On the roles of correlation and abstraction in cross-modal multimedia retrieval

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

[†]These authors contributed equally to this work
[1]Dept. of Electrical and Computer Engineering,
[2]Dept. of Linguistics
University of California, San Diego
{ecoviell,josecp,gdoyle,nikux,rlevy}@ucsd.edu,{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 text to search for images. A mathematical formulation is proposed, equating the design of cross-modal retrieval systems to that of isomorphic feature spaces for different content modalities. Two hypothesis 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), which models cross-modal correlations, semantic matching (SM), which 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, multimodal, cross-modal, image and text, retrieval model, semantic spaces, kernel correlation, logistic regression**

I. Introduction

Classical approaches to information retrieval are of a *unimodal* nature [25], [36], [40]. 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. Recently, there has been a surge of interest in *multimodal* modeling, representation, and retrieval [6], [8], [18], [32], [39], [42], [44]. Multimodal retrieval relies on queries combining multiple content modalities (e.g. the images and sound of a music video-clip) to retrieve database entries with the same combination of modalities (e.g. other music video-clips). These efforts have, in part, been spurred by a variety of large-scale research and evaluation experiments, such as TRECVID [39] and ImageCLEF [32], [44], involving datasets that span multiple data modalities. However, much of this work has focused on the straightforward extension of methods shown successful in the unimodal 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 unimodal techniques are then applied to the low-dimensional representation. This limits the applicability of the resulting multimedia models and retrieval systems.

In this work, we consider a richer interaction paradigm, which is denoted *cross-modal* retrieval. The goal is to build multimodal 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., the best accompanying audio-track). This form of retrieval can be seen as a generalization of current content labeling systems, where one dominant modality is augmented with simple information from another, which can be subsequently searched. Examples include keyword-based image [1], [4], [30] and song [5] retrieval systems. One property of cross-modal retrieval is that, by definition, it requires *representations that 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 much deeper content understanding than the classical matching of unimodal 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 *abstract* 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 unimodal retrieval for the study of fundamental hypotheses on multimedia modeling.

We exploit this property 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 of importance for the success of the joint models. The second, denoted the *abstraction hypothesis*, is that the modeling 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 [16], and abstraction, e.g., hierarchical topic models for text clustering [2] or hierarchical semantic representations for image retrieval [35], improve performance on unimodal 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., to 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 [21] and a novel dataset based on Wikipedia's featured articles. These experiments show independent benefits to both correlation modeling and abstraction. In particular, best results are obtained by a model that accounts for both low-level correlations — by performing a kernel canonical correlation analysis (KCCA) [37], [47] — and semantic abstraction — by projection of images and texts into a common semantic space [35] designed with logistic regression. This suggests that the abstraction and correlation hypotheses are complementary, each improving the modeling in a different manner. Individually, the gains of abstraction are larger than those of correlation modeling.

The paper is organized as follows. Section II discusses previous work in multimodal 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 discusses the experimental setup used to evaluate the hypotheses. Model validation and parameter tunning are detailed in Section VI. The hypotheses are finally tested on Section VII and conclusions presented in Section VIII. A preliminary version of this work appeared in [34].

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 [6], [28], [32], [39], [40]. In all these areas, the emphasis has been on *unimodal* approaches, where query and retrieved documents share a single modality [6], [40], [45]. This is not effective for all problems. For example, the existence of a well known *semantic gap*, between current image representations and those adopted by humans, severely limits the performance of unimodal image retrieval systems [40]. 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 [32], [39], [44]. When this metadata is available, the retrieval operation tends to be unimodal and ignore the images — a text query is simply matched to the available text metadata. Because manual image labeling is labor-intensive, recent research has addressed the problem of automatic image labeling [1], [4], [12], [19], [23], [29]. Although not commonly perceived as *multi-modal*, these systems support cross-modal retrieval, by returning images in response to text queries. However, the ability to bridge the gap between the two modalities is limited, since all queries are restricted to the keywords in the concept vocabulary used to train the image labeling system. These vocabularies are usually small, rarely containing more than a few hundred words.

A solution to this problem is to design a *semantic space*, where each dimension is a semantic concept [35], [41]. Statistical models of the distribution of low-level image features are first learned for each of the concepts in the vocabulary. The probability of the features extracted from each image, under each of the concept models, is then computed. Bayes rule is finally used to compute the posterior probabilities of the image under each concept, and the image is represented by the vector of these *posterior concept probabilities*. As illustrated in Figure 1, this can be seen as a *semantic image descriptor,* which maps the image into a semantic feature space. This descriptor is commonly denoted as a *semantic multinomial* (SMN) distribution. All standard image analysis/classification tasks can then be conducted in semantic space, at a higher level of abstraction than that supported by low-level feature spaces. For example, image retrieval can be formulated as retrieval by *semantic similarity,* by combining the semantic space with a suitable similarity function [35]. 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 [46].

In parallel, advances have been reported in the area of *multi-modal* retrieval systems [6], [8], [18], [32], [39], [42], [44]. These are extensions of the classic unimodal 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 [10], [33], [50], or by learning different models for different modalities and fusing their predictions [22], [49]. One popular approach is to 1) concatenate features from different modalities into a common vector and 2) rely on unsupervised structure discovery algorithms, such as latent semantic analysis (LSA), to find statistical patterns that span the different modalities. A good overview of

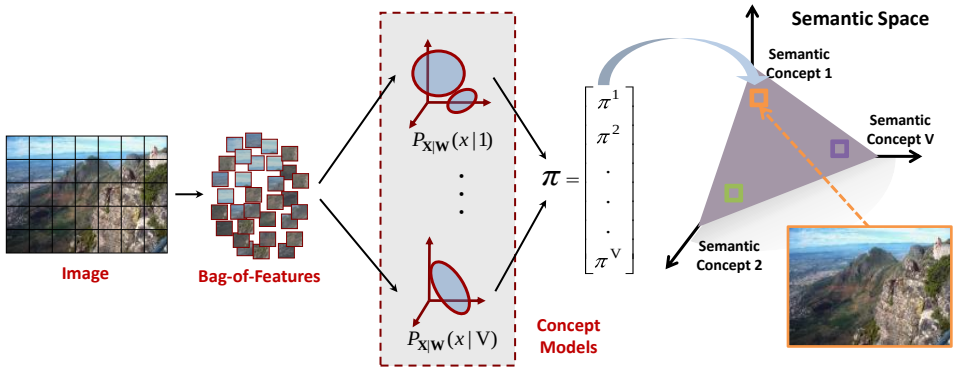


Fig. 1. Image representation in semantic space. Images are decomposed into bags-of-features. A vocabulary of visual concepts establishes the dimensions of the semantic space. A model of the feature distribution is learned for each concept. Each image is finally represented by the vector of posterior concept probabilities, given its features.

these methods is given in [10], which also discusses the combination of unimodal and multimodal retrieval systems. Multimodal integration has also been applied to retrieval tasks including audio-visual content [13], [31]. In general, the inability to access each data modality individually (after the fusion of modalities) limits the applicability of these systems to cross-modal retrieval.

Recently, there has been progress towards multimodal systems that do not suffer from this limitation. These include retrieval methods for corpora of images and text [8], images and audio [24], [53], text and audio [38], or images, text, and audio [51]–[55]. One popular approach is to rely on graph-based manifold learning techniques [51]–[55]. These methods learn a manifold from a matrix of distances between multimodal objects. Retrieval then consists of finding the nearest document, on the manifold, to a multimedia query. The main limitation of methods in this class 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 solution is to learn correlations between different modalities [24], [48], [53]. For example, [24] 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 correlated features in the two data modalities. A kernelized version of CCA was also proposed in [48] 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 [16].

Despite these advances in multi-modal modeling, current approaches tend to rely on a limited textual representation, in the form of keywords, captions, or small text snippets. 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 that of *light annotation*, weakening the one-to-one mapping between textual words and class labels. For example, Figure 2 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 3, from the TVGraz dataset [21]. 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

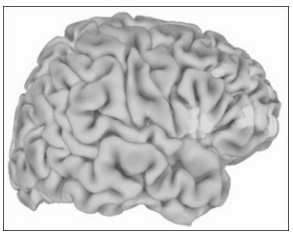
In this section, we present a novel multi-modal content modeling framework, which is flexible and applicable to rich content modalities. Although the fundamental ideas are applicable to any combination of modalities we restrict the discussion to documents containing images and text.

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. In practice, these documents can be quite diverse: from documents where a single text is 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 on feature spaces $\Re^I$ and $\Re^T$, respectively, as illustrated in Figure 4. In this way, documents establish a one-to-one mapping between $\Re^I$ 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 $\Re^I$ ($\Re^T$).

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. Reverend Wyatt Tee Walker, one of the SCLC founders and the executive director from 19601964, planned the tactics of the direct action protests, specifically targeting Bull Connor's tendency to react to demonstrations with violence. "My theory was that if we mounted a strong nonviolent movement, the opposition would surely do something to attract the media, and in turn induce national sympathy and attention to the everyday segregated circumstance of a person living in the Deep South," Walker said. He headed the planning of what he called Project C, which stood for "confrontation". (...)

Fig. 2. A section from the Wikipedia article on the Birmingham campaign, belonging to the "History" category.



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. Knowledge of how the human brain works is important for all citizens, and the lessons to be learned have enormous implications for public policy makers and educators. The course will cover the regional anatomy of the brain and provide an introduction to the cellular function of neurons, synapses and neurotransmitters. Commonly used drugs that alter brain function can be understood through a knowledge of neurotransmitters. Along similar lines, common diseases that illustrate normal brain function will be discussed. (...)

Fig. 3. Part of a document-image pair from the TVGraz dataset, in the "brain" category.

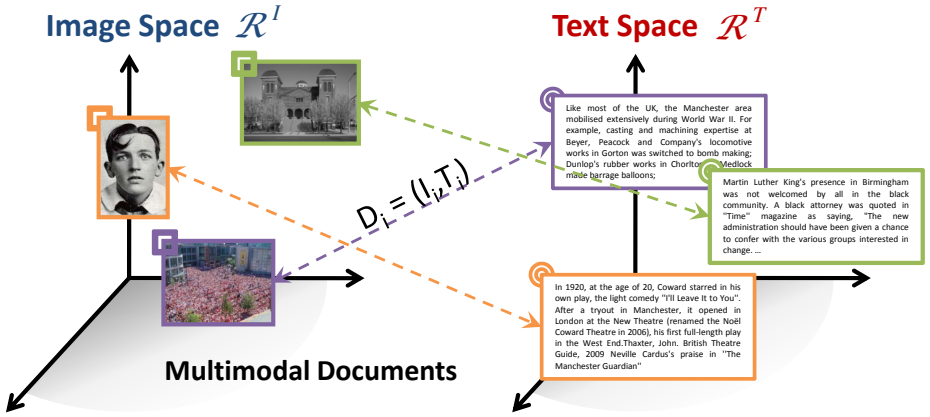


Fig. 4. Each document consists of an *image* and accompanying *text*, i.e. $D_i = (I_i, T_i)$, which are represented as vectors on feature spaces $\Re^I$ and $\Re^T$, respectively. Documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$.

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. Let

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

be invertible mappings from each of the image and text spaces to two isomorphic spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ such that there is an invertible mapping

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

Given a query T_q in $\Re^T$, cross-modal retrieval reduces to finding the nearest neighbor of

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

in $\Re^I$. Similarly, given a query I_q in $\Re^I$, the goal is to find the nearest neighbor of

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

in $\Re^T$. 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 $\mathcal{V}^T$.

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. While 1) a unimodal retrieval system can successfully retrieve images of "swans" because they are the only white objects in the database, 2) a text retrieval system can successfully retrieve documents about "swans" because they are the only containing the "swan" word, and 3) a multimodal retrieval system can just match "white" to "white" and "swan" to "swan", a cross-modal retrieval system cannot solve the task without *abstracting* 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 unimodal or multimodal retrieval. In this work, we exploit the cross-model retrieval problem to objectively 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 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 when important hypotheses are met, the resulting models are more effective, and cross-modal retrieval performance improves.

TABLE I
TAXONOMY OF THE PROPOSED APPROACHES TO CROSS-MODAL RETRIEVAL.

	correlation hypothesis	abstraction hypothesis
CM	√	
SM		√
SCM	√	√

IV. CROSS-MODAL RETRIEVAL

In this section, we present each of 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) [7] and principal component analysis (PCA) [20]. 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 multimodal setting and kernelized. This has motivated the introduction of a number of multimodal subspace methods in the literature. 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 5 illustrates this process. Cross-modal retrieval is finally performed within the low-dimensional subspaces.

1) Linear subspace learning: CFA [24] finds the orthonormal transformations Ω_I and Ω_T that project the two modalities onto the shared space, $\mathcal{U}^I = \mathcal{U}^T = \mathcal{U}$, where the projections have minimum distance

$$\left\| X_I\Omega_I - X_T\Omega_T \right\|_F^2. \quad (1)$$

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

$$trace(X_I' X_T) \quad (2)$$

and the optimal matrices Ω_I, Ω_T can be obtained by a singular value decomposition of the matrix $X_I' X_T$ [24],

$$X_I' X_T = \Omega_I \Lambda \Omega_T \quad (3)$$

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

CCA [17] 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

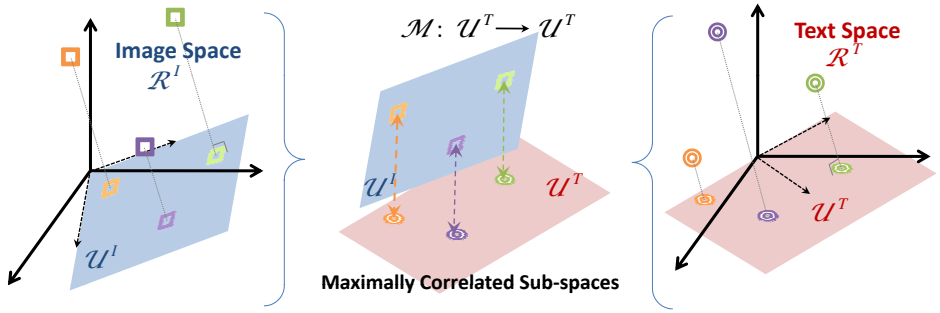


Fig. 5. Correlation matching (CM) performs joint feature selection in the text and image spaces, projecting them onto two maximally correlated subspaces $\mathcal{U}_T$ and $\mathcal{U}_I$.

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. The canonical components in the image space are the eigenvectors of $\Sigma_I^{-1/2}\Sigma_{IT}\Sigma_T^{-1}\Sigma_{TI}\Sigma_I^{-1/2}$, and in the text space 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 $\mathcal{K}_I(\cdot,\cdot)$, whose ranges are inner product spaces such that $\mathcal{K}_T(T_m, T_n) = \langle \phi(T_m), \phi(T_n) \rangle$ respectively $\mathcal{K}_I(I_m, I_n) = \langle \phi(I_m), \phi(I_n) \rangle$.

KCCA [37], [47] 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, i.e. $w_i = \Phi_I(X_I)^T \alpha_i$ and $w_t = \Phi_T(X_T)^T \alpha_t$. This is shown in equation (5). Where $\kappa \in [0, 1]$ is a regularization factor, K_I and K_T are the kernel matrices of the two representations, e.g., $(K_I)_{mn} = \mathcal{K}_I(I_m, I_n)$, and $\Phi_T(X_T)$ ($\Phi_I(X_I)$) is the matrix whose rows contain the high-dimensional image of the text (image) features. In our implementation, the optimal (α_i, α_t) are learned with the software package of [47]. The first d solutions, $\{\alpha_{i,k}\}_{k=1}^d$ and $\{\alpha_{t,k}\}_{k=1}^d$, are weight vectors for the linear combination of the training examples $\{\phi_I(I_k)\}_{k=1}^{|\mathcal{D}|}$, and $\{\phi_T(T_k)\}_{k=1}^{|\mathcal{D}|}$, so as to form 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 $\mathcal{U}^T \subset \mathcal{F}^T$, where $1 \leq d \leq |\mathcal{D}|$.

3) Image and text projections: Images and text are represented by their projections into subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$. For CFA and CCA, these are dot-products p_I between image features x_I and image basis vectors, and p_T between text features x_T and text basis vectors. For CFA, basis vectors are the columns of Ω_I, Ω_T, for CCA they are the eigenvectors $\{w_{i,k}\}_{k=1}^d, \{w_{t,k}\}_{k=1}^d$. In the case of KCCA, an image is mapped to its projection $p_I = \mathcal{P}_I(\phi_I(I))$ onto $\{w_{i,k}\}_{k=1}^d$ with

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

where $k = 1, \ldots, d$. Analogously, a text $T \in \Re^T$ is mapped into its projection $p_T = \mathcal{P}_T(\phi_T(T))$ onto $\{w_{t,k}\}_{k=1}^d$, 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_T and p_I are coordinates in two isometric d dimensional subspaces, as shown in Figure 5. 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 of 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 7(a).

B. Semantic matching (SM)

An alternative to subspace learning is to represent documents at a higher level of abstraction, so that there is a natural correspondence between the text and image spaces. This is obtained by augmenting the database $\mathcal{D}$ with a vocabulary $\mathcal{V} = \{v_1, \ldots, v_K\}$ of semantic concepts, such as "History" or "Biology". Individual documents are grouped into these classes. 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 of posterior probabilities $P_{V|T}(v_j|T), j \in \{1, \ldots, K\}$ with

$$\max_{\alpha_i \neq 0,\, \alpha_t \neq 0} \frac{\alpha_i' K_I K_T \alpha_t}{\sqrt{(1-\kappa)\alpha_i' K_I^2 \alpha_i + \kappa \alpha_i' K_I \alpha_i}\sqrt{(1-\kappa)\alpha_t' K_T^2 \alpha_t + \kappa \alpha_t' K_T \alpha_t}}, \tag{5}$$

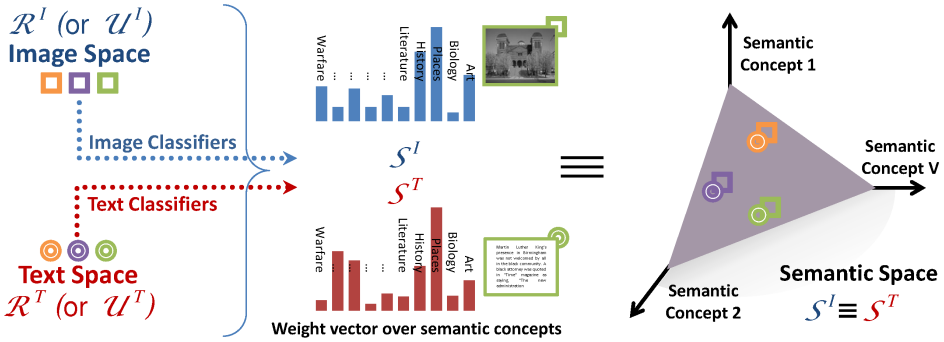


Fig. 6. Semantic matching (SM) maps text and image features onto a semantic space. Classifiers for the text (image) modality are used to represent text (image) as semantic text (image) descriptors, i.e., weight vectors over semantic concepts.

respect to each of the classes in $\mathcal{V}$. The space $\mathcal{S}^T$ of these vectors is referred to as the *semantic space for text*, and the probabilities $P_{V|T}(v_j|T)$ as *semantic text features*. Similarly, $\mathcal{L}_I$ maps image I into a vector of *semantic image features* $P_{V|I}(v_j|I), j \in \{1, \ldots, K\}$ in a semantic image space $\mathcal{S}^I$.

Semantic models have two advantages for cross-modal retrieval. First, they provide a higher level of abstraction. While standard features in $\Re^T$ and $\Re^I$ are the result of unsupervised learning, and frequently have no obvious interpretation (e.g. image features tend to be edges, edge orientations or frequency bases), the features in $\mathcal{S}^I$ and $\mathcal{S}^T$ are semantic concept probabilities (e.g. the probability that the image belongs to the "History" or "Biology" document classes). Previous work has shown that this increased semantic abstraction can lead to substantially better generalization for tasks such as image retrieval [35]. Second, the semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$ are isomorphic: in both cases, images and text are represented as vectors of posterior probabilities with respect to the *same* document classes. Hence, the spaces can be treated the same, i.e. $\mathcal{S}^T = \mathcal{S}^I$, leading to the schematic representation of Figure 6.

In this work, the posterior probability distributions are computed through multi-class logistic regression. This produces a linear classifier with a probabilistic interpretation. Logistic regression computes the posterior probability of class j, by fitting image or text features to a logistic function,

$$P_{V|X}(j|x; w) = \frac{1}{Z(x, w)} \exp(w_j' x) \tag{8}$$

where $Z(x, w) = \sum_j \exp(w_j' x)$ is a normalization constant, V the class label, X the vector of features in the input space, and $w = \{w_1, \ldots, w_K\}$, with w_j a vector of parameters for class j. A multi-class logistic regression is learned for the text and image modalities, by making X the image and text representation $I \in \Re^I$ and $T \in \Re^T$ respectively. In our implementation we use the software package Liblinear [11]. Given a query image I (text T), represented by a probability vector $\pi_I \in \mathcal{S}^I$ ($\pi_T \in \mathcal{S}^T$), retrieval consists of finding the text T (image I), represented by a probability vector $\pi_T \in \mathcal{S}^T$ ($\pi_I \in \mathcal{S}^I$), that minimizes

$$D(I, T) = d(\pi_I, \pi_T), \tag{9}$$

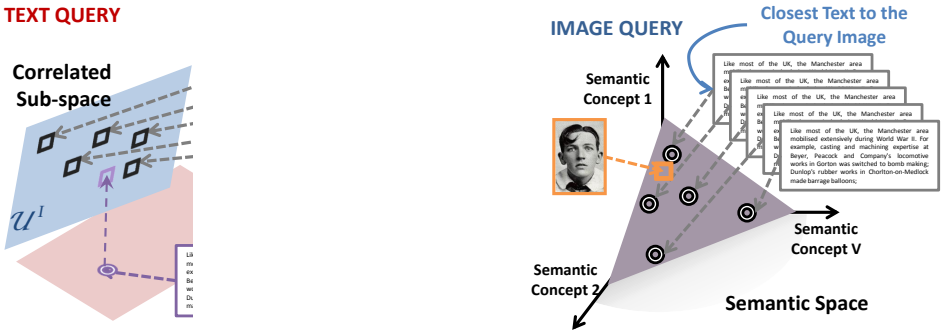
for some suitable distance measure d between probability distributions. An illustration of cross-modal retrieval using SM is given in Figure 7(b).

C. Semantic Correlation Matching (SCM)

Although CM and SM operate on different principles, they are not mutually exclusive. In fact, a corollary to the two hypotheses discussed above is that there may 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 $\mathcal{U}^I \subset \mathcal{F}^I$ and $\mathcal{U}^T \subset \mathcal{F}^T$ are first learned with correlation modeling. Logistic regressors $\mathcal{L}_I$ 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 (9), based on the semantic mappings $\pi_I = \mathcal{L}_I(\mathcal{P}_I(\phi_I(I)))$ and $\pi_T = \mathcal{L}_T(\mathcal{P}_T(\phi_T(T)))$ after projecting onto $\mathcal{U}^I$ and $\mathcal{U}^T$ respectively.

V. Experimental Setup

In this section, we describe an extensive experimental evaluation of the proposed framework. Two tasks were considered: text retrieval from an image query, and image retrieval from a text query. The cross-modal retrieval performance is measured with *precision-recall* (PR) curves and



(a) Correla (b) Semantic matching

Fig. 7. Cross-modal retrieval using CM and SM. On the left, CM is used to find the images that best match a query text. On the right, SM is used to find the texts that best match a query image.

mean average precision (MAP) scores. The standard 11-point interpolated PR curves [27] are used. The MAP score is the average precision at the ranks where recall changes. Both metrics are evaluated at the level of *in-* or *out-of-category*, which is a popular choice in the information retrieval literature [35]. We start by detailing the datasets used in the experiments.

A. Datasets

The evaluation of a cross-modal retrieval system requires a dataset which pairs pictures to rich text. While many datasets have been proposed in the literature for either of these modalities, few are rich in both modalities. In this work, we focus on two datasets, the "Text and Vision Graz" dataset collected by Khan et at [21], and a novel dataset composed of Wikipedia's featured articles [34]. In the following, we refer to these datasets as TVGraz and Wikipedia.

1) TVGraz: The TVGraz dataset is a collection of web-pages compiled by Khan et al [21]. The Google Image search engine was used to retrieve $1,000$ web-pages for each of ten categories from the Caltech-256 [15] dataset. The results were filtered into a set of $2,592$ positive web-pages, containing both text and image data, for which the image belonged to the query category. Due to copyright issues, the TVGraz database is stored as a list of URLs, and must be recompiled by each new user. We collected $2,058$ image-text pairs, since some URLs were defunct and we discarded web-pages that did not contain at least 10 words and one image. The median text length, per web-page, is 289 words. A random split was used to produce $1,558$ training and 500 test documents, as summarized in Table II.

2) Wikipedia: A novel dataset was assembled from the "Wikipedia featured articles", a continually updated collection of Wikipedia articles, which contained $2,669$ entries when the data was collected, in October 2009. These articles, which are selected and reviewed for style and quality by Wikipedia's editors, are often accompanied by one or more pictures from the Wikimedia Commons, supplying a text-image pairing. The Wikipedia featured articles are divided into 29 categories, but some contain very few entries. We considered only articles from the 10 most populated categories, which were used as a semantic vocabulary. 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 a total of $7,114$ 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, as summarized in Table III.

Dataset Comparison: The two datasets have important differences. On TVGraz, the images are archetypal members of the categories, due to the collection procedure [21]. The dataset is imminently visual, since its categories (e.g., "Harp", "Dolphin") are specific objects or animals, and the classes are semantically well-separated, with little or no semantic overlap. However, the texts are small and can be less representative of the categories to which they are associated. For example, the syllabus of a Neuroscience class can be attached to a picture of a brain. In Wikipedia, on the other hand, the category membership is assessed based on text content. Hence, texts are mostly of good-quality and representative of the category, while the image categorization is more ambiguous. For example, a portrait of an historical figure can appear in the class "War". The Wikipedia categories (e.g., "History", "Biology") are more abstract concepts, and have much broader scope. Frequently, documents can be classified into one or more categories. Individually, the images can be difficult to classify, even for a human. Together, the two datasets represent an important subset of the diversity of practical cross-modal retrieval scenarios: applications where there is more uniformity of text than images, and vice-versa.

B. Image and text representation

The base representation of the two modalities is the bag-of-words (BOW). Text words were obtained by stemming

TABLE II
SUMMARY OF THE TVGRAZ DATASET.

Category	Training set		Test set	Total
		Validation		
Brain	109	32	47	156
Butterfly	195	39	51	246
Cactus	137	27	37	174
Deer	223	51	51	274
Dice	169	36	50	219
Dolphin	163	24	59	222
Elephant	120	22	54	174
Frog	215	35	67	282
Harp	131	27	42	173
Pram	96	20	42	138
total	1558	(313)	500	2058

TABLE III
SUMMARY OF THE WIKIPEDIA DATASET

Category	Training set		Test set	Total
		Validation		
Art & architecture	138	29	34	172
Biology	272	50	88	360
Geography & places	244	50	96	340
History	248	54	85	333
Literature & theatre	202	37	65	267
Media	178	32	58	236
Music	186	35	51	237
Royalty & nobility	144	44	41	185
Sport & recreation	214	41	71	285
Warfare	347	63	104	451
total	2173	(435)	693	2866

the text with the Python Natural Language Toolkit[1]. Direct word histograms was unsuitable for text because the large lexicon made the correlation analysis intractable. Instead, a latent Dirichlet allocation (LDA) [2] model was learned from the text features, using the implementation of [9]. LDA summarizes a document as a mixture of topics. More precisely, a text is modeled as a multinomial distribution over topics, each of which is in turn modeled as a multinomial distribution over words. Each word in a text is generated by first sampling a topic from the document-specific topic distribution, and then sampling a word from that topic's multinomial. This serves two purposes; it reduces dimensionality and increases feature abstraction, by converting documents from distributions over words to distributions over topics.

Image words were learned with the scale invariant feature transformation (SIFT) [26]. A bag of SIFT descriptors was first extracted from each image in the training set, using the SIFT implementation of LEAR[2]. A codebook, or dictionary of visual words was then learned with the K-means clustering algorithm. The SIFT descriptors extracted from each image were vector quantized with this codebook, producing a vector of visual word counts per image. For compatibility with the text, we also considered a lower-dimensional representation, by fitting an LDA model to visual word histograms. Preliminary experiments indicated that this outperformed a principal component analysis (PCA). In all experiments, model parameters were learned with cross-validation, using a random split of 80% - 20% of the training set, for training and validation respectively. In TVGraz, the training set $(1,558)$ was divided into $1,245$ training and 313 validation examples. In Wikipedia $(2,173)$ the split was made at $1,738$ training and 435 validation documents. Codebook sizes ranged from 128 to $8,192$ visual words. LDA models were learned with a number of topics ranging from 5 to $4,000$.

[1]http://www.nltk.org/
[2]https://lear.inrialpes.fr/people/dorko/downloads.html

VI. PARAMETER SELECTION

The combination of three retrieval modes (CM, SM, and SCM), three correlation matching approaches (CFA, CCA, KCCA), two image representations (BOW, LDA), and various distance measures generates a large number of possibilities for the implementation of cross-modal retrieval. Since each configuration has a number of parameters to tune, it is difficult to perform an exhaustive comparison of all possibilities. Instead, we pursued a sequence of preliminary comparisons, on the validation set, to prune the configuration space. The top performing approaches were then compared on the test set, as explained in the following section.

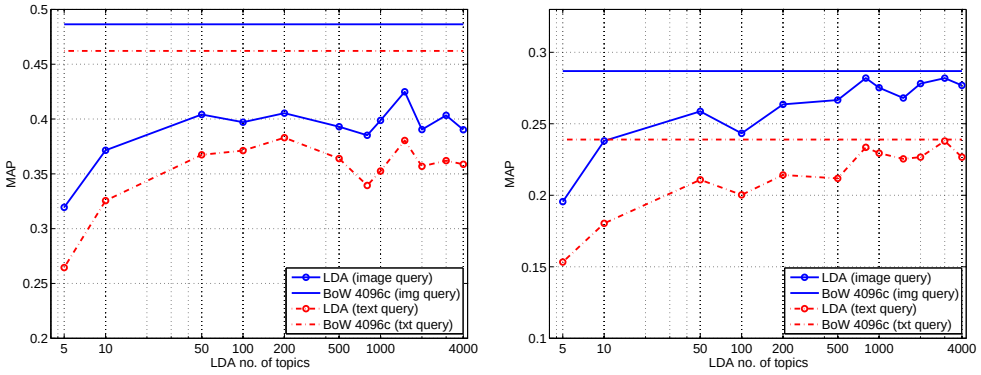
1) Distance Measures: We started by comparing a number of distance measures, for the evaluation of (7) and (9), in CM, SM, and SCM retrieval experiments. The subspaces used for CM were produced by KCCA. The measures are listed in Table IV, and include the Kullback-Leibler divergence (KL), ℓ_1 and ℓ_2 norms, normalized correlation (NC), and centered normalized correlation (NC_c). The KL divergence was not used with CM because this technique does not produce a probability simplex. Table IV presents the MAP scores achieved with each measure, on the validation set. NC_c achieved the best average performance in all experiments other than CM-based retrieval on TVGraz, where it was outperformed by NC. Since the difference was small even in this case, NC_c was adopted as distance measure in all remaining experiments.

2) Text and image representation: Due to the intractability of word counts, we considered only the LDA representation for text. For each experiment – CM, SM, SCM – and dataset – TVGraz, Wikipedia – the number of topics with maximum retrieval performance, on the validation set, was adopted. This is detailed in Table VII. In the image domain, we compared the performance of the BOW and LDA representations, using an SCM system based on KCCA subspaces. Figure 8 presents the MAP scores for both text and image queries on TVGraz and Wikipedia. Since the retrieval performance of LDA was inferior to that of BOW, for all topics cardinalities, BOW was adopted as the image representation of the remaining experiments.

3) Correlation matching: The next set of experiments were designed to compare the different CM methods. These

TABLE IV
CROSS-MODAL RETRIEVAL PERFORMANCE (MAP) ON THE VALIDATION SET USING DIFFERENT DISTANCE METRICS, FOR BOTH TVGRAZ AND WIKIPEDIA. μ_p AND μ_q ARE THE SAMPLE AVERAGES FOR p AND q.

			TVGraz			Wikipedia		
Experiment	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	$\sum_i p_i \log \frac{p_i}{q_i}$	0.296	0.546	0.421	0.188	0.276	0.232
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.412	0.548	0.480	0.232	0.276	0.254
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.380	0.550	0.465	0.211	0.278	0.245
	NC	$\frac{p^T q}{\lvert\lvert p\rvert\rvert \lvert\lvert q\rvert\rvert}$	0.533	0.560	0.546	0.315	0.278	0.296
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p\rvert\rvert \lvert\lvert q-\mu_q\rvert\rvert}$	0.579	0.556	**0.568**	0.354	0.272	**0.313**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.576	0.636	0.606	0.287	0.282	0.285
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.637	0.645	0.641	0.329	0.286	0.308
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.614	0.63	0.622	0.307	0.286	0.296
	NC	$\frac{p^T q}{\lvert\lvert p\rvert\rvert \lvert\lvert q\rvert\rvert}$	0.669	0.646	0.658	0.375	0.288	0.330
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p\rvert\rvert \lvert\lvert q-\mu_q\rvert\rvert}$	0.678	0.641	**0.660**	0.388	0.285	**0.337**



(a) TVGraz (b) Wikipedia

Fig. 8. MAP performance (cross-modal retrieval, validation set) of SCM using two image models: BOW (flat lines) and LDA.

methods have different degrees of freedom and thus require different amounts of parameter tuning. The most flexible representation is KCCA, whose performance varies with the choice of kernel and regularization parameter κ of (5). We started by comparing various kernel combinations. Best results were achieved with the combination of a *chi-square radial basis function* kernel[3] for images and a *histogram intersection* kernel [3], [43] for text. Combinations involving other kernels (e.g., linear, Gaussian, exponential) achieved inferior validation set performance. Regarding regularization, best results were obtained with $\kappa = 10\%$ on TVGraz and $\kappa = 50\%$ on Wikipedia. The need for a stronger regularizer in Wikipedia suggests that there are more spurious correlations on this dataset, which could lead to over-fitting. This is sensible, given the greater diversity and abstraction of the concepts in this dataset.

For CCA (CFA), the only free parameter is the number of canonical components (dimensionality of the shared space) used for both image and text representation. A grid search was performed to find the parameter of best retrieval performance under each method. In all cases, KCCA yields top performance. On TVGraz, the average gain (for text and

TABLE V
CM: CROSS-MODAL MAP TVGRAZ (VALIDATION SET).

Experiment	Image Query	Text Query	Average	Average Gain
KCCA	**0.486**	**0.462**	**0.474**	-
CCA	0.284	0.254	0.269	76%
CFA	0.195	0.179	0.187	153%

TABLE VI
CM: CROSS-MODAL MAP ON WIKIPEDIA (VALIDATION SET).

Experiment	Image Query	Text Query	Average	Average Gain
KCCA	**0.287**	**0.239**	**0.263**	-
CCA	0.210	0.174	0.192	37%
CFA	0.195	0.156	0.176	50%

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

image queries) is 153% over CFA and 76% over CCA. On Wikipedia, the gain over CFA is 50% and over CCA 37%. KCCA was chosen to implement the correlation hypothesis in the remaining experiments.

4) Overall optimization: The experiments above resulted in the selection of a cross-modal retrieval architecture combining KCCA to learn correlation subspaces, the centered normalized correlation distance measure, and a combination of the BOW representation for images and LDA representation for text. A final set of parameter tuning experiments was conducted to select the codebook size for image representation, the number of topics for text representation and the number of KCCA components. This was based on a grid search over the parameter space, on the validation set, which was repeated for CM, SM, and SCM.

As an example, Figure 9 presents the retrieval performance achieved with different parameter settings on the CM experiments. Note that the best MAP scores are obtained with a small number of KCCA components (< 10). For the image representation, best performance was achieved with codebooks of $4,096$ visual words, on both datasets. For text, 200 topics performed the best on TVGraz and 20 on Wikipedia. These results, and those of similar experiments for SM and SCM, are summarized in table VII. Note that in the test set experiments of Section VII, the number of KCCA components of Table VII is scaled by the ratio between the numbers of training points of the test experiments (Tables II and III) and that of the validation experiments (Section V-B), so that a comparable fraction of correlation is preserved after dimensionality reduction[4].

VII. Testing the Fundamental Hypotheses

TABLE VIII
CROSS-MODAL MAP ON TVGRAZ (TEST SET)

Experiment	Image Query	Text Query	Average	Average Gain
SCM	**0.693**	**0.696**	**0.694**	-
SM	0.625	0.618	0.622	11.6%
CM	0.507	0.486	0.497	39.6%
Random	0.114	0.114	0.114	509%

TABLE IX
CROSS-MODAL MAP ON WIKIPEDIA (TEST SET)

Experiment	Image Query	Text Query	Average	Average Gain
SCM	**0.372**	**0.268**	**0.320**	-
SM	0.362	0.252	0.307	4.2%
CM	0.282	0.225	0.253	26.5%
Random	0.119	0.119	0.119	170%

[4]KCCA seeks directions of maximum correlation in $\text{span}\{\phi_I(I_1), \ldots, \phi_I(I_{|\mathcal{D}|})\}$ and $\text{span}\{\phi_T(T_1), \ldots, \phi_T(T_{|\mathcal{D}|})\}$, where $|\mathcal{D}|$ is the training set size. This is larger for test than for validation experiments ($2,173$ v.s. $1,738$ on Wikipedia and $1,558$ v.s. $1,245$ on TVGraz). Hence, in average, a KCCA component will explain less correlation in the test than in the validation experiments. It follows that a larger number of KCCA components are needed to capture the same fraction of the total correlation.

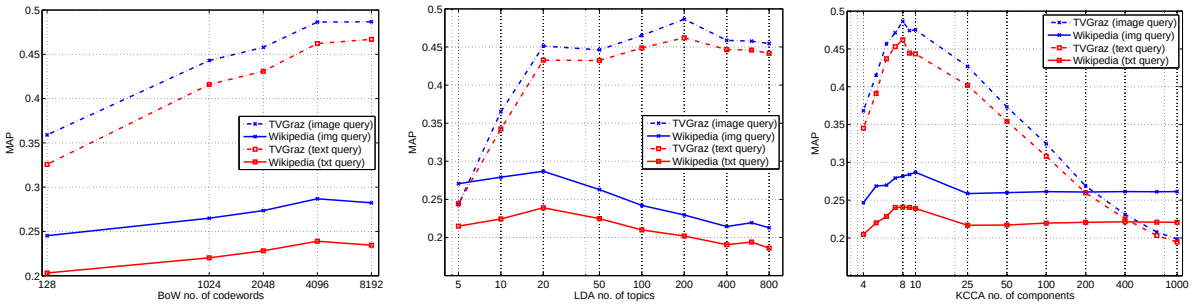
In this section, we compare the performance of CM, SM, and SCM on the test set. In all cases the parameter configurations are those that achieved best cross-validation performance in the previous section. Table VIII compares the MAP scores of cross-modal retrieval — text-to-image, image-to-text, and their average — using CM, SM and SCM, on TVGraz, to chance-level performance[5]. Two distinct observations can be made from the table. First, it provides evidence in support of the two hypotheses of Section III-C. Both joint dimensionality reduction (CM) and semantic abstraction (SM) are beneficial for multimodal modeling, leading to a non-trivial improvement over chance-level performance. For example, in TVGraz, CM achieves an average MAP score of 0.497, over four times the random retrieval performance of 0.114. SM yields an even greater improvement, attaining a MAP score of 0.622. Second, combining correlation modeling with semantic abstraction (SCM) is desirable, leading to higher MAP scores. On TVGraz, SCM improves about 12% over SM and 40% over CM, achieving an average MAP score of 0.694. This suggests that the contributions of cross-modal correlation and semantic abstraction are *complementary*: not only there is an independent benefit to both correlation modeling and abstraction, but the *best performance is achieved when the two hypothesis are combined*. The gains hold for both cross-modal retrieval tasks, i.e. image and text queries.

The corresponding results on Wikipedia are reported in Table IX, where similar conclusions can be drawn — both SM and CM perform far above chance level and their combination yields further improvements. However, the improvement of SCM over SM is less substantial than in TVGraz. In fact, the retrieval performances on Wikipedia are generally lower than those on TVGraz. As discussed in Section V-A, this is likely due to the broader scope of the Wikipedia categories. In this dataset, a significant fraction of documents could be classified into multiple categories, making the data harder to model. This explanation is supported by the confusion matrices of Figure 10. These were built by assigning each text and image query to the class of highest MAP in the raking produced by SCM[6]. Note, for example, the significant confusion between the categories "Architecture" and "Places", or "Royalty" and "Warfare".

Figure 10 presents PR curves of cross-modal retrieval with CM, SM and SCM. All methods yield non trivial precision improvements, at all levels of recall, when compared to the random baseline. On TVGraz, SM has higher precision than CM, and SCM has higher precision than SM, at all levels of recall. On Wikipedia, SCM improves over CM, at all levels of recall, but the improvement over SM is small. Figure 11 shows the MAP scores achieved per category by all approaches. SCM has a significantly higher MAP than CM and SM on all classes of TVGraz, and is either comparable or better than CM and SM on the majority of

[5]Random documents returned in response to the query.

[6]Note that this is not ideal for classification, since the MAP is computed over a ranking of the test set.



(a) no. of codewords (b) no. of topics (c) no. of KCCA components

Fig. 9. Cross-modal MAP of CM on TVGraz and Wikipedia (validation set), as a function of (a) number of image codewords, (b) number of text LDA topics, and (c) number of KCCA components.

TABLE VII
BEST PARAMETERS SETTINGS FOR CM, SM AND SCM, ON BOTH TVGRAZ AND WIKIPEDIA.

	CM	SM	SCM	
MAP image query / text query	0.49 / 0.46	0.59 / 0.56	0.68 / 0.64	TVGraz
BOW no. of codewords	4096			
LDA no. of topics	200	100	400	
KCCA no. of components	8	-	1125	
MAP image query / text query	0.29 / 0.24	0.35 / 0.27	0.39 / 0.29	Wikipedia
BOW no. of codewords	4096			
LDA no. of topics	20	600	200	
KCCA no. of components	10	-	38	

classes of Wikipedia.

Figure 12 presents two examples of text queries, by SCM, on TVGraz. In each case, the text query is shown at the top, along with its probability vector π_T and the ground truth image. The top five image matches are shown on the bottom, along with their probability vectors π_I. Note that SCM assigns these images the highest ranks in the retrieved list because their semantic vectors (π_I) most closely match that of the text (π_T). This can be verified by noting the common concentration of probability mass around "Cactus" (first example), and "Butterfly" (second example). Figure 13 presents similar examples from Wikipedia. Finally, Figure 14 shows examples of image-to-text retrieval. The queries are framed on the left column, and the images associated with the four best text matches are shown on the right.

VIII. CONCLUSION

The increasing availability of multimodal information demands the development of novel representations for content-based retrieval. In this work, we proposed models applicable to the task of 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 multimodal content modeling.

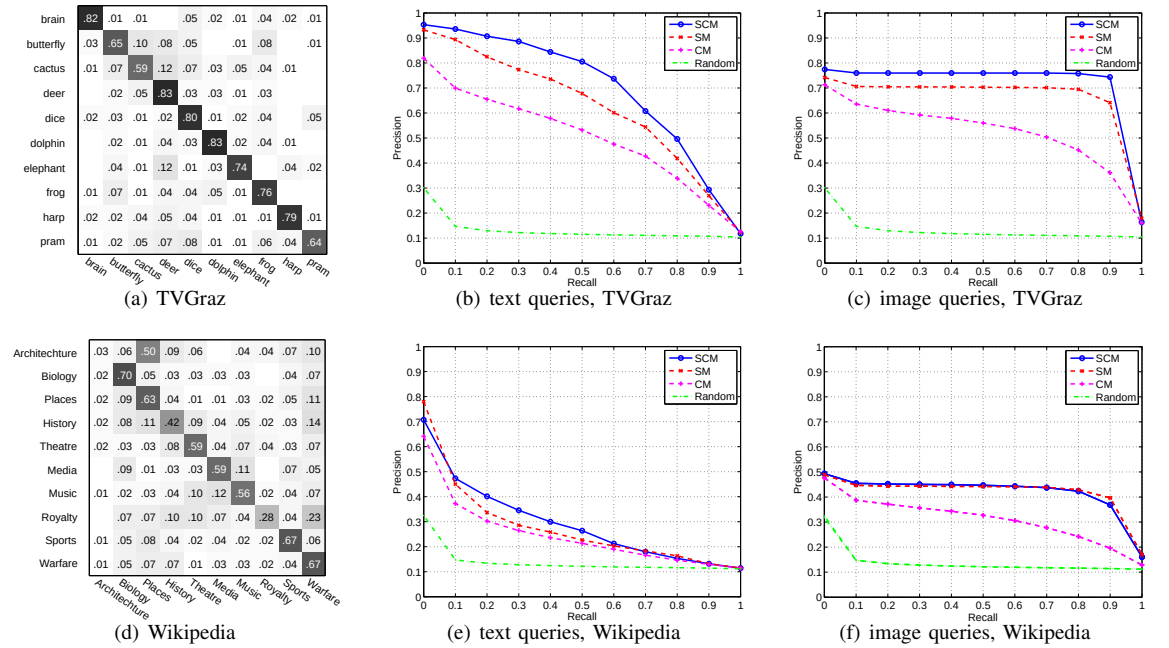
The hypotheses were objectively tested by comparing the performance of three new approaches to cross-modal retrieval: 1) correlation matching, based on the correlation hypothesis, 2) semantic matching, based on the abstraction hypothesis, and 3) semantic correlation matching, based on the combination 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 extensively tested on two datasets, "Wikipedia" and "TVGraz", containing documents that combine images and rich text. While the two fundamental hypotheses were shown to hold for the two datastets, where both CM and SM achieved significant improvements over chance retrieval, SM achieved overall better performance than CM. This implies stronger evidence for the abstraction than for the correlation hypothesis. 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 NSF award CCF-0830535.

REFERENCES

[1] *Learning Joint Statistical Models for Audio-Visual Fusion and Segregation*, 2001.
[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.



(a) TVGraz (b) text queries, TVGraz (c) image queries, TVGraz

(d) Wikipedia (e) text queries, Wikipedia (f) image queries, Wikipedia

Fig. 10. Confusion matrices on the test set, for both TVGraz (top) and Wikipedia (bottom). Rows refer to true categories, and columns to category predictions. The more confusion on Wikipedia motivates the lower retrieval performance. PR curves for cross-modal retrieval using both text and image queries.

[4] S. Boughorbel, J.P. Tarel, and N. Boujemaa. Generalized histogram intersection kernel for image recognition. In *IEEE International Conference on Image Processing*, volume 3. IEEE, 2005.

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

[6] E. Coviello, A.B. Chan, and G. Lanckriet. Time series models for semantic music annotation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):602–612, 2010.

[7] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[8] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[9] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. *Information Processing & Management*, 40(5):807–827, 2004.

[10] G. Doyle and C. Elkan. Accounting for word burstiness in topic models. In *Proceedings 26th International Conference on Machine Learning*, 2009.

[11] H.J. Escalante, C.A. Hérnadez, L.E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings 1st ACM International Conference on Multimedia Information Retrieval*, pages 172–179. ACM, 2008.

[12] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[13] SL Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition*, volume 2, 2004.

[14] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical report, Caltech, 2006.

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

[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[17] J. Iria, F. Ciravegna, and J. Magalhães. Web news categorization using a cross-media document graph. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8. ACM, 2009.

[18] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings 26th Annual International ACM SIG Information Retrieval*, page 126. ACM, 2003.

[19] IT Jolliffe. *Principal Component Analysis*. Springer, 2002.

[20] I. Khan, A. Saffari, and H. Bischof. Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In *Proceedings 33rd Workshop of the Austrian Association for Pattern Recognition*, 2009.

[21] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. Combining image captions and visual analysis for image concept classification. In *Proceedings 9th International Workshop on Multimedia Data Mining at ACM SIG Knowledge Discovery and Data Mining*, pages 8–17. ACM New York, NY, USA, 2008.

[22] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Neural Information Processing Systems*, 2003.

[23] D. Li, N. Dimitrova, M. Li, and I.K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings 11th ACM International Conference on Multimedia*, pages 604–611. ACM, 2003.

[24] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *IEEE Conference on Multimedia and Expo*, page 190, 2001.

[25] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

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

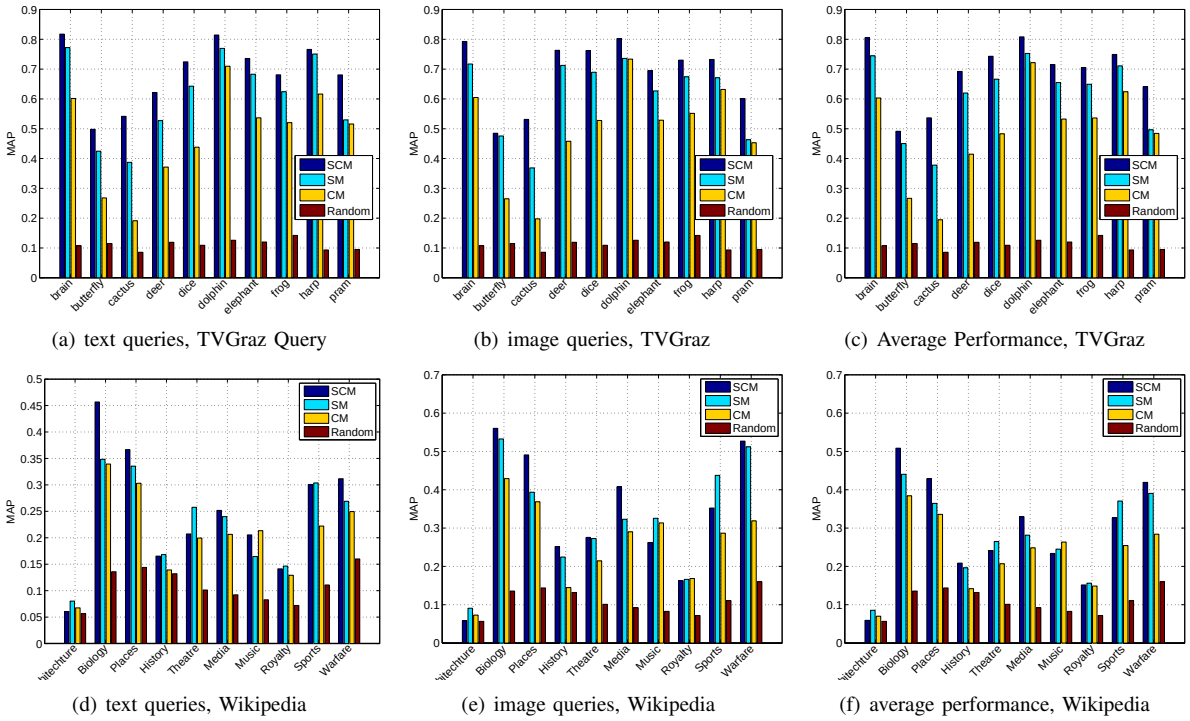
[27] C.T. Meadow, B.R. Boyce, D.H. Kraft, and C.L. Barry. *Text Information Retrieval Systems*. Emerald Group Pub Ltd, 2007.

[28] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007.

[29] Y. Mori, H. Takahashi, and R. Oka. Automatic word assignment to images based on image division and vector quantization. In *Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO)*. Citeseer, 2000.

[30] S. Nakamura. Statistical multimodal integration for audio-visual speech processing. *IEEE Transactions on Neural Networks*, 13(4):854–866, 2002.

[31] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo



(a) text queries, TVGraz Query (b) image queries, TVGraz (c) Average Performance, TVGraz

(d) text queries, Wikipedia (e) image queries, Wikipedia (f) average performance, Wikipedia

Fig. 11. Per category cross-modal MAP, on TVGraz and Wikipedia.

retrieval: Overview of the imageclef 2009 photo task. *CLEF working notes*, 2009.

[32] T.T. Pham, N.E. Maillot, J.H. Lim, and J.P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings 16th ACM International Conference on Information and Knowledge Management*, pages 439–444. ACM, 2007.

[33] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings 18th ACM International Conference on Multimedia*, 2010.

[34] N. Rasiwasia, PJ Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007.

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

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

[37] M. Slaney. Semantic-audio retrieval. In *IEEE International Conference on Acoustics Speech and Signal Processing*, volume 4, pages 4108–4111. IEEE, 2002.

[38] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proceedings 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[39] A.W.M. 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*, 22(12):1349–1380, 2000.

[40] John R. Smith. The real problem of bridging the "semantic gap". In *International Conference on Multimedia Content Analysis and Mining*, MCAM'07, pages 16–17, 2007.

[41] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[42] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[43] T. Tsikrika and J. Kludas. Overview of the wikipediamm task at imageclef 2009. In *Working Notes for the CLEF Workshop*, 2009.

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

[45] N. Vasconcelos. From pixels to semantic spaces: Advances in content-based image retrieval. *IEEE Computer*, 40(7):20–26, 2007.

[46] A. Vinokourov, D.R. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *Proceedings 4th International Symposium on Independent Component Analysis and Blind Source Separation*. Citeseer, 2003.

[47] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. *Advances in Neural Information Processing Systems*, pages 1497–1504, 2003.

[48] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *Proceedings of 19th International Conference on Pattern Recognition*, 2009.

[49] T. Westerveld. Image retrieval: Content versus context. *Content-Based Multimedia Information Access*, pages 276–284, 2000.

[50] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings 17th ACM International Conference on Multimedia*, pages 175–184. ACM, 2009.

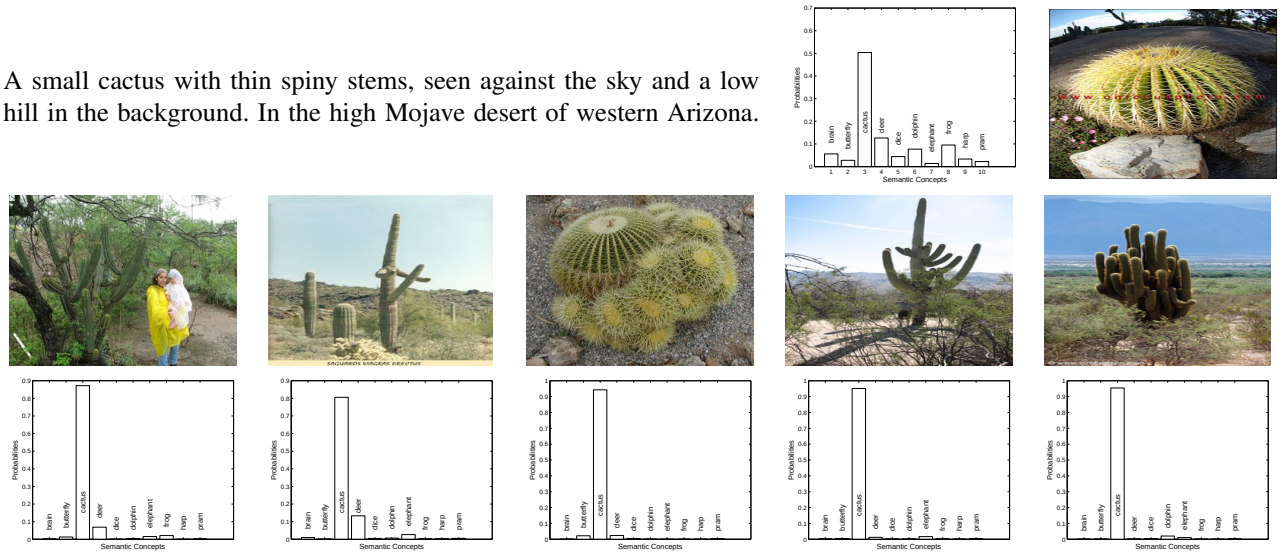
[51] Y. Yang, Y.T. Zhuang, F. Wu, and Y.H. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008.

[52] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *Proceedings 15th ACM International Conference on Multimedia*, page 276. ACM, 2007.

[53] 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*, 46(2):153–164, 2007.

[54] Y.T. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.

A small cactus with thin spiny stems, seen against the sky and a low hill in the background. In the high Mojave desert of western Arizona.



On the Nature Trail behind the Bathabara Church, there are numerous wild flowers and plants blooming, that attract a variety of insects, bees and birds. Here a beautiful Butterfly is attracted to the blooms of the Joe Pye Weed.

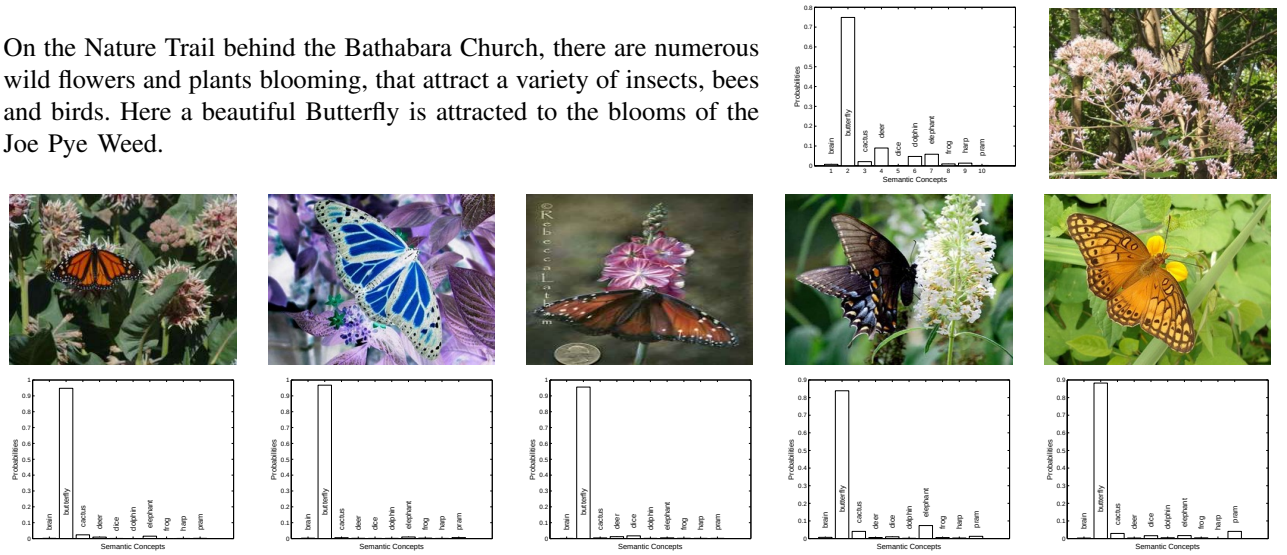
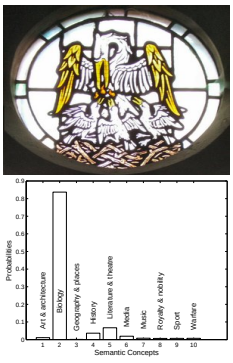
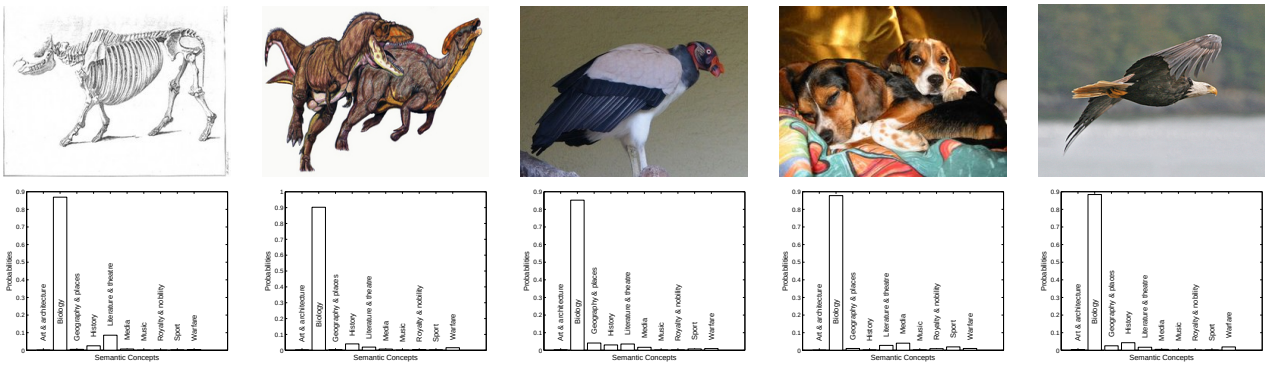


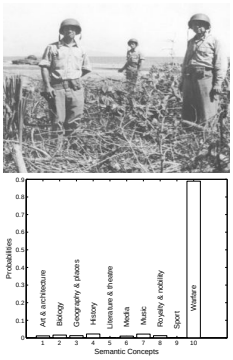
Fig. 12. Two examples of text-based cross-modal retrieval, on TVGraz, using SCM. The query text, associated probability vector and ground truth image are shown on the top. Retrieved images are presented at the bottom.

Many seabirds are little studied and poorly known, due to living far out to sea and breeding in isolated colonies. However, some seabirds, particularly, the albatrosses and gulls, have broken into popular consciousness. The albatrosses have been described as "the most legendary of birds",Carboneras, C. (1992) "Family Diomedeidae (Albatrosses)" in "Handbook of Birds of the World" Vol 1. Barcelona:Lynx Edicions, ISBN 84-87334-10-5 and have a variety of myths and legends associated with them, and today it is widely considered unlucky to harm them, although the notion that sailors believed that is a myth Cocker, M., & Mabey, R., (2005) "Birds Britannica" London:Chatto & Windus, ISBN 0-7011-6907-9 which derives from Samuel Taylor Coleridge's famous poem, "The Rime of the Ancient Mariner", in which a sailor is punished for killing an albatross by having to wear its corpse around his neck.
"Instead of the Cross the Albatross" "About my neck was hung"
Sailors did, however, consider it unlucky to touch a storm-petrel, especially one that has landed on the ship.Carboneras, C. (1992) "Family Hydrobatidae (Storm-petrels)" in "Handbook of Birds of the World" Vol 1. Barcelona:Lynx Edicions, ISBN 84-87334-10-5
Gulls are one of the most commonly seen seabirds, given their use of human-made habitats (such as cities and dumps) and their often fearless nature. They therefore also have made it into the popular consciousness - they have been used metaphorically, as in "Jonathan Livingston Seagull" by Richard Bach, or to denote a closeness to the sea, such as their use in the "The Lord of the Rings" both in the insignia of Gondor and therefore Namenor (used in the design of the films), and to call Legolas to (and across) the sea. Other species have also made an impact; pelicans have long been associated with mercy and altruism because of an early Western Christian myth that they split open their breast to feed their starving chicks.





Between October 1 and October 17, the Japanese delivered 15,000 troops to Guadalcanal, giving Hyakutake 20,000 total troops to employ for his planned offensive. Because of the loss of their positions on the east side of the Matanikau, the Japanese decided that an attack on the U.S. defenses along the coast would be prohibitively difficult. Therefore, Hyakutake decided that the main thrust of his planned attack would be from south of Henderson Field. His 2nd Division (augmented by troops from the 38th Infantry Division), under Lieutenant General Masao Maruyama and comprising 7,000 soldiers in three infantry regiments of three battalions each was ordered to march through the jungle and attack the American defences from the south near the east bank of the Lunga River.Shaw, "First Offensive", p. 34, and Rottman, "Japanese Army", p. 63. To distract the Americans from the planned attack from the south, Hyakutake's heavy artillery plus five battalions of infantry (about 2,900 men) from the 4th and 124th Infantry Regiments under the overall command of Major General Tadashi Sumiyoshi were to attack the American defenses from the west along the coastal corridor.Rottman, "Japanese Army", p. 61, Frank, "Guadalcanal", p. 289340, Hough, "Pearl Harbor to Guadalcanal", p. 32230, Griffith, "Battle for Guadalcanal", p. 18687, Dull, "Imperial Japanese Navy", p. 22630, Morison, "Struggle for Guadalcanal", p. 14971. The Japanese troops delivered to Guadalcanal during this time comprised the entire 2nd (Sendai) Infantry Division, two battalions from the 38th Infantry Division, and various artillery, tank, engineer, and other support units. Kawaguchi's forces also included what remained of the 3rd Battalion, 124th Infantry Regiment which was originally part of the 35th Infantry Brigade commanded by Kawaguchi during the Battle of Edson's Ridge.



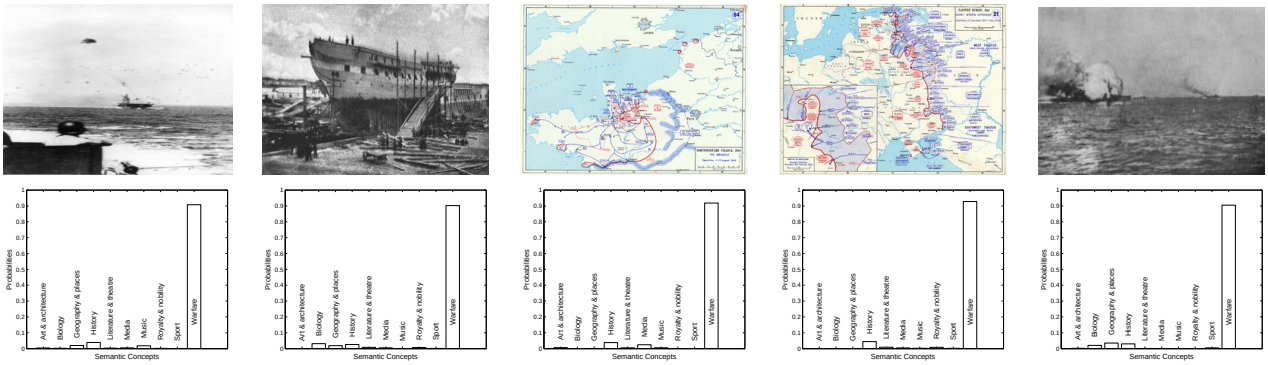


Fig. 13. Two examples of text-based cross-modal retrieval, on Wikipedia, using SCM. The query text, associated probability vector and ground truth image are shown on the top. Retrieved images are presented at the bottom.

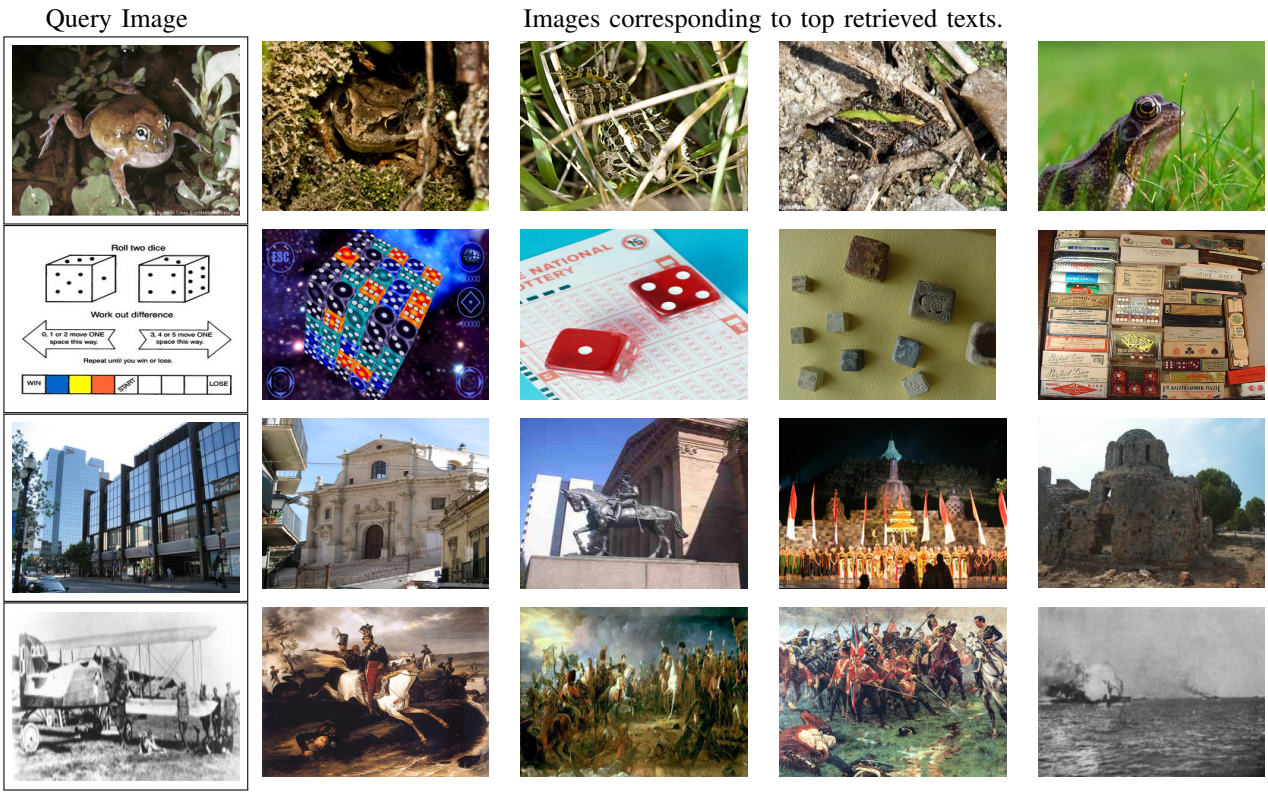


Fig. 14. Image-to-text retrieval on TVGraz (top two rows) and Wikipedia (remaining two). Query images are framed on the far-left column. The four most relevant texts, represented by their ground truth images, are shown on the remaining columns.