A New Approach to Cross-Modal Multimedia Retrieval

Nikhil Rasiwasia[1], Jose Costa Pereira[1], Emanuele Coviello[1], Gabriel Doyle[2],
Gert R.G. Lanckriet[1], Roger Levy[2], Nuno Vasconcelos[1]
[1]Dept. of Electrical and Computer Engineering, [2]Dept. of Linguistics,
University of California, San Diego
{nikux,josecp,ecoviell,gdoyle,rlevy}@ucsd.edu,{gert,nuno}@ece.ucsd.edu

ABSTRACT

The problem of joint modeling the *text* and *image* components of multimedia documents is studied. The text component is represented as a sample from a hidden topic model, learned with latent Dirichlet allocation, and images are represented as bags of visual (SIFT) features. Two hypotheses are investigated: that 1) there is a benefit to explicitly modeling *correlations* between the two components, and 2) this modeling is more effective in feature spaces with higher levels of *abstraction*. Correlations between the two components are learned with canonical correlation analysis. Abstraction is achieved by representing text and images at a more general, semantic level. The two hypotheses are studied in the context of the task of cross-modal document retrieval. This includes retrieving the text that most closely matches a query image, or retrieving the images that most closely match a query text. It is shown that accounting for cross-modal correlations and semantic abstraction both improve retrieval accuracy. The cross-modal model is also shown to outperform state-of-the-art image retrieval systems on a unimodal retrieval task.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Design

Keywords

Cross-media, cross-modal, retrieval, document processing, image and text, multimedia

1. INTRODUCTION

Over the last decade there has been a massive explosion of multimedia content on the web. This explosion has not been matched by an equivalent increase in the sophistication of multimedia content modeling technology. Today, the prevailing tools for searching multimedia repositories are still text-based, e.g. search engines such as Google or Bing. To address this problem, the academic community has devoted itself to the design of models that can account for multiple content modalities. In computer vision, substantial effort has been devoted to the problem of image annotation [4, 12, 9, 16, 19, 1]. The multimedia community has started a number of large-scale research and evaluation efforts, such as TRECVID [26] and imageCLEF [21, 29], involving image or video data complemented with annotations, close-caption information, or speech recognition transcripts. Many techniques have been proposed in this literature to automatically augment images with captions or labels and to retrieve and classify imagery augmented with information from these modalities [26, 21, 29, 28, 6, 31, 14, 8, 32].

An important requirement for further progress in these areas is the development of sophisticated joint models for multiple content modalities. Particularly important is the development of models that support inference with respect to content that is *rich* in multiple modalities. These include models that do not simply consider the text accompanying an image as a source of keywords for image classification, but make use of the full structure of *documents* that pair a body of text with a number of images or video-clips. The availability of such documents, which include web-pages, newspaper articles, and technical articles, has blossomed with the explosion of internet-based information. In this work, we consider the design of these multimedia models. We concentrate on documents containing text and images, although many of the ideas would be applicable to other modalities. We start from the extensive literature available on text and image analysis, including the representation of documents as bags of features (word histograms for text, SIFT histograms [5] for images), and the use of topic models (such as latent Dirichlet allocation [3]) to extract low-dimensionality generalizations from document corpora. We build on these representations to design a joint model for images and text.

The performance of this model is evaluated on a *cross-modal* retrieval problem that includes two 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 tasks are central to many applications of practical interest, such as finding on the web the picture that best illustrates 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

MM'10, October 25–29, 2010, Firenze, Italy.

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, speci↓cally 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". According to this historians Isserman and Kazin, the demands on the city authorities were straightforward: desegregate the economic life of Birmingham its restaurants, hotels, public toilets, and the unwritten policy of hiring blacks for menial jobs only Maurice Isserman and Michael Kazin, America Divided: The Civil War of the 1960s, (Oxford, 2008), p.90. Organizers believed their phones were tapped, so to prevent their plans from being leaked and perhaps in'uencing the mayoral election, they used code words for demonstrations. The plan called for direct nonviolent action to attract media attention to "the biggest and baddest city of the South".Hampton, p. 126. In preparation for the protests, Walker timed the walking distance from the Sixteenth Street Baptist Church, headquarters for the campaign, to the downtown area. He surveyed the segregated lunch counters of department stores, and listed federal buildings as secondary targets should police block the protesters' entrance into primary targets such as stores, libraries, and all-white churches.

Figure 1: A section from Wikipedia article on the Birmingham campaign `http://en.wikipedia.org/wiki/Birmingham_campaign`

a combination of text and images. We only briefly discuss these applications, concentrating in this study on the problem of model design. We use performance on the retrieval tasks as an indirect measure of the model quality, under the intuition that the best model should produce the highest retrieval accuracies.

With regards to model design, we investigate two hypotheses. The first is that explicit modeling of correlations between images and text is important. We propose models that explicitly account for cross-modal correlations using canonical correlation analysis (CCA), and compare their performance to models where the two modalities are modeled independently. The second is that a useful role can be played by *abstraction*—defined here as hierarchical inference across layers of increasingly general semantics. Various results have shown that such representations improve performance in multimedia tasks, e.g. the use of hierarchical topic models for text clustering [3] or hierarchical semantic representations for image retrieval [25]. The retrieval problems we consider here are amenable to the design of such abstraction hierarchies: for example, features group into documents, which themselves group into classes or topics, which form corpora. Abstract representations are proposed for both vision and text, by modeling images and documents as vectors of posterior probabilities with respect to a set of pre-defined document classes, computed with logistic regression [10].

We investigate the retrieval performance of various combinations of image and text representations, which cover all possibilities in terms of the two guiding hypotheses. Our results show that there is a benefit to both abstraction and cross-modal correlation modeling. In particular, our best results are obtained by a model that combines semantic abstraction for both images and text with explicit modeling of cross-correlations in a joint space. We also demonstrate the benefits of joint text and image modeling by comparing the performance of a state-of-the-art image retrieval system to an image retrieval system that accounts for the text which accompanies each image, using the proposed joint model. It is shown that the latter has substantially higher retrieval accuracy.

2. PREVIOUS WORK

The problems of image and text retrieval have been the subject of extensive research in areas such as information retrieval, computer vision, and multimedia [6, 27, 26, 21, 18]. In all these areas, the main emphasis of the retrieval literature has been on *unimodal* approaches, where query and retrieved documents share a single modality [30, 6, 27]. More recently, there has been some interest in image retrieval systems that rely on collateral text metadata. The latter is usually provided by human annotators, typically in the form of a few keywords, a small caption or a brief image description [21, 29, 26]. We do not refer to such systems as *cross-modal* since the retrieval operation itself is unimodal, simply matching a text query to available text metadata. However, because manual image labeling is a labor intensive process, these systems inspired research on the problem of automatic extraction of semantic descriptors from images [4, 12, 9, 16, 19, 1]. This enabled a first generation of truly cross-modal systems, which support text-based queries of image databases that do not contain text metadata. However, these models are limited by their inherently impoverished textual information. Images are simply associated with keywords, or class labels, and there is no explicit modeling of free-form text. Two notable exceptions are the works of [2, 19], where separate "latent-space" models are learned for images and text, in a form suitable for cross-media image annotation and retrieval.

In parallel, advances have been reported in the area of *multi-modal* retrieval systems [21, 29, 26, 28, 6]. These are extensions of the classic unimodal systems, where a single retrieval model is applied to information from various modalities. This can be done by fusing features from different modalities into a single vector [32, 22], or by learning different models for different modalities and fusing their outputs [31, 14]. An excellent overview of these approaches is given in [8], which also presents an approach to further combine unimodal and multimodal retrieval systems. However, most of these approaches require *multimodal queries*, queries composed of both image and text features. An alternative paradigm is to improve the models of one modality (say image) using information from other modalities (e.g., image captions) [23, 20]. Lastly, it is possible to design multimodal systems by using text annotations to construct a semantic space [25]. Images are also represented on this space, and similarity measured in this space can be used for retrieval. This representation uses text to perform image retrieval at a higher level of abstraction than simple image matching.

In spite of the these developments in cross- and multimodal retrieval, current systems rely on a limited textual representation, in the form of keywords, captions, or small

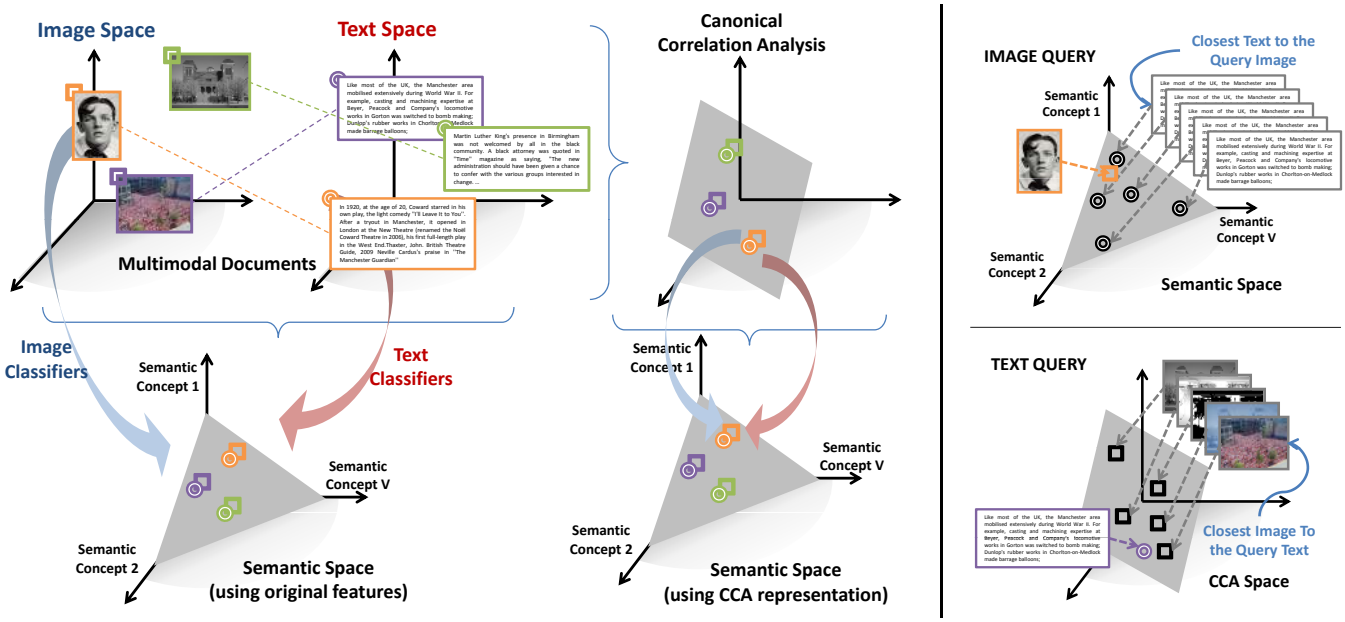


Figure 2: Schematic of the proposed cross-modal retrieval system. Left) Mapping of the text and image from their respective natural spaces to a CCA space, Semantic Space and a Semantic space learned using CCA representation. Right) Example of cross-modal query in the proposed system. At the top is shown an example of retrieving text in response to an image query, where both text and images are mapped to a common semantic space. At the bottom is shown an example of retrieving images in response to a text query, where both text and images are mapped to a common subspace using CCA.

text snippets. On the other hand, with the ongoing explosion of Web-based multimedia content, it is now possible to collect large datasets of more richly annotated data. Examples include news archives, or Wikipedia pages, where pictures are related to complete text articles, not just a few keywords. In these datasets, the connection between images and text is much less direct than that provided by light annotation, weakening the one-to-one mapping between textual words and class labels. For example, Fig 1 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. A major long-term goal of modeling in this domain should be to recover this *latent* relationship between the text and image components of a document, and put it to practical use in applications.

3. CROSS-MODAL RETRIEVAL

In this section, we present a novel approach to cross-modal retrieval. Although the fundamental ideas are applicable to any combination of content modalities, we restrict the discussion to documents containing images and text. The goal is to support truly cross-modal queries: to retrieve text articles in response to query images and vice-versa.

3.1 The problem

We consider the problem of information retrieval from a database $\mathcal{D} = \{D_1, \ldots, D_{|D|}\}$ of *documents* which contain components of *images* and *text*. In practice, these components 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 an *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. In this way, each document establishes a one-to-one mapping between points in the text and image spaces. 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 $\Re^I$ ($\Re^T$).

3.2 Matching images and text

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 a I_q in $\Re^I$, it suffices to find the nearest neighbor to $\mathcal{M}^{-1}(I_q)$. In this case, the design of a cross-modal retrieval system reduces to the design of an effective similarity function for the determination of nearest neighbors.

Since different representations tend to be adopted for images and text, there is typically no natural correspondence between $\Re^I$ and $\Re^T$. In this case, the mapping $\mathcal{M}$ has to be learned from examples. One possibility, that we pursue in this work, is to map the two representations into two intermediate spaces $\mathcal{U}^I$ and $\mathcal{U}^T$ that have a natural correspondence. Let

$$\mathcal{M}_I : \Re^I \rightarrow \mathcal{U}^I$$

and

$$\mathcal{M}_T : \Re^T \rightarrow \mathcal{U}^T$$

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

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

Given a query T_q in $\Re^T$ the cross-modal retrieval operation 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 approach, the main problem in the design of a cross-modal retrieval system is to learn the intermediate spaces $\mathcal{U}^I$ and $\mathcal{U}^T$. In this work, we consider three possibilities that result from the combination of two main procedures. In the first case, two linear projections

$$\mathcal{P}_T : \Re^T \rightarrow \mathcal{U}^T$$

and

$$\mathcal{P}_I : \Re^T \rightarrow \mathcal{U}^I$$

are learned to map $\Re^I$ respectively $\Re^T$ onto *correlated* d-dimensional *subspaces* $\mathcal{U}^I$ and $\mathcal{U}^T$. This maintains the level of abstraction of the representation. In the second case, a pair of non-linear transformations

$$\mathcal{L}_T : \Re^T \rightarrow \mathcal{S}^T$$

and

$$\mathcal{L}_I : \Re^I \rightarrow \mathcal{S}^I$$

are used to map the image and text spaces into a pair of *semantic* spaces $\mathcal{S}^T$, $\mathcal{S}^I$ such that $\mathcal{S}^T = \mathcal{S}^I$. This increases the semantic abstraction of the representation. We next describe the two approaches in greater detail.

3.3 Correlation matching

Learning $\mathcal{U}^T, \mathcal{U}^I$ requires some notion of an optimal correspondence between the representations in the text and image spaces. One possibility is to rely on subspace learning. This is a learning framework that underlies some extremely popular dimensionality reduction approaches in both the text and vision literatures, such as latent semantic indexing [7] or principal component analysis (PCA) [13]. Subspace learning methods are typically efficient from a computational point of view, and produce linear transformations which are easy to conceptualize, implement, and deploy. In this case, a natural measure of correspondence between the image and text subspaces is their correlation. This suggests canonical correlation analysis (CCA) as a natural subspace representation for cross-modal modeling.

Canonical correlation analysis (CCA) [11] is a data analysis and dimensionality reduction method similar to PCA. While PCA deals with only one data space, CCA is a technique for joint dimensionality reduction across two (or more) spaces that provide heterogeneous representations of the same data. The assumption is that the representations in these two spaces contain some joint information that is reflected in correlations between them. CCA learns d-dimensional subspaces $\mathcal{U}^I \subset \Re^I$ and $\mathcal{U}^T \subset \Re^T$ that maximize the correlation between the two modalities.

Similar to principal components in PCA, CCA learns a basis of canonical components, i.e., directions $w_i \in \Re^I$ and $w_t \in \Re^T$ along which the data is maximally correlated, i.e.,

$$\max_{w_i \neq 0,\, w_t \neq 0} \frac{w_i^T \Sigma_{IT} w_t}{\sqrt{w_i^T \Sigma_{II} w_i} \sqrt{w_t^T \Sigma_{TT} w_t}}, \quad (1)$$

where Σ_{II} and Σ_{TT} represent the empirical covariance matrices for images $\{I_1, \ldots, I_{|D|}\}$ and text $\{T_1, \ldots, T_{|D|}\}$ respectively, while $\Sigma_{IT} = \Sigma_{TI}^T$ represents the cross-covariance matrix between them. The optimization of (1) can be solved as a generalized eigenvalue problem (GEV) [24]

$$\begin{pmatrix} 0 & \Sigma_{IT} \\ \Sigma_{TI} & 0 \end{pmatrix} \begin{pmatrix} w_i \\ w_t \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{II} & 0 \\ 0 & \Sigma_{TT} \end{pmatrix} \begin{pmatrix} w_i \\ w_t \end{pmatrix}.$$

The generalized eigenvectors determine a set of uncorrelated canonical components, with the corresponding generalized eigenvalues indicating the explained correlation. GEVs can be solved as efficiently as regular eigenvalue problems [15].

The first d canonical components $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$ define a basis for projecting $\Re^I$ respectively $\Re^T$ on a subspace $\mathcal{U}^I$ respectively $\mathcal{U}^T$. A natural invertible mapping between these two projections follows from the correspondence between the d-dimensional bases of maximal cross-modal correlation, as $w_{i,1} \leftrightarrow w_{t,1}$, ..., $w_{i,d} \leftrightarrow w_{t,d}$. For cross-modal retrieval, every text $T \in \Re^T$ is mapped into its projection $p_T = \mathcal{P}_T(T)$ onto $\{w_{t,k}\}_{k=1}^d$, and every image into its projection $p_I = \mathcal{P}_I(I)$ onto $\{w_{i,k}\}_{k=1}^d$. This results in a compact, efficient representation of both modalities. Since the vectors p_T and p_I are coordinates in two isometric d dimensional subspaces $\mathcal{U}^T$ respectively $\mathcal{U}^I$, they can be thought as belonging to a single space $\mathcal{U}$, obtained by overlaying $\mathcal{U}^I$ and $\mathcal{U}^T$. This leads to the schematic representation of Figure 2, where CCA defines a common subspace ($\mathcal{U}$) for cross-modal retrieval.

Given an image query I_q with projection $p_I = \mathcal{P}(I_q)$, the text $T \in \Re^T$ that most closely matches it is that for which $p_T = \mathcal{P}(T)$ minimizes

$$D(I, T) = d(p_I, p_T) \quad (2)$$

for some suitable measure of distance $d(\cdot, \cdot)$ in a d-dimensional vector space. Similarly, given a query text T_q with projection $p_T = \mathcal{P}(T_q)$, the image $I \in \Re^I$ that most closely matches it is that for which $p_I = \mathcal{P}_I(I)$ minimizes $d(p_I, p_T)$. We refer to this type of retrieval as *correlation matching*.

3.4 Semantic matching

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 accomplished by augmenting the database $\mathcal{D}$ with a vocabulary $\mathcal{V} = \{v_1, \ldots, v_K\}$ of semantic concepts. These are broad document classes, such as "History" or "Biology", into which individual documents are grouped. Two mappings $\mathcal{L}_T$ and $\mathcal{L}_I$ are then implemented with recourse to two classifiers of text and images, respectively. $\mathcal{L}_\mathcal{T}$ maps a text $T \in \Re^T$ into a vector of posterior probabilities $P_{V|T}(v_i|T), i \in \{1, \ldots, K\}$ with respect to each of the classes in $\mathcal{V}$. The space $\mathcal{S}^T$ of these posterior vectors is referred to as the *semantic space for text*, and the probabilities $P_{V|T}(v_i|T)$ as *semantic text features*. Similarly, $\mathcal{L}_I$ maps an image I into a vector of *semantic image features* $P_{V|I}(v_i|I), i \in \{1, \ldots, K\}$ in a semantic image space $\mathcal{S}^I$.

One possibility to compute posterior probability distributions is 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 data to a logistic function,

$$P_{V|X}(j|x;w) = \frac{1}{Z(x,w)} \exp(w_j^T x) \quad (3)$$

where $Z(x,w) = \sum_j \exp(w_j^T x)$ 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.

Semantic modeling has two advantages for cross-modal retrieval. First, it provides 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 abstraction can lead to substantially better generalization for tasks such as image retrieval [25]. 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 thought as the same, i.e. $\mathcal{S}^T = \mathcal{S}^I$, leading to the schematic representation of Figure 2.

Given a query image I_q, represented by a probability vector $\pi_I \in \mathcal{S}^I$, retrieval consists of finding the text T, represented by a probability vector $\pi_T \in \mathcal{S}^T$, that minimizes

$$D(I,T) = d(\pi_I, \pi_T), \quad (4)$$

for some suitable measure of distance d between probability distributions. We refer to this type of retrieval as *semantic matching*.

3.5 Semantic correlation matching

It is also possible to combine subspace and semantic modeling. In this case, logistic regression is performed within two maximally correlated subspaces. The CCA modeling of Section 3.3 is first applied to learn the maximally correlated subspaces $\mathcal{U}^I \subset \Re^I$ and $\mathcal{U}^T \subset \Re^T$. 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 (4), based on the semantic mappings $\pi_I = \mathcal{L}_I(\mathcal{P}_I(I))$ and $\pi_T = \mathcal{L}_T(\mathcal{P}_T(T))$ after projections onto $\mathcal{U}^I$ and $\mathcal{U}^T$ respectively. We refer to this type of retrieval as *semantic correlation matching*.

3.6 Text and Image Representation

In this work, the representation of text on $\Re^T$ is derived from a latent Dirichlet allocation (LDA) model [3]. LDA is a generative model for a text corpus, where the semantic content or "gist" of a text is summarized as a mixture of *topics*. More precisely, a text is modeled as a multinomial distribution over K topics, each of which is in turn modeled as a multinomial distribution over words. Each word in a text D_i is generated by first sampling a topic z from the text-specific topic distribution, and then sampling a word from that topic's multinomial. In $\Re^T$ text documents are represented by their topic assignment probability distributions.

Table 1: Summary of the Wikipedia dataset.

Category	Training	Query/ Retrieval	Total documents
Art & architecture	138	34	172
Biology	272	88	360
Geography & places	244	96	340
History	248	85	333
Literature & theatre	202	65	267
Media	178	58	236
Music	186	51	237
Royalty & nobility	144	41	185
Sport & recreation	214	71	285
Warfare	347	104	451

In $\Re^I$, image representation is based on the popular scale invariant feature transformation (SIFT) [17]. A bag of SIFT descriptors is first extracted from each image in the training set (using the SIFT implementation of LEAR [1]). A codebook, or dictionary, of visual words is then learned with the k-means clustering algorithm. The SIFT descriptors extracted from each image are vector quantized with this codebook, and images are represented by the SIFT descriptor histograms that result from this quantization [5].

4. EXPERIMENTS

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

4.1 Dataset

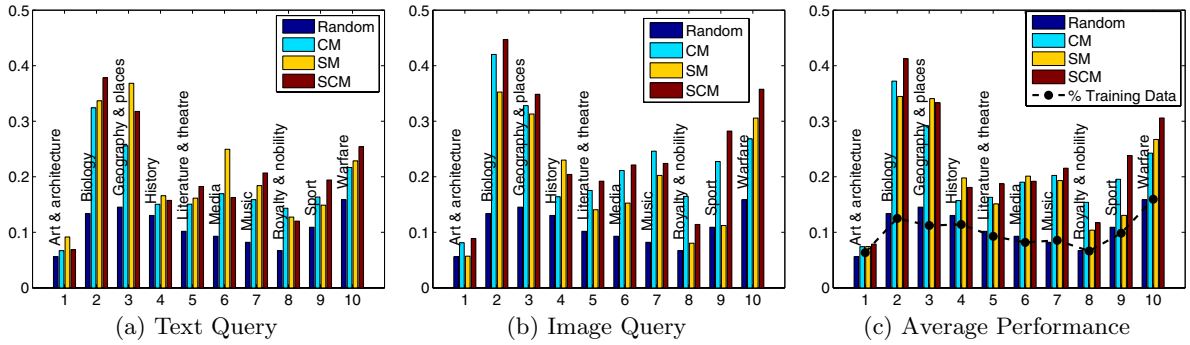
The evaluation of a cross-modal retrieval system requires a document corpus with paired texts and images. To design such a corpus we relied on Wikipedia's "featured articles." This is a continually updated collection of 2700 articles that have been selected and reviewed by Wikipedia's editors, since 2009. The articles are accompanied by one or more pictures from the Wikimedia Commons, supplying a pairing of the desired kind. In addition, each featured article is categorized by Wikipedia into one of 29 categories. These category labels were assigned to both the text and image components of each article. Since some of the categories are very scarce, we considered only the 10 most populated ones.

Each article was split into sections, based on its section headings, and each image in the article assigned to the section in which it was placed by the article author(s). This produced a set of short and focused articles, usually containing a single image. The dataset was finally pruned by removing the sections without any image. The final corpus contains a total of 2866 documents. These are *text - image* pairs, annotated with a label from the vocabulary of 10 semantic classes. A random split was used to produce a training set of 2173 documents, and a test set of 693 documents, as summarized in Table 1.

4.2 The fundamental hypotheses

In Section 3, we have introduced three approaches to cross-modal retrieval, which we denoted by *correlation matching* (CM), *semantic matching* (SM), and *semantic correlation*

[1]https://lear.inrialpes.fr/people/dorko/downloads.html



(a) Text Query (b) Image Query (c) Average Performance

Figure 3: MAP performance for each category

Table 2: Taxonomy of the proposed approaches to cross-modal retrieval.

	correlation hypothesis	abstraction hypothesis
CM	√	
SM		√
SCM	√	√

matching (SCM). These approaches are illustrated in Figure 2. They represent different choices with respect to two fundamental hypothesis in the modeling of joint images and text:

- $\mathcal{H}_1$ (**correlation** hypothesis): that explicit modelling of correlations between the two modalities is important
- $\mathcal{H}_2$ (**abstraction** hypothesis): that the use of an abstract representation for images and text is desirable.

Table 2 summarizes which hypotheses hold for each of the three approaches. The comparative evaluation of the performance of the latter on cross-modal retrieval experiments provides indirect evidence for the importance of the former to the joint modelling 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. To evaluate the importance of the two hypotheses, we conducted a number of experiments which are discussed in the remainder of this section.

4.3 Experimental Protocol

The training set of Table 1 was used to learn all mappings of Section 3. The performance of CM, SM, and SCM, was evaluated on the test set. Two tasks were considered: text retrieval using an image query, and image retrieval using a query text. In the first case, each image was used as a query, producing a ranking of all texts. In the second, the roles of images and text were reversed. In all cases, performance was measured with precision-recall (PR) curves and mean average precision (MAP). The MAP score is the average precision at the ranks where recall changes. It is widely used in the image retrieval literature [25].

An LDA text model of 10 topics, and a SIFT codebook of 128 codewords were learned in an unsupervised manner, from a large corpus of images and texts[2]. They were used to compute a topic or SIFT histogram for each text or image, respectively, i.e. the document representation in the spaces $\Re^T$ and $\Re^I$. For CM, the projection these spaces onto maximally correlated subspaces was learned with a CCA of 9 components. For SM and SCM, the semantic spaces $\mathcal{S}_T$ and $\mathcal{S}_I$ produced by logistic regression were 10-dimensional probability simplexes. Since, in these spaces, each point represents a multinomial distribution over the Wikipedia categories, we refer these points as *semantic multinomials* (SMN). Each image and text is thus represented as a SMN.

[2]The performance of the system was not very sensitive to the choice of these parameters

Table 3: Different Distance Metric (MAP Scores)

Experiment	Distance Metric	Image Query	Text Query
CM	NC	**0.249**	**0.196**
CM	L2	0.235	0.181
CM	L1	0.226	0.179
SM	NC	**0.225**	**0.223**
SM	L1	0.165	0.221
SM	KL	0.177	0.221
SCM	NC	**0.277**	**0.226**
SCM	KL	0.241	**0.226**
SCM	L1	0.228	**0.226**

Table 4: Retrieval Performance (MAP Scores)

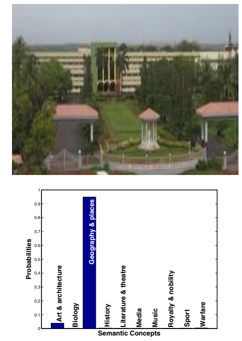
Experiment	Image Query	Text Query	Average
Random	0.118	0.118	0.118
CM	0.249	0.196	0.223
SM	0.225	0.223	0.224
SCM	**0.277**	**0.226**	**0.252**

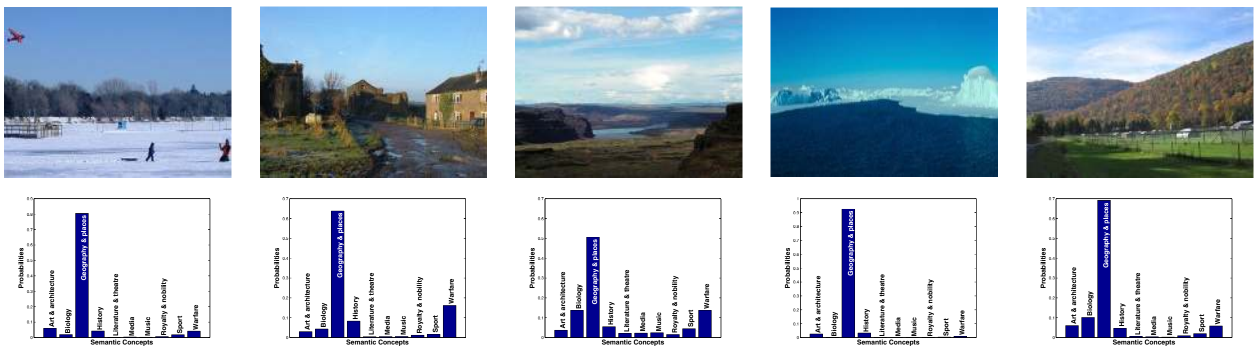
Various distance functions were used in (2) and (4). Table 3 presents the MAP scores obtained with the $L1$ distance, normalized correlation (NC), Kullback-Leibler divergence (KL), - for SM and SCM - and $L2$ distance, NC, KL - for CM. Since, in all cases, NC has the best performance, it was adopted in the remaining experiments.

4.4 Cross-modal Retrieval

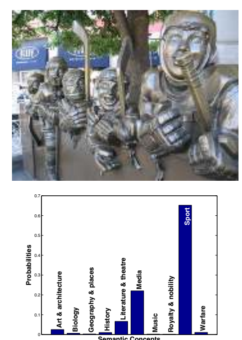
Table 4 summarizes the MAP scores obtained for the three cross-modal retrieval approaches, as well the chance-level performance. The table includes scores for both text retrieval from an image query, and image retrieval from a text

The pre-collegiate medium of instruction in schools is predominantly Kannada, while English and Kannada are predominant languages in private schools. Additionally, other media of instruction exist in Mangalore. The medium of instruction in educational institutions after matriculation in colleges is English. Recently, a committee of experts constituted by the Tulu Sahitya Academy recommended the inclusion of Tulu (in Kannada script) as a medium of instruction in education.
Schools and colleges in Mangalore are either government-run or run by private trusts and individuals. The schools are a¡liated with either the Karnataka State Board, Indian Certi↓cate of Secondary Education (ICSE), or the Central Board for Secondary Education (CBSE) boards. After completing 10 years of schooling in secondary education, students enroll in Higher Secondary School, specializing in one of the three streams â Arts, Commerce or Science. Since the 1980s, there have been a large number of professional institutions established in a variety of ↓elds including engineering, medicine, dentistry, business management and hotel management. The earliest schools established in Mangalore were the Basel Evangelical School (1838) and Milagres School (1848). The Kasturba Medical College established in 1953, was India's ↓rst private medical college. Popular educational institutions in the city are National Institute of Technology (Karnataka), KS Hegde Medical Academy,Father Muller Medical College, St. Aloysius College, Canara College,Canara Engineering College, S.D.M. College and St. Joseph Engineering College. The Bibliophile's Paradise, a hi-tech public library run by the Corporation Bank, is located at Mannagudda in Mangalore. Mangalore University was established on September 10, 1980. It caters to the higher educational needs of Dakshina Kannada, Udupi and Kodagu districts and is a National Assessment and Accreditation Council (NAAC) accredited four-star level institution.





Frank J. Selke served as Chairman of the selection committee from 1960 until 1971, when he resigned because of the induction of Harvey "Busher" Jackson. Jackson, known for his o↑-ice lifestyle, had died in 1966 of liver failure. Selke would not condone the induction and even tried to block it because he considered Jackson a poor role model."Honoured members: the Hockey Hall of Fame", p. 91
On March 30, 1993, it was announced that Gil Stein, who at the time was the president of the National Hockey League, would be inducted into the Hall of Fame. There were immediate allegations that he had engineered his election through manipulation of the hall's board of directors. Due to these allegations, NHL commissioner Gary Bettman hired two independent lawyers, Arnold Burns and Yves Fortier, to lead an investigation. They concluded that Stein had "improperly manipulated the process" and "created the false appearance and illusion" that his nomination was the idea of Bruce McNall. They concluded that Stein pressured McNall to nominate him and had refused to withdraw his nomination when asked to do so by Bettman. There was a dispute over McNall's role and Stein was "categorical in stating that the idea was Mr. McNall's." They recommended that Stein's selection be overturned, but it was revealed Stein had decided to turn down the induction before their announcement.
In 1989, Alan Eagleson, a long time executive director of the National Hockey League Players Association, was inducted as a builder. He resigned nine years later from the Hall after pleading guilty to mail fraud and embezzling hundreds of thousands of dollars from the NHL Players Association pension funds."Honoured members: the Hockey Hall of Fame", p. 167 His resignation came six days before a vote was scheduled to determine if he should be expelled from the Hall. Originally, the Hall of Fame was not going to become involved in the issue, but was forced to act when dozens of inductees, including Bobby Orr, Ted Lindsay and Brad Park, campaigned for Eagleson's expulsion, even threatening to renounce their membership if he was not removed. He became the ↓rst member of a sports hall of fame in North America to resign.



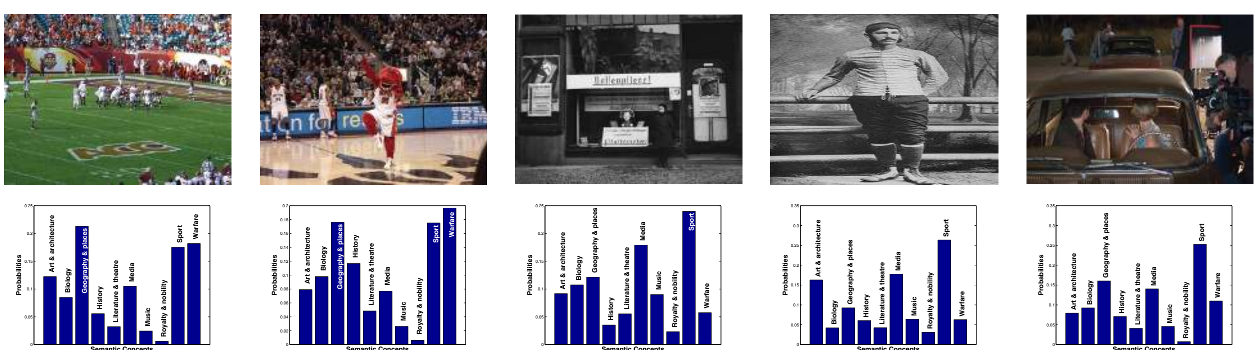
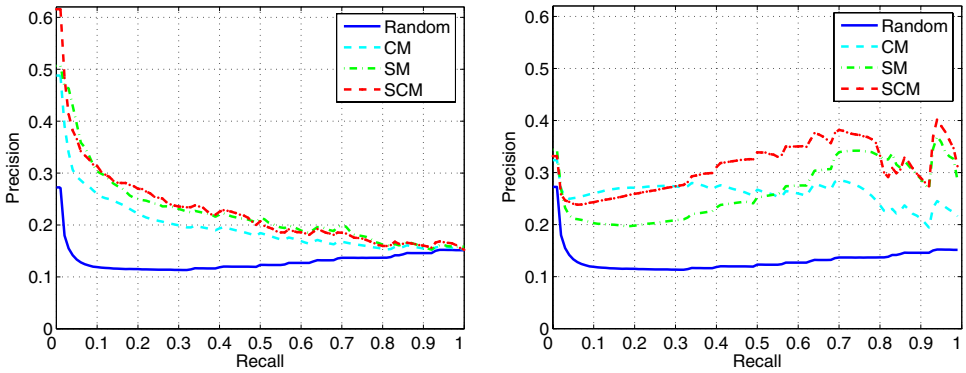


Figure 4: Two examples of text queries and the top images retrieved by SCM.



(a) Text Query (b) Image Query

Figure 5: Precision recall curves

query, and their average. The table provides evidence in support of the two fundamental hypotheses. First, both modeling of correlations (CM) and abstraction (SM) lead to a non-trivial improvement over chance-level performance. In both cases, the average MAP score, ~ 0.22, is close to double that of random retrieval ($\sim .12$). Second, the combination of correlation modeling and abstraction (SCM) further improves the MAP score to 0.252. This suggests that the hypothesis are *complementary*: not only there is a benefit to correlation modeling and abstraction, but the *best performance is achieved when the two are combined*.

Table 4 also shows that the gain of SCM holds for both forms of cross-modal retrieval, i.e. image and text query. Figure 3 shows the MAP scores achieved per category by all approaches. Notice that, for most categories, the average MAP of SCM is at least competitive with those of CM and SM. Figure 4 shows two examples of text queries and the top images retrieved by SCM. In each case, the query text and groundtruth image are shown at the top, together with the semantic multinomial computed from the query. The top five retrieved images are shown at the bottom, along with their semantic multinomials. Note that SCM retrieves these images because they are perceived by the retrieval system as belonging to the category of the query text ("Geography & Places" at the top, "Sports" at the bottom). This can be seen from the semantic multinomials, which SCM uses to find the closest matches.

Further analysis of the results is presented in Figure 5, which shows the PR of all approaches for both image and text queries. It is clear that SM, CM, and SCM all consistently improve on random retrieval at all levels of recall. These plots also provide some insight into the improvements due to the semantic representation. Comparing the PR curves of SCM and CM, shows that SCM attains higher precision at all levels of recall for text queries (left). However, for image queries (right), the improvement is only substantial at higher levels of recall. It appears that, for the latter, abstraction effectively widens the scope of matches. Initially, both methods tend to retrieve texts that are related specifically to the query image. However, whereas CM fails to generalize this to the query category, SCM is able to also retrieve texts whose relation to the query is only categorical. This seems to indicate that abstraction is especially important for exploratory tasks.

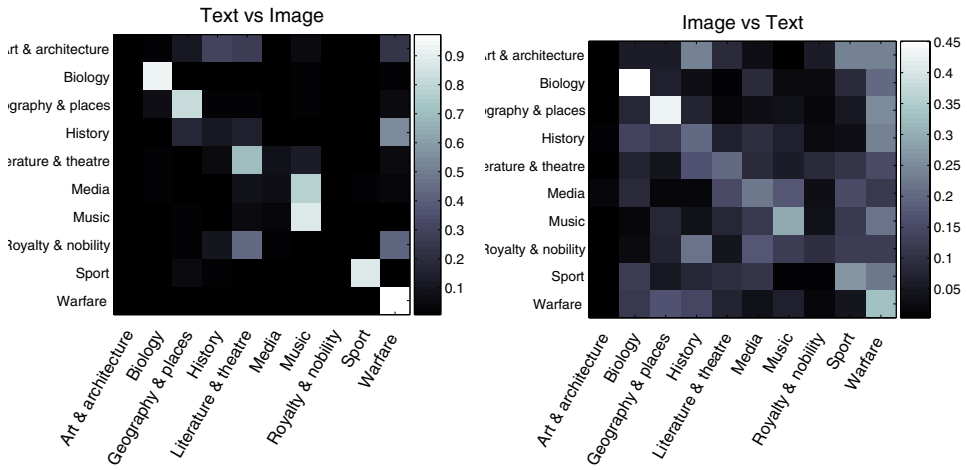
Figure 6 presents confusion matrices for illustrative purposes. The queries are classified into the class of highest MAP[3] as computed by SCM. Rows refer to true categories, and columns to category predictions. Text queries are considered on the left and image queries on the right. Note that the confusion matrix for the former is cleaner than that of the latter. This is likely due to the fact that category labels apply more directly to text than images. Consider the example article of Figure 1, where the text suggests a clear fit to the "History" category, but the image of a church could also fit into the "Art & Architecture" or "Geography & Places" categories. Note that for many of the misclassified queries, the misclassification is reasonable. This is especially true for text queries. For instance, the non-visual arts of "Literature", "Media", and "Music" tend to be confused with each other, but not to unrelated categories, like "Sports". Similarly, the fact that "History" and "Royalty" are commonly misclassified as "Warfare" is not surprising, since these three categories share similar words and imagery.

4.5 Comparison with image retrieval systems

Since the comparison to chance-level retrieval is somewhat unsatisfactory, the performance of SCM-based cross-modal retrieval was also compared to familiar benchmarks on the problem of unimodal image retrieval (where both query and retrieved documents are images). To perform this comparison, we note that a cross-modal retrieval system can be adapted to perform unimodal retrieval in two ways. In the first, a query image is complemented with a text article, and the latter is used to rank the images in the retrieval set. Note that the query image is not used, and the query text serves as its proxy. In the second, images in the retrieval set are complemented by text articles, which are ranked by similarity to the query image. In this case, the images in the retrieval set are not used, and the text articles serves as a proxy for them.

Table 5 compares the retrieval performances of these two cross-modal approaches with a number of state-of-art unimodal image retrieval methods. These include the method of [30], which represents images as distributions of SIFT features, and that of [25], where the images are projected on a semantic space similar to that used in this work. In both cases, images are retrieved by similarity of the associated probability distributions. Note that, as noted in [25], the

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



(a) Text Query (b) Image Query

Figure 6: Category-level confusion matrices as computed by classifying query to the class of highest MAP.

Table 5: Content based image retrieval.

Experiment	Model	MAP Score
CCA + SMN (Proxy Text Ranking)	Log. Regression	**0.277**
CCA + SMN (Proxy Text Query)	Log. Regression	**0.226**
Image SMN [25]	Gauss Mixture	0.161
Image SMN	Log. Regression	0.152
Image SIFT Features [30]	Gauss Mixture	0.135
Image SIFT Features	Histogram	0.140
Random	-	0.117

more abstract semantic representation also improves performance on the unimodal image matching task: using SMNs over SIFT features improves the MAP score from 0.140 to 0.161. However, all unimodal retrieval approaches have very limited gains over chance-level performance. This illustrates the difficulty of image matching on the dataset used in this work, where there is a wide variety of images in each class. The two cross-modal approaches significantly improve the retrieval performance, achieving MAP scores of 0.226 and 0.277 for the proxy text query and proxy text ranking respectively[4]. This indicates that there is a significant benefit in approaching the classical image retrieval problem from a cross-modal point of view, whenever images are part of larger documents (e.g. web pages). Figure 7 presents some examples, for the case where the query image was used to rank the text articles in the retrieval set. Shown in the figure is the query image and the images corresponding to the top retrieved text articles.

5. REFERENCES

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

[2] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 127–134. ACM, 2003.

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

[4] 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*, 29(3):394–410, 2007.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, page 22. Citeseer, 2004.

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

[7] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[8] H. Escalante, C. Hérnadez, L. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179. ACM, 2008.

[9] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, volume 2, 2004.

[10] D. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley-Interscience, 2000.

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

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference*, page 126. ACM, 2003.

[4]These numbers are slightly different from those of the previous section, as the retrieval set no longer includes the query image itself or the corresponding text article.

Figure 7: Some examples of image queries (framed image on the far-left column is the query object) and the corresponding top retrieved images (as ranked by text similarity).

[13] I. Jolliffe. *Principal component analysis*. Springer verlag, 2002.

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

[15] A. Laub. *Matrix analysis for scientists and engineers*. Siam, 2005.

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

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] C. Meadow, B. Boyce, D. Kraft, and C. Barry. *Text information retrieval systems*. Emerald Group Pub Ltd, 2007.

[19] 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.

[20] A. Nakagawa, A. Kutics, K. Tanaka, and M. Nakajima. Combining words and object-based visual features in image retrieval. In *Proceedings 12th International Conference on Image Analysis and Processing*, pages 354–359, 2003.

[21] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. *CLEF working notes*, 2009.

[22] T. Pham, N. Maillot, J. Lim, and J. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 439–444. ACM, 2007.

[23] A. Quattoni, M. Collins, T. Darrell, and C. MIT. Learning visual representations using images with captions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[24] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 1997.

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

[26] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[27] 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*, 22(12):1349–1380, 2000.

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

[29] T. Tsikrika and J. Kludas. Overview of the wikipediaMM task at ImageCLEF 2009. In *Working Notes for the CLEF 2009 Workshop*, 2009.

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

[31] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *Proceedings of 19th international conference on pattern recognition*, 2009.

[32] T. Westerveld. Probabilistic multimedia retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference*, page 438. ACM, 2002.