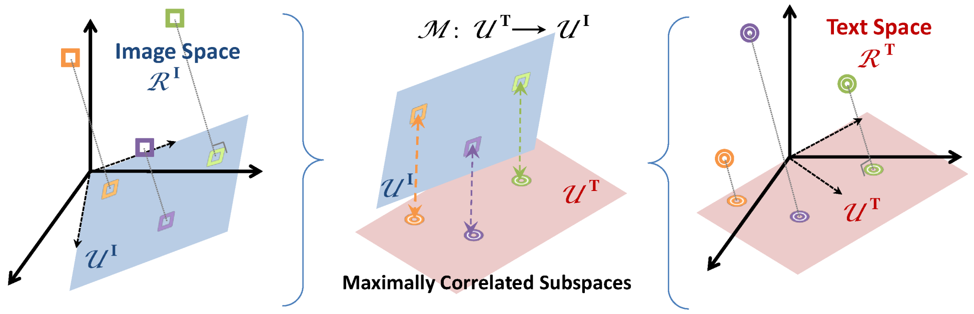


Fig. 4: Correlation matching. Text ($\Re^T$) and images ($\Re^I$) are projected onto two maximally correlated isomorphic subspaces $\mathcal{U}_T$ and $\mathcal{U}_I$, respectively.

TABLE I: Taxonomy of proposed approaches to cross-modal retrieval.

	correlation hypothesis	abstraction hypothesis
CM	√	
SM		√
SCM	√	√

and text spaces into a pair of semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, which then act as $\mathcal{V}^I$ and $\mathcal{V}^T$. This increases the semantic abstraction of the representation without directly seeking correlation maximization. The spaces $\mathcal{S}^I$ and $\mathcal{S}^T$ are made isomorphic by using the same set of semantic concepts for both modalities. We refer to this as *semantic matching* (SM). Finally, a third approach combines the previous two techniques: project onto maximally correlated subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$, and then project again onto a pair of semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, which act as $\mathcal{V}^I$ and $\mathcal{V}^T$. We refer to this as *semantic correlation matching* (SCM).

Table I summarizes which hypotheses hold for each of the three approaches. The comparative evaluation of the performance of these approaches on cross-modal retrieval experiments provides indirect evidence for the importance of the above hypotheses to the joint modeling of images and text. The intuition is that a better cross-modal retrieval performance results from a more effective joint modeling.

IV. Cross-modal Retrieval

In this section, we present the three approaches in detail.

A. Correlation matching (CM)

The design of a mapping from $\Re^T$ and $\Re^I$ to the correlated spaces $\mathcal{U}^T$ and $\mathcal{U}^I$ requires a combination of dimensionality reduction and some measure of correlation between the text and image modalities. In both text and vision literatures, dimensionality reduction is frequently accomplished with methods such as latent semantic indexing (LSI) [56] and principal component analysis (PCA) [57]. These are members of a broader class of learning algorithms, denoted subspace learning, which are computationally efficient, and produce linear transformations that are easy to conceptualize, implement, and deploy. Furthermore, because subspace learning is usually based on second order statistics, such as correlation, it can be easily extended to the multi-modal setting and kernelized. This has motivated a number of multi-modal subspace methods. In this work, we consider *cross-modal factor analysis* (CFA), *canonical correlation analysis* (CCA), and *kernel canonical correlation analysis* (KCCA). All these methods include a training stage, where the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are learned, followed by a projection stage, where images and text are projected into these spaces. Figure 4 illustrates this process. Cross-modal retrieval is performed in the low-dimensional subspaces.

1) Linear subspace learning: CFA seeks transformations that best represent coupled patterns between different subsets of features (*e.g.* different modalities) describing the same objects [45]. It finds the orthonormal transformations Ω_I and Ω_T that project the two modalities onto a shared space, $\mathcal{U}^I = \mathcal{U}^T = \mathcal{U}$, where the projections have minimum distance

$$\left\| X_I \Omega_I - X_T \Omega_T \right\|_F^2. \tag{1}$$

X_I and X_T are matrices containing corresponding features from the image and text domains, and $|| \cdot ||_F^2$ is the Frobenius norm. It can be shown that this is equivalent to maximizing

$$trace(X_I \Omega_I \Omega_T' X_T'), \tag{2}$$

and the optimal matrices Ω_I, Ω_T can be obtained by a singular value decomposition of the matrix $X_I' X_T$, *i.e.*,

$$X_I' X_T = \Omega_I \Lambda \Omega_T, \tag{3}$$

where Λ is the matrix of singular values of $X_I' X_T$ [45].

CCA [58] learns the d-dimensional subspaces $\mathcal{U}^I \subset \Re^I$ (image) and $\mathcal{U}^T \subset \Re^T$ (text) where the correlation between the two data modalities is maximal. It is similar to principal components analysis (PCA), in the sense that it learns a basis of canonical components, directions $w_i \in \Re^I$ and $w_t \in \Re^T$, but seeks directions along which the data is maximally correlated

$$\max_{w_i \neq 0,\, w_t \neq 0} \frac{w_i' \Sigma_{IT} w_t}{\sqrt{w_i' \Sigma_I w_i} \sqrt{w_t' \Sigma_T w_t}} \tag{4}$$

where Σ_I and Σ_T are the empirical covariance matrices for images $\{I_1, \ldots, I_{|D|}\}$ and text $\{T_1, \ldots, T_{|D|}\}$ respectively, and $\Sigma_{IT} = \Sigma_{TI}'$ the cross-covariance between them. Repeatedly solving (4) for directions that are orthogonal to all previously obtained solutions, provides a series of canonical components. It can be shown that the canonical components in the image space can be found as the eigenvectors of $\Sigma_I^{-1/2}\Sigma_{IT}\Sigma_T^{-1}\Sigma_{TI}\Sigma_I^{-1/2}$, and in the text space as the eigenvectors of $\Sigma_T^{-1/2}\Sigma_{TI}\Sigma_I^{-1}\Sigma_{IT}\Sigma_T^{-1/2}$. The first d eigenvectors $\{w_{i,k}\}_{k=1}^{d}$ and $\{w_{t,k}\}_{k=1}^{d}$ define a basis of the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$.

2) Non-linear subspace learning: CCA and CFA can only model linear dependencies between image and text features. This limitation can be avoided by mapping these features into high-dimensional spaces, with a pair of non-linear transformations $\phi_T : \Re^T \rightarrow \mathcal{F}^T$ and $\phi_I : \Re^I \rightarrow \mathcal{F}^I$. Application of CFA or CCA in these spaces can then recover complex patterns of dependency in the original feature space. As is common in machine learning, the transformations $\phi_T(\cdot)$ and $\phi_I(\cdot)$ are computed only implicitly, by the introduction of two kernel functions $\mathcal{K}_T(\cdot,\cdot)$ and $\mathcal{K}_I(\cdot,\cdot)$, specifying the inner products in $\mathcal{F}^T$ and $\mathcal{F}^I$, *i.e.*, $\mathcal{K}_T(T_m, T_n) = \langle\phi_T(T_m), \phi_T(T_n)\rangle$ and $\mathcal{K}_I(I_m, I_n) = \langle\phi_I(I_m), \phi_I(I_n)\rangle$, respectively.

KCCA [24, 25] implements this type of extension for CCA, seeking directions $w_i \in \mathcal{F}^I$ and $w_t \in \mathcal{F}^T$, along which the two modalities are maximally correlated in the transformed spaces. The canonical components can be found by solving

$$\max_{\alpha_i \neq 0,\, \alpha_t \neq 0} \frac{\alpha_i' K_I K_T \alpha_t}{V(\alpha_i, K_I)V(\alpha_t, K_T)}, \tag{5}$$

where $V(\alpha, K) = \sqrt{(1-\kappa)\alpha' K^2 \alpha + \kappa\alpha' K\alpha}$, $\kappa \in [0,1]$ is a regularization parameter, and K_I and K_T are the kernel matrices of the image and text representations, *e.g.*, $(K_I)_{mn} = \mathcal{K}_I(I_m, I_n)$. Given optimal α_i and α_t for (5), w_i and w_t are obtained as linear combinations of the training examples $\{\phi_I(I_k)\}_{k=1}^{|\mathcal{D}|}$, and $\{\phi_T(T_k)\}_{k=1}^{|\mathcal{D}|}$, with α_i and α_t as weight vectors, *i.e.*, $w_i = \Phi_I(X_I)^T\alpha_i$ and $w_t = \Phi_T(X_T)^T\alpha_t$, where $\Phi_I(X_I)$ ($\Phi_T(X_T)$) is the matrix whose rows contain the high-dimensional representation of the image (text) features. To optimize (5), we solve a generalized eigenvalue problem using the software package of [25]. The first d generalized eigenvectors, where $1 \leq d \leq |\mathcal{D}|$, are the d weight vectors $\{\alpha_{i,k}\}_{k=1}^{d}$ and $\{\alpha_{t,k}\}_{k=1}^{d}$ that define the bases $\{w_{i,k}\}_{k=1}^{d}$ and $\{w_{t,k}\}_{k=1}^{d}$ of the two maximally correlated d-dimensional subspaces $\mathcal{U}^I \subset \mathcal{F}^I$ and $\mathcal{U}^T \subset \mathcal{F}^T$.

3) Image and text projections: Images and text are represented by their projections p_I and p_T onto the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively. p_I (p_T) is obtained by computing the dot-products between the vector representing the image (text) $I \in \Re^I$ ($T \in \Re^T$) and the image (text) basis vectors spanning $\mathcal{U}^I$ ($\mathcal{U}^T$). For CFA, the basis vectors are the columns of Ω_I and Ω_T, respectively. For CCA, they are $\{w_{i,k}\}_{k=1}^{d}$ and $\{w_{t,k}\}_{k=1}^{d}$. In the case of KCCA, an image $I \in \Re^I$ is first mapped into $\mathcal{F}^I$ and subsequently projected onto $\{w_{i,k}\}_{k=1}^{d}$, *i.e.*, $p_I = \mathcal{P}_I(\phi_I(I))$ with

$$\begin{aligned} p_{I,k} &= \langle\phi_I(I), w_{i,k}\rangle \\ &= \langle\phi_I(I), \left[\phi_I(I_1), \ldots, \phi_I(I_{|\mathcal{D}|})\right]\alpha_{i,k}\rangle \\ &= \left[\mathcal{K}_I(I, I_1), \ldots, \mathcal{K}_I\left(I, I_{|\mathcal{D}|}\right)\right]\alpha_{i,k}, \end{aligned} \tag{6}$$

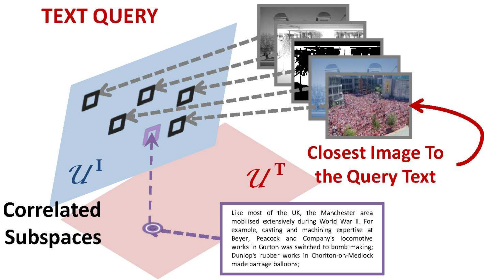


Fig. 5: Example of cross-modal retrieval using CM. Here, CM is used to find the images that best match a query text.

where $k = 1, \ldots, d$. Analogously, a text $T \in \Re^T$ is mapped into $\mathcal{F}^T$ and then projected onto $\{w_{t,k}\}_{k=1}^{d}$, *i.e.*, $p_T = \mathcal{P}_T(\phi_T(T))$, using $\mathcal{K}_T(.,.)$.

4) Correlation matching: For all methods, a natural invertible mapping between the projections onto $\mathcal{U}^I$ and $\mathcal{U}^T$ follows from the correspondence between the d-dimensional bases of the subspaces, as $w_{i,1} \leftrightarrow w_{t,1}$, ..., $w_{i,d} \leftrightarrow w_{t,d}$. This results in a compact, efficient representation of both modalities, where vectors p_I and p_T are coordinates in two isomorphic d-dimensional subspaces, as shown in Figure 4. Given an image query I with projection p_I, the text $T \in \Re^T$ that most closely matches it is that for which p_T minimizes

$$D(I,T) = d(p_I, p_T), \tag{7}$$

for some suitable distance measure $d(\cdot,\cdot)$ in a d-dimensional vector space. Similarly, given a query text T with projection p_T, the closest image match $I \in \Re^I$ is that for which p_I minimizes $d(p_I, p_T)$. An illustration of cross-modal retrieval using CM is given in Figure 5.

B. Semantic matching (SM)

An alternative to subspace learning is to map images and text to representations at a higher level of abstraction, where a natural correspondence can be established. This is obtained by augmenting the database $\mathcal{D}$ with a vocabulary $\mathcal{V} = \{v_1, \ldots, v_K\}$ of semantic concepts. These can be generic or application dependent, ranging from generic document attributes, such as “Long” or “Short”, to specific topics such as “History” or “Biology”, or any other categories that are deemed relevant. Individual documents are grouped into these semantic concepts. Two mappings $\mathcal{L}_T$ and $\mathcal{L}_I$ are then implemented using classifiers of text and images, respectively. $\mathcal{L}_T$ maps a text $T \in \Re^T$ into a vector π_T of posterior probabilities $P_{V|T}(v_j|T), j \in \{1,\ldots,K\}$ with respect to each of the concepts in $\mathcal{V}$. The space $\mathcal{S}^T$ of these vectors is referred to as the *semantic space for text*, and the probabilities in π_T as the *semantic text features*. Similarly, $\mathcal{L}_I$ maps an image I into a vector π_I of *semantic image features* in a *semantic space for images* $\mathcal{S}^I$.

Semantic representations have two advantages for cross-modal retrieval. First, they provide a higher level of abstraction. While features in $\Re^T$ and $\Re^I$ frequently have no obvious interpretation (*e.g.*, image features tend to be edges, edge orientations or frequency bases), the features

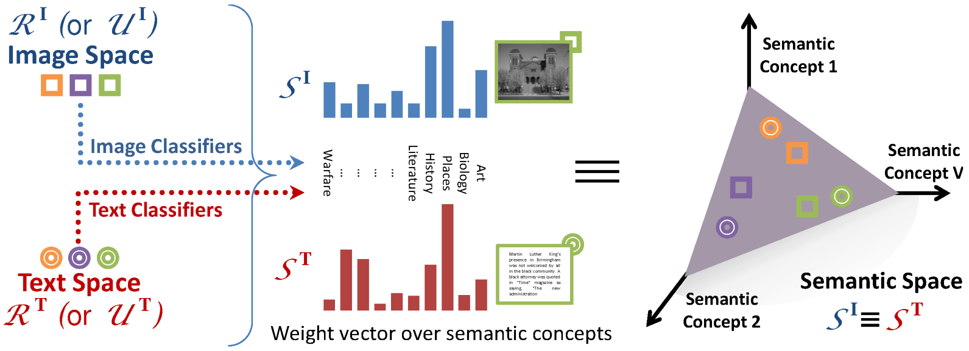


Fig. 6: Semantic matching (SM). Text and images are mapped into a common semantic space, using the posterior class probabilities produced by a multi-class text or image classifier.

in $\mathcal{S}^T$ and $\mathcal{S}^I$ are (semantic) concept probabilities (*e.g.*, the probability that the image belongs to the "History" or "Biology" document classes). Previous work has shown that increased feature abstraction can lead to substantially better generalization for tasks such as image retrieval [22]. Second, the semantic spaces $\mathcal{S}^T$ and $\mathcal{S}^I$ are isomorphic, since both images and text are represented as vectors of posterior probabilities with respect to the *same* set of semantic concepts. Hence, the spaces can be treated as being the same, *i.e.*, $\mathcal{S}^T = \mathcal{S}^I$, leading to the representation of Figure 6.

1) Learning: Many classification techniques can be used to learn the mappings $\mathcal{L}_T$ and $\mathcal{L}_I$. In this work, we consider three popular methods. Logistic regression computes the posterior probability of a particular class by fitting image (text) features to a logistic function. Parameters are chosen to minimize the loss function,

$$\min_w \frac{1}{2} w'w + C \sum_i \log(1 + \exp(-y_i w' x_i)) \quad (8)$$

where y_i is the class label, x_i the feature vector in the input space, and w a vector of parameters. A multi-class logistic regression can be learned for the image and text modalities, by making x_i the image and text representation, $I \in \Re^I$ and $T \in \Re^T$, respectively. In our implementation this is done with the Liblinear software package [59].

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

$$\min_{w,b,\xi} \frac{1}{2} w'w + C \sum_i \xi_i \quad (9)$$

$$\begin{aligned} s.t. \quad & y_i(w'x_i + b) \geq 1 - \xi_i, \ \forall i \\ & \xi_i \geq 0 \end{aligned}$$

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. Although the SVM output does not have a probabilistic interpretation, a sigmoidal transformation of the SVM scores $y_i w' x_i$ is often taken as a proxy for the posterior class probabilities. This is, for example, supported by the LibSVM [60] package, which we use in our implementation.

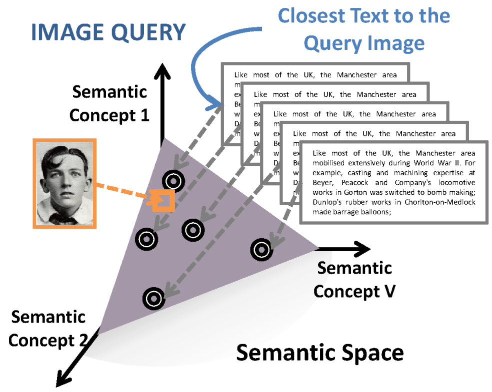


Fig. 7: An example of cross-modal retrieval using SM. Here SM is used to find the texts that best match a query image.

Boosting methods combine weak learners into a strong decision rule. Many boosting algorithms have been proposed in the literature. In this work we adopt the multi-class boosting method of [61]. This is based on multi-dimensional codewords (y^k) and predictors (f). Each class k is mapped to a distinct class label y^k, and the strong classifier, $F(x)$, is a mapping from examples $x_i \in \mathcal{X}$ into class labels y^k

$$F(x) = \arg\max_k y^k f^*(x) \quad (10)$$

where $f^*(x) : \mathcal{X} \to \mathbb{R}$ is the continuous valued predictor that maximizes the classification margin. Posterior class probabilities can then be recovered by applying a non-linear transformation to the classifier output. In our implementation this is done with recourse to the multi-class boosting software package of [61].

2) Retrieval: Given a query image I (text T), represented by $\pi_I \in \mathcal{S}^I$ ($\pi_T \in \mathcal{S}^T$), SM-based cross-modal retrieval returns the text T (image I), represented by $\pi_T \in \mathcal{S}^T$ ($\pi_I \in \mathcal{S}^I$), that minimizes

$$D(I, T) = d(\pi_I, \pi_T), \quad (11)$$

for some suitable distance measure d between probability distributions. An illustration of cross-modal retrieval using SM is given in Figure 7.

TABLE II: MAP scores (validation set) of different distance measures. μ_p and μ_q are the sample averages for p and q, respectively.

			TVGraz			Wikipedia		
	measure	$d(p,q)$	img query	txt query	avg	img query	txt query	avg
CM	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.376	0.418	0.397	0.193	0.234	0.214
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.391	0.444	0.417	0.199	0.243	0.221
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.498	0.476	**0.487**	0.288	0.239	**0.263**
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \lvert\lvert q-\mu_q \rvert\rvert}$	0.486	0.462	0.474	0.287	0.239	**0.263**
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.362	0.564	0.463	0.206	0.274	0.240
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.525	0.573	0.549	0.220	0.274	0.247
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.492	0.570	0.531	0.205	0.276	0.241
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.582	0.581	0.582	0.301	0.276	0.289
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \lvert\lvert q-\mu_q \rvert\rvert}$	0.598	0.578	**0.588**	0.352	0.272	**0.312**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.560	0.623	0.592	0.311	0.270	0.291
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.623	0.633	0.628	0.334	0.273	0.304
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.605	0.615	0.610	0.315	0.267	0.291
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.665	0.632	0.649	0.371	0.279	0.325
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \lvert\lvert q-\mu_q \rvert\rvert}$	0.669	0.633	**0.651**	0.382	0.281	**0.332**

C. Semantic Correlation Matching (SCM)

CM and SM are not mutually exclusive. In fact, a corollary to the two hypotheses discussed above is that there may be a benefit in combining CM and SM. CM extracts maximally correlated features from $\Re^T$ and $\Re^I$. SM builds semantic spaces using original features to gain semantic abstraction. When the two are combined, by building semantic spaces using the feature representation produced by correlation maximization, it may be possible to improve on the individual performances of both CM and SM. To combine the two approaches, the maximally correlated subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are first learned and the projections (p_I, p_T) of each image-text pair (I, T) computed, as discussed in Section IV-A. The transformations $\mathcal{L}_I$ and $\mathcal{L}_T$ are then learned in each of these subspaces to produce the semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, respectively. Retrieval is finally based on the image-text distance $D(I, T)$ of (11), based on the semantic mappings $\pi_I = \mathcal{L}_I(p_I)$ and $\pi_T = \mathcal{L}_T(p_T)$.

V. Experiments

In this section, we describe an extensive experimental evaluation of the proposed cross-modal retrieval framework.

A. Experimental set-up

We start with a brief review of the adopted datasets, performance metrics, and image and text representations.

1) Datasets: The performance of cross-modal retrieval was evaluated with two datasets that pair pictures and text, *TVGraz* and *Wikipedia*. TVGraz is a collection of web-pages compiled by Khan *et al.* [23]. Google image search was used to retrieve $1,000$ web-pages for each of the 10 categories of the Caltech-256 [62] dataset listed in the second column of Table VII. This dataset is provided as a list of URLs, which we used to collect $2,058$ image-text pairs (defunct URLs and web-pages without at least 10 words and one image were discarded). The median text length, per web-page, is 289 words. A random split was used to produce $1,558$ training and 500 test documents.

Wikipedia is a novel dataset, assembled from the "Wikipedia featured articles". The latter are divided into 30 categories. Since some of these contain very few entries we considered only articles from the 10 top most populated ones, which are listed in the second column of Table VIII. Since the featured articles tend to have multiple images and span multiple topics, each article was split into sections, based on its section headings. Each image was assigned to the section in which it was placed by the author(s). This produced $7,114$ sections, which are internally more coherent and usually contain a single picture. The dataset was then pruned, by keeping only sections with exactly one image and at least 70 words. The final corpus contains a total of $2,866$ documents. The median text length is 200 words. A random split was used to produce a training set of $2,173$ documents and a test set of 693 documents.

The two datasets have different properties. TVGraz images are archetypal members of the categories. The dataset is eminently visual, since its categories (*e.g.*, "Harp", "Dolphin") are specific objects or animals. The texts are small and can be less representative of the categories. In Wikipedia, on the other hand, category membership is mostly driven by text. Texts are mostly of good quality and representative of the category, while the image categorization is more ambiguous. For example, a portrait of a historical figure can appear in the class "War". The Wikipedia categories (*e.g.*, "History", "Biology") are more abstract concepts, and have much broader scope. Individually, the images can be difficult to classify, even for a human. Together, the two datasets illustrate the potential diversity of cross-modal retrieval: applications where there is more uniformity of text than images, and vice-versa.

2) Performance metrics: Two tasks were considered: text retrieval from an image query, and image retrieval from a text query. All text queries were based on full text

TABLE III: MAP scores (validation set) under the CM hypothesis.

	img q.	txt q.	avg.
TVGraz			
LD-KCCA	0.428	**0.471**	0.450
KCCA	**0.486**	0.462	**0.474**
CCA	0.284	0.254	0.269
CFA	0.195	0.179	0.187
PCA	0.162	0.144	0.153
Wikipedia			
LD-KCCA	0.242	**0.241**	0.242
KCCA	**0.287**	0.239	**0.263**
CCA	0.210	0.174	0.192
CFA	0.195	0.156	0.176
PCA	0.208	0.132	0.170

TABLE IV: MAP scores (validation set) under the SM hypothesis.

	img q.	txt q.	avg.
TVGraz			
Log. Reg.	**0.598**	**0.578**	**0.588**
SVM	0.556	0.548	0.552
Boosting	0.567	0.476	0.522
Wikipedia			
Log. Reg.	**0.352**	**0.272**	**0.312**
SVM	0.318	0.237	0.278
Boosting	0.322	0.207	0.265

TABLE V: Optimal parameters (validation set) for best retrieval architecture.

	Cbk size	LDA topics	KCCA comps
TVGraz			
SCM	4096	400	700
SM		100	n/a
CM		200	8
Wikipedia			
SCM	4096	200	100
SM		600	n/a
CM		20	10

documents. Retrieval performance was evaluated using 11-point interpolated *precision-recall* (PR) curves [63]. These were then used to compute mean average precision (MAP) scores. The MAP score is the average precision at the ranks where recall changes. The set of classes used to evaluate precision and recall is denoted *ground-truth semantics*. Unless otherwise noted, this is the set of classes that make up each dataset, as listed in the second column of Tables VII and VIII. All results were compared to a baseline established by a recently published cross-modal retrieval approach, the *Text-To-Image* (TTI) translator of [44]. This was implemented with code provided by its authors.

3) Image and text representation: For both modalities, the base representation is a bag-of-words (BOW). Text words (extracted by stemming the text with the Python Natural Language Toolkit[2]) were fit by a latent Dirichlet allocation (LDA) [21] model, using the implementation of [64]. For images, a bag of SIFT descriptors was first extracted per training image[3] and a visual word codebook learned with K-means clustering. SIFT descriptors extracted from each image were finally vector quantized with this codebook, to produce a vector of visual word counts.

B. Preliminary experiments

Various preliminary experiments were conducted to identify the best parameter configurations for the cross-modal retrieval architecture. They were based on a random 80/20 split of the training sets: $1,245$ training and 313 validation examples on TVGraz, $1,738$ training and 435 validation on Wikipedia. The validation sets were used to determine the best parameter configurations. When a semantic representation is used, the semantic vocabulary $\mathcal{V}$ used to design the semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$ consists of the ground-truth semantics.

1) Distance Measures: A number of distance measures, listed in Table II, were considered for the evaluation of (7) and (11): Kullback-Leibler divergence (KL), ℓ_1 and ℓ_2 norms, normalized correlation (NC), and centered normalized correlation (NC_c). KL was not used in correlation matching because this technique does not produce a probability simplex. Table II presents the MAP scores achieved with each measure. Since NC_c had the best average performance in nearly all experiments, it was adopted as distance measure.

2) Correlation matching: A set of experiments was performed to compare the performance of CFA, CCA, and KCCA. In all cases, the number of canonical components was validated in each retrieval experiment. As shown in Table III, KCCA had the top performance. Best results were achieved with a chi-square radial basis function kernel[4] for images, a histogram intersection kernel for text [65, 66], and regularization constants $\kappa = 10\%$ on TVGraz and $\kappa = 50\%$ on Wikipedia. To verify the importance of modeling correlations, we considered two alternative representations. The first implemented dimensionality reduction but no correlation modeling. The two modalities were independently projected into subspaces of the same dimension, learned with PCA. The second investigated the benefits of complementing correlation with discriminant modeling, by introducing a linear discriminant analysis on the correlated subspaces discovered by KCCA. It is denoted *linear discriminant kernel canonical correlation analysis* (LD-KCCA). As shown in Table III, neither alternative improved on the average MAP scores of KCCA. This shows that there are benefits to correlation matching beyond dimensionality reduction and that further gains are not trivial to achieve, supporting the hypothesis that correlation modeling is important for cross-model retrieval. Given its good performance, KCCA was used in all remaining CM experiments.

3) Semantic matching: A set of experiments was performed to evaluate the impact of the classification architecture used to design the semantic space on retrieval accuracy. Three architectures were compared: logistic regression,

[2]http://www.nltk.org/

[3]SIFT from https://lear.inrialpes.fr/people/dorko/downloads.html

[4]$\mathcal{K}(x, y) = \exp\left(\frac{d_{\chi^2}(x, y)}{\gamma}\right)$ where $d_{\chi^2}(x, y)$ is the chi-square distance between x and y and γ is the average chi-square distance among training points.

TABLE VI: MAP scores (test set) of CM, SM, SCM, and TTI, on TVGraz and Wikipedia.

	img. query	txt. query	avg.	gain
TVGraz				
SCM	**0.664**	**0.649**	**0.657**	–
SM	0.619	0.585	0.602	9%
CM	0.460	0.450	0.455	44%
TTI [44]	0.216	0.153	0.185	255%
Wikipedia				
SCM	**0.362**	**0.273**	**0.318**	–
SM	0.350	0.249	0.300	6%
CM	0.267	0.219	0.243	31%
TTI [44]	0.237	0.137	0.187	70%

boosting, and SVMs. As shown in Table IV, the semantic space obtained with logistic regression performed best for both cross-modal retrieval tasks. It was thus chosen to implement SM in all remaining experiments.

4) Optimization: The experiments above lead to a retrieval architecture that combines KCCA for learning correlated subspaces, logistic regression to learn the semantic space, and the centered normalized correlation NC_c distance measure to evaluate (7) and (11). Using this architecture, a final round of experiments was used to determine the best combination of 1) BOW codebook size for image representation, 2) number of LDA topics for text representation, and 3) number of KCCA components, for each of the CM, SM, and SCM retrieval regimes and dataset. Table V summarizes the optimal parameter configuration, which was used in the remaining experiments.

C. *Testing the fundamental hypotheses*

This architecture was used on a set of experiments aimed to test the fundamental hypotheses of Section III. In these experiments, MAP scores were measured on the test set.

1) Overall performance: Table VI compares the scores of cross-modal retrieval with CM, SM, SCM and the baseline TTI method. The table provides evidence in support of the two hypotheses of Section III-C: both joint dimensionality reduction and semantic abstraction are beneficial for multi-modal modeling, leading to a non-trivial improvement over TTI. For example, in TVGraz, the average MAP score of CM is more than double that of TTI. For SM the improvement is more than threefold. Overall, the best performance is achieved by SCM. Similar conclusions can be drawn for Wikipedia, although the average gains of SCM are slightly lower than in TVGraz. This is not surprising, since the retrieval scores are generally lower on Wikipedia than on TVGraz. As discussed in Section V-A, this is explained by the broader scope of the Wikipedia categories.

Figure 8 presents a more detailed analysis of the retrieval performance, in the form of PR curves. CM, SM, and SCM again achieve large improvements over TTI. These improvements tend to occur at all levels of recall, indicating better generalization, and often involve substantial increases in precision, indicating higher accuracy. Overall, these results suggest that the contributions of cross-modal correlation and semantic abstraction are *complementary*: not only is there an independent benefit to both correlation modeling and abstraction, but the *best performance is achieved when the two are combined*.

2) Per-class performance: Figure 8 shows the per-class MAP scores of all methods. SCM has higher MAP than CM and SM on all classes of TVGraz, and is either comparable to, or better than CM and SM on the majority of Wikipedia classes. TTI does very poorly in general, and seems biased towards one class. This is evident from Figure 8 (c) and (f), where it achieves a very high score on one class – "Frog" on TVGraz and "Warfare" on Wikipedia – and very low scores in the remaining. In both cases, the favored class has a large number of training examples.

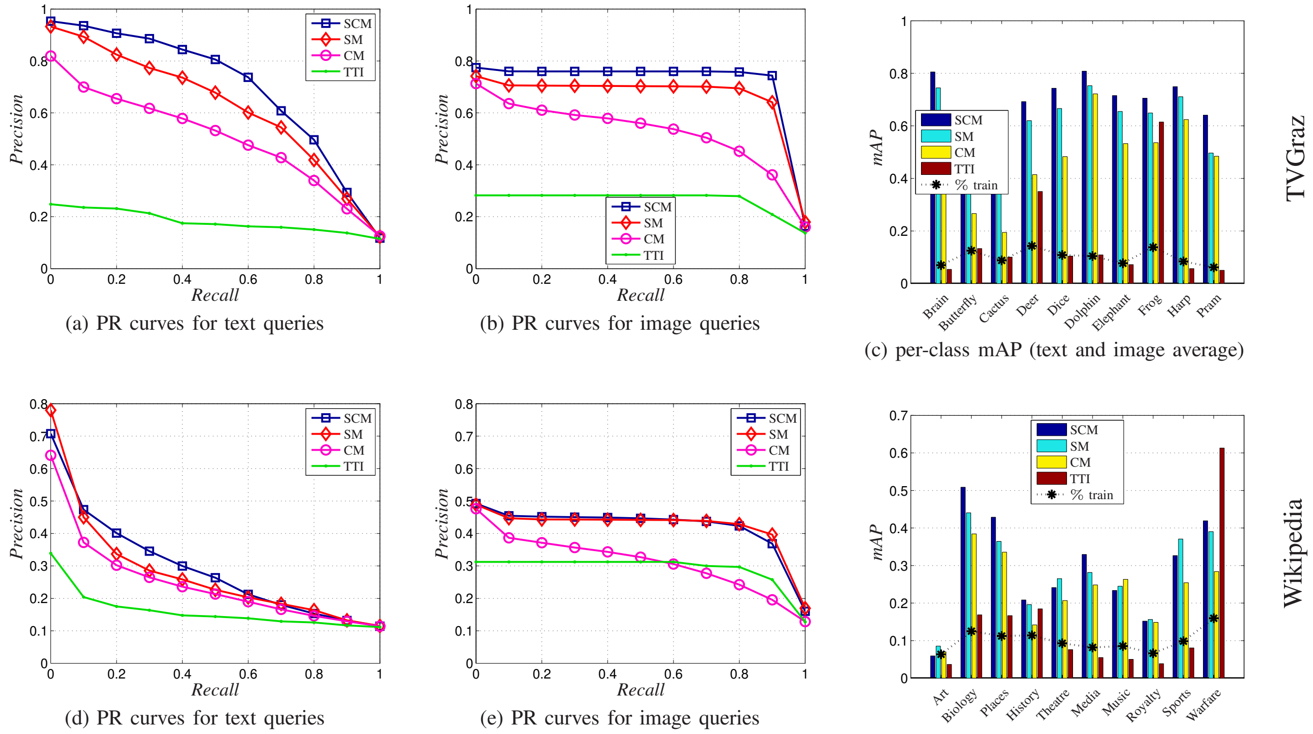
Two examples of text queries and corresponding retrieval results, using SCM, are shown in Figure 10. The text query is presented along with its probability vector π_T and the ground truth image. The top five image matches are shown below the text, along with their probability vectors π_I. Finally, Figure 11 shows some examples of image-to-text retrieval. Since displaying the retrieved texts would require too much space, we present the associated ground-truth images instead. The query images are framed in the left column, and the images associated with the four best text matches are shown on the right.

D. *Robustness*

The previous experiments indicate that semantic spaces are beneficial for cross-modal retrieval. However, in each experiment, the semantic space was designed with a vocabulary $\mathcal{V}$ identical to the ground truth semantics. This could be argued to give an unfair advantage to SM and SCM. To evaluate this possibility, we performed a number of additional experiments that evaluated the robustness of SM to mismatches between semantic vocabulary and ground-truth semantics. Two classes of experiments were performed.

1) Extended semantics: This set of experiments tested the impact of the size of the vocabulary $\mathcal{V}$ on SM performance. It was based on an *extended vocabulary* $\mathcal{V}'$, which was *shared by the two datasets*. This contained the 10 classes from TVGraz, the 10 classes of Wikipedia, the 20 classes of Wikipedia featured articles that were not used in the Wikipedia dataset, and 20 categories from the *Pascal-Sentences* [67] dataset (50 image/text pairs per class). Overall, $\mathcal{V}'$ contained 60 classes. The ground-truth semantics were as before, i.e. the classes in the second column of Tables VII and VIII.

To evaluate the impact of the composition of the semantic space on retrieval scores, we repeated the retrieval experiment using multiple subsets of $\mathcal{V}'$ as vocabulary $\mathcal{V}$. Starting with $\mathcal{V}$ containing the 10 ground-truth classes, we sequentially added one of the remaining classes in $\mathcal{V}'$ to $\mathcal{V}$. This produced a sequence of semantic spaces with between 11 and 60 dimensions. To introduce randomness, the whole experiment was repeated five times, using a sequence of randomly selected classes to add at each step. Figure 9



(a) PR curves for text queries (b) PR curves for image queries (c) per-class mAP (text and image average)

(d) PR curves for text queries (e) PR curves for image queries (f) per-class mAP (text and image average)

Fig. 8: PR curves of cross-modal retrieval using both text (a,d) and image (b,e) queries on TVGraz (top) and Wikipedia (bottom). Average (across image & text queries) per-class MAP scores also shown on (c) and (f).

TABLE VII: TVGraz semantics.

Alternative	Vocabulary
Anatomy	1. Brain
Pollination	2. Butterfly 3. Cactus
Land Animals	4. Deer 7. Elephant
Marine Animals	6. Dolphin 8. Frog
Objects	5. Dice 9. Harp 10. Pram

TABLE VIII: Wikipedia semantics.

Alternative	Vocabulary
Humanities	1. Art & architecture 3. Geography & places 4. History 5. Literature & theatre
Nature	2. Biology
Entertainment	6. Media 7. Music 9. Sport & recreation
Honor	8. Royalty & nobility 10. Warfare

TABLE IX: Average MAP scores (test set) under the original ("Vocabulary") and alternative semantics.

	Retrieval Evaluation	
	Alternative	Vocabulary
TVGraz		
SCM	**0.584**	**0.657**
SM	0.568	0.602
CM	0.492	0.455
TTI [44]	0.292	0.185
Wikipedia		
SCM	**0.448**	**0.318**
SM	0.436	0.300
CM	0.413	0.243
TTI [44]	0.347	0.187

presents the MAP scores as a function of the vocabulary size, for image and text queries on the two datasets. The straight horizontal lines are the scores obtained when $\mathcal{V}$ contained the 10 original classes. The image query task appears to be slightly more affected than its text counterpart, this is a natural consequence of the noisier semantic descriptor of images when compared to that of texts [68]. While there is some degradation of performance as the vocabulary grows, the effect is small. This indicates that the performance of SM is fairly insensitive to the size of the vocabulary $\mathcal{V}$.

2) Alternative semantics: In the previous experiments, the vocabulary $\mathcal{V}$ always included the ground-truth semantics. To further test the robustness of SM to the make-up of the semantic space, a final set of experiments was performed with ground-truth semantics that are only loosely related to the vocabulary $\mathcal{V}$. For this, we defined a new set of ground-truth semantics for each dataset, according to Tables VII and VIII. In all experiments, the vocabulary $\mathcal{V}$ consisted of the original dataset classes, also shown in the tables.

Table IX presents a comparison of the average MAP scores achieved with the alternative ground-truth semantics of Tables VII-VIII (denoted "alt. semantics") and with the original dataset classes (denoted "vocabulary"). Since there are fewer classes in the alternative semantics, the retrieval

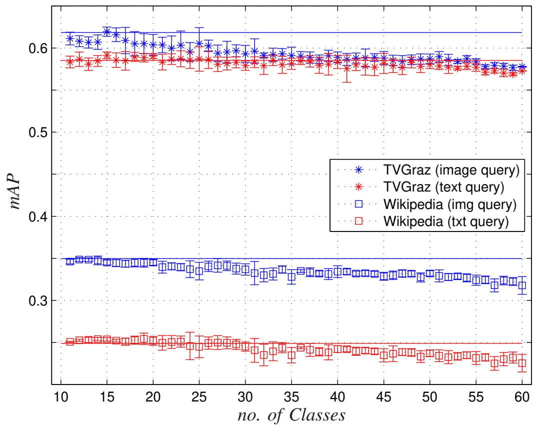


Fig. 9: MAP scores under SM. The solid horizontal line is the score obtained with the 10 original dataset categories.

performance is expected to improve. However, the fact that these classes are more abstract could also lead to a degradation. The two behaviors are visible in the table. On Wikipedia, where the original classes are already quite abstract, all methods have improved performance under the alternative semantics. On TVGraz, where the alternative semantics are much more abstract than the vocabulary classes, performance decreases for SM and SCM. Note, however, that these variations do not affect the relative performance of the different methods. In both cases, CM and SM achieve significant improvements over TTI and the best overall performance is obtained when they are combined (SCM). In summary, this experiment confirms all the conclusions reached above.

VI. Conclusion

The increasing availability of multi-modal information demands novel representations for content-based retrieval. In this work, we proposed models applicable to cross-modal retrieval. This entails the retrieval of database entries from one content modality in response to queries from another. While the emphasis was on cross-modal retrieval of images and rich text, the proposed models support many other content modalities. By requiring representations that can generalize across modalities, cross-modal retrieval establishes a suitable context for the objective investigation of fundamental hypotheses in multimedia modeling.

We have considered two such hypotheses, regarding the importance of low-level cross-modal correlations and semantic abstraction in multi-modal modeling. The hypotheses were objectively tested by comparing the performance of three methods: 1) CM, based on the correlation hypothesis, 2) SM, based on the abstraction hypothesis, and 3) SCM, based on the combination of the two. All of these map objects from different native spaces (*e.g.*, rich text and images) to a pair of isomorphic spaces, where a natural correspondence can be established for cross-modal retrieval purposes. The retrieval performance of the three solutions was tested on two datasets, "Wikipedia" and "TVGraz", which combine images and rich text, and compared to a *state-of-the-art* cross-modal retrieval method (TTI).

While the two fundamental hypotheses were shown to hold for the two datastets, where both CM and SM achieved significant improvements over TTI, SM achieved overall better performance than CM. This implies stronger evidence for the abstraction than for the correlation hypothesis. However, the two hypotheses were also found to be complementary, with SCM achieving the best results of all methods considered.

Acknowledgments

This work was funded by FCT graduate Fellowship SFRH/BD/40963/2007 and NSF grant CCF-0830535. The authors thank Malcolm Slaney for helpful discussions.

References

[1] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE International Conference on Multimedia and Expo*, 2001, pp. 745–748.

[4] S. Sclaroff, M. Cascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Journal of Computer Vision and Image Understanding, Elsevier*, vol. 75, no. 1, pp. 86–98, 1999.

[5] C. Frankel, M. Swain, and V. Athitsos, "Webseer: An image search engine for the world wide web," University of Chicago, Computer Science Department, Tech. Rep., 1996.

[6] W. Li, K. Candan, and K. Hirata, "SEMCOG: an integration of semantics and cognition-based approaches for image retrieval," in *ACM Symposium on Applied Computing*, 1997, pp. 136–143.

[7] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 408–415.

[8] L. Denoyer and P. Gallinari, "Bayesian network model for semi-structured document classification," , *Elsevier*, vol. 40, no. 5, pp. 807–827, 2004.

[9] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Journal of Multimedia Tools and Applications, Springer*, vol. 25, no. 1, pp. 5–35, 2005.

[10] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[11] J. Iria, F. Ciravegna, and J. Magalhães, "Web news categorization using a cross-media document graph," in *Proc. ACM International Conference on Image and Video Retrieval*, 2009, pp. 1–8.

[12] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Workshop on Multimedia Information Retrieval at Proc. ACM International Conference on Multimedia*, 2006, pp. 321–330.

[13] T. Tsikrika and J. Kludas, "Overview of the wikipedia multimedia task at ImageCLEF 2008," *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 539–550, 2009.

[14] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *Journal of Machine Learning Research, MIT Press*, vol. 3, pp. 1107–1135, 2003.

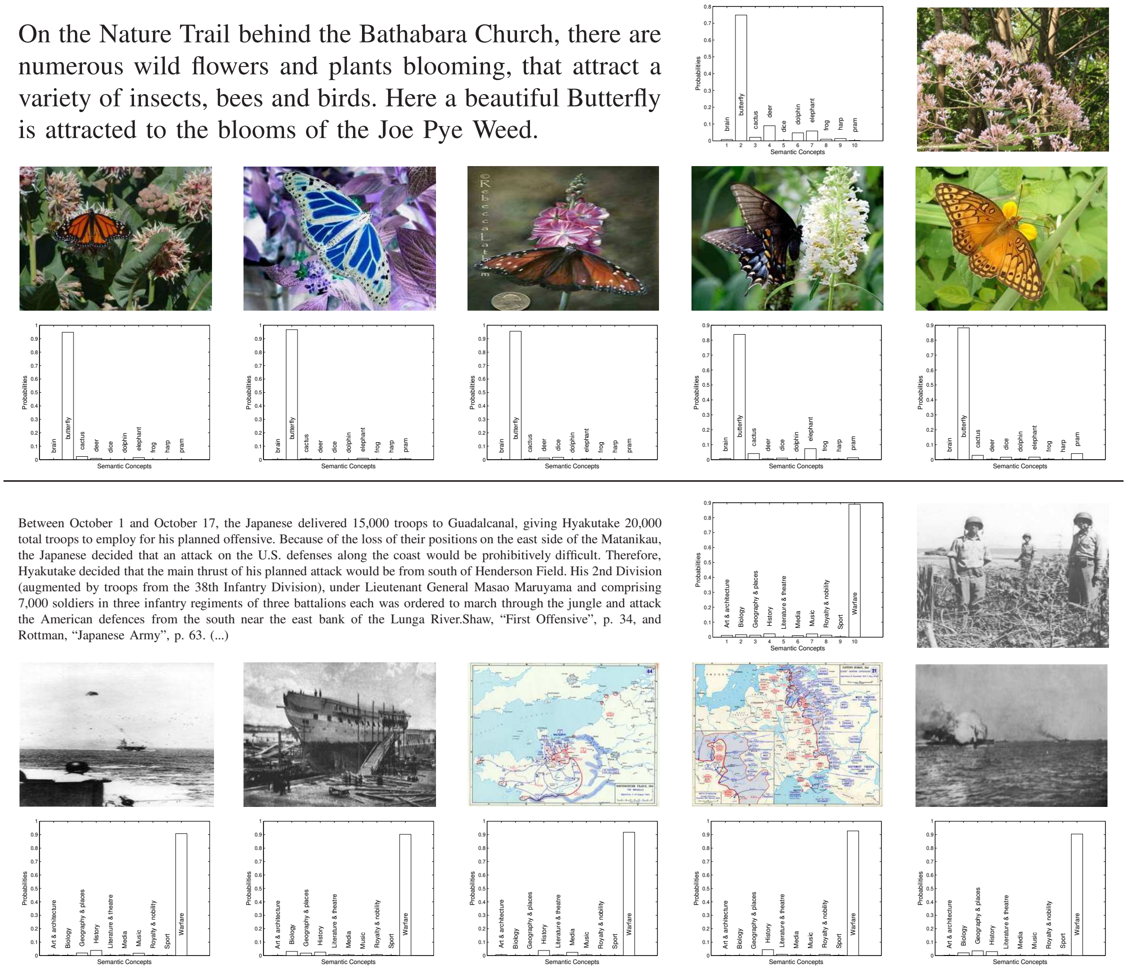


Fig. 10: Two examples of text-based cross-modal retrieval using SCM. The first example is from TVGraz and the second example from Wikipedia. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

[15] Y. Mori, H. Takahashi, and R. Oka, "Automatic word assignment to images based on image division and vector quantization," in *Recherche d'Information Assistée par Ordinateur*, 2000.

[16] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[17] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[18] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[19] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," in *Advances in Neural Information Processing Systems*, 2008, vol. 20, pp. 385–392.

[20] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Journal of Neural Computation, MIT Press*, vol. 16, no. 12, pp. 2639–2664, 2004.

[21] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research, MIT Press*, vol. 3, pp. 993–1022, 2003.

[22] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[23] I. Khan, A. Saffari, and 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.

[24] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[25] A. Vinokourov, D. Hardoon, and J. Shawe-Taylor, "Learning the semantics of multimedia content with application to web image retrieval and classification," in *Symposium on Independent Component Analysis and Blind Source Separation*, 2003.

[26] 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 *Proc. ACM International Conference on Multimedia*, 2010, pp. 251–260.

[27] M. Paramita, M. Sanderson, and P. Clough, "Diversity in photo retrieval: Overview of the ImageCLEF 2009 photo task," *Multilingual Information Access Evaluation: Multimedia Experiments*, pp. 45–59, 2010.

[28] C. Meadow, B. Boyce, D. Kraft, and C. Barry, *Text Information Retrieval Systems*. Emerald Group Pub Ltd, 2007.

[29] G. Salton, *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

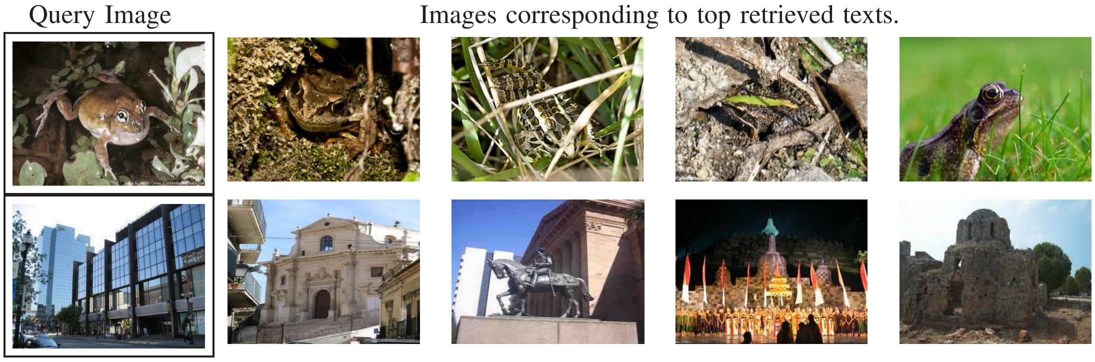


Fig. 11: Image-to-text retrieval on TVGraz (top row) and Wikipedia (bottom). Query images are framed in the far-left column. The four most relevant texts, represented by their ground truth images, are shown in the remaining columns.

[30] N. Vasconcelos, “Minimum probability of error image retrieval,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004.

[31] F. Monay and D. Gatica-Perez, “Modeling semantic aspects for cross-media image indexing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, 2007.

[32] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *Proc. ACM SIG International Conference Information Retrieval*, 2003, pp. 119–126.

[33] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *Advances in Neural Information Processing Systems*, 2004, vol. 16.

[34] S. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, vol. 2, 2004, pp. 1002–1009.

[35] J. Z. Wang and J. Li, “Learning-based linguistic indexing of pictures with 2-D MHMMs,” in *Proc. ACM International Conference on Multimedia*, 2002, pp. 436–445.

[36] N. Vasconcelos, “From pixels to semantic spaces: Advances in content-based image retrieval,” *IEEE Transactions on Computers*, vol. 40, no. 7, pp. 20–26, 2007.

[37] T. Westerveld, “Image retrieval: Content versus context,” in *Content-Based Multimedia Information Access at Recherche d’Information Assistée par Ordinateur*, 2000, pp. 276–284.

[38] T. Pham, N. Maillot, J. Lim, and J. Chevallet, “Latent semantic fusion model for image retrieval and annotation,” in *Proc. ACM International Conference on Information and Knowledge Management*, 2007, pp. 439–444.

[39] H. Escalante, C. Hérnadez, L. Sucar, and M. Montes, “Late fusion of heterogeneous methods for multimedia image retrieval,” in *Proc. ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 172–179.

[40] G. Wang, D. Hoiem, and D. Forsyth, “Building text features for object image classification,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2009, pp. 1367–1374.

[41] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo, “Combining image captions and visual analysis for image concept classification,” in *Workshop on Neural Networks for Signal Processing at Proc. ACM SIG International Conference Knowledge Discovery and Data Mining*, 2008, pp. 8–17.

[42] S. Nakamura, “Statistical multimodal integration for audio-visual speech processing,” *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 854–866, 2002.

[43] J. Fisher III, T. Darrell, W. Freeman, and P. Viola, “Learning joint statistical models for audio-visual fusion and segregation,” in *Advances in Neural Information Processing Systems*, 2001, pp. 772–778.

[44] G. Qi, C. Aggarwal, and 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.

[45] D. Li, N. Dimitrova, M. Li, and I. Sethi, “Multimedia content processing through cross-modal association,” in *Proc. ACM International Conference on Multimedia*, 2003, pp. 604–611.

[46] H. Zhang, Y. Zhuang, and F. Wu, “Cross-modal correlation learning for clustering on image-audio dataset,” in *Proc. ACM International Conference on Multimedia*, 2007, pp. 273–276.

[47] M. Slaney, “Semantic-audio retrieval,” in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 4, 2002, pp. 4108–4111.

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

[49] Y. Zhuang, Y. Yang, and F. Wu, “Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval,” *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, 2008.

[50] Y. Zhuang, Y. Yang, F. Wu, and Y. Pan, “Manifold learning based cross-media retrieval: a solution to media object complementary nature,” *Journal of VLSI Signal Processing, Springer*, vol. 46, no. 2, pp. 153–164, 2007.

[51] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, “Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval,” *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, 2008.

[52] V. Mahadevan, C. Wah Wong, J. Costa Pereira, T. T. Liu, N. Vasconcelos, and L. K. Saul, “Maximum covariance unfolding: Manifold learning for bimodal data,” in *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 918–926.

[53] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini, “Inferring a semantic representation of text via cross-language correlation analysis,” in *Advances in Neural Information Processing Systems*, 2003, vol. 15, pp. 1473–1480.

[54] W. Hsu, T. Mei, and R. Yan, “Knowledge discovery over community-sharing media: From signal to intelligence,” in *Proc. IEEE International Conference on Multimedia and Expo*, 2009, pp. 1448–1451.

[55] T. Mei, W. Hsu, and J. Luo, “Knowledge discovery from community-contributed multimedia,” *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 16–17, 2010.

[56] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Jour-*

nal of the American Society for Information Science, vol. 41, no. 6, pp. 391–407, 1990.

[57] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2005.

[58] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[59] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research, MIT Press*, vol. 9, pp. 1871–1874, 2008.

[60] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[61] M. J. Saberian and N. Vasconcelos, "Multiclass boosting: Theory and algorithms," in *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 2124–2132.

[62] G. Griffin, A. Holub, and P. Perona, "The Caltech-256," Caltech, Tech. Rep., 2006.

[63] C. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge University Press, 2008.

[64] G. Doyle and C. Elkan, "Accounting for word burstiness in topic models," in *Proc. ACM International Conference on Machine Learning*, 2009, pp. 281–288.

[65] M. Swain and D. Ballard, "Color indexing," *International Journal on Computer Vision, Springer*, vol. 7, no. 1, pp. 11–32, 1991.

[66] S. Boughorbel, J. Tarel, and N. Boujemaa, "Generalized histogram intersection kernel for image recognition," in *Proc. IEEE International Conference on Image Processing*, vol. 3, 2005, pp. III – 161–164.

[67] 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 Amazon's Mechanical Turk*. NAACL HLT, 2010, pp. 139–147.

[68] J. Costa Pereira and 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.

Jose Costa Pereira received the Licenciatura in computer science and engineering from the Universidade do Porto, Portugal, in 2000, and the MS in computational methods in 2003 from the same university. He is a PhD candidate at the University of California San Diego, Electrical and Computer Engineering Department, in the Statistical Visual Computing Laboratory. From 2008 to 2012 he was the recipient of a PhD Fellowship from the Ministry of Sciences and Education, Portugal. His research interests are in computer vision, multimedia and machine learning. He is a student member of IEEE.

Emanuele Coviello received the Laurea Triennale in information engineering from the University of Padova, Italy, in 2006, and the Laurea Specialistica in telecommunications engineering in 2008 from the same university. He his working towards his PhD in Electrical and Computer Engineering at UC San Diego, in the Computer Audition Laboratory. His main research interests are in machine learning applied to (music) information retrieval and multimedia data modeling.

Gabriel Doyle received the AB degree in mathematics from Princeton University in 2005 and received the MA degree in linguistics from the University of California, San Diego in 2011. He currently is working toward the PhD degree, also in linguistics, at UCSD. His primary research area is computational psycholinguistics, especially models of language acquisition that integrate multiple sources of information.

Nikhil Rasiwasia received his B.Tech degree in electrical engineering from Indian Institute of Technology Kanpur (India) in 2005, and the MS and PhD degrees from the University of California, San Diego in 2007 and 2011 respectively. He is currently a scientist at Yahoo Labs! Bangalore. His research interests are in the areas of computer vision and machine learning. He was recognized as an 'Emerging Leader in Multimedia' in 2008 by IBM T. J. Watson Research. He was also awarded with the best student paper award at ACM Multimedia 2010.

Roger Levy is Associate Professor in the Department of Linguistics at UC San Diego, where he heads the Computational Psycholinguistics Laboratory. He received the BS degree in Mathematics at the University of Arizona in 1996, the MS degree in Anthropological Sciences from Stanford University in 2002, and his PhD in Linguistics from Stanford University in 2005. In 2005–2006 he was a postdoctoral fellow in Informatics at the University of Edinburgh and joined UC San Diego in 2006. His research focuses on theoretical and applied questions in the processing of natural language. He is the recipient of a Hellman Fellowship, an NSF Career award, and an Alfred P. Sloan Fellowship, and is presently an Associate Editor for the journal Cognitive Science. In 2013–2014 he will be a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University.

Gert R.G. Lanckriet received the MS degree in electrical engineering from the Katholieke Universiteit Leuven, Belgium, in 2000 and the MS and PhD degrees in electrical engineering and computer science from the University of California, Berkeley, in 2001 and 2005, respectively. In 2005, he joined the Department of Electrical and Computer Engineering, University of California, San Diego, where he heads the Computer Audition Laboratory. His research focuses on the interplay of convex optimization, machine learning, and signal processing, with applications in computer audition, music information retrieval, and personalized health. He was awarded the SIAM Optimization Prize in 2008 and is the recipient of a Hellman Fellowship, an IBM Faculty Award, an NSF CAREER Award, and an Alfred P. Sloan Foundation Research Fellowship. In 2011, MIT Technology Review named him one of the 35 top young technology innovators in the world (TR35). He is a senior member of the IEEE.

Nuno Vasconcelos received the Licenciatura in electrical engineering and computer science from the Universidade do Porto, Portugal, in 1988, and the MS and PhD degrees from the Massachusetts Institute of Technology in 1993 and 2000, respectively. From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which in 2002 became the HP Cambridge Research Laboratory. In 2003, he joined the Electrical and Computer Engineering Department at the University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. He is the recipient of a US National Science Foundation (NSF) CAREER award, a Hellman Fellowship, and has authored more than 150 peer-reviewed publications. His work spans various areas, including computer vision, machine learning, signal processing and compression, and multimedia systems. He is a senior member of the IEEE.