UNIVERSITY OF CALIFORNIA, SAN DIEGO

Adaptation of Visual Models with Cross-modal Regularization

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Jose Costa Pereira

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Serge Belongie
Professor Sanjoy Dasgupta
Professor Kenneth Kreutz-Delgado
Professor Gert Lanckriet

2015

The dissertation of Jose Costa Pereira is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2015

DEDICATION

To my Mother.

EPIGRAPH

Use a picture. It's worth a thousand words.

—Arthur Brisbane

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Nuno Vasconcelos for his teachings and guidance in my, yet short, research career. I would like to extended my gratitude to the members of my doctoral committee, Professors Serge Belongie (now at Cornell University), Sanjoy Dasgupta, Kenneth Kreutz-Delgado, and Gert Lanckriet, for their exciting lectures and seminars, as well as support and active involvement in my research progress.

A heart-felt word of gratitude to the person that triggered all this: my late mentor, Professor André Puga. I was fortunate enough to know him well. He believed in me even before I did, his strong conviction that I should pursue a Ph.D. was an enormous motivation and a true leverage in my return to school. He was involved in everything; selecting a University, apply for funding, choosing an advisor. His example in life was, and still is, inspirational to me. Thank you for everything, André.

I gratefully acknowledge the support given by my country, Portugal, in the form of a Fellowship SFRH/BD/40963/2007 awarded by the Foundation for Science and Technology (FCT)[1]. I also acknowledge the support from NSF[2] given to my advisor through grant CCF-0830535 that partially funded this work.

In the Statistical and Visual Computing Laboratory (SVCL), where I've developed my work, I belong to a generation of students that has known everyone all the way back to "*patient zero*": Dashan Gao, Antoni Chan, Hamed Masnadi-Shirazi, Sunhyoung Han, Vijay Mahadevan, Nikhil Rasiwasia, Kritika Muralidharan, Mulloy Morrow, Ehsan Saberian, Weixin Li and Mandar Dixit. To all I would like to thank their help and support. To the more recent generation of students:

[1] *Fundação para a Ciência e Tecnologia* (portuguese), a Foundation for Science and Technology from the Portuguese government.

[2] National Science Foundation, an independent US federal agency.

Can Xu, Song Lu, Si Chen, Pedro Morgado, Yingwei Li and Bo Liu, I extend my thanks and wish them safe travels in their path. Many people worked at the ECE department in the course of these years. In the person of Travis Spackman I would like to address my gratitude to the staff for their help with all administrative issues coming my way. At University of Porto, where I have spent a few months on a research project, I would like to thank Professor Maria Cristina Ribeiro and, through her, extend my gratitude to all the researchers at the InfoLab in the School of Engineering.

It may seem like a common place, but graduation is never easy. Fortunately for me, a handful of instrumental people have kept me on track and allowed this day to happen. For different reasons but equally important, I would like to thank all of them. Nikhil Rasiwasia was my research mentor when I first came to the lab, my dissertation builds on his Ph.D. work, and I have authored my first paper at UCSD with him. Vijay Mahadevan, who was patient enough to bear with my multiple complaints, and gave me valuable advice on how to succeed in the program. Ankit Srivastava one of the most brilliant and humble students I've met, and to whom I have the pleasure of calling friend; we have had some of the deepest and most entertaining conversations at 'The art of espresso' and 'The blind lady ale house'; many times in the company of our good friend Adarsh Krishnamurthy. Fr. John Paul Forte who was always so patiently available to listen to my problems and questions. Least but certainly not least my adorable roommates over the years; in particular, Amirali Shayan who took me in when I first arrived, this was just the beginning of a beautiful friendship who will endure for life, and Jefy Jayamon for endless conversations about life in general; thank you both for being who you are and always being there for me. To all my Portuguese friends in San Diego, André Barbosa and his wonderful family who provided all the support in my first few days in San Diego; Rui Moreira with whom I share roots in Leça da Palmeira, he was always there, good and bad times alike; Nuno Camboa, Filipe Jacinto, Bibiana Ferreira and Pedro Resende who welcomed me into their group; and Rui Carlos Sá for all great advice about life as a researcher. They all made distance seem a little easier to tolerate. My gratitude extends to many other wonderful friends I was fortunate to meet in San Diego and surroundings, they all made my life a marvelous experience.

Certainly not less important, all the friends I left in Portugal; who have unconditionally supported me in this challenge, and who have always been there for me, in the course of these years lending a helping hand. Miguel Lopo, Manel Gil Fernandes, Miguel Novaes, João Maria Fonseca, António Maria Novaes, Francisco Taveira, João Azeredo, Alexandre Garrett and Miguel Lencastre. You may not realize this, but I've appreciated every single conversation we've had. They were instrumental in providing the necessary motivation to continue strong and focused on my goals. I will see you all, hopefully, very soon.

To Mercedes who has so patiently supported me in the course of these last two years, I have but words of profound gratitude.

Lastly, I would like to thank my family. My grandmother Mizé with whom I've became pen-friend, she has always had a kind word of comfort to me in spite of all the hurdles life would lay in her path. To Mom and Dad, who I owe life itself, they too have had a similar experience of living abroad – Norwich (UK) – 37 years ago. I can only imagine how they suffered with my departure, yet they were always thoughtful and supportive making it possible for me to reach my goals. My brothers, Duarte and Alexandre, for their example in life and unconditional support, and my little sister Mizé for her strength and unwavering faith in me. They mean the world to me. I could never have made it without them.

The text of Chapters 2 and 3 is, in part, based on the material as it appears in: IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 36(3), pp. 521-535, March 2014, ‘On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval’, J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 4 is, in part, based on the material as it appears in: IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 36(3), pp. 521-535, March 2014, ‘On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval’, J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos, and Elsevier, Computer Vision and Image Understanding, Vol. 124, pp. 123-135, July 2014, ‘Cross-modal Domain Adaptation for Text-based Regularization of Image Semantics’, J. Costa Pereira, N. Vasconcelos. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 5 is, in part, based on the material as it appears in: Elsevier, Computer Vision and Image Understanding, Vol. 124, pp. 123-135, July 2014, ‘Cross-modal Domain Adaptation for Text-based Regularization of Image Semantics’, J. Costa Pereira, N. Vasconcelos. The dissertation author was a primary researcher and an author of the cited material.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

GOVERNO DA REPÚBLICA PORTUGUESA

POPH
PROGRAMA OPERACIONAL POTENCIAL HUMANO

QREN
QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL
PORTUGAL 2007.2013

UNIÃO EUROPEIA
Fundo Social Europeu

VITA

2000	Licenciatura in Computer Science and Engineering, University of Porto, Portugal
2000–2005	Telecom Engineer, Vodafone, Portugal
2003	M.A. Applied Mathematics (Computational Methods in Science and Engineering), University of Porto, Portugal
2005–2008	Telecom Engineer, Alcatel-Lucent, Portugal
2009–2015	Research Assistant, Statistical and Visual Computing Laboratory, Department of Electrical and Computer Engineering, University of California San Diego, USA
2011	M.S. Electrical and Computer Engineering (Intelligent Systems, Robotics and Control), University of California San Diego, USA
2015	Ph.D. Electrical and Computer Engineering (Intelligent Systems, Robotics and Control), University of California San Diego, USA

PUBLICATIONS

J. Costa Pereira and N. Vasconcelos, *'Cross-modal Domain Adaptation for Text-based Regularization of Image Semantics'*, Computer Vision and Image Understanding (CVIU), Elsevier, vol. 124, pp. 123-135, 2014.

J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy and N. Vasconcelos, *'On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval'*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, n. 3, pp. 521-535, 2014.

J. Costa Pereira, J. Luque and X. Anguera, *'Sentiment Retrieval on Web Reviews using Spontaneous Natural Speech'*, IEEE Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014.

J. Costa Pereira and N. Vasconcelos, *'On the Regularization of Image Semantics by Modal Expansion'*, IEEE Proc. of International Conference on Computer Vision and Pattern Recognition (CVPR), Providence (RI), USA, 2012.

V. Mahadevan, C-W. Wong, J. Costa Pereira, T.T. Liu, N. Vasconcelos and L. Saul, *'Maximum Covariance Unfolding: Manifold Learning for Bimodal Data'*, Advances in Neural Information Processing Systems (NIPS), Granada, Spain, 2011.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy and N. Vasconcelos, *‘A New Approach to Cross-Modal Multimedia Retrieval’*, ACM Proc. of International Conference on Multimedia (ACMMM), Florence, Italy, 2010.

J. Costa Pereira and A. Puga, *‘Blind Source Separation’*, In Proceedings of the Computational Methods Conference, Lisbon, Portugal, 2004.

J. Costa Pereira and A. Puga, *‘Blind Source Separation using Independent Component Analysis’*, M.A. Thesis, Porto, Portugal, 2003.

J. Costa Pereira, H. Necho and F. Restivo, *’Voice over IP in Vodafone-Telecel’* Licenciatura Report, Porto, Portugal, 2000.

ABSTRACT OF THE DISSERTATION

Adaptation of Visual Models with Cross-modal Regularization

by

Jose Costa Pereira

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics,
and Control)

University of California, San Diego, 2015

Professor Nuno Vasconcelos, Chair

Semantic representations of images have been widely adopted in Computer Vision. A vocabulary of concepts of interest is first identified and classifiers are learned for the detection of those concepts. Images are classified and mapped to a space where each feature is a score for the detection of a concept. This representation brings several advantages. First, the generalization from low-level features to concept-level enables similarity measures that correlate much better with user expectations. Second, because semantic features are, by definition, discriminant for tasks like image categorization, the semantic representation enables a solution

for such tasks with low-dimensional classifiers. Third, the semantic representation is naturally aligned with recent interest on contextual modeling. This is of importance for tasks such as object recognition, where detection of contextually related objects has been shown to improve detection of certain objects of interest, or semantic segmentation, where the coherence of segment semantics can be exploited to achieve more robust segmentations. Lastly, due to their abstract nature, semantic spaces enable a unified representation for data from different content modalities, *e.g.* images, text, or audio. This opens up a new set of possibilities for multimedia processing, enabling operations such as cross-modal retrieval, or image de-noising by text regularization. This unified representation for multi-modal data is the starting point of the proposed framework on adaptation of visual models with cross-modal regularization.

We start by pointing the problems in computing similarity on heterogeneous data, proposing two fundamental hypotheses to deal with those issues. One, learning a space that maximizes the correlation on the (heterogeneous) data; two, learning a representation where data lies at a higher level of abstraction. Empirical evidence is shown in favor of each hypothesis; furthermore the hypotheses are shown to be complementary. We follow on the (semantic) abstraction hypothesis for a deeper understanding on the robustness of these representations and to study the richness of this space, as it highly influences the discriminative power of such descriptors.

It has been shown that categories unknown to the semantic space, when represented in it, exhibit a pattern of co-occurring concepts that describe them accurately and sensibly; *e.g.* the concept of fishing might not belong to the semantic space and instead be represented by the set water, boat, people and gear. Even though the amount of labeled data continues to increase with ongoing efforts from different research communities, it is a challenging task to build a semantic space that is universal. We show evidence towards robustness of representations in the semantic space.

Noting that images are frequently published on the web together with loosely related text, we use the semantic representations described above to introduce the theoretical principles to a feature regularizer for image semantic representations based on auxiliary data. This proves very effective on improving retrieval precision and recall in the task of content-based image retrieval (CBIR). Its results are compared to recently proposed methods, achieving significant gains in three benchmark datasets, raising the bar of state-of-the-art performance for image retrieval.

Chapter 1

Understanding Images

"*Use a picture. It's worth a thousand words*", newspaper editor Arthur Brisbane once said in a discussion on journalism and publicity [1]. It's uncontroversial to say that a diagram or a picture help conveying complex ideas to an audience. These pictorial descriptions are used frequently in different scenarios: from science to daily newspapers, from commercials to books. As an extreme example, one could argue that movies are nothing but a set of images that tell a story previously written in a novel or an original screenplay.

Going as far back as the Royal Library of Alexandria (*circa* 300 B.C.), civilization has always craved for knowledge organization. Even though there is little consensus about the crumble of that particular stronghold of the ancient world, the intuition is that knowledge which is not organized cannot be retrieved and is, therefore, as good as ignorance. Modern Libraries are no longer housed in magnificent buildings, and knowledge is certainly no longer stored in papyrus scrolls. Rather, information is scattered across digital platforms in many different places, through an intricate mesh of connections that we know as the *web*. Search engines are the modern version of librarians, and it is imperative that information can be properly understood so that it can be archived and accessed efficiently.

Visual contents are particularly troublesome to organize for two main reasons, that we now elaborate. One; they have become increasingly popular in recent years, either through social media (*e.g.* Facebook, Instagram), photo sharing sites (*e.g.* Flickr, Imgur), or mainstream media channels (*e.g.* BBC, NYTimes). The potential sources from where visual information can emerge today is, nothing less than massive. A few numbers reported by a statistical volume on the handset industry [2] are impressive. In 2014 alone, a total of 1.8 billion new cameras for consumers were sold worldwide (excluding security cameras and webcams). The demand for stand-alone cameras is decreasing (5%) and the vast majority of the new cameras (95%) are part of a mobile phone. These have therefore inherent portability and connectivity to the *web*. Combined with the global existing portfolio, operable cameras today sum up to an estimate of some impressive 5.8 billion, out of which 4 billion are in use by unique owners. From this global base of installation, between the stand-alone digital cameras (11%) and cellphone cameras (89%) it is estimated that, in 2014 for the first time, more than 1 trillion photos were taken. This brings mankind's cumulative picture production to a total 5.7 trillion photographs taken since the first camera was invented. Second; automatic description of visual contents still falls short when compared to human annotation. Therefore, proper storage and retrieval of multimedia contents still requires a fair amount of human intervention. Considering the amount of pictures proliferating on the *web* as described above, makes this a very concerning issue.

It becomes clear, from the above landscape, the pressing need to understand

and organize visual information in a way that makes these contents manageable.

1.1 Contribution of the thesis

We know that images have a high degree of variability. A single concept can have very different visual appearances. In Figure 1.1 two simple concepts, “sun” and “trees”, are shown in great visual diversity. Understanding the content of scenes such as those, is a very challenging problem for machines, yet humans can do it almost instantaneously. In Computer Vision, pattern analysis and classification try to mimic this behavior. In a somewhat broad sense, we can identify two major branches of study. On one hand, people have devoted significant efforts to (increasingly) sophisticated algorithms to address the problem of classification in scenarios where data exhibits this much variability. It’s the search for generalization that is always present whether in discriminative or generative models. On the other hand, there is a sense that classification algorithms and models are already “*good enough*”, and there is just a need to craft better features. Something that captures the essence of visual contents. This is noticeably the case with the advent of deep neural networks. By simply using activations of these networks as features, significant improvements are achieved in the performance of almost any classification architecture.

Going back to how humans seemingly perform the task of (visual) classification with great ease, it is important to note that the human visual system is trained for years, during childhood and adolescence. And a large number of diverse (training) examples is used to achieve this level of performance. Humans also make use of more than just visual sensory to learn about concepts. We recognize water to be refreshing because we’ve tasted it; fire has burned us the first time we tried to

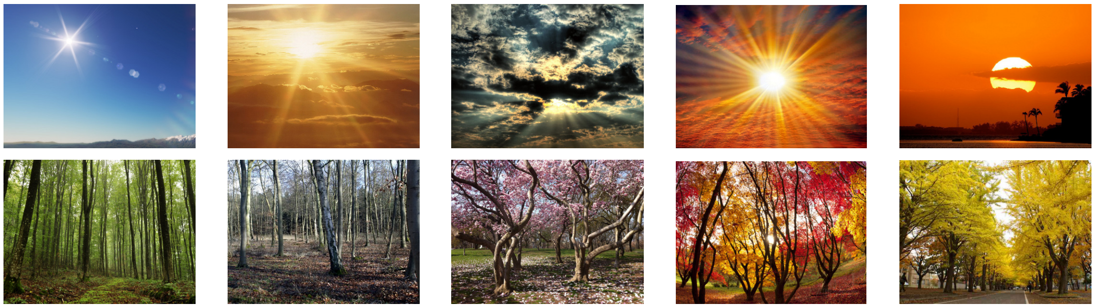


Figure 1.1: The diversity of visual contents makes scene understanding a very challenging problem. Two simples concepts – "sun" and "trees" (one per line) – are shown in the pictures above with very diverse appearances.

touch it; and we all know that the thorns on a beautiful rose can be a displeasing experience. Following this line of reasoning, recognizing an object doesn't necessarily come only from seeing it multiple times. In this dissertation I propose a framework that achieves the incorporation of knowledge from other sources of information (*i.e.* senses) into the visual representation of objects or scenes. Evidence is shown in favor of a certain type of image representations, pointing two important aspects of it: it is an abstract representation that easily allows for computation of similarity to other sources of information (*e.g.* audio or text), and it provides the tools for a novel approach of regularization of image features for improved retrieval accuracy.

1.1.1 Similarity on Heterogeneous Data

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, for example, using an image to search for texts. A mathematical formulation is proposed, equating the design of cross-modal retrieval systems to that of isomorphic feature spaces for different content modalities. Two hypotheses are then investigated regarding the fundamental attributes of these spaces. The first is that low-level cross-modal correlations should be accounted for.

The second is that the space should enable semantic abstraction. Three new solutions to the cross-modal retrieval problem are then derived from these hypotheses: correlation matching (CM), an unsupervised method which models cross-modal correlations, semantic matching (SM), a supervised technique that relies on semantic representation, and semantic correlation matching (SCM), which combines both. An extensive evaluation of retrieval performance is conducted to test the validity of the hypotheses. All approaches are shown successful for text retrieval in response to image queries and vice versa. It is concluded that both hypotheses hold, in a complementary form, although evidence in favor of the abstraction hypothesis is stronger than that for correlation.

1.1.2 Robustness of Semantic Representations

In the query-by-semantic-example image retrieval paradigm, images are ranked by similarity of semantic descriptors. Much like in the SM cross-modal retrieval, these descriptors are obtained by classifying each image with respect to a pre-defined vocabulary of semantic concepts. Context and multi-modality are introduced as potential sources to further improve semantic descriptors of images. We thoroughly test variations of semantic representations in the scenario of cross-modal retrieval. These are shown to be robust to design decisions made in the semantic space. This leads to the next contribution of this thesis, where the problem of improving the accuracy of image semantic descriptors through cross-modal regularization is considered.

1.1.3 Regularization of Image Semantics

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

1.2 Organization of the thesis

Chapter 2 presents an introduction to multimedia information retrieval. It is an overview of related work, covering different retrieval paradigms. In Chapter 3 the problem of measuring similarity between different modalities is formulated. Solutions to this problem are proposed based on two fundamental hypotheses: (i) a projection onto a common space that maximizes correlation between data from different sources, and (ii) a mapping to a space of higher abstraction, where data from different sources can be represented. The mehrits of all solutions are discussed and a semantic representation is adopted for images. This representation is further studied in Chapter 4. Finally on Chapter 5 we propose a learning framework, for regularization of image representations on the semantic space, using auxiliary sources of information. The algorithms proposed in this chapter heavily rely on the cross-modal similarity fundaments introduced on Chapter 3.

Chapter 2

Multimedia Information Retrieval

Classical approaches to information retrieval are *uni-modal* [1] in nature [96, 102, 66]; *i.e.* text repositories are searched with text queries, image databases with image queries, and so forth. This (uni-modal) retrieval paradigm is of limited use in the modern information landscape, where multimedia content is ubiquitous. Due to this; *multi-modal* modeling, representation, and retrieval have been extensively studied in the multimedia literature [97, 39, 65, 6, 25, 105, 22, 50]. In multi-modal retrieval systems, queries combining multiple content modalities (*e.g.* images and sound of a music video-clip) are used to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). These efforts have become increasingly widespread, due in part to large-scale research and evaluation efforts, such as TRECVID [101] and ImageCLEF [110], involving datasets that

[1] In the present context, the word modality is used to refer to a particular source of information (*e.g.* text, music, image).

span multiple data modalities. However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario. Typically, the different modalities are fused into a representation that does not allow individual access to any of them, *e.g.* some form of dimensionality reduction of a large feature vector that concatenates measurements from two (or more) modalities. Classical uni-modal techniques are then applied to the low-dimensional representation.

In the remainder of this chapter, we will briefly overview the most relevant related work, covering different retrieval paradigms and systems proposed in a recent past.

2.1 Unimodal Retrieval

The problems of image or text retrieval, for example, have been the subject of extensive research in the fields of information retrieval, computer vision, and multimedia [22, 102, 101, 77, 72]. In all these areas, the emphasis has been on uni-modal approaches, where query and retrieved documents share a single modality [96, 95, 116]. For example, in [95] a query text, and in [116] a query image is used to retrieve similar text documents and images, based on low-level text (*e.g.* words) and image (*e.g.* DCTs) representations, respectively. However, this is not effective for all problems. The existence of a well known *semantic gap,* between machine image representations and those adopted by humans, severely hampers the performance of uni-modal image retrieval systems [102].

2.1.1 Annotations

In general, successful retrieval from large-scale image collections requires that the latter be augmented with text meta-data provided by human annotators. These manual annotations are typically in the form of a few keywords, a small caption, or a brief image description [77, 110, 101]. When this meta-data is available, the retrieval operation tends to be uni-modal, ignoring the images — the text meta-data of the query image is simply matched to the text meta-data available for images in the database. Because manual image labeling is labor-intensive, recent research has addressed the problem of automatic image labeling[2].

2.1.2 Labeling

A common assumption is that images can be segmented into regions, which can be described by a small (word) vocabulary. The focus is then on learning a probability model that relates image regions and words. This can be done by learning a joint probability distribution for words and visual features, *e.g.* using latent Dirichlet allocation models [5], probabilistic latent semantic analysis [74], histograming methods [51], or a combination of Bernoulli distributions for text and kernel-based models for visual features [60, 36]. Alternatively, it is possible to use categorized images to train a dictionary of concept models, *e.g.* Gaussian mixtures [13] or two-dimensional hidden Markov models [127], in a weakly supervised manner. The extent of association between images and concepts or words is measured by the likelihood of each image under these models. All these methods assume that each image or image region is associated with a single word.

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

2.1.3 Semantic Space

An alternative representation, where images are modeled as weighted combinations of concepts in a pre-defined vocabulary, is proposed in [85]. Statistical models of the distribution of low-level image features are first learned for each concept. The posterior probability of the features extracted from each image, under each of the concept models, is then computed. The image is finally represented by the vector of these posterior concept probabilities. This can be interpreted as a vector of semantic features, establishing a semantic feature space where each dimension is associated with a vocabulary concept. Figure 2.1 illustrates how this descriptor, denoted *semantic multinomial* (SMN), maps the image into the *semantic space*. All standard image analysis/classification tasks can then be conducted in the latter space, at a higher level of abstraction than that supported by low-level feature spaces. For example, image retrieval is formulated as retrieval by *semantic similarity,* by combining the semantic space with a suitable similarity function [85]. This allows assessments of image similarity in terms of weighted combinations of vocabulary words, and substantially extends the range of concepts that can effectively

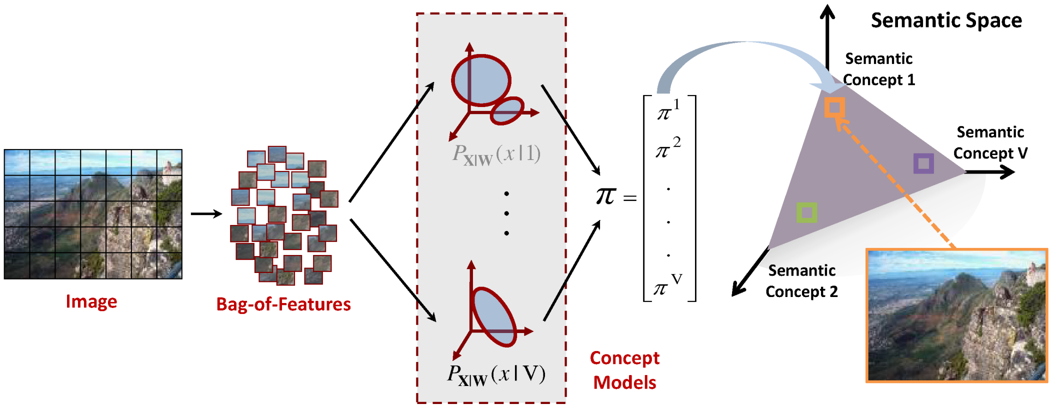


Figure 2.1: Semantic space representation of images. An image is decomposed into a bag-of-features, and represented by the vector of its posterior probabilities with respect to the concepts in a semantic vocabulary $\mathcal{V}$.

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 [118].

2.2 Multi-modal Retrieval

In parallel with these developments, advances have been reported in multimodal retrieval systems [77, 110, 101, 105, 22, 50, 25]. These are extensions of the classic uni-modal systems, where a common retrieval system integrates information from various modalities. This can be done by fusing features from different modalities into a single vector [128, 79, 30], or by learning different models for different modalities and fusing their predictions [126, 54]. One popular approach is to concatenate features from different modalities and rely on unsupervised structure discovery algorithms, such as latent semantic analysis (LSA), to find multi-modal statistical regularities. A good overview of these methods is given in [30], which also discusses the combination of uni-modal and multi-modal retrieval systems. Multimodal integration has also been applied to retrieval tasks including audio-visual content [76, 37]. In general, the inability to access each data modality individually (after the fusion of modalities) prevents the use of these systems for cross-modal retrieval.

2.3 Cross-modal Retrieval

To overcome these difficulties, progress has been made towards cross-modal systems. This includes retrieval methods for corpora of images and text [25, 80], images and audio [63, 136], text and audio [100], images, text, and audio [132, 136, 141, 140, 133], or even other sources of data like EEG and fMRI [68]. One popular approach is to rely on manifold learning techniques [68, 132, 136, 141, 140, 133]. These methods learn a manifold from a matrix of distances between multi-modal objects. The multi-modal distances are formulated as a function of the distances between individual modalities, which allows to single out particular modalities or ignore missing ones. Retrieval then consists of finding the nearest document, on the manifold, to a multimedia query (which can be composed of any subset of modalities). The main limitation of these methods is the lack of out-of-sample generalization. Since there is no computationally efficient way to project the query into the manifold, queries are restricted to the training set used to learn the latter. Hence, all unseen queries must be mapped to their nearest neighbors in this training set, defeating the purpose of manifold learning.

An alternative is to learn correlations between modalities [63, 125]. For example, [63] 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 [125] 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 [46]. Another approach is *re-ranking*: uni-modal retrieval is first performed using the query modality, and a second modality is used to re-rank the results [49, 73].

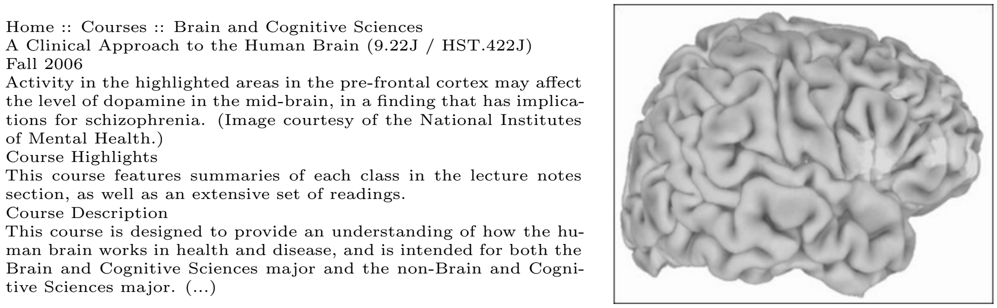
2.3.1 Rich Annotation

Despite all these advances, current retrieval systems (of any kind) tend to rely on a limited textual representation, in the form of keywords, captions, or small text snippets. We refer to these as forms of lighter annotation. This is at odds with the ongoing explosion of multimedia content on the *web*, where it is now possible to collect large sets of extensively annotated data. Examples include news archives, blog posts, or Wikipedia pages, where pictures are related to *complete* text articles, not just a few keywords. We refer to these datasets as *richly annotated*. While potentially more informative, rich annotation establishes a much more nuanced connection between images and text than light annotation. While keywords tend to be explicit image labels, many of the words in a rich text can be unrelated to the image used to illustrate it. For example, Figure 2.2 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 show in Figure 2.3. This is a course syllabus that, beyond the pictured brain, includes course information and other unrelated matters.

Martin Luther King’s presence in Birmingham was not welcomed by all in the black community. A local black attorney complained in *Time* that the new city administration did not have enough time to confer with the various groups invested in changing the city’s segregation policies. Black hotel owner A. G. Gaston agreed. A white Jesuit priest assisting in desegregation negotiations attested the “demonstrations [were] poorly timed and misdirected”.
Protest organizers knew they would meet with violence from the Birmingham Police Department and chose a confrontational approach to get the attention of the federal government. Wyatt Tee Walker, one of the SCLC founders and the executive director from 1960 to 1964, 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 (...)”

Figure 2.2: A section from the Wikipedia article on the Birmingham campaign (“History” category).

A major long-term goal of modeling richly annotated data is to recover this *latent* relationship, here exemplified between the text and image components of two different types of documents, and exploit it in benefit of practical applications. In the next chapter, we consider a richer interaction paradigm, which is denoted *cross-modal* retrieval. The goal is to build content models that enable interactivity with content *across* modalities. Such models can then be used to



Home :: Courses :: Brain and Cognitive Sciences
A Clinical Approach to the Human Brain (9.22J / HST.422J)
Fall 2006
Activity in the highlighted areas in the pre-frontal 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. (...)

Figure 2.3: Part of a Cognitive Science class syllabus from the TVGraz dataset ("Brain" category).

design cross-modal retrieval systems, where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.* audio tracks). This form of retrieval can be seen as a generalization of current content labeling systems, where a primary modality is augmented with keywords, which can be subsequently searched. Examples include keyword-based image [5, 75, 13] and song [112, 111, 29] retrieval systems. Furthermore, as will be empirically shown, some modalities have more (Shannon) entropy than others. In this thesis, we propose a regularization procedure to adapt visual models typically more prone to higher entropy levels.

Acknowledgements

The text of Chapter 2, in part, is based on the material as it appears in: 'On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval', J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 36(3), pp. 521-535, March 2014. The dissertation author was a primary researcher and an author of the cited material.

Chapter 3

Cross-Modal Similarity

In this chapter, the problem of cross-modal retrieval from multimedia repositories is considered. This problem addresses the design of retrieval systems that support queries *across* content modalities, *e.g.*, using an image to search for texts. A mathematical formulation is proposed, equating the design of cross-modal retrieval systems to that of isomorphic feature spaces for different content modalities. Two hypotheses are then investigated, regarding the fundamental attributes of these spaces. The first is that low-level cross-modal correlations should be accounted for. The second is that the space should enable semantic abstraction. Three new solutions to the cross-modal retrieval problem are then derived from these hypotheses: correlation matching (CM), an unsupervised method which models cross-modal correlations, semantic matching (SM), a supervised technique that relies on semantic representation, and semantic correlation matching (SCM), which combines both. An extensive evaluation of cross-modal 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.

3.1 Introduction

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

We exploit representations that generalize across content modalities to study two hypotheses on the joint modeling of images and text. The first, denoted the *correlation hypothesis*, is that explicit modeling of low-level correlations between the different modalities is important for the success of the joint models. The second, denoted the *abstraction hypothesis*, is that model benefits from semantic abstraction, *i.e.*, the representation of images and text in terms of semantic (rather than low-level) descriptors. These hypotheses are partly motivated by previous evidence that correlation, *e.g.*, correlation analysis on fMRI [46], and abstraction, *e.g.*, hierarchical topic models for text clustering [11] or hierarchical semantic representations for image retrieval [85], improve performance on uni-modal retrieval tasks. Three joint image-text models that exploit low-level correlation, denoted *correlation matching*, semantic abstraction, denoted *semantic matching*, and both, denoted *semantic correlation matching*, are introduced.

The correlation and abstraction hypotheses are then tested by measuring the retrieval performance of these models on two reciprocal cross-modal retrieval tasks: 1) the retrieval of text documents in response to a query image, and 2) the retrieval of images in response to a query text. These are basic cross-modal retrieval problems, central to many applications of practical interest, such as finding pictures that effectively illustrate a given text (*e.g.* illustrate a page of a story book), finding the texts that best match a given picture (*e.g.* a set of vacation accounts about a given landmark), or searching using a combination of text and images. Model performance on these tasks is evaluated with two datasets: TVGraz [53] and a novel dataset based on Wikipedia's featured articles. These experiments show that correlation modeling and abstraction yield independent benefits. In particular, the best results are obtained by a model that accounts for both low-level correlations — by performing a kernel canonical correlation analysis (KCCA) [98, 124] — and semantic abstraction — by projecting images and texts into a common semantic space [85]. This suggests that the hypotheses of abstraction and correlation are complementary, each improving the modeling in a different manner.

3.2 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.

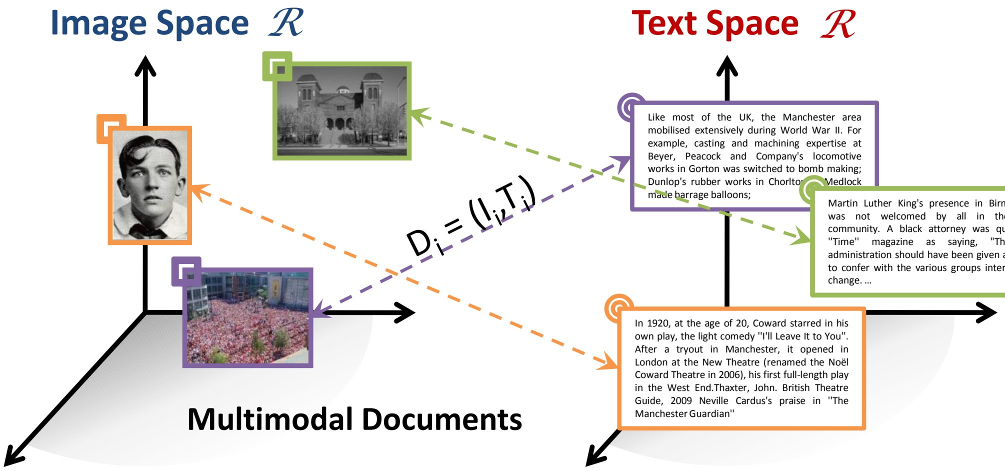
3.2.1 The problem

We consider the problem of information retrieval from a database $\mathcal{D} = \{D_1, \ldots, D_{|D|}\}$ of *documents* comprising *image* and *text* components. Such documents can be quite diverse: from a single text complemented by one or more images (*e.g.* a newspaper article) to documents containing multiple pictures and text sections (*e.g.* a Wikipedia page). For simplicity, we consider the case where each document consists of a single image and its accompanying text, *i.e.*, $D_i = (I_i, T_i)$. Images and text are represented as vectors in feature spaces $\Re^I$ and $\Re^T$, respectively, as illustrated in Figure 3.1. In this way, documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$. Given a text (image) query $T_q \in \Re^T$ ($I_q \in \Re^I$), the goal of *cross-modal retrieval* is to return the closest match in the image (text) space $\Re^I$ ($\Re^T$).

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

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

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



Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlto [illegible] Medlock made barrage balloons;

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. ...

In 1920, at the age of 20, Coward starred in his own play, the light comedy "I'll Leave It to You". After a tryout in Manchester, it opened in London at the New Theatre (renamed the Noël Coward Theatre in 2006), his first full-length play in the West End.Thaxter, John. British Theatre Guide, 2009 Neville Cardus's praise in "The Manchester Guardian"

Figure 3.1: A document (D_i) is a pair of an image (I_i) and a text (T_i) represented as vectors in feature spaces $\Re^I$ and $\Re^T$, respectively. Documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$.

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

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

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

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

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

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

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

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

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

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

3.2.2 The fundamental hypotheses

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

We exploit the cross-modal retrieval problem to test two such hypotheses regarding the joint modeling of images and text.

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

The hypotheses are tested by comparing three possibilities for the design of the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ of cross-modal retrieval. In the first case, two feature transformations map $\Re^I$ and $\Re^T$ onto *correlated* d-dimensional *subspaces* denoted as $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively, which act as $\mathcal{V}^I$ and $\mathcal{V}^T$. This maintains the level of semantic abstraction of the representation while maximizing the correlation between the two spaces. We refer to this matching technique as *correlation matching* (CM). In the second case, a pair of transformations are used to map the image 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 3.1: Taxonomy of proposed approaches to cross-modal retrieval.

	correlation hypothesis	abstraction hypothesis
CM	$\surd$	
SM		$\surd$
SCM	$\surd$	$\surd$

Table 3.1 summarizes which hypotheses hold for each of the three approaches. The comparative evaluation of the performance of these approaches on cross-modal retrieval experiments provides indirect evidence for the importance of the above hypotheses to the joint modeling of images and text. The intuition is that a better cross-modal retrieval performance results from a more effective joint modeling.

3.3 Cross-modal Retrieval

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

3.3.1 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) [24] and principal component analysis (PCA) [52]. These are members of a broader class of learning algorithms, denoted subspace learning, which are computationally efficient, and produce linear transformations that are easy to conceptualize, implement, and deploy. Furthermore, because subspace learning is usually based on second order statistics, such as correlation, it can be easily extended to the multi-modal setting and kernelized. This has motivated a number of multi-modal subspace methods. In this work, we consider *cross-modal factor analysis* (CFA), *canonical correlation analysis* (CCA), and *kernel canonical correlation analysis* (KCCA). All these methods include a training stage, where the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are learned, followed by a projection stage, where images and text are projected into these spaces. Figure 3.2 illustrates this process. Cross-modal retrieval is performed in the low-dimensional subspaces, $\mathcal{U}^I$ and $\mathcal{U}^T$.

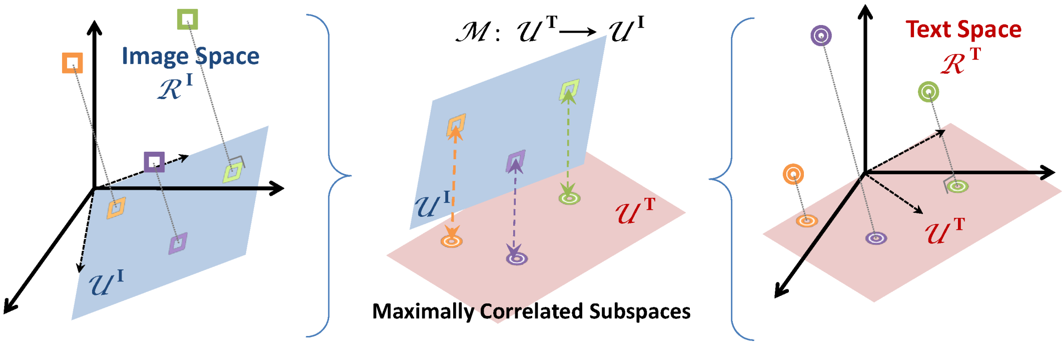


Image Space
$\mathcal{R}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{M}: \mathcal{U}^{\mathrm{T}} \longrightarrow \mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{T}}$
Maximally Correlated Subspaces
Text Space
$\mathcal{R}^{\mathrm{T}}$
$\mathcal{U}^{\mathrm{T}}$

Figure 3.2: Correlation matching. Text ($\Re^T$) and images ($\Re^I$) are projected onto two maximally correlated isomorphic subspaces $\mathcal{U}_T$ and $\mathcal{U}_I$, respectively.

Linear subspace learning

CFA seeks transformations that best represent coupled patterns between different subsets of features (*e.g.* different modalities) describing the same objects [63]. It finds the orthonormal transformations $\mathbf{\Omega}_I$ and $\mathbf{\Omega}_T$ that project the two modalities onto a shared space, $\mathcal{U}^I = \mathcal{U}^T = \mathcal{U}$, where the projections have minimum distance

$$\left|\left|X_I\Omega_I - X_T\Omega_T\right|\right|_F^2. \tag{3.1}$$

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

$$trace(X_I \Omega_I \Omega'_T X'_T), \tag{3.2}$$

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

$$X'_I X_T = \Omega_I \Lambda \Omega_T, \tag{3.3}$$

where Λ is the matrix of singular values of $X'_I X_T$ [63].

CCA [48] learns the d-dimensional subspaces $\mathcal{U}^I \subset \Re^I$ (image) and $\mathcal{U}^T \subset \Re^T$ (text) where the correlation between the two data modalities is maximal. It is similar to principal components analysis (PCA), in the sense that it learns a basis of canonical components, directions $w_i \in \Re^I$ and $w_t \in \Re^T$, but seeks directions along which the data is maximally correlated

$$\max_{w_i \neq 0, w_t \neq 0} \frac{w'_i \Sigma_{IT} w_t}{\sqrt{w'_i \Sigma_I w_i} \sqrt{w'_t \Sigma_T w_t}} \tag{3.4}$$

where Σ_I and Σ_T are the empirical covariance matrices for images $\{I_1, \ldots, I_{|D|}\}$ and text $\{T_1, \ldots, T_{|D|}\}$ respectively, and $\Sigma_{IT} = \Sigma'_{TI}$ the cross-covariance between them. Repeatedly solving (3.4) for directions that are orthogonal to all previously obtained solutions, provides a series of canonical components. It can be shown that the canonical components in the image space can be found as the eigenvectors of $\Sigma_I^{-1/2} \Sigma_{IT} \Sigma_T^{-1} \Sigma_{TI} \Sigma_I^{-1/2}$, and in the text space as the eigenvectors of $\Sigma_T^{-1/2} \Sigma_{TI} \Sigma_I^{-1} \Sigma_{IT} \Sigma_T^{-1/2}$. The first d eigenvectors $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$ define a basis of the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$.

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 $\phi_T : \Re^T \to \mathcal{F}^T$ and $\phi_I : \Re^I \to \mathcal{F}^I$. Application of CFA or CCA in these spaces can then recover complex patterns of dependency in the original feature space. As is common in machine learning, the transformations $\phi_T(\cdot)$ and $\phi_I(\cdot)$ are computed only implicitly, by the introduction of two kernel functions $\mathcal{K}_T(\cdot,\cdot)$ and $\mathcal{K}_I(\cdot,\cdot)$, specifying the inner products in $\mathcal{F}^T$ and $\mathcal{F}^I$, *i.e.*, $\mathcal{K}_T(T_m,T_n) = \langle\phi_T(T_m),\phi_T(T_n)\rangle$ and $\mathcal{K}_I(I_m,I_n) = \langle\phi_I(I_m),\phi_I(I_n)\rangle$, respectively.

KCCA [98, 124] implements this type of extension for CCA, seeking directions $w_i \in \mathcal{F}^I$ and $w_t \in \mathcal{F}^T$, along which the two modalities are maximally correlated in the transformed spaces. The canonical components can be found by solving

$$\max_{\alpha_i \neq 0, \alpha_t \neq 0} \frac{\alpha_i' K_I K_T \alpha_t}{V(\alpha_i, K_I)V(\alpha_t, K_T)}, \tag{3.5}$$

where $V(\alpha,K) = \sqrt{(1-\kappa)\alpha' K^2 \alpha + \kappa\alpha' K\alpha}$, $\kappa \in [0,1]$ is a regularization parameter, and K_I and K_T are the kernel matrices of the image and text representations, *e.g.*, $(K_I)_{mn} = \mathcal{K}_I(I_m,I_n)$. Given optimal α_i and α_t for (3.5), w_i and w_t are obtained as linear combinations of the training examples $\{\phi_I(I_k)\}_{k=1}^{|\mathcal{D}|}$, and $\{\phi_T(T_k)\}_{k=1}^{|\mathcal{D}|}$, with α_i and α_t as weight vectors, *i.e.*, $w_i = \Phi_I(X_I)^T\alpha_i$ and $w_t = \Phi_T(X_T)^T\alpha_t$, where $\Phi_I(X_I)$ $(\Phi_T(X_T))$ is the matrix whose rows contain the high-dimensional representation of the image (text) features. To optimize (3.5), we solve a generalized eigenvalue problem using the software package of [124]. The first d generalized eigenvectors, where $1 \leq d \leq |\mathcal{D}|$, are the d weight vectors $\{\alpha_{i,k}\}_{k=1}^d$ and $\{\alpha_{t,k}\}_{k=1}^d$ that define the bases $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$ of the two maximally correlated d-dimensional subspaces $\mathcal{U}^I \subset \mathcal{F}^I$ and $\mathcal{U}^T \subset \mathcal{F}^T$.

Image and text projections

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

$$\begin{aligned} p_{I,k} &= \langle \phi_I(I), w_{i,k} \rangle \\ &= \langle \phi_I(I), \left[\phi_I(I_1), \ldots, \phi_I(I_{|\mathcal{D}|})\right] \alpha_{i,k} \rangle \\ &= \left[\mathcal{K}_I(I, I_1), \ldots, \mathcal{K}_I\left(I, I_{|\mathcal{D}|}\right)\right] \alpha_{i,k}, \end{aligned} \tag{3.6}$$

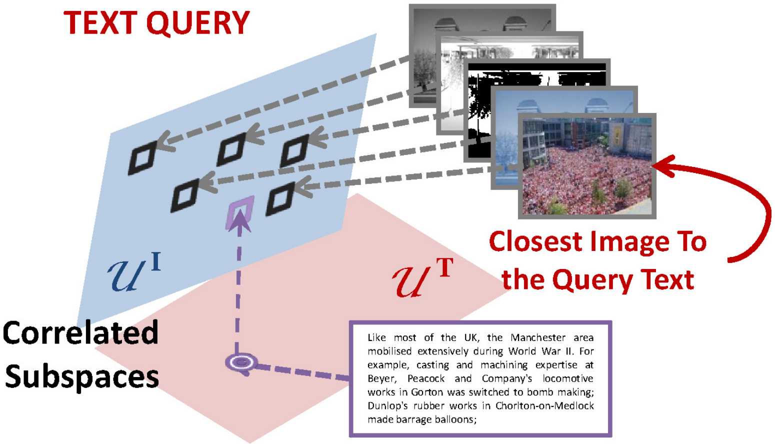
where $k = 1, \ldots, d$. Analogously, a text $T \in \Re^T$ is mapped into $\mathcal{F}^T$ and then projected onto $\{w_{t,k}\}_{k=1}^d$, *i.e.*, $p_T = \mathcal{P}_T(\phi_T(T))$, using $\mathcal{K}_T(.,.)$.

Correlation-based retrieval

For all methods, a natural invertible mapping between the projections onto $\mathcal{U}^I$ and $\mathcal{U}^T$ follows from the correspondence between the d-dimensional bases of the subspaces, as $w_{i,1} \leftrightarrow w_{t,1}$, ..., $w_{i,d} \leftrightarrow w_{t,d}$. This results in a compact, efficient representation of both modalities, where vectors p_I and p_T are coordinates in two isomorphic d-dimensional subspaces, as shown in Figure 3.2. 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

$$D(I,T) = d(p_I, p_T), \tag{3.7}$$

for some suitable distance measure $d(\cdot,\cdot)$ in a d-dimensional vector space. Similarly, given a query text T with projection p_T, the closest image match $I \in \Re^I$ is that for which p_I minimizes $d(p_I, p_T)$. An illustration of cross-modal retrieval using CM is given in Figure 3.3.


TEXT QUERY
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{T}}$
Closest Image To the Query Text
Correlated Subspaces

Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlton-on-Medlock made barrage balloons;

Figure 3.3: Example of cross-modal retrieval using CM. Here, CM is used to find the images that best match a query text.

3.3.2 Semantic matching (SM)

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

of these vectors is referred to as the *semantic space for text*, and the probabilities in $\boldsymbol{\pi}_T$ as the *semantic text features*. Similarly, $\mathcal{L}_I$ maps an image I into a vector $\boldsymbol{\pi}_I$ of *semantic image features* in a *semantic space for images* $\mathcal{S}^I$.

Semantic representations have two advantages for cross-modal retrieval. First, they provide a higher level of abstraction. While features in $\Re^T$ and $\Re^I$ frequently have no obvious interpretation (*e.g.*, image features tend to be edges, edge orientations or frequency bases), the features in $\mathcal{S}^T$ and $\mathcal{S}^I$ are (semantic) concept probabilities (*e.g.*, the probability that the image belongs to the "History" or "Biology" document classes). Previous work has shown that increased feature abstraction can lead to substantially better generalization for tasks such as image retrieval [85]. Second, the semantic spaces $\mathcal{S}^T$ and $\mathcal{S}^I$ are isomorphic, since both images and text are represented as vectors of posterior probabilities with respect to the *same* set of semantic concepts. Hence, the spaces can be treated as being the same, *i.e.*, $\mathcal{S}^T = \mathcal{S}^I$, leading to the representation of Figure 3.4.

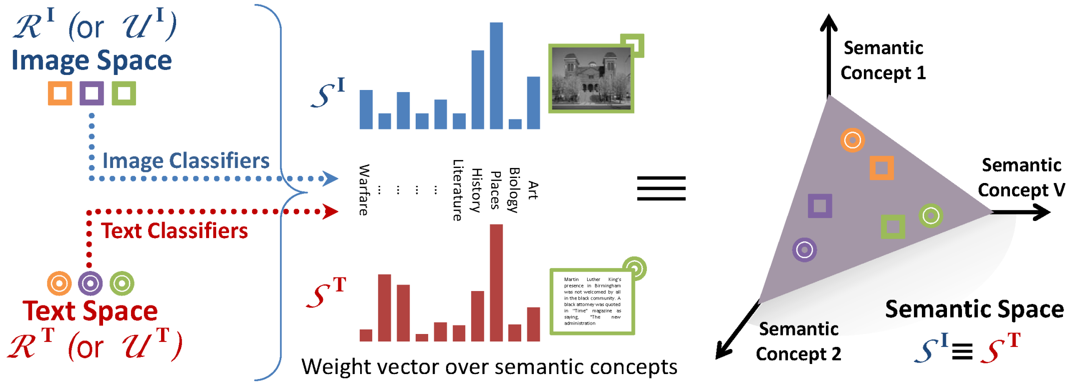

$\mathcal{R}^{\mathrm{I}}$ (or $\mathcal{U}^{\mathrm{I}}$)
Image Space
Image Classifiers
Text Classifiers
Text Space
$\mathcal{R}^{\mathrm{T}}$ (or $\mathcal{U}^{\mathrm{T}}$)
$\mathcal{S}^{\mathrm{I}}$
Art
Biology
Places
History
Literature
Warfare
$\mathcal{S}^{\mathrm{T}}$
Weight vector over semantic concepts
Semantic Concept 1
Semantic Concept V
Semantic Concept 2
Semantic Space
$\mathcal{S}^{\mathrm{I}} \equiv \mathcal{S}^{\mathrm{T}}$

Figure 3.4: Semantic matching (SM). Text and images are mapped into a common semantic space, using the posterior class probabilities produced by a multi-class text or image classifier.

Learn the mappings

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

$$\min_{w} \frac{1}{2} w'w + C \sum_{i} \log(1 + \exp(-y_i w' x_i)) \tag{3.8}$$

where y_i is the class label, x_i the feature vector in the input space, and w a vector of parameters. A multi-class logistic regression can be learned for the image and text modalities, by making x_i the image and text representation, $I \in \Re^I$ and $T \in \Re^T$, respectively. In our implementation this is done with the Liblinear software package [32].

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

$$\min_{w,b,\xi} \frac{1}{2} w'w + C \sum_{i} \xi_i \tag{3.9}$$

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

where w and b are the hyperplane parameters, y_i the class label, x_i input feature vectors, ξ_i slack variables that allow outliers, and $C > 0$ a penalty on the number of outliers. Although the SVM output does not have a probabilistic interpretation, a sigmoidal transformation of the SVM scores $y_i w' x_i$ is often taken as a proxy for the posterior class probabilities. This is, for example, supported by the LibSVM [15] package, which we use in our implementation.

Boosting methods combine weak learners into a strong decision rule. Many boosting algorithms have been proposed in the literature; we adopt the multi-class boosting method of [92]. This is based on multi-dimensional codewords (y^k) and predictors (f). Each class k is mapped to a distinct class label y^k, and the strong classifier, $F(x)$, is a mapping from examples $x_i \in \mathcal{X}$ into class labels y^k

$$F(x) = \arg\max_k y^k f^*(x) \tag{3.10}$$

where $f^*(x) : \mathcal{X} \to \mathbb{R}$ is the continuous valued predictor that maximizes the classification margin. Posterior class probabilities can then be recovered by applying a non-linear transformation to the classifier output. In our implementation this is done with recourse to the multi-class boosting software package of [92].

Abstraction-based retrieval

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

$$D(I,T) = d(\pi_I, \pi_T), \tag{3.11}$$

for some suitable distance measure d between probability distributions. An illustration of cross-modal retrieval using SM is given in Figure 3.5.

3.3.3 Semantic Correlation Matching (SCM)

CM and SM are not mutually exclusive. In fact, a corollary to the two hypotheses discussed above is that there may be a benefit in combining CM and SM.

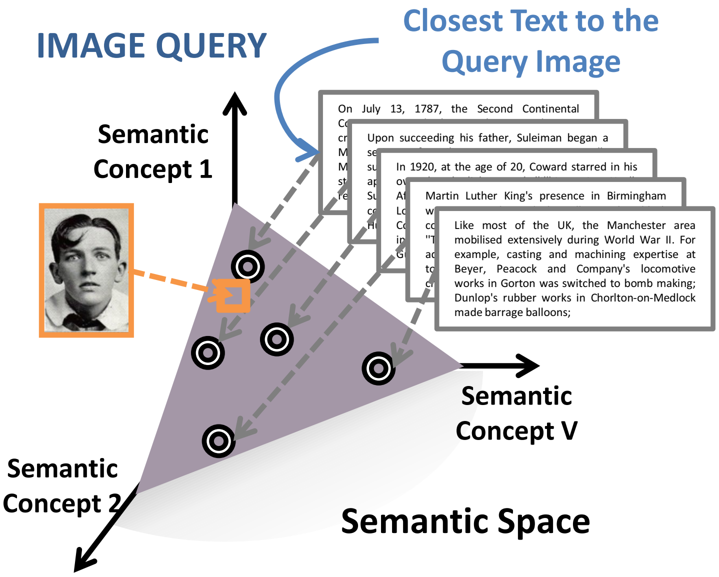


IMAGE QUERY
Closest Text to the Query Image
Semantic Concept 1
Semantic Concept 2
Semantic Concept V
Semantic Space
On July 13, 1787, the Second Continental
Upon succeeding his father, Suleiman began a
In 1920, at the age of 20, Coward starred in his
Martin Luther King's presence in Birmingham
Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlton-on-Medlock made barrage balloons;

Figure 3.5: An example of cross-modal retrieval using SM. Here SM is used to find the texts that best match a query image.

CM extracts maximally correlated features from $\mathfrak{R}^T$ and $\mathfrak{R}^I$. SM builds semantic spaces using original features to gain semantic abstraction. When the two are combined, by building semantic spaces using the feature representation produced by correlation maximization, it may be possible to improve on the individual performances of both CM and SM. To combine the two approaches, the maximally correlated subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are first learned and the projections (p_I, p_T) of each image-text pair (I,T) are computed as discussed in Section 3.3.1. The transformations $\mathcal{L}_I$ and $\mathcal{L}_T$ are then learned in each of these subspaces to produce the semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, respectively. Retrieval is finally based on the image-text distance $D(I,T)$ of (3.11), based on the semantic mappings $\pi_I = \mathcal{L}_I(p_I)$ and $\pi_T = \mathcal{L}_T(p_T)$.

3.4 Experiments

In this section, we describe an extensive experimental evaluation of the proposed cross-modal retrieval framework. First, through validation, we go over a number of experiments to determine model parameters and performance metrics. Secondly, we test each of the two fundamental hypotheses individually, CM and SM, and their combination. There are two different tasks to consider: 1) search in a repository of text articles using an image, we refer to this as an image query task; and 2) the converse, *i.e.* using a full text article to query a repository of images, we refer to this as a text query task.

In what follows the essential details for the experiments are given, but the experimental set-up is discussed in greater detail in Appendix A.

3.4.1 Experimental set-up

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

Datasets

Two datasets are used in all cross-modal retrieval experiments: “TVGraz” [53], which contains $2,058$ image/text pairs of 10 semantic categories, and “Wikipedia” [84] with $2,866$ pairs from 10 categories. Table 3.2 enumerates the set of classes for each dataset. They have different characteristics that are important to point out. TVGraz images are archetypal members of the categories. The dataset is eminently visual, since its categories (*e.g.*, “Harp”, “Dolphin”) are specific objects or animals. The texts are small and can be less representative of the categories. In Wikipedia, on the other hand, category membership is mostly driven by text.

Table 3.2: Dataset semantic classes for the two cross-modal retrieval tasks considered. This is used as the vocabulary, $\mathcal{V}$, whenever a semantic space is involved.

	TVGraz	Wikipedia
1.	Brain	Art & architecture
2.	Butterfly	Biology
3.	Cactus	Geography & places
4.	Deer	History
5.	Dice	Literature & theatre
6.	Dolphin	Media
7.	Elephant	Music
8.	Frog	Royalty & nobility
9.	Harp	Sport & recreation
10.	Pram	Warfare

Texts are mostly of good quality and representative of the category, while the image categorization is more ambiguous. For example, a portrait of a historical figure can appear in the class “War”. The Wikipedia categories (*e.g.*, “History”, “Biology”) are more abstract concepts, and have much broader scope. Individually, the images can be difficult to classify, even for a human. Together, the two datasets illustrate the potential diversity of cross-modal retrieval: applications where there is more uniformity of text than images, and its converse.

Performance metrics

In both tasks considered – text retrieval from an image query, and image retrieval from a text query – text is always based on a full text documents, rather than just a handful of labels. Retrieval performance is evaluated using standard information retrieval metrics, from which mean average precision (MAP) is the most common. We also make use of the 11-point interpolated *Precision-Recall*

(PR) curves [70] to better assess retrieval performance at different levels of recall. All results are compared to a baseline established by a recently published cross-modal retrieval approach, the *Text-To-Image* (TTI) translator of [80]. This is implemented with code provided by the authors.

Image and text representation

For both modalities, the base representation is a bag-of-words (BOW). Text BOW is then fitted by a latent Dirichlet allocation (LDA) [11] model. For images, SIFT descriptors are first extracted to learn a visual-word codebook, through K-means clustering. And finally SIFT descriptors extracted from each image are vector quantized, to produce a vector representation of visual word counts.

For the learning of text and image models, a training set is randomly sampled from the datasets. Table 3.3 shows the train/test split for these experiments.

Table 3.3: Data split among training and test sets.

Dataset	train set	test set
TVGraz	1558	500
Wikipedia	2173	693

3.4.2 Validation experiments

Various preliminary experiments were conducted to identify the best models and parameter configurations for the cross-modal retrieval architecture. These are based on a random split (80/20) of the training sets (from Table 3.3). Yielding $1,245$ training and 313 validation examples on TVGraz; and $1,738$ training and 435 validation on Wikipedia. The validation sets were used to determine the best parameter configurations. In the case where the abstraction hypothesis is involved

(SM and SCM), the semantic vocabulary $\mathcal{V}$ consists of the ground-truth semantics, therefore the semantic spaces from both modalities, $\mathcal{S}^I$ and $\mathcal{S}^T$, are one and the same.

Distance Measures

A number of distance measures, listed in Tables 3.4 and 3.5, were considered for the evaluation of (3.7) and (3.11): Kullback-Leibler divergence (KL), ℓ_1 and ℓ_2 norms, normalized correlation (NC), and centered normalized correlation (NC_c).

Table 3.4: MAP scores for TVGraz data (validation set) of different distance measures. μ_p and μ_q are the sample averages for p and q, respectively.

	measure	$d(p,q)$	image queries	text queries	average
CM	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.376	0.418	0.397
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.391	0.444	0.417
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.498	0.476	**0.487**
	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
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.362	0.564	0.463
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.525	0.573	0.549
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.492	0.570	0.531
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.582	0.581	0.582
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \lvert\lvert q-\mu_q \rvert\rvert}$	0.598	0.578	**0.588**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.560	0.623	0.592
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.623	0.633	0.628
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.605	0.615	0.610
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \lvert\lvert q \rvert\rvert}$	0.665	0.632	0.649
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \lvert\lvert q-\mu_q \rvert\rvert}$	0.669	0.633	**0.651**

Tables 3.4 and 3.5 present MAP scores achieved with each measure, for both image and text query tasks individually and their average representing the overall perfor-

mance of the system (across the two tasks). KL was not considered in correlation matching because this technique does not produce a probability simplex.

Since NC_c had the best average performance in nearly all experiments, it was adopted as distance measure.

Table 3.5: MAP scores for Wikipedia data (validation set) of different distance measures. μ_p and μ_q are the sample averages for p and q, respectively.

	measure	$d(p,q)$	image queries	text queries	average
CM	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.193	0.234	0.214
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.199	0.243	0.221
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \, \lvert\lvert q \rvert\rvert}$	0.288	0.239	**0.263**
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \, \lvert\lvert q-\mu_q \rvert\rvert}$	0.287	0.239	**0.263**
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.206	0.274	0.240
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.220	0.274	0.247
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.205	0.276	0.241
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \, \lvert\lvert q \rvert\rvert}$	0.301	0.276	0.289
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \, \lvert\lvert q-\mu_q \rvert\rvert}$	0.352	0.272	**0.312**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.311	0.270	0.291
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.334	0.273	0.304
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.315	0.267	0.291
	NC	$\frac{p^T q}{\lvert\lvert p \rvert\rvert \, \lvert\lvert q \rvert\rvert}$	0.371	0.279	0.325
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lvert\lvert p-\mu_p \rvert\rvert \, \lvert\lvert q-\mu_q \rvert\rvert}$	0.382	0.281	**0.332**

Correlation matching

A set of experiments was performed to compare the performance of CFA, CCA, and KCCA. In all cases, the number of canonical components was validated in each retrieval experiment. As shown in Table 3.6, KCCA had the top performance. Best results were achieved with a chi-square radial basis function kernel[1] for images, a histogram intersection kernel[2] for text [108, 12], and regularization constants $\kappa = 10\%$ on TVGraz and $\kappa = 50\%$ on Wikipedia.

Furthermore, to verify the importance of modeling correlations, we considered two alternative representations. The first implemented dimensionality reduction but no correlation modeling. The two modalities were independently projected into subspaces of the same dimension, learned with PCA. The second investigated the benefits of complementing correlation with discriminant modeling, by introducing a linear discriminant analysis on the correlated subspaces discovered by KCCA. It is denoted *linear discriminant kernel canonical correlation analysis* (LD-KCCA).

Table 3.6: MAP scores (validation set) under the CM hypothesis.

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

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

[2] $\mathcal{K}(x,y) = \sum_{i=1}^{n} \min(x_i, y_i)$.

As shown in Table 3.6, neither alternative improved on the average MAP scores of KCCA. This shows that there are benefits to correlation matching beyond dimensionality reduction and that further gains are not trivial to achieve, supporting the hypothesis that correlation modeling is important for cross-model retrieval. Given its good performance, KCCA was used in all remaining experiments involving CM.

Semantic matching

Another set of experiments was performed to evaluate the impact of the classification architecture used to design the semantic space on retrieval accuracy. Three architectures were compared: logistic regression, boosting, and SVMs. As shown in Table 3.7, the semantic space obtained with logistic regression performed best for both cross-modal retrieval tasks. It was thus chosen to implement SM in all remaining experiments.

Table 3.7: MAP scores (validation set) under the SM hypothesis.

	image queries	text queries	average
TVGraz			
Log. Regression	**0.598**	**0.578**	**0.588**
SVM	0.556	0.548	0.552
Boosting	0.567	0.476	0.522
Wikipedia			
Log. Regression	**0.352**	**0.272**	**0.312**
SVM	0.318	0.237	0.278
Boosting	0.322	0.207	0.265

Optimization

The experiments above lead to a retrieval architecture that combines KCCA for learning correlated subspaces, logistic regression to learn the semantic space, and the centered normalized correlation NC_c distance measure to evaluate (3.7) and (3.11). Using this architecture, a final round of experiments was used to determine the best combination of 1) BOW codebook size for image representation, 2) number of LDA topics for text representation, and 3) number of KCCA components, for each of the CM, SM, and SCM retrieval regimes and dataset. Table 3.8 summarizes the optimal parameter configuration, which was used in the remaining experiments.

Table 3.8: Optimal parameters (validation set) for best retrieval architecture.

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

3.4.3 Testing the fundamental hypotheses

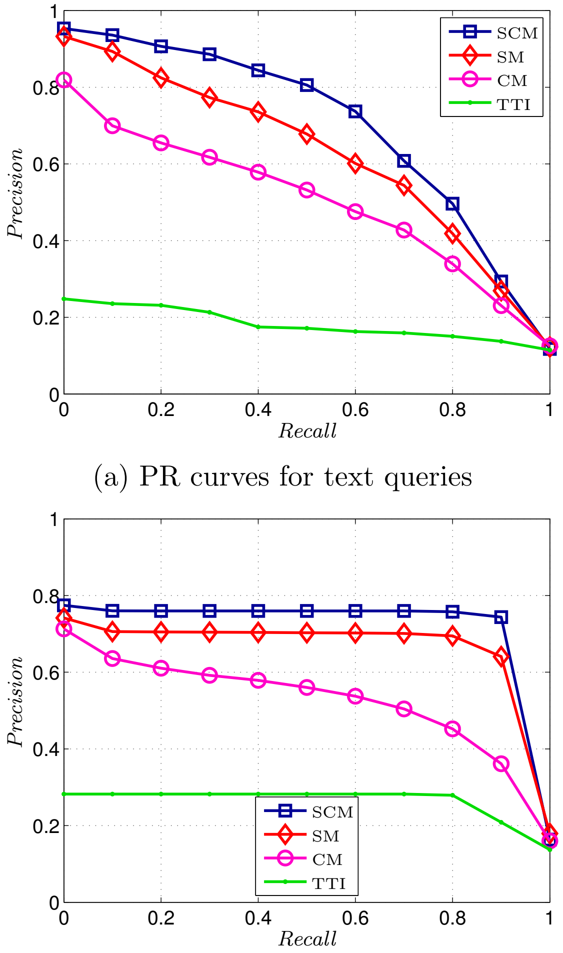
The previous validation experiments, defined the complete set of paramenters that are now used on a new set of experiments aimed to test the fundamental hypotheses of Section 3.2. In what follows, all MAP scores refer to performance measured on the test set.

Table 3.9: MAP scores (test set) of CM, SM, SCM, and TTI, on TVGraz and Wikipedia.

	image queries	text queries	average	gain
		TVGraz		
SCM	**0.664**	**0.649**	**0.657**	–
SM	0.619	0.585	0.602	9%
CM	0.460	0.450	0.455	44%
TTI [80]	0.216	0.153	0.185	255%
		Wikipedia		
SCM	**0.362**	**0.273**	**0.318**	–
SM	0.350	0.249	0.300	6%
CM	0.267	0.219	0.243	31%
TTI [80]	0.237	0.137	0.187	70%

Overall performance

Table 3.9 compares the scores of cross-modal retrieval with CM, SM, SCM and the baseline TTI method. The table provides evidence in support of the two hypotheses of Section 3.2.2: both joint dimensionality reduction and semantic abstraction are beneficial for multi-modal modeling, leading to a non-trivial improvement over TTI. For example, in TVGraz, the average MAP score of CM is more than double that of TTI. For SM the improvement is more than threefold. Overall, the best performance is achieved by SCM. Similar conclusions can be drawn for Wikipedia, although the average gains of SCM are slightly lower than in TVGraz. This is not surprising, since the retrieval scores are generally lower on Wikipedia than those found on TVGraz. As discussed in Section 3.4.1, this is explained by the broader scope of the Wikipedia categories.

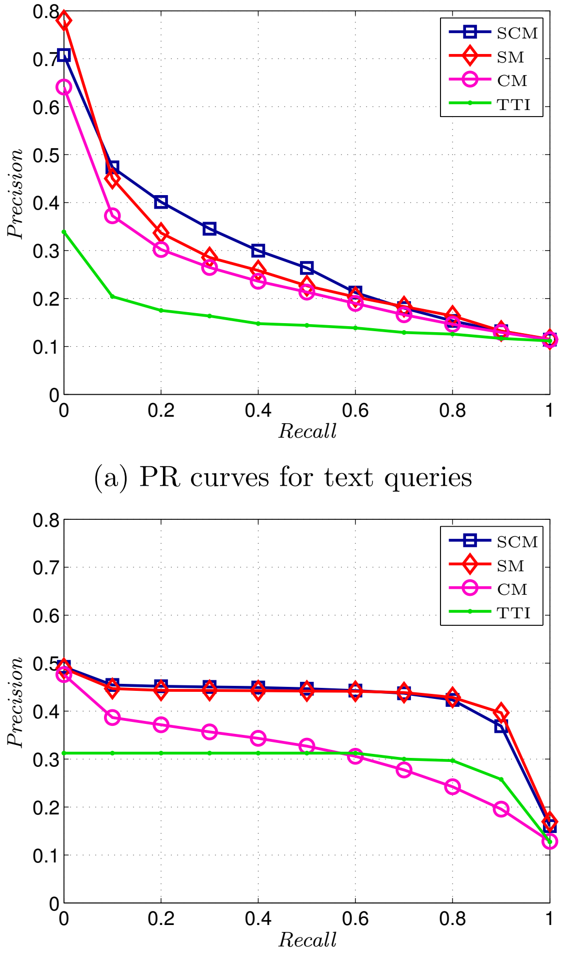


(a) PR curves for text queries

(b) PR curves for image queries

Figure 3.6: PR curves of cross-modal retrieval on TVGraz dataset, using (a) text queries and returning images, while the converse is shown on (b).

Figures 3.6 and 3.7 present a more detailed analysis on the retrieval performance, in the form of PR curves. CM, SM, and SCM again achieve large improvements over TTI. These improvements tend to occur at all levels of recall, indicating better generalization, and often involve substantial increases in precision, indicating higher accuracy. Overall, these results suggest that the contributions of cross-modal correlation and semantic abstraction are *complementary*: not only is there an independent benefit to both correlation modeling and abstraction, but



(a) PR curves for text queries

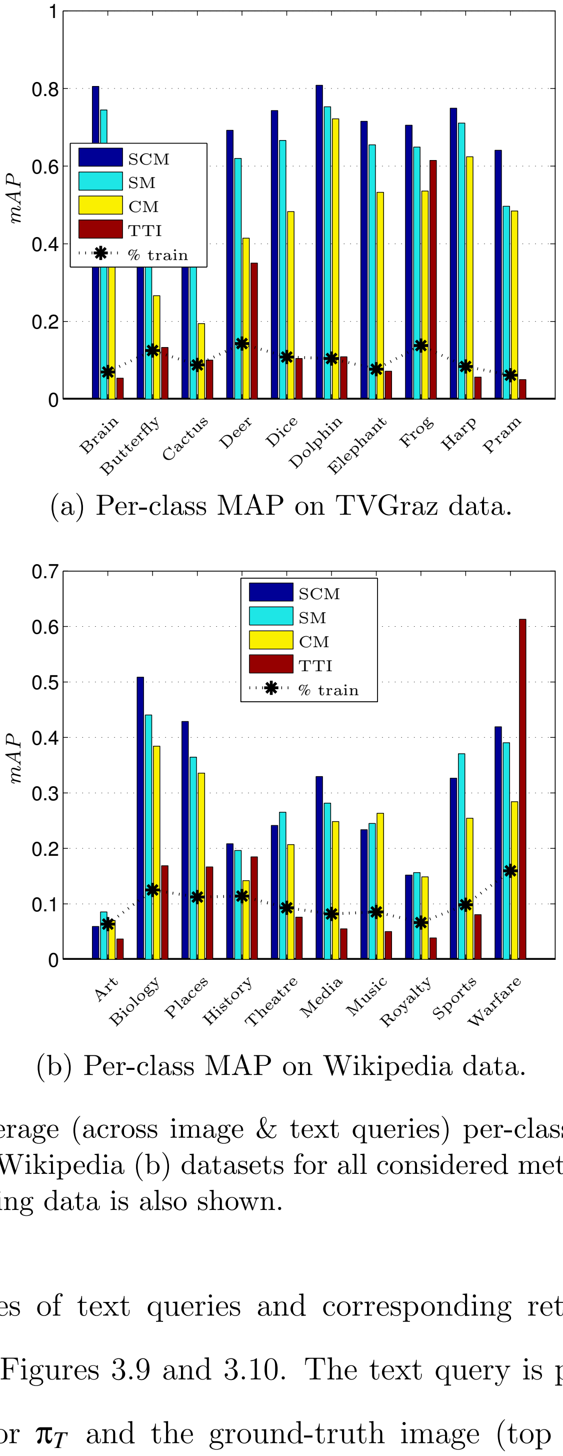
(b) PR curves for image queries

Figure 3.7: PR curves of cross-modal retrieval on the Wikipedia dataset, using (a) text queries and returning images, while the converse is shown on (b).

the *best performance is achieved when the two are combined*.

Per-class performance

Figure 3.8 shows the per-class MAP scores of all methods on both datasets. SCM has higher MAP than CM and SM on all classes of TVGraz, and is either comparable to, or better than CM and SM on the majority of Wikipedia classes. TTI does very poorly in general, and seems biased towards one class. This is evident from Figures 3.8a and 3.8b, where it achieves a very high score on one class – “Frog” on TVGraz and “Warfare” on Wikipedia – and very low scores in the remaining classes. In both cases, the favored class has the larger number of training examples.

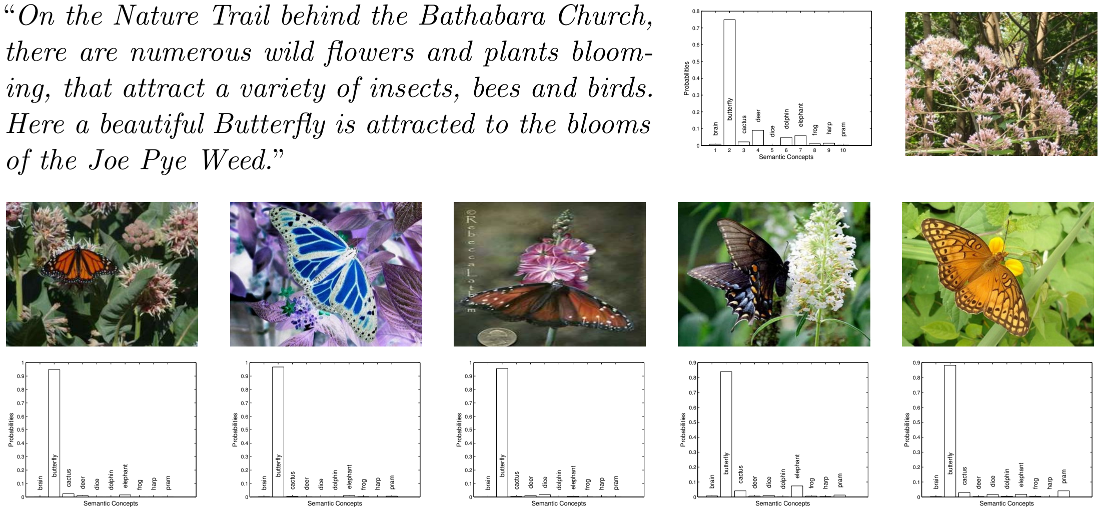


(a) Per-class MAP on TVGraz data.

(b) Per-class MAP on Wikipedia data.

Figure 3.8: Average (across image & text queries) per-class MAP scores for TVGraz (a) and Wikipedia (b) datasets for all considered methods. Percentage of per-class training data is also shown.

Two examples of text queries and corresponding retrieval results, using SCM, are shown in Figures 3.9 and 3.10. The text query is presented along with its probability vector $\boldsymbol{\pi}_T$ and the ground-truth image (top row). The top five image matches are shown below, along with their probability vectors π_I. Finally, Figure 3.11 shows some examples of image-to-text retrieval. Since displaying the retrieved texts would require too much space, we present the associated ground-truth images instead. The query images are framed in the left column, and the images associated with the four best text matches are shown on the right.

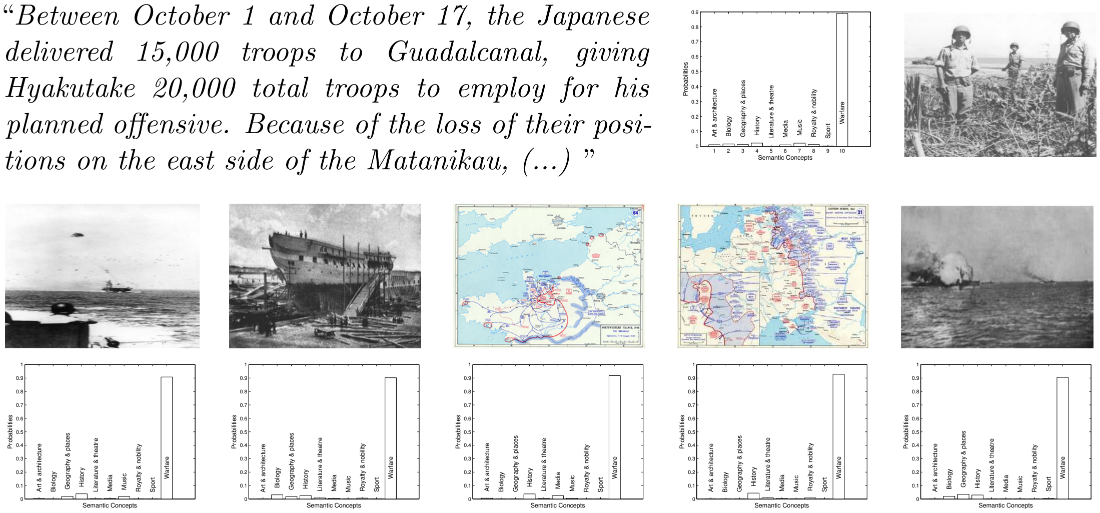


"*On the Nature Trail behind the Bathabara Church, there are numerous wild flowers and plants blooming, that attract a variety of insects, bees and birds. Here a beautiful Butterfly is attracted to the blooms of the Joe Pye Weed.*"

Figure 3.9: An example of text-based cross-modal retrieval from TVGraz dataset, using SCM. The query text, associated probability vector, and ground-truth image are shown on the top; retrieved images are presented at the bottom.

3.5 Discussion

The increasing availability of multi-modal information demands novel representations for content-based retrieval. We proposed models applicable to cross-modal retrieval scenarios. This entails the retrieval of database entries from one content modality in response to queries from another modality. 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.



"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, (...) "

Figure 3.10: Another example of text-based SCM cross-modal retrieval. The query text is a Wikipedia article (shown at the top); retrieved images are presented at the bottom.

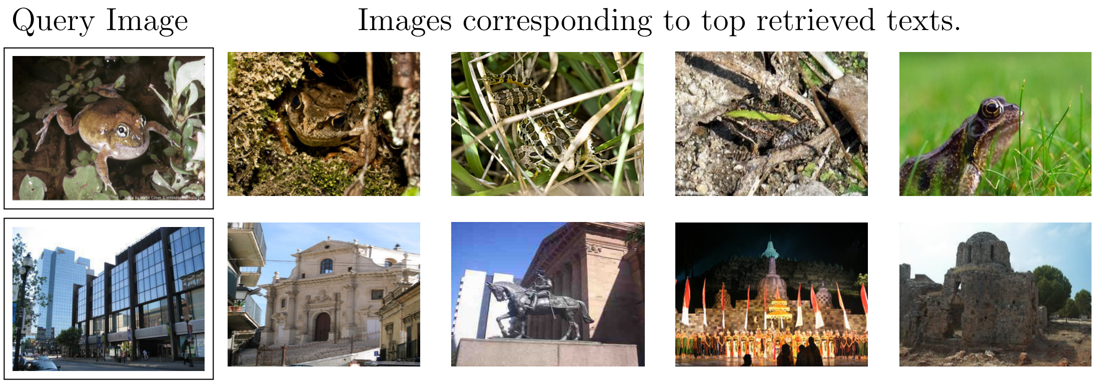


Figure 3.11: Image-to-text retrieval on TVGraz (top row) and Wikipedia (bottom). Query images are framed in the far-left column. The four most relevant texts, represented by their ground-truth images (for practical reasons), are shown in the remaining columns.

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

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

Acknowledgements

The text of Chapter 3, in part, is based on the material as it appears in: ‘On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval’, J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 36(3), pp. 521-535, March 2014. The dissertation author was a primary researcher and an author of the cited material.

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

Chapter 4

Semantic Space Robustness

Cross-modal retrieval tasks are interesting in many aspects; particularly, it opens possibilities to interpret and understand information sourced from multiple modalities. But ultimately, we are interested in building better content-based image retrieval (CBIR) systems. These are retrieval systems where search is based not on the visual patterns of the query image, but on the actual contents of the image (*e.g.* "sun", "kids playing", "water", "beach"). It is therefore imperative that such systems *understand* the images. Image representation plays a key role in scene understanding. In this context, the design of visual features has been a subject of substantial interest in the research community. Early representations relied on explicit representation of low-level image properties such as color, texture, or shape, through color histograms [108], color moments [107, 106], Gabor wavelets [69], Fourier features [123], stochastic models [71], or shape contexts [8], among others. More recently, substantial efforts have been devoted to the extension and robustification of these representations, through operations like normalization and spatial pooling, leading to modern descriptors such as SIFT [67], HoG [21], SURF [7], spatial pyramids [61], or Fisher vectors [78]. However, it was realized early on, that one of the limitations of these representations is a *semantic gap* [102] between strict visual similarity, *i.e.* similarity in terms of patterns of color or texture, and human judgments of image similarity. This spurred significant interest in the development of image representations that account for *semantic abstraction* [120, 121, 122, 104, 86, 34, 33, 109, 64]. Such representations are designed by identifying a vocabulary of concepts of interest and learning classifiers for the detection of these concepts. Images are classified and mapped to a space where each feature is a score for the detection of each concept. These are the representations used in the cross-modal retrieval scenario of SM (semantic matching) for both images and text. Several methods have been proposed to implement this. For images a popular framework is the one of [86], which relies on the vector of *posterior probabilities* of the image under the concepts in the vocabulary, as a semantic feature vector. This feature vector is denoted *semantic multinomial* (SMN). Other implementations have been proposed in the literature, *e.g.* the query-by-example semantic retrieval method of [104], the *classeme* representation of [109], or the *object bank* of [64].

In what follows, a detailed description of semantic representation for images and their applications in content-based image retrieval (CBIR), is given. The pros and cons are highlighted, with particular relevance for context and multi-modality. This will lay the foundations for the regularization algorithms of Chapter 5. We conclude with a set of experiments on the robustness of semantic representations using the cross-modal retrieval tasks introduced in the previous chapter.

4.1 Semantic Image retrieval

CBIR has been a subject of research for many years. Popular retrieval systems such as QBIC [38] and Virage [4], sprung the first efforts for Internet-scale image search engines such as Visualseek [103] and Webseer [39]. These systems were based on similarity of low-level descriptors accounting for properties such as image color and texture. Semantic representations were first introduced in the video classification literature [120, 121, 122] and then extended to the CBIR literature. In this context, one of the first and most comprehensive efforts towards semantic representation was the ImageScape system [62]. [104] extended the popular *query-by-example* retrieval paradigm to the realm of semantic representations. Many other proposals have since been made in the CBIR, scene classification, object recognition, and video understanding literatures [86, 109, 64]. Some of these apply to special domains or specific sets of semantic concepts. For example, the space of attributes [34, 59, 33] is a mid-level semantic representation that has enjoyed substantial popularity in recent years [135, 9, 91, 99, 134]. More recently, deep convolutional neural networks have gained substantial popularity in the task of large scale visual recognition [55], including scenarios of joint image-text embeddings [129, 40].

The *query-by-semantic-example* retrieval paradigm of [86] introduced the SMN image representation and extended the minimum probability of error retrieval framework of [117] to the semantic domain. It consists of retrieving images by similarity of the associated SMNs. This was demonstrated to significantly improve the performance of the classical *query-by-visual-example,* where images are matched by similarity of visual descriptors [86].

4.1.1 Semantic space

A semantic space is an abstract space for data representation where each dimension is associated with a certain *word* or *concept*. The way to obtain such a space is as follows. First, a vocabulary of L semantic concepts is defined, $\mathcal{L} = \{z_1, z_2, \ldots, z_L\}$; these can be broad classes, such as "indoors"', "sports", "forest", or a finer grained set like object classes or attributes, such as "tall", "furry", "four-legged", or any other concept of interest. Given a set of images labeled with such concepts, $\mathcal{G} = \{I_1, \ldots, I_G\}$, a classifier is designed to assign a score $\pi_{i,j}$ to each image I_i under each concept z_j. The vector of all such scores, π_i, constitutes image I_i semantic features. This can be seen as the projection of the image into a space $\mathcal{S}$ where each dimension corresponds to a concept in the vocabulary $\mathcal{L}$. The space $\mathcal{S}$ is usually denoted as the semantic space. To train the classifiers, images are first represented in a low-level feature space $\mathcal{X}$, *e.g.* the space of SIFT descriptors sampled over a pre-defined image grid. These features might be fed directly to the classifiers, or, an intermediate low-level representation such as bag of descriptors $I_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n}\}$ can be used.

4.1.2 Semantic representations

In essence, semantic representations proposed in the literature, fall in one of two types. In the first, the semantic space $\mathcal{S}$ consists of a set of mutually exclusive classes [86]. For example, the classes in a taxonomy used to organize an image database, where each image is placed in one and only one folder. In this case, the class label is a categorical random variable $Z \in \{1, \ldots, L\}$ and the semantic representation of image I_i is a vector of class probabilities $\pi_{i,j}$ that add up to one. For the second type of semantic representations, $\mathcal{S}$ consists of a set of non-exclusive classes. For example, a set of binary attributes [34, 33] that can be simultaneously active for any I_i. In this case, the class label is a multivariate Bernoulli random variable $Z \in \{0,1\}^L$, *i.e.* a vector of independent binary random variables, and the entries of π_i do not necessarily add up to one. This distinction is somewhat artificial, since the first representation can be extended into a hierarchical taxonomy, where higher levels in the hierarchy are composed of broader images classes, containing images that belong to different classes in the subsequent levels. Attribute-based classes could be implemented at these higher levels [114, 115]. Similarly, a retrieval system that adopts the second representation must always have access to a disjoint set of classes, namely the classes used as ground-truth to optimize and evaluate the retrieval operation. This may only be used non-parametrically, *e.g.* retrieval may be based on a nearest-neighbor search, but must exist. Otherwise, no claims can be made about the optimality of the system, it is not clear what the system attempts to do, and no claims can be made that the system is preferable to any other system. The two representations can probably be best seen as alternative semantic views of an image database. One view based on generic semantics (attributes) that can be shared by all images, the other view based on categorical semantics that can be used to organize images into disjoint sets. The two views can also be combined, *e.g.* by expressing images as attribute vectors, mapping these vectors into a categorical variable (*e.g.* things that have “fur”, and “ears” belong to the class “dog” if they also “eat meat” or to the class “cat” if they instead “eat fish”), and using the resulting probabilities as dimensions of $\mathcal{S}$.

4.1.3 Implementation

Under the categorical representation of [13] image descriptors are considered samples from a random variable $\mathbf{X}$, and concepts are samples from a random variable $Z \in \{1,\ldots,L\}$. Each concept is assumed to induce a probability density, $P_{\mathbf{X}|Z}(\mathbf{x}|z)$ on $\mathcal{X}$. Bayes rule then enables the representation of image I_i as a vector of posterior probability scores

$$\pi_{i,j} = P_{Z|\mathbf{X}}(j|I_i). \tag{4.1}$$

An illustration is shown in Figure 4.1. The vector π_i defines a multinomial distribution, denoted as semantic multinomial (SMN) [86], and the semantic space $\mathcal{S}$ is a probability simplex, *i.e.* all dimensions of π_i are positive and add to one. Given a set of manually labeled training examples per concept, the posterior probabilities $\pi_{i,j}$ can be learned in several manners. One possibility is to learn the concept distributions $P_{\mathbf{X}|Z}(\mathbf{x}|z), \forall z$ using the training set, and apply Bayes rule to compute the posteriors of (4.1). Another possibility is to learn a discriminative multi-class classifier[1], which produces estimates of the posterior probabilities directly.

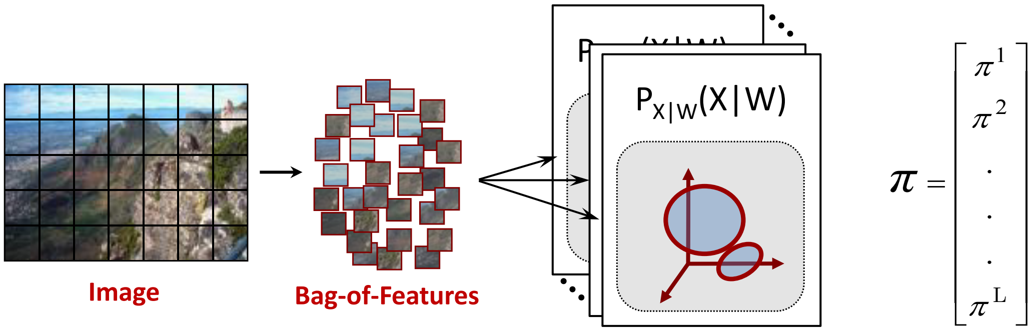


Figure 4.1: Example of the categorical semantic representation of an image. The vector π of posterior concept probabilities is the SMN image descriptor.

[1] The experiments in this work – published in TPAMI and CVIU – use multi-class logistic regression from the package of [32] because they produced the best results among several alternatives. More recently generative models were tested, using activations (conv5) from deep neural networks, that also produce competitive results.

4.1.4 Query by semantic example

Given a semantic space $\mathcal{S}$, we implement image retrieval with the query-by-semantic example procedure of [86]. This consists of mapping all images I_i in a database into $\mathcal{S}$, by computing the associated SMNs π_i, and measuring image similarity with a suitable measure of similarity between SMNs. Given a query image I_q, and the associated SMN π_q, the database images are ranked by increasing values of $d(\pi_q, \pi_i)$ where $d(.,.)$ is the measure of SMN distance. Several such measures can be used, as investigated in the previous chapter. Here we restrict to widely used Kullback-Leibler divergence

$$d(\pi_q, \pi_i) = \sum_{j=1}^{L} \pi_{q,j} \log\left(\frac{\pi_{q,j}}{\pi_{i,j}}\right). \tag{4.2}$$

In all experiments so far, classes used to construct the semantic space $\mathcal{S}$ are also the ground-truth classes inherent to the optimality criterion. This is why the terms *semantics* or *classes* are frequently used interchangeably. We note that this choice of semantics makes the mapping from $\mathcal{X}$ to $\mathcal{S}$ a discriminant feature transformation for the retrieval operation. Discriminant transformations, *i.e.* transformations informed by the ground-truth classes, are a commonly used feature extraction procedure in machine learning. As to assess robustness of semantic representations we investigate alternative semantic configurations. An attempt to build a universal semantic space is made using an expanded set of classes derived from various datasets, this is denoted extended semantics. In another scenario, an alternative ground-truth semantics is constructed that is only loosely related with the concepts included in the semantic space. With this two scenarios we which to assess the robustness of SM retrieval, and in particular how it compares to the alternative CM retrieval.

4.1.5 Context and multi-modality

The semantic representation discussed above has three properties of particular relevance to this work. First, it is a representation that encodes contextual dependencies between different concepts. For example, because most images of the "outdoors" class include "vegetation," the presence of the "vegetation" concept is a clue for image assignment to the "outdoors" class. The semantic representation encodes this contextual relationship by assigning image I_i to the two concepts with some probability. This enables image retrieval and classification systems to take contextual cues into account [86, 87]. Second, unlike any low-level feature space $\mathcal{X}$, the semantic space $\mathcal{S}$ offers a unified representation for information from multiple modalities. As a direct consequence, we were able to use the semantic space to perform the cross-modal retrieval tasks from chapter 3 (under the SM hypothesis). For example, in Figure 4.1, replacing the image descriptors of $\mathcal{X}$, with features extracted from text documents produces a semantic representation for text. This enables a broader representation of context than that possible from images alone: by augmenting the training set with text, it is possible to learn contextual dependencies from the latter. One immediate benefit is that, because text classification is less ambiguous than image classification, the probabilities of (4.1) tend to be much more accurate for the former. This is illustrated in Figure 4.2, where data from both modalities is mapped to the semantic probability simplex. On the left, representation of images from several different articles in a semantic space limited to three concepts ("History", "Royalty" and "Warfare"). Different colors correspond to different article classes (black for "Warfare", blue for "Royalty", and red for "History"); on the right, similar representation for the text components. Note that the concept probability estimates are much noisier for images than text. In result, the image semantics are substantially more ambiguous than the text semantics. This motivates what we denote cross-modal regularization – to be discussed in chapter 5 – where a regularizer learned from a corpus of images and text is used to de-noise the semantic representation of subsequent images. Third, by projecting images and text in the same semantic space, translation between modalities becomes automatic. It is a side-effect of this representation, and there is no need to learn a translator for any modality represented in this space. This reduces the cross-modal regularization problem to one of domain adaptation between two homogeneous domains. In this way, domain adaptation is decoupled from translation, and considerably simpler than in the low-level space $\mathcal{X}$, where a translator must always be learned [81, 19].

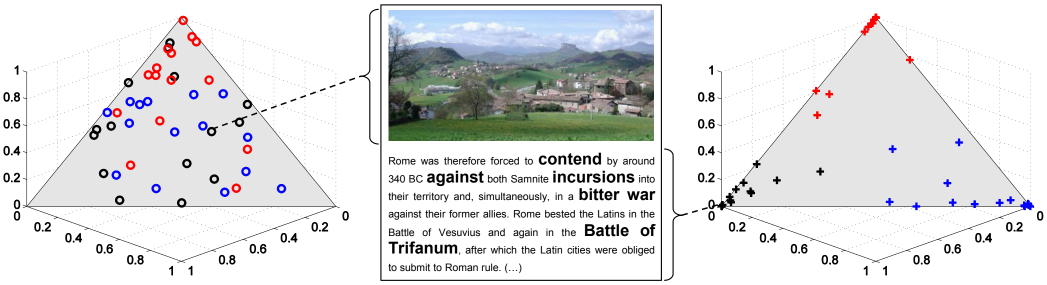

Rome was therefore forced to **contend** by around 340 BC **against** both Samnite **incursions** into their territory and, simultaneously, in a **bitter war** against their former allies. Rome bested the Latins in the Battle of Vesuvius and again in the **Battle of Trifanum**, after which the Latin cities were obliged to submit to Roman rule. (…)

Figure 4.2: An excerpt from an article of the "Warfare" class from the Wikipedia dataset, with the corresponding image (middle). Left and right, semantic representation for the image and text, respectively.

4.2 Experiments

The experiments from the previous chapter, section 3.4, show evidence that that semantic spaces are beneficial for cross-modal retrieval tasks. However, in each experiment, the semantic space was designed with a vocabulary, $\mathcal{V}$, identical to the ground-truth semantics. It could be argued that this gives an unfair advantage to SM retrieval when compared to CM. We perform a number of additional experiments tailored to investigate this.

4.2.1 Experimental set-up

The experimental set-up is similar to that of chapter 3 and detailed in appendix A; two cross-modal retrieval tasks are considered (*i.e.* search a text repository using an image query, and the converse), and two experiments are performed: (1) extended semantics – to assess the impact of dimensionality on semantic space matching; and (2) alternative semantics – to test a semantic representation that does not overlap with the performance evaluation criteria.

4.2.2 Extended semantics

This first set of experiments tests the impact of the size of the vocabulary $\mathcal{V}$ on SM retrieval performance. It was based on an *extended vocabulary* $\mathcal{V}''$, which is shared across the two datasets. It contains the 10 classes from TVGraz, 30 classes from Wikipedia[2], and 20 categories from a smaller multi-modal dataset called *Pascal-Sentences* [83] (50 image/text pairs per class). Overall, $\mathcal{V}''$ contains up to 60 classes. The ground-truth semantics for each dataset remains unchanged.

To evaluate the impact of the composition of the semantic space on retrieval scores, we repeat the cross-modal retrieval experiment using multiple subsets of $\mathcal{V}''$ as the vocabulary $\mathcal{V}$. Starting with $\mathcal{V}$ containing the 10 ground-truth classes, we sequentially added, without repetition, one of the remaining classes in $\mathcal{V}''$ to $\mathcal{V}$. This produces a sequence of semantic spaces with between 11 and 60 dimensions. To introduce randomness, the whole experiment is repeated five times, using a sequence of randomly selected classes to add at each step.

[2] Both the 10 previously used classes and the remaining 20 classes of Wikipedia featured articles are used in $\mathcal{V}''$

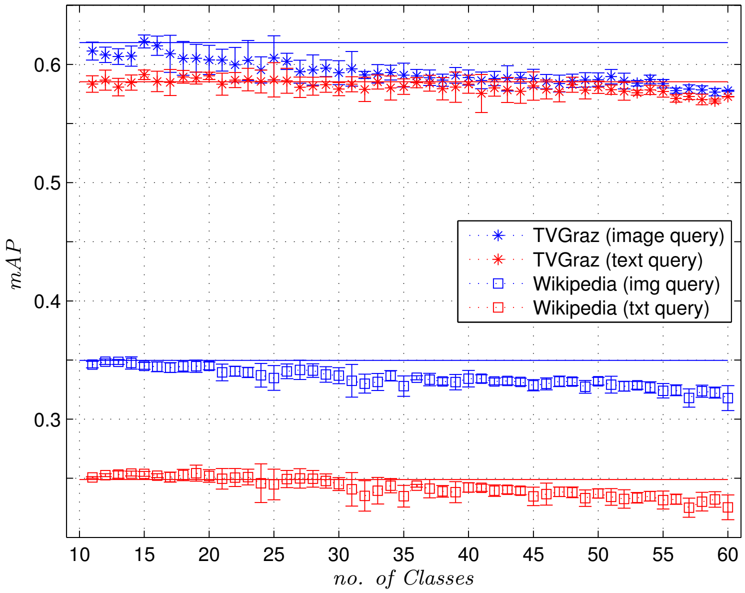


Figure 4.3: MAP scores under SM. The solid horizontal line is the score obtained with the 10 original dataset categories.

Figure 4.3 presents the MAP scores as a function of the vocabulary size, for image and text queries on the two datasets. The straight horizontal lines are the scores obtained for the two cross-modal retrieval tasks considered when $\mathcal{V}$ contains the 10 original classes (*i.e.* SM retrieval results from chapter 3). The image query task appears to be slightly more affected than its text counterpart, this is a natural consequence of the noisier semantic descriptor of images when compared to those of texts [16]. While there is some degradation of performance as the vocabulary grows, the effect is small. This indicates that the performance of SM is fairly insensitive to the size of the vocabulary $\mathcal{V}$.

4.2.3 Alternative semantics

In the experiments from chapter 3, the vocabulary $\mathcal{V}$ used in SM retrieval was equivalent to the set of classes that constitute the ground-truth semantics. While in the previous section, the vocabulary used contains all ground-truth semantics, but it also includes unrelated classes. Thus, in all experiments so far, the vocabulary $\mathcal{V}$ always included the ground-truth semantics. To further test the robustness of SM, a final set of experiments is performed where ground-truth semantics are not explicitly included in the design of the semantic space ($\mathcal{V}$), but are only loosely related to that vocabulary. For this, we define a new set of ground-truth semantics for each dataset, as shown in leftmost column of Tables 4.1a and 4.1b. For all experiments, the vocabulary $\mathcal{V}$ of the semantic space remains unaltered with respect to the original dataset classes. This is also shown in the aforementioned tables for convenience (right column).

Table 4.1: Alternative semantics for both datasets, TVGraz and Wikipedia, and their relationship with the original semantics.

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

(a) TVGraz

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

(b) Wikipedia

Table 4.2 shows a comparison of the average MAP scores achieved with the alternative ground-truth semantics of Tables 4.1a and 4.1b (denoted *"Alternative"*) and with the original dataset classes (denoted *"Vocabulary"*). Since there are fewer classes in the alternative semantics ground-truth, the retrieval performance is expected to improve. However, the fact that these classes are more abstract could also lead to retrieval performance degradation. The two behaviors are visible in the table. On Wikipedia, where the original classes are already quite abstract, all methods exhibit improved performance under the alternative semantics. On TVGraz, where the alternative semantics are much more abstract than the vocabulary classes, performance decreases for SM and SCM. Note, however, that these variations do not affect the relative performance of the different methods. In both cases, CM and SM achieve significant improvements over TTI and the best overall performance is obtained when they are combined (SCM). SM continues to achieve the better performance when directly compared to the CM hypothesis. In summary, this experiment using an alternative ground-truth semantics, confirms the conclusions described in chapter 3.

Table 4.2: Average MAP scores (test set) under the original ("Vocabulary") and alternative semantics.

	Alternative	Vocabulary
TVGraz		
SCM	**0.584**	**0.657**
SM	0.568	0.602
CM	0.492	0.455
TTI [80]	0.292	0.185
Wikipedia		
SCM	**0.448**	**0.318**
SM	0.436	0.300
CM	0.413	0.243
TTI [80]	0.347	0.187

4.3 Discussion

The representation of images in a semantic space has several advantages. The generalization from low-level features to semantic concepts enables a similarity measure that correlates much better with the expectations of CBIR users [121, 86, 119]. Furthermore, because semantic features are, by definition, discriminant for tasks like image categorization, they enable solutions for such tasks with low-dimensional classifiers [88, 58]. They also enable hierarchical representations with increasing complexity for finer grained classification or retrieval; *e.g.* we might be interested in “furry domestic animals” or in a specific breed of dogs. Due to their abstract nature, semantic spaces also enable unified representation for data from different content modalities, *e.g.* images, text, or audio. In this chapter, we have shown the robustness of semantic spaces under two cross-modal retrieval tasks involving images and text. This opens up a new set of possibilities for multimedia processing, enabling operations such as the cross-modal similarity as well as cross-modal regularization to be discussed in the next chapter.

Acknowledgements

The text of Chapter 4, in part, is based on the material as it appears in: ‘On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval’, J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R.Levy, N. Vasconcelos, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 36(3), pp. 521-535, March 2014. ‘Cross-modal Domain Adaptation for Text-based Regularization of Image Semantics’, J. Costa Pereira, N. Vasconcelos, Elsevier, Computer Vision and Image Understanding, Vol. 124, pp. 123-135, July 2014. The dissertation author was a primary researcher and an author of the cited material.

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

Chapter 5

Cross-Modal Regularization

In query-by-semantic-example retrieval systems, images are ranked by similarity of semantic descriptors. These descriptors are obtained by classifying each image with respect to a pre-defined vocabulary of semantic concepts. This results in a probability vector, with a high level of entropy. Based on the principles introduced in chapters 3 and 4, we now consider the problem of improving the accuracy of image semantic descriptors through cross-modal regularization; using auxiliary text, a cross-modal regularizer is proposed. Training image-text pairs are first mapped to a common semantic space. A regularization operator is then learned for each concept in the semantic vocabulary. This is an operator which maps the semantic descriptors of images labeled with that concept to the descriptors of the associated texts. A convex formulation of the learning problem is introduced, enabling the efficient computation of concept-specific regularization operators. The final step is the selection of the most suitable operator for a particular image we wish to regularize. This is implemented through a quantization of the semantic space, where a regularization operator is associated with each quantization cell. Overall, the proposed regularizer is a non-linear mapping, implemented as a piecewise linear transformation of the semantic image descriptors. This transformation is a form of cross-modal domain adaptation, and it is shown to achieve better performance than recent proposals in the domain adaptation literature, while requiring much simpler optimization.

5.1 Introduction

The semantic matching (SM) hypothesis used in the scope of cross-modal retrieval provides the necessary tools to compare and/or combine data from multiple sources in a unified semantic space. In this space, text and image representations exhibit different levels of uncertainty, with the highest levels of noise observed in the latter modality. While this effect allows to uncover co-occurrences of semantic concepts in a data driven fashion, it also comes with a drawback: hampering precision on CBIR systems. Given the fact that textual information – label, text snippet or full-sized article – is intrinsicly semantic, it has more discrimintive power in the semantic space. We would like to use this as to reduce the level of noise in the semantic representations of images. The basic idea is to leverage the fact that most images exist in a rich multi-modal context, *e.g.* web-pages, which provide contextual information about the image content. In fact, some of this information may be much easier to model or classify than the image itself. For example, text classifiers tend to have higher accuracy than state-of-the-art image classifiers. Due to this, an SMN inferred from an image is likely to be more noisy than an SMN derived from an associated text document.

A question that arises naturally is whether it is possible to exploit the presence of text associated to an image, to de-noise the semantic representation of the latter. One possibility would be to simply replace the image SMN with the associated text SMN. This would reduce to the cross-modal retrieval scheme of chapter 3, where a query image is matched to a database of texts. While effective, this solution is not fully general, since it assumes the availability of text for all images in the CBIR database. A more general solution is to collect a dataset of image-text pairs and *learn a transformation* that maps the ambiguous image semantics to the less ambiguous text semantics. This transformation can then be applied to images that have no complementary text. Because this de-noising operation is likely to enable better generalization for all retrieval operations we denote it as a *regularization* of the semantic image representation. Since text information is used to regularize visual information, the process is denoted *cross-modal regularization.* The de-noised semantic representation is denoted as *regularized image semantics*.

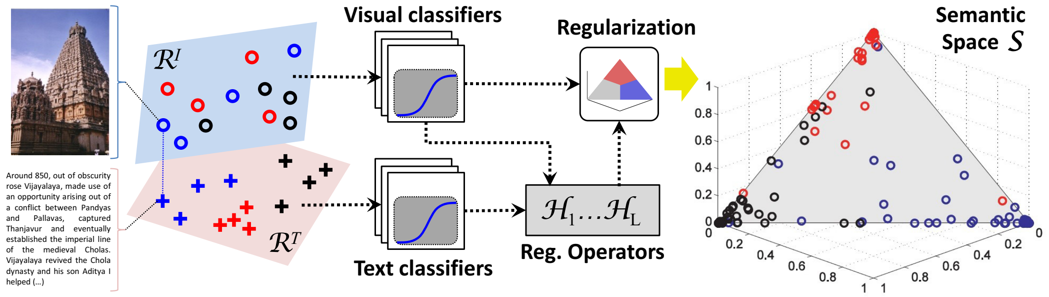

Around 850, out of obscurity rose Vijayalaya, made use of an opportunity arising out of a conflict between Pandyas and Pallavas, captured Thanjavur and eventually established the imperial line of the medieval Cholas. Vijayalaya revived the Chola dynasty and his son Aditya I helped (...)

Figure 5.1: Proposed cross-modal regularizer of image semantics: (i) uni-modal classification, (ii) regularization learning and (iii) image de-noising. As before, different colors correspond to images labeled with different concepts.

The proposed cross-modal regularizer of image semantics (RIS) is composed of three steps, illustrated in Figure 5.1. Training images and texts are first mapped to the semantic space, using a set of uni-modal classifiers. A regularization operator is then learned for each concept in the semantic vocabulary. This operator maps

SMNs of images labeled with that concept to the SMNs of the associated texts. In practice this results in a quantization of the probability simplex, where each of these operators is assigned to a quantization cell. After regularization, images labeled with the same concept tend to cluster in a subspace of the simplex. This contains the vertices of the simplex associated with that concept and its contextually related concepts. Because the transformation is linear on an affine space (probability simplex), and the objective function is to minimize the mean squared error of the mapping, the problem can be framed in a convex formulation, which lends itself to efficient optimization. The process results in a set of concept-specific regularization operators. The final step is a procedure for the selection of the most suitable regularization operator for the image to regularize. In this quantization of the probability simplex, each cell is associated with a regularization operator. Overall, the proposed regularizer is a non-linear mapping, implemented as a piecewise linear transformation of the image SMN to regularize. This is shown to enable better performance than other recent proposals in the domain adaptation literature [93, 94, 43, 47, 81], and requires a much simpler optimization.

It should be noted that, the ideas now proposed could be applied to other semantic image representations in the literature. The question that we investigate is whether it is possible to improve any such semantic representations by taking advantage of additional data modalities. In particular whether, given a training set of images and texts, it is possible to learn a transformation that de-noises the semantic representation of unseen images. This is expected to further improve QBSE [86] performance. Since it leverages text to improve image retrieval, cross-modal regularization is a form of *transfer learning.* This consists of transferring information from an *auxiliary* dataset to regularize a learning operation on a *target* dataset. Transfer learning is useful when learning is poorly constrained in the target domain, *e.g.* when too little training data is available. Several forms of transfer learning have been proposed. The most popular is probably *semi-supervised learning* [139], where a small set of labeled target data is augmented by a large auxiliary corpus of unlabeled data. These methods assume that the statistics of the target and auxiliary datasets are similar and are not directly applicable to cross-modal regularization. A second form is *multi-task* learning [14], where a common model and training data are shared for the solution of two or more learning tasks, *e.g.* the simultaneous classification of images and text. This is again unlike cross-modal regularization, where the goal is to learn improved image classifiers only. No text classification is performed after learning. A third form of transfer learning is *model adaptation*, where auxiliary data is used to regularize the parameters of a target model, which can be either generative [90, 130, 26, 137, 35, 82] or discriminative [27, 131, 10, 20, 3]. Although this is sometimes denoted *domain adaptation,* the latter usually refers to methods that regularize the target feature space, rather than the models themselves. This is frequently implemented by learning a feature transformation that maximizes the similarity of feature vectors from target and auxiliary domains [23, 57, 44, 43, 28, 138]. Some methods have also been proposed to implement both domain and model adaptation [47]. The proposed approach to cross-modal regularization can be seen as a form of domain adaptation, although it has significant differences with respect to previous implementations of the former. First, while domain adaptation assumes more auxiliary than target data, this is not the case for cross-modal regularization. Here, the problem is instead that data from the two modalities has different degrees of *semantic ambiguity:* cross-modal regularization is useful even if there is infinite image data. Second, most domain adaptation methods assume that auxiliary and target domains produce data of the same type, *e.g.* images taken under different views or from different datasets.

This simplifies the problem in two ways. One, it enables simplifying assumptions, *e.g.* the existence of a smooth path through a sequence of subspaces between the auxiliary and target domains [44, 43], that does not hold for cross-modal regularization. Another, it implies the absence of a semantic gap between the two domains, leading to a simpler correspondence problem than that of cross-modal regularization. This assumption contradicts the essence of cross-modal regularization, where the goal is to leverage the smaller semantic ambiguity of text to regularize image classification.

Perhaps due to the issues pointed above, the notion of performing regularization in a semantic space has received little attention in the literature. Instead, domain adaptation is usually implemented through a global transformation between low-level features in the auxiliary and target domains. This is the case even for the few approaches previously proposed for cross-modal domain adaptation using images and text [81, 19]. These methods simply learn a feature transformation between the two spaces, denoted a *translator,* from co-occurrence counts of visual and text words. While global low-level transformations can be used for cross-modal regularization, our experiments show that they have weaker performance than the now proposed combination of semantic-specific regularization operators.

This chapter is organized as follows. The proposed regularization operator is introduced in Section 5.2. These operators are learned by formulating a convex problem, that is efficiently solved. In the final part of this section two regularization algorithms are introduced – interpolation and classification-base methods. Section 5.3 presents an extensive experimental evaluation of the regularizer in the context of CBIR. Finally, a discussion is presented in Section 5.4.

5.2 Cross-modal Regularization

The regularization procedures proposed here can be applied to the two types of semantic representations discussed in Section 4.1.2. However, for clarity of presentation, we limit the discussion to the categorical view, and adopt the approach of [13]. The modifications needed to extend the regularization procedure to the multivariate Bernoulli representation are discussed at the very end of Section 5.2.2.

In all following equations d-dimensional vectors are represented as *column* vectors ($d \times 1$) and lowercase font, where matrices are in uppercase font.

5.2.1 Regularization on the probability simplex

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

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

that makes the noisy target observations as "similar as possible" to the cleaner observations from the auxiliary source. This is implemented as a convex combination of class-specific linear regularizers. We start by discussing the linear regularizers and the procedure to learn them, to then discuss their combination in Section 5.2.4.

5.2.2 Linear regularizers

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

$$A = TH, \tag{5.1}$$

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

$$\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{pmatrix} = \begin{pmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{pmatrix} \begin{pmatrix} h_1 & h_2 & \cdots & h_L \end{pmatrix} \tag{5.2}$$

and h_i are the columns of H. It is assumed that $N > L$ and (5.1) has no analytical solution. We seek the best H in the least squares sense, under the constraint that the transformed vector lies in $\mathcal{S}$, *i.e.*

$$t_i^T h_k \geq 0, \quad \forall i = 1 \ldots N, \forall k = 1 \ldots L \tag{5.3}$$

and

$$t_i^T H\mathbf{1} = 1, \quad \forall i = 1 \ldots N, \tag{5.4}$$

where **1** is the vector of all ones. This least squares problem can be written in the canonical form

$$\begin{aligned} x^* &= \arg\min_x \parallel Mx - b \parallel_2^2 \\ \text{subject to:} \quad & Mx \succeq \mathbf{0} \\ & Sx = \mathbf{1}. \end{aligned} \tag{5.5}$$

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

$$S = \begin{pmatrix} t_1^T & t_1^T & \cdots & t_1^T \\ t_2^T & t_2^T & \cdots & t_2^T \\ \vdots & \vdots & & \vdots \\ t_N^T & \cdots & t_N^T & t_N^T \end{pmatrix} \tag{5.6}$$

and rewrite the transformation of (5.1) as

$$b = Mx, \tag{5.7}$$

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

$$\underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} t_1^T & 0 & \cdots & 0 \\ 0 & t_1^T & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & t_1^T \\ t_2^T & 0 & \cdots & 0 \\ \vdots & & & \\ 0 & \cdots & 0 & t_N^T \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_L \end{pmatrix}}_{x}. \tag{5.8}$$

Since the constraints are affine the feasible set is convex, and the optimization problem of (5.5) is convex whenever $M^T M$ is positive definite.

Positive definiteness of $M^T M$

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

$$M^T M = \begin{pmatrix} B & 0 & \cdots & 0 \\ 0 & B & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & B \end{pmatrix}, \tag{5.9}$$

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

$$B = \sum_{i=1}^{N} (t_i t_i^T), \tag{5.10}$$

it has full-rank if there are at least L linearly independent t_i in this summation. In this case, $B \succ 0$, $M^T M \succ 0$, and the solution of (5.5) is a global minimum. Making $N \gg L$ yields $rank(B) = L$ almost surely. In practice, the stochastic nature of t_i makes it sufficient to have $N = L$[1].

Multivariate Bernoulli representation

In this section, we briefly discuss the extension of the regularization procedure presented above to the case of multivariate Bernoulli semantic representations. The only modification needed, is to replace the constraint that regularized semantic descriptors must add up to one ($Sx = \mathbf{1}$) by a constraint that each concept probability must be less or equal to one ($Mx \preceq \mathbf{1}$). The optimization problem of (5.5) then becomes

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

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

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

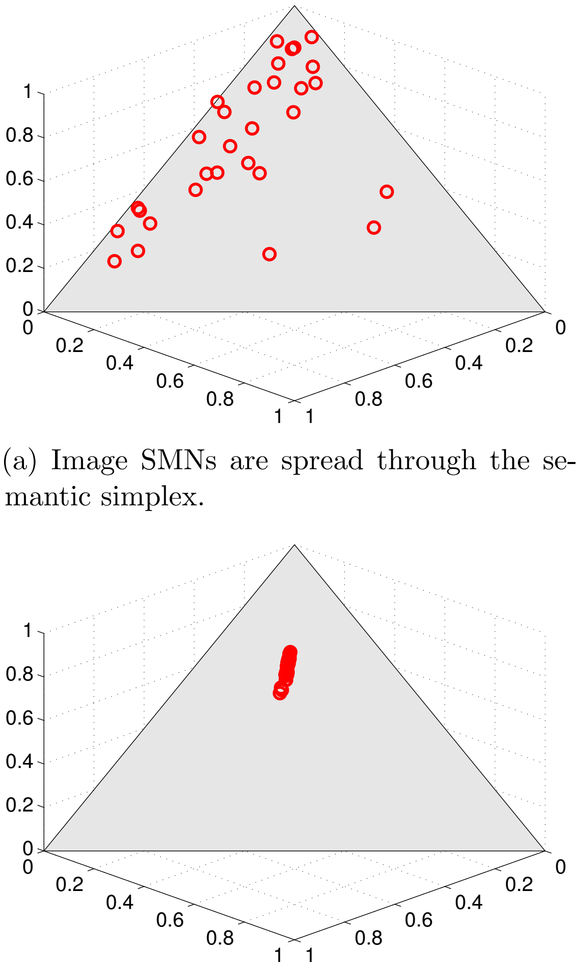
5.2.3 Learning

The optimization of (5.5) is a quadratic programming problem and can be solved by many standard optimization procedures. In our implementation, we use an active-set strategy (also known as a projection method) similar to that of [41, 42]. In all experiments, the matrix $M^T M$ was found to be positive definite, making the solution a global minimum. From (5.8), the regularization matrix H can be assembled by sequential extraction of the columns h_i from x^*. The procedure is summarized in Algorithm 1.

Algorithm 1 compute regularization operators (5.5)

```
input: train set of images and auxiliary data $\forall$ classes $i = 1, 2, \ldots, L$
    $\mathcal{T}_i = \{I_1, I_2, \ldots, I_N\}$
    $\mathcal{A}_i = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N\}$
1   compute vectors of posterior probabilities
    $t_k \leftarrow \Psi(I_k)$
    $a_k \leftarrow \Theta(\mathcal{X}_k)$
2   for each concept: $i = 1, \ldots, L$
        solve: $x^* = \arg\min_x \| Mx - b \|_2^2$
                  s.t. $Mx \succeq \mathbf{0}$
                       $Sx = \mathbf{1}$
    where $M, b$ are defined in (5.8) and $S$ in (5.6).
output: set of regularization operators: $\mathcal{H} = \{H_1, H_2, \ldots, H_L\}$
```

A conceptual illustration of the regularization is given in Figure 5.2. The figure shows the outcome of the regularization on a small sample of images from the "Warfare" class of the Wikipedia dataset, using a semantic space of three concepts ("Warfare", "History", and "Royalty"). The images are represented by their



(a) Image SMNs are spread through the semantic simplex.

(b) After regularization the SMNs cluster towards one corner of the simplex.

Figure 5.2: Image SMNs before (a) and after (b) class-specific regularization.

SMNs, shown in Figure 5.2a, which, due to the ambiguity of image classification, are scattered throughout the probability simplex. The auxiliary source is text. Figure 5.2b shows the result of the regularization of the image SMNs, t, with the transformation

$$\Phi(t) = H^T t. \tag{5.11}$$

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

5.2.4 Class-adaptive regularization

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

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

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

The two procedures are summarized by Algorithm 2-(i) and (ii), respectively.

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

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

Algorithm 2-(i) classification-based regularization

```
input: set of regularization operators $\mathcal{H}$, and image-text pair $(t,a)$,
where $t$ is the image to regularize and $a$ its auxiliary information.
1  $j^* = \arg\max_j P(j|a), \quad \forall j = \{1,2,\dots,L\}$
2  $\Phi(t) \leftarrow H_{j^*}^T t$
output: regularized image $\Phi(t)$
```

Algorithm 2-(ii) interpolation-based regularization

```
input: set of regularization operators $\mathcal{H}$, and image-text pair $(t,a)$,
where $t$ is the image to regularize and $a$ its auxiliary information.
1  $w_j(t) \leftarrow f_j(a), \quad \forall j = \{1,2,\dots,L\}$
           $f_j()$ is a regression function for class $j$
2  $w \leftarrow \sigma(w)$
3  $\Phi(t) \leftarrow \sum_i w_i(t) H_i^T t$
output: regularized image $\Phi(t)$
```

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

Regularizing in the absence of auxiliar modality

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

5.2.5 Classification and regression functions

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

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

$$\min_{w} \frac{1}{2} w^T w + C \sum_i \log(1 + \exp(-y_i w^T x_i)) \tag{5.12}$$

where y_i is the class label, x_i the input feature vector, and w a parameter vector. A multi-class LR returns a vector of posterior probabilities that can be used as weights in the interpolation scenario. For classification, we select the class of largest posterior probability. Our implementation of LR is based on the Liblinear package of [32].

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

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_i \xi_i \tag{5.13}$$

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

where w and b are the hyperplane parameters, y_i the class label, x_i input feature vectors, ξ_i slack variables that allow outliers, and $C > 0$ a penalty on the number of outliers. SVM classification can be used directly to select the regularization operator. For interpolation, the SVM scores $y_i w^T x_i$ can be converted into class probabilities through a calibration function. Our SVM implementation is based on the LibSVM [15] package.

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

$$f(x) \sim \mathcal{GP}(m(x), k(x, x^T)), \tag{5.14}$$

which is specified by a mean and covariance functions

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

In this work, we adopt a squared-exponential covariance and affine mean, with a Gaussian likelihood function. This combination enables an exact inference procedure, which is implemented with the GPML [89] package.

5.3 Experiments

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

5.3.1 Experimental set-up

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

Datasets

Three datasets are used in all experiments: “TVGraz” [53] contains 2,058 image/text pairs of 10 semantic categories, “Wikipedia” [84] 2,866 pairs from 10 categories, and “Pascal sentences” [83] 1,000 pairs from 20 categories. Table 5.1 shows the train/test split used in the following experiments.

Table 5.1: Data split for training and test sets.

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

Performance metrics

Retrieval performance is assessed with Precision-Recall curves. To facilitate comparisons of different methods, mean average precision (MAP) and R-Precision are also shown.

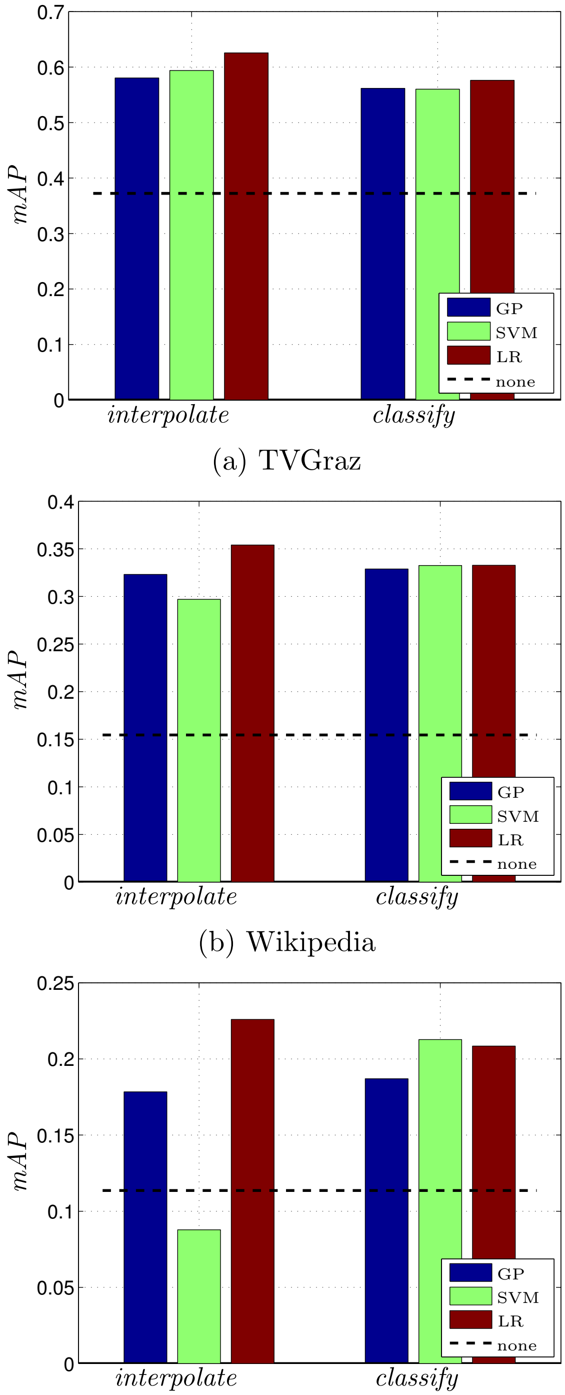
Image and text representation

All images are represented as a *bag-of-words* (BOW) [18], using SIFT descriptors quantized with a 1,024 visual word codebook. Text representation is based on *latent Dirichlet allocation* [11]. An LDA model is learned from all texts, and used to compute the probability of each text under 100 hidden topics. This probability vector is used for text representation. Both this and the image representation are mapped into a semantic space whose features are the classes that compose the dataset. This is implemented by designing a classifier $\mathbf{\Psi}$ of visual word histograms and a classifier $\mathbf{\Theta}$ of hidden topic probabilities. In both cases, the classifier is a multi-class logistic regressor [32] and the semantic descriptor the vector of posterior probabilities of equation (4.1).

5.3.2 Regularization methods

A first set of experiments is designed to evaluate the effectiveness of various regularizer configurations. This includes classification vs. interpolation based regularization (Algorithm 2-(i) vs. 2-(ii)) and the choice of classification or interpolation function (GP, SVM or LR). In these experiments all database images have auxiliary text. Figure 5.3 compares the MAP of all regularization methods. In each graph, the dashed line labeled "none" represents QBSE without regularization. Since it makes use of no auxiliary information, this lower-bound can be seen as measure of the visual complexity of each dataset. It confirms that both Wikipedia and Pascal are significantly more challenging than TVGraz.

The figure shows that the benefits of regularization are substantial for all datasets. In some cases, the regularized MAP is more than double of that achieved without regularization. With the exception of SVM-based interpolation, all methods achieve significant gains in all datasets. In general, the relative gains over *normal* QBSE are largest for the more difficult datasets. Concerning the relative performances of the different regularizers, the two regularization strategies have similar performance, with a slight advantage for interpolation in TVGraz and Wikipedia and a slight advantage for classification in Pascal. With regards to the choice of regularization functions, SVMs tended to be weaker than GPs and LR



(a) TVGraz

(b) Wikipedia

(c) Pascal-sentences

Figure 5.3: Retrieval performance (MAP) of the various regularizer configurations on the three datasets. The dashed line denoted "none" indicates the score without regularization (*i.e.* QBSE).

for interpolation, but perform well under the classification strategy. Overall, the best performance was achieved by the LR implementation of interpolation-based regularization.

5.3.3 Coping with absent text

A second set of experiments is designed to evaluate the robustness of the regularization to missing auxiliary data. In these experiments only a percentage of the database images are complemented by text. These images are regularized with the interpolation-based regularizer as detailed in Algorithm 2-(ii), Section 5.2.4. For the regularization of the remaining images different weighting functions (w) are tested. Denoted $w_{\langle \mathit{function} \rangle}(\langle \mathit{feature} \rangle)$, where the possible values for ⟨*function*⟩ and ⟨*feature*⟩ are listed in Table 5.2.

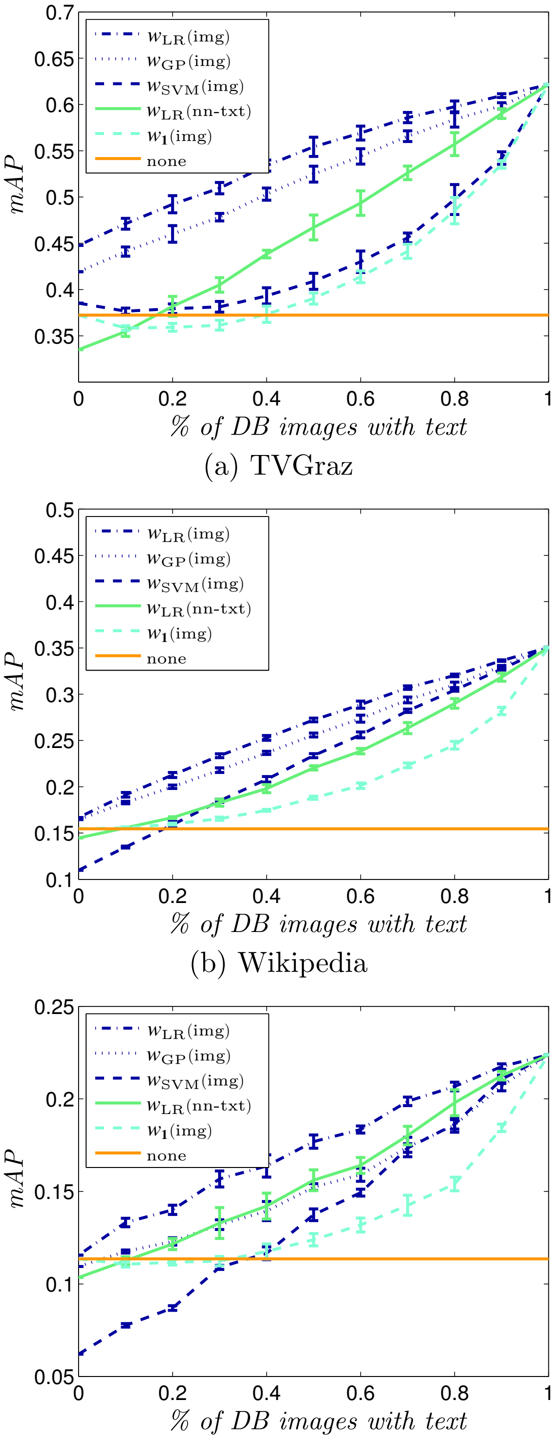
Table 5.2: Functions and features used to obtain the regularization weights for an image with no text.

⟨*function*⟩	⟨*feature*⟩
LR	image
GP	
SVM	*NN*-text
1	

All ⟨*function*⟩-⟨*feature*⟩ pairs are admissible combinations to obtain regularization weights. Logistic regression (LR), support vector machines (SVM) and Gaussian processes (GP) are interpolation functions detailed in Section 5.2.5. Another possible function is the *identity* (denoted **1**) that maps the ⟨*feature*⟩ vector directly to act as the weights. For image features all functions are tested: w_{LR}, w_{GP}, w_{SVM} and $w_{\mathbf{1}}$. When the image has no text of its own, we look for a *surrogate* text, and since the superiority of LR-based interpolation has already been established for text features in the previous section (Figure 5.3), when using these features (*NN*-text) we test only: w_{LR}. These experiments are repeated for various percentages of images with text. Each experiment is repeated five times, each using a different random set of such images.

Figure 5.4 shows plots of MAP vs. the % of images complemented by text. As before, we present the lower-bound of QBSE without regularization – labeled "none". A second lower bound was computed by regularizing only the images that are complemented by text while applying the identity weights to the remaining images – labeled "$w_{\mathbf{1}}$(img)". While superior to standard QBSE, this approach is not very robust. Its MAP degrades quickly as the percentage of text decreases. Better results are achieved by using a surrogate text to weigh the regularization operators applied to images without text. For clarity we only present the implementation of LR-based regularization for surrogate text – labeled "$w_{\mathbf{LR}}$(nn-txt)". As mentioned, this method achieved superior performance when compared to GP and SVM. However, the surrogate text features under-perform the image-driven selection of regularization operators. The remaining curves in each plot correspond to the implementation of this strategy with LR, GP, and SVM – labelled "$w_{\mathbf{LR}}$(img)", "$w_{\mathbf{GP}}$(img)" and "$w_{\mathbf{SVM}}$(img)" respectively. Among these, LR achieves the best results on all datasets.

Overall, the experiments of this and the previous section provide strong evidence for the benefits of regularization. Best results are obtained with an interpolation-based regularizer, using class-probabilities inferred with LR to weigh the class-specific regularization operators. This strategy proved quite robust to the absence of auxiliary text in the retrieval set. For images without text, good results are obtained by simply using the class probabilities derived from the image itself to weigh regularization operators. For example, on the harder Wikipedia



(a) TVGraz

(b) Wikipedia

(c) Pascal-sentences

Figure 5.4: MAP of the different regularizers vs. the percentage of database images complemented with text.

and Pascal datasets, the MAP achieved with regularization was double that of baseline QBSE when only 60% of the images contained text. On the less challenging TVGraz dataset, where image-based estimates of class probability are more reliable, it improved on QBSE even when *no* images have auxiliary text. Interestingly, in all datasets, this regularization strategy also led to a nearly-linear increase in MAP with the percentage of database images complemented by text. For all these reasons, we only considered the LR implementation of interpolation-based regularization in the remaining experiments.

5.3.4 Alternative regularization methods

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

Previous approaches to cross-modal adaptation, *e.g.* [81, 19], represent images and text in low-level feature spaces and attempt to learn a translator function that maps text into the image domain. This is done by measuring co-occurrences of visual and text words on image-text pairs. To compare the proposed regularization approach with these methods, we implemented an extension of the Text-To-Image translator (TTI) method of [81]. The implementation was based on code provided by the authors, which learns a translator function that assigns a confidence value to image/text pairs. This is a measure of how relevant the text is for the image. Preliminary experiments showed that best results were obtained by learning one translator per semantic class. In all experiments, each image/text pair in the retrieval set is represented by concatenation of the scores computed for all classes. Since queries have no text, query images were paired with the average text computed from the training set.

Previous approaches to both cross-modal and image-specific domain adaptation have proposed global transformations between the auxiliary and target domains. For example, the (DT) method of [93] learns the linear transformation, W, that minimizes the regularization cost $tr(W) - \log\det(W)$ subject to constraints that enforce (positive) similarity for a random sample of same-class object pairs. The choice of regularizer and constraints had been previously proposed in [56], where it is denoted Information Theoretic Metric Learning (ITML). Since the learned transformation is always symmetric positive definite, the method is denoted $\mathrm{DT}_{\mathrm{Symm}}$. A variant of this method, proposed in [94], uses a different objective function that does not enforce positive definitness. This is referred to as $\mathrm{DT}_{\mathrm{Asymm}}$. Max-Margin Domain Transforms (MMDT) was later proposed in [47]. This is a combination of domain and model adaptation that optimizes an objective function of a discriminant classifier rather than the similarity measure used in [57]. Finally, we also consider the Geodesic Flow Kernel (GFK) method of [43] (GFK). This method models domain shift by integrating an infinite number of subspaces that establish a path between the auxiliary and target domains. It determines the optimal dimensionality of the subspaces in which to embed the two domains and constructs the geodesic curve connecting them through the Grassmann manifold. The geodesic distance is used to define a kernel that measures similarity between auxiliary and target data. For more details on these methods the reader is referred to the original publications.

Table 5.3: MAP scores of the proposed regularizer with those of previous approaches. Relative gains with respect to the latter are shown in (%).

Method		TVGraz		Wikipedia		Pascal	
		mAP	%	mAP	%	mAP	%
RIS		**0.622**	-	**0.356**	-	**0.224**	-
TTI [81]		0.531	*17*	0.323	*10*	0.220	*2*
MMDT [47]		0.405	*53*	0.155	*129*	0.115	*95*
GFK [43]		0.384	*62*	0.155	*129*	0.131	*71*
DT	Symm. [93]	0.375	*65*	0.153	*133*	0.101	*122*
	Asymm. [94]	0.425	*46*	0.152	*134*	0.118	*90*
QBSE [86]		0.372	*67*	0.155	*129*	0.114	*97*
Random		0.1	*522*	0.1	*256*	0.05	*348*

All methods were implemented with the code provided by the authors. Table 5.3 summarizes all results in the MAP scores over all queries[2]. These results support several conclusions. First, the class-specific transformation used by both the proposed regularizer and our extension of TTI achieves better regularization than the holistic transformation of the space used by the other methods. This seems to be

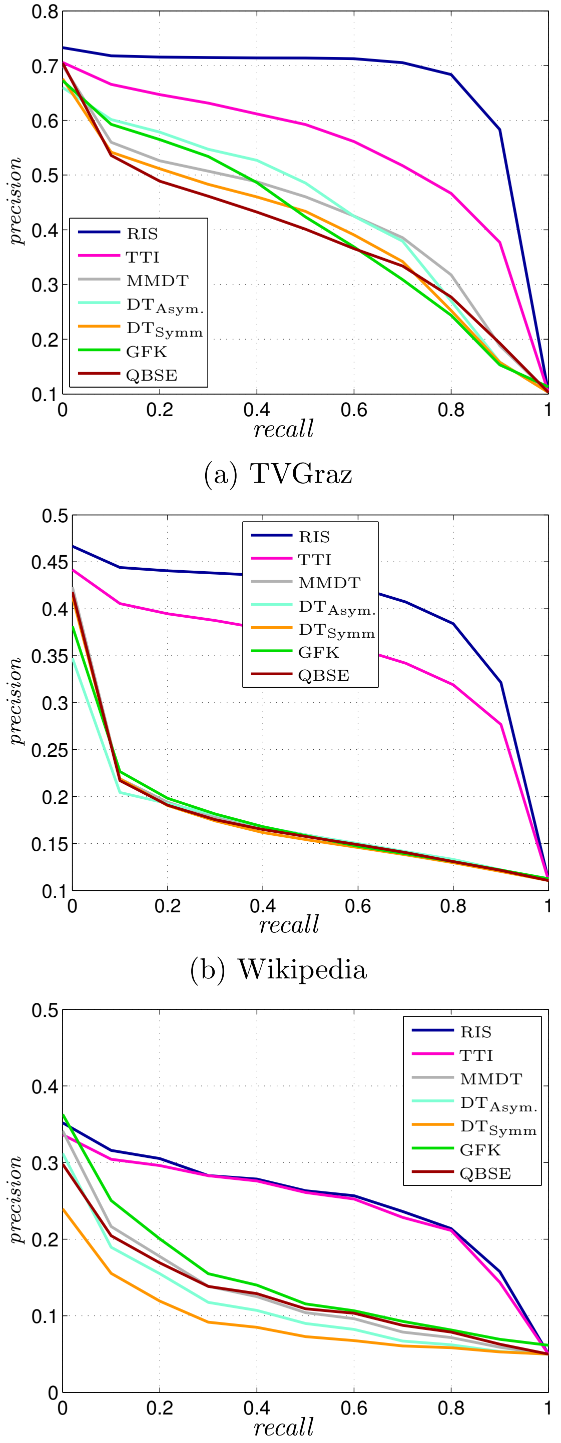
[2] We note that some of the results reported in the table for TTI and QBSE are weaker than those reported in our similar experiment in [17]. This is due to the fact that the similarity functions used for the image retrieval operation are different. The centered normalized correlation was used in [17], while we use the Kullback-Leibler divergence of (4.2) in this work. These functions yield slight variations in the MAP for certain dataset/method combinations. However, the differences are small and do not affect the conclusions of this work.

particularly important on the datasets (Wikipedia and Pascal) where image classification is most ambiguous. Second, the simpler learning problem inherent to the representation in semantic space (no need to learn a translator function) enables further improvements. This is visible both by 1) the better performance of the proposed regularizer than TTI, and 2) the better performance of the global transform methods in the semantic space (observed in our preliminary experiments). Third, all methods outperformed QBSE in at least some datasets, with significant gains for the proposed regularizer.

Table 5.4: R-precision scores of the proposed regularizer with those of previous approaches. All methods were implemented with code provided by the authors.

Method		TVGraz	Wikipedia	Pascal
		R-Precision		
RIS		**0.554**	**0.272**	**0.182**
TTI [81]		0.476	0.259	0.168
MMDT [47]		0.400	0.158	0.114
GFK [43]		0.372	0.159	0.135
DT	Symm. [93]	0.377	0.157	0.102
	Asymm. [94]	0.396	0.148	0.120
QBSE [86]		0.368	0.156	0.107
Random		0.1	0.1	0.05

Table 5.4 shows R-precision scores, which are in line with the conclusions mentioned for MAP, and Figure 5.5 plots the behavior of the system (*i.e.* Precision) at different levels of Recall by showing the 11-point interpolated PR curves for all methods. Overall, these results confirm that the regularization of image semantics is beneficial (improvements over QBSE), and show that both the semantic representation and the class-adaptive nature of the proposed regularizer are beneficial



(a) TVGraz

(b) Wikipedia

(c) Pascal-sentences

Figure 5.5: Precision-recall curves of different regularizers on the three datasets. The proposed method is denoted as “RIS”.

for image retrieval. Finally, these gains tend to be most significant when the ambiguity of image classification is largest, as in the Wikipedia and Pascal datasets. Figure 5.6 illustrates the robustness of the retrieval operation after semantic regularization, by presenting the top four matches for various query images from the three datasets. Each query is shown in a different row, displaying the query image on the left and the top matches on the right.

5.4 Discussion

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

This robustness was justified by two properties of the proposed regularizer. The first is the semantic nature of the underlying image and text representation. This enables the modeling of contextual relationships between semantic concepts and establishes a unified space for image and text data. In result, the cross-modal regularization problem is reduced to one of adaptation between two homogeneous domains, *i.e.* there is no need to learn a translator between images and text. It was shown that, when compared to previous proposals to cross-modal regularization, this significantly simplifies the learning problem, enabling better generalization. The second is the implementation of the regularizer as a combination of class-

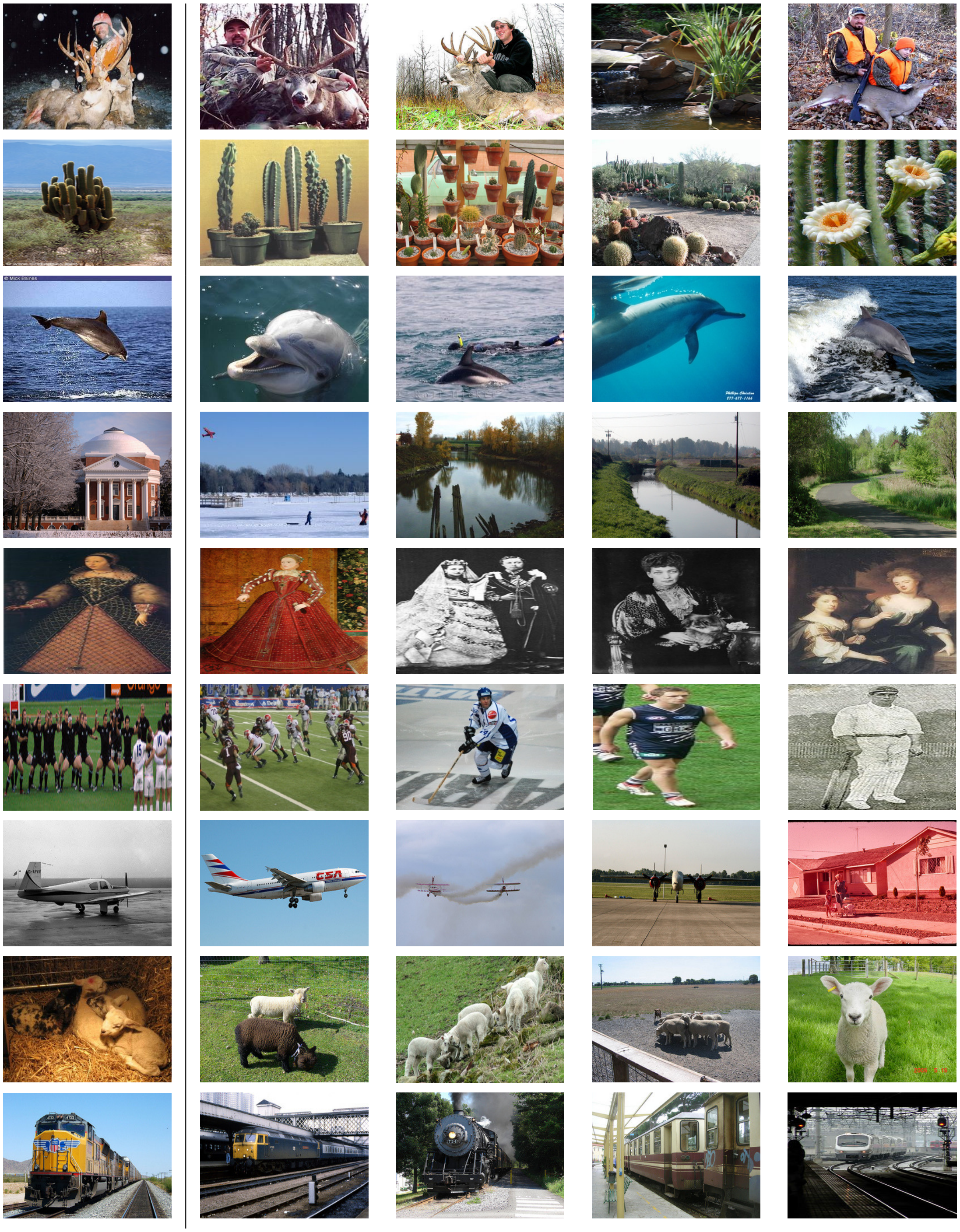


Figure 5.6: Retrieval examples, three queries of TVGraz (top), Wikipedia (middle) and Pascal-sentences (bottom). In all cases the query image is shown on the leftmost column and top four database matches on the right.

specific regularizers. This leads to a piecewise-linear transformation of the image descriptors to regularize, which is highly non-linear but can be learned efficiently. When compared to previous approaches to domain adaptation in computer vision, the resulting regularizer is both more flexible and naturally aligned to the semantics of images and text. This was shown to enable significant gains in regularization performance.

Acknowledgements

The text of Chapter 5, in part, is based on the material as it appears in: 'Cross-modal Domain Adaptation for Text-based Regularization of Image Semantics', J. Costa Pereira, N. Vasconcelos, Elsevier, Computer Vision and Image Understanding, Vol. 124, pp. 123-135, July 2014. The dissertation author was a primary researcher and an author of the cited material.

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

Chapter 6

Conclusions and Future Work

Image representation is at the very core of Computer Vision. Tasks like classification, detection or retrieval, in a large part, depend on good image representations for their success. Semantic representations gained popularity when it became obvious that simple visual signatures where not capable of encoding the complexity and variability so common in visual scenes. In particular, the disparity between strict visual similarity and the human notion of image similarity is a well known problem and is referred in the literature as the *semantic gap.* This has spurred significant interest in image representations, particularly those that have higher levels of abstraction. Depending on the task, it is probably not difficult to find a problem where enconding edges or color blobs in a representation is less important than knowing wether the scene refers to an indoor or outdoor scenario. A representation that is capable of enconding these high-level characteristics – not necessarily identifiable by a single visual cue – is likely to allow the minimization of the (semantic) gap between the two notions of similarity.

Traditionally retrieval on the visual space is made using some visual signature extracted directly from the images. The existance of the semantic gap lead to the adoption of signatures that are semantic in nature, leading to significant improvements in retrieval performance [86]. But the sheer diversity of images still makes content-based image retrieval (CBIR) a very challenging problem; possibly made even more dificult by the ambiguity of what is to be considered relevant in the query (image) provided by the end user. Traditional solutions for CBIR, usually involve new ranking algorithms or improved image representations. We extend the subject of image representations to one of multi-modal scene understanding. Making an analogy on how humans understand images using full sensory capabilities, we argue that images can also be understood using different sources of informative data about their contents, amenable to produce better semantic representations. The *rationale* is that more information yields more robust image representations, leading to improved retrieval accuracy.

We proposed a method that works on the semantic space and acts as a cross-modal regularization function for images. In many ways this is closely related to subjects like domain adaptation or transfer learning. Loosely speaking, they both consist of transferring information from an auxiliary dataset to regularize a learning operation on a target dataset. In domain adaptation it is usually assumed that the input distribution – from auxiliary to target data – changes, but the labels remain the same; while in transfer learning, the input distributions – for both target and auxiliary data – stay the same, but the labels change. Methods that have been proposed to address this problem can be split into three major groups:

1) those that follow the reasoning that heterogenous data can only have highly non-linear mappings, such is the different nature of target and auxiliary domains.

This class of methods focus on learning complex transformations that capture the non-linearities trying to explain the existing relationship between the two low-level spaces, where the data is originally represented. These methods usually involve kernels and learning algorithms of substantial complexity. They are hampered by two factors: one, there is no attempt to introduce some generalization, *e.g.* semantic abstraction; two, they try to explain the heterogeneity between target and auxiliar data with a single transformation.

2) the second class of methods, tackles the latter problem (of learning a single transformation). It attempts to find a piecewise transfer learning between the two domains. All methods that follow this line of strategy have shown to behave comparatively better to other that opt for a single transformation. To some degree, however, these methods too fall short. A signal that translation might still be required (between domains).

3) semantic-based methods achieve this translation. In particular, our proposal maps both domains to isomorphic spaces where data shows different levels of semantic ambiguity. Inherently semantic, and therefore more accurate, the auxiliary domain is used to learn appropriate piecewise transformations for regularization in the target domain.

The advantages of the latter approach are two-folded: one, semantic abstraction avoids difficult translations while keeping complexity low, and second, piecewise rather than holistic transformations can better account for class-specific variations. Those are the strong points of the proposed framework. Domain adaptation and transfer learning techniques are frequently used to cope with absent or insufficient data, *e.g.* in situations where it is deemed difficult to obtain more data on a certain target domain. For cross-modal regularization this is not a major concern.

Here, the problem is instead that (the same) data is being described in two different ways (modalities) with different degrees of semantic ambiguity: cross-modal regularization is useful even if there is infinite image data.

6.1 Future Work

This work is a preliminary effort showing how multi-modal modeling can be useful in traditional computer vision problems. We have shown the benefits of using cross-modal regularization in semantic representations to aid solving problems where auxiliary data is present. For traditional uni-modal problems, such as CBIR, one method was proposed – piecewise linear transformation in the semantic space – achieving great success; with substantial improvements in retrieval accuracy for three benchmark datasets. This is but one application for cross-modal regularization, but the paradigm of learning using privileged information is gaining notoriety.

Recently, famous mathematician Vladimir Vapnik has developed a theory of Learning Using Privileged Information (LUPI) [113], also known as the SVM+ method. Privileged information is data that can be accessed during training but not during test time. This paradigm raises interesting questions and unveils new paths of research. Scenarios of privileged information are present in many situations of our daily life. For example, in the problem of CBIR with cross-modal regularization the textual data is the privileged inforamtion. In medical imaging, learning from complementary means of diagnostic can improve detection on certain diseases; a case using fMRI and EEG data is shown in [68]. Another interesting and challenging problem is to be able to interpret the queries; *e.g.* a user might be searching for a certain product, service or person that he is able to describe but for which he does not know the description or name. This immediately suggests hierarchical semantic representations, with fine grained search at the lower levels of the hierarchy.

A universal semantic space is an ultimate gooal, but something that we see difficult to attain in a near future. But to be able to learn semantic spaces of successive increasing complexity would be an extremely valuable contribution, and would lead the way to this greater goal.

Appendix A

Experimental set-up

The research presented in this dissertation posed many challenges. Even though new problems were presented, such as cross-modal similarity, we also address classical vision problems such as content-based image retrieval (CBIR). The proposed approach to CBIR includes scenarios where multi-modal data is needed. This fact prevents the usage of most, if not all, benchmark datasets. An inital challenge in the course of this work was to setup a new benchmark corpus for vision; one that included auxiliary textual data This was achieved by collecting a few thousand articles from a popular online encycolpedia, which was later made public to the research community. In less than five years of that publication, it has received great enthusiasm and acceptance from researchers with well over 200 citations.

In this appendix we provide a brief overview of the adopted datasets, image and text representations used, and a general description of the experiments and the evaluation metrics used.

A.1 Datasets

Wikipedia [84] is a novel dataset, assembled from the "Wikipedia featured articles". These are typical Wikipedia articles that are reviewed by their editors, for superior quality. They fall into one of 30 possible categories. The complete list of categories can be found on Table A.1a. Since some of these contain very few entries, in most of the experiments involving this dataset, we considered only articles from the 10 top most populated classes. Featured articles tend to have multiple images and span multiple topics, for this reason each article was split into sections, based on its section headings; and each image was assigned to the section in which it was placed by the author(s). This produced $7,114$ sections, which are internally more coherent and usually contain a single picture. The dataset was then pruned, by keeping only sections with exactly one image and at least 70 words. The final corpus of top-10 most populated classes contains a total of $2,866$ documents. The median text length is 200 words.

On this dataset, classes are very broad themes (*e.g.* "Media", "Music", "Biology"). Therefore intra-class image variability is quite large, and image classification tends to have low accuracy. In fact, in the absence of additional information (*i.e.* text), many of the images are difficult to classify even for a human subject. On the other hand, the text that accompanies the images is fairly unambiguous, and it yileds a much higher classification accuracy.

TVGraz [53] contains narrow object classes (*i.e.* "Caltech-like"). It is a collection of web-pages compiled by Khan *et al.*. Google image search was used to retrieve 1,000 web-pages for each of the 10 categories they selected from Caltech-

256 [45]. The classes used are listed in Table A.1b. This dataset is provided as a

Table A.1: Class names for the three datasets used in the experiments

#	Name
1.	Art, Architecture and Archaeology
2.	Biology
3.	Geography and Places
4.	History
5.	Literature and Theatre
6.	Media
7.	Music
8.	Royalty, Nobility and Heraldry
9.	Sport and Recreation
10.	Warfare
11.	Awards, Decorations and Vexillology
12.	Business, Economics and Finance
13.	Chemistry and Mineralogy
14.	Computing
15.	Culture and Society
16.	Education
17.	Engineering and Technology
18.	Food and Drink
19.	Geology, Geophysics and Meteorology
20.	Health and Medicine
21.	Language and Linguistics
22.	Law
23.	Mathematics
24.	Philosophy and Psychology
25.	Physics and Astronomy
26.	Politics and Government
27.	Religion, Mysticism and Mythology
28.	Transport
29.	Video gaming
30.	Miscellaneous

(a) Wikipedia

#	Name
1.	Brain
2.	Butterfly
3.	Cactus
4.	Deer
5.	Dice
6.	Dolphin
7.	Elephant
8.	Frog
9.	Harp
10.	Pram

(b) TVGraz

#	Name
1.	Aeroplane
2.	Bicycle
3.	Bird
4.	Boat
5.	Bottle
6.	Bus
7.	Car
8.	Cat
9.	Chair
10.	Cow
11.	Dining-table
12.	Dog
13.	Horse
14.	Motorbike
15.	Person
16.	Potted plant
17.	Sheep
18.	Sofa
19.	Train
20.	TV monitor

(c) Pascal (VOC) sentences

list of URLs, which we used to collect $2,058$ image-text pairs (defunct URLs and web-pages without at least 10 words and one image were discarded). The median text length, per web-page, is 289 words.

TVGraz images are archetypal members of the categories, making this dataset eminently visual. Its categories (*e.g.*, “Harp”, “Dolphin”) are specific objects or animals. Their text counterpart can be less representative of the categories, since the web-pages are sometimes only loosely related to the image. However, the text, although less stylistic than that of Wikipedia, is still informative of the class. This leads to fairly high classification accuracies for both images and text classifiers.

Pascal-sentences [83] originates from a subset of Pascal VOC [31] images that were augmented with five sentences written by human annotators. There are $1,000$ image-text pairs from 20 different categories, listed in Table A.1c. The added text provides some context for each picture, but is not a semantically rich document. On both Wikipedia and TVGraz, the text is much more extensive and informative. The Pascal-sentences was released in 2010. Therefore this dataset was not used in the cross-modal retrieval experiments of Chapter 3. The dataset was made public in the same year that our work on that subject was released, hence we were unaware of its existance.

The three datasets described above, exhibit different properties. The Wikipedia categories are abstract concepts, and have a broad scope. Individually, images from this dataset can be difficult to classify, even for a human. The class label is mostly driven by text. These are of high quality and representative of the category, while image categorization is more ambiguous. For example, a portrait of a historical figure can appear in both class “War” or “History”. TVGraz is an object-driven dataset, it’s images are those of well-known objects with different appearances. The text is only loosely related to the image. Pascal-sentences dataset can contain multiple objects in a single image, but only one ground-truth label. The associated text relates closely to the images but it is fairly limited in its descriptive power, *e.g.* one sentence describing the image.

Train and test splits

For all experiments there was a learning and a testing stage. In cross-modal retrieval experiments, classifiers for each modality had to be learned; and for content-based image retrieval (CBIR), the learning of cross-modal regularization operators and interpolation functions must also be done. As usual, the datasets were split in two; one for training and another for testing, in the range of 70-80% for the former and 30-20% for the latter. When validation is necessary, to determine best model parameters, the training set is itself divided to include a small validation cut.

In each case, the training set is used to learn all (uni-modal) semantic classifiers and regularization operators (both classification or interpolation functions and linear regularizers). On Wikipedia, a random split of $2,173$ documents for training and and 693 documents for testing was made. In TVGraz another random split produced $1,558$ training and 500 test documents. In Pascal-sentences 700 documents are used for the training stage and the remaining 300 are used only for testing. This is summarized in Table A.2.

Table A.2: Data split among training and test sets.

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

A.2 Image and text representation

For both modalities, the base representation is a bag-of-words (BOW) [18]. Text words (extracted by stemming the text with the Python Natural Language Toolkit[1]) were fit by a latent Dirichlet allocation (LDA) [11] model. The probability of each text BOW under the LDA-discovered topics is used for text representation. For images, a bag of SIFT descriptors was first extracted per training image[2] and a visual word codebook learned with K-means clustering. SIFT descriptors extracted from each image were finally vector quantized with this codebook, to produce a vector of visual word counts which, ultimately, is used as representation for the images.

A.3 Retrieval experiments

There are essentially two kinds of retrieval experiments performed in this dissertation: 1) based on cross-modal similarity, where the query object is from a different nature than that of retrieved results; and 2) classical content-based image retrieval (CBIR), where an image is provided as a query and images from the database are ranked according to highest similarity.

For all retrieval operations, unless otherwise stated, only the test portion of the dataset is used. A query refers to the act of selecting one object from the database and using it to rank the remaining ones. One by one, all objects in the test set are used to query the database containing the remaining test objects; *i.e.* experiments are carried out in a leave-one-out setting, averaging the results over all queries. The terms retrieval set and database are used interchangeably when

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

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

referring to the repository of objects being ranked (images or texts).

Performance metrics

Cross-modal retrieval experiments of Chapters 3 and 4 was evaluated with the two datasets available in that moment: *TVGraz* and *Wikipedia*. In this cross-modal retrieval setting, two tasks were considered: text retrieval from an image query, and image retrieval from a text query. All text queries were based on full text documents. Chapter 5 where classical CBIR experiments are performed, uses the three datasets, *i.e.* also including *Pascal-sentences*. In both scenarios, retrieval performance is evaluated using several information retrieval metrics: mean average precision (MAP), R-Precision and Precision-Recall curves. All these metrics are based on two fundamental concepts: Precision and Recall.

$$\text{Precision} = \frac{|\{\text{relevant}\}\bigcap\{\text{retrieved}\}|}{|\{\text{retrieved}\}|} \tag{A.1}$$

$$\text{Recall} = \frac{|\{\text{relevant}\}\bigcap\{\text{retrieved}\}|}{|\{\text{relevant}\}|} \tag{A.2}$$

The MAP score is the standard measure for evaluating information retrieval systems. It averages the precision at the ranks where recall changes, and reports this in a single number. Evaluation of relevancy is made based on the usual *s*ame/different-class paradigm, using the label information provided in Table A.1. For certain experiments, the MAP scores are computed on a per-class basis, to assess whether class information has influence on the obtained overall scores. *R-Precision* is a measure similar to the more conventional *Precision@k*. But rather than computing *Precision* at a fix level of k, it requires to know all relevant documents (r) for every query beforehand, and computes *Precision* for *Recall* level of r

(*Precision@r*). This measure strongly correlates with MAP. We also make use of the 11-point interpolated *precision-recall* (PR) curves [70] that allow visualization of precision at different levels of recall.

All results are compared to competitive methods using code provided by the authors.

Bibliography

[1] "Speakers give sound advice," *Syracuse Post Standard*, p. 18, March 28, 1911. 1

[2] T. T. Ahonen, *TomiAhonen Phone Book - Statistical Review of Handset Industry*, 2014. 2

[3] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 2252–2259. 67

[4] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C.-F. Shu, "Virage image search engine: an open framework for image management," in *Electronic Imaging: Science & Technology*, 1996, pp. 76–87. 50

[5] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *Journal of Machine Learning Research, MIT Press*, vol. 3, pp. 1107–1135, 2003. 10, 15

[6] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 408–415. 8

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Journal of Computer Vision and Image Understanding, Elsevier*, vol. 110, no. 3, pp. 346–359, 2008. 49

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002. 48

[9] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *European Conference on Computer Vision, Springer*, 2010, pp. 663–676. 50

[10] A. Bergamo and L. Torresani, “Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach,” in *Advances in Neural Information Processing Systems*, 2010, pp. 181–189. 67

[11] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research, MIT Press*, vol. 3, pp. 993–1022, 2003. 18, 35, 81, 105

[12] S. Boughorbel, J. Tarel, and N. Boujemaa, “Generalized histogram intersection kernel for image recognition,” in *Proc. IEEE International Conference on Image Processing*, vol. 3, 2005, pp. III – 161–164. 38

[13] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007. 10, 15, 52, 69

[14] R. Caruana, “Multitask learning: A knowledge-based source of inductive bias,” *Machine Learning*, vol. 28, pp. 41–75, 1997. 67

[15] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm. 30, 79

[16] J. Costa Pereira and N. Vasconcelos, “On the regularization of image semantics by modal expansion,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2012, pp. 3093–3099. 58

[17] ——, “On the regularization of image semantics by modal expansion,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2012, pp. 3093–3099. 89

[18] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *Workshop on Stat. Learn. in Comp. Vision, European Conference on Computer Vision*, vol. 1, 2004, pp. 1–22. 81, 105

[19] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, “Translated learning: Transfer learning across different feature spaces,” in *Advances in Neural Information Processing Systems*, 2008, pp. 353–360. 56, 68, 87

[20] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in *Proc. ACM International Conference on Machine Learning*, 2007, pp. 193–200. 67

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, vol. 1, 2005, pp. 886–893. 49

[22] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008. 8, 9, 12

[23] H. Daumé III, "Frustratingly easy domain adaptation," in *Association of Computational Linguistics*, vol. 1785, 2007, pp. 256–263. 67

[24] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. 23

[25] L. Denoyer and P. Gallinari, "Bayesian network model for semi-structured document classification," *, Elsevier*, vol. 40, no. 5, pp. 807–827, 2004. 8, 12

[26] M. Dixit, N. Rasiwasia, and N. Vasconcelos, "Adapted gaussian models for image classification," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2011, pp. 937–943. 67

[27] C. Do and A. Ng, "Transfer learning for text classification," in *Advances in Neural Information Processing Systems*, 2005. 67

[28] L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," *arXiv:1206.4660*, 2012. 67

[29] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," in *Advances in Neural Information Processing Systems*, 2008, vol. 20, pp. 385–392. 15

[30] H. Escalante, C. Hérnadez, L. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proc. ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 172–179. 12

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 103

[32] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research, MIT Press*, vol. 9, pp. 1871–1874, 2008. 30, 53, 79, 82

[33] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2010, pp. 2352–2359. 49, 50, 52

[34] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2009, pp. 1778–1785. 49, 50, 52

[35] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian approach to unsupervised one-shot learning of object categories," in *Proc. IEEE International Conference on Computer Vision*, 2003, pp. 1134–1141. 67

[36] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, vol. 2, 2004, pp. 1002–1009. 10

[37] J. Fisher III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems*, 2001, pp. 772–778. 12

[38] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the QBIC system," *IEEE Transactions on Computers*, vol. 28, no. 9, pp. 23–32, 1995. 50

[39] C. Frankel, M. Swain, and V. Athitsos, "Webseer: An image search engine for the world wide web," University of Chicago, Computer Science Department, Tech. Rep., 1996. 8, 50

[40] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, 2013, pp. 2121–2129. 50

[41] P. Gill, W. Murray, M. Saunders, and M. Wright, "Procedures for optimization problems with a mixture of bounds and general linear constraints," *ACM Transactions on Mathematical Software*, vol. 10, no. 3, pp. 282–298, 1984. 74

[42] P. Gill, W. Murray, and M. Wright, *Numerical Linear Algebra and Optimization.* Addison-Wesley, 1991, vol. 1. 74

[43] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2012, pp. 2066–2073. 66, 67, 68, 88, 89, 90

[44] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 999–1006. 67, 68

[45] G. Griffin, A. Holub, and P. Perona, "The Caltech-256," Caltech, Tech. Rep., 2006. 102

[46] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Journal of Neural Computation, MIT Press*, vol. 16, no. 12, pp. 2639–2664, 2004. 13, 18

[47] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell, "Efficient learning of domain-invariant image representations," in *International Conference on Learning Representations*, 2013. 66, 67, 88, 89, 90

[48] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936. 25

[49] W. Hsu, T. Mei, and R. Yan, "Knowledge discovery over community-sharing media: From signal to intelligence," in *Proc. IEEE International Conference on Multimedia and Expo*, 2009, pp. 1448–1451. 13

[50] J. Iria, F. Ciravegna, and J. Magalhães, "Web news categorization using a cross-media document graph," in *Proc. ACM International Conference on Image and Video Retrieval*, 2009, pp. 1–8. 8, 12

[51] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. ACM SIG International Conference Information Retrieval*, 2003, pp. 119–126. 10

[52] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2005. 23

[53] I. Khan, A. Saffari, and H. Bischof, "TVGraz: Multi-modal learning of object categories by combining textual and visual features," in *Workshop of the Austrian Association for Pattern Recognition*, 2009. 18, 33, 81, 101

[54] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo, "Combining image captions and visual analysis for image concept classification," in *Workshop on Neural Networks for Signal Processing at Proc. ACM SIG International Conference Knowledge Discovery and Data Mining*, 2008, pp. 8–17. 12

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in *Advances in Neural Information Processing Systems*, vol. 1, no. 2, 2012, p. 4. 50

[56] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2143–2157, 2009. 88

[57] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2011, pp. 1785–1792. 67, 88

[58] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, “Scene recognition on the semantic manifold,” in *Computer Vision–ECCV 2012*. Springer Berlin Heidelberg, 2012, pp. 359–372. 61

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

[60] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *Advances in Neural Information Processing Systems*, 2004, vol. 16. 10

[61] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, vol. 2, 2006, pp. 2169–2178. 49

[62] M. S. Lew, “Next-generation web searches for visual content,” *IEEE Transactions on Computers*, vol. 33, no. 11, pp. 46–53, 2000. 50

[63] D. Li, N. Dimitrova, M. Li, and I. Sethi, “Multimedia content processing through cross-modal association,” in *Proc. ACM International Conference on Multimedia*, 2003, pp. 604–611. 12, 13, 24, 25

[64] L. Li, H. Su, E. Xing, and L. Fei-Fei, “Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification,” *Advances in Neural Information Processing Systems*, 2010. 49, 50

[65] W. Li, K. Candan, and K. Hirata, “SEMCOG: an integration of semantics and cognition-based approaches for image retrieval,” in *ACM Symposium on Applied Computing*, 1997, pp. 136–143. 8

[66] B. Logan and A. Salomon, “A music similarity function based on signal analysis,” in *Proc. IEEE International Conference on Multimedia and Expo*, 2001, pp. 745–748. 8

[67] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal on Computer Vision, Springer, Springer*, vol. 60, no. 2, pp. 91–110, 2004. 49

[68] V. Mahadevan, C. Wah Wong, J. Costa Pereira, T. T. Liu, N. Vasconcelos, and L. K. Saul, “Maximum covariance unfolding: Manifold learning for bimodal data,” in *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 918–926. 12, 13, 98

[69] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996. 48

[70] C. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval.* Cambridge University Press, 2008. 35, 107

[71] J. Mao and A. K. Jain, “Texture classification and segmentation using multiresolution simultaneous autoregressive models,” *Pattern recognition*, vol. 25, no. 2, pp. 173–188, 1992. 48

[72] C. Meadow, B. Boyce, D. Kraft, and C. Barry, *Text Information Retrieval Systems.* Emerald Group Pub Ltd, 2007. 9

[73] T. Mei, W. Hsu, and J. Luo, “Knowledge discovery from community-contributed multimedia,” *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 16–17, 2010. 13

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

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

[76] S. Nakamura, “Statistical multimodal integration for audio-visual speech processing,” *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 854–866, 2002. 12

[77] M. Paramita, M. Sanderson, and P. Clough, “Diversity in photo retrieval: Overview of the ImageCLEF 2009 photo task,” *Multilingual Information Access Evaluation: Multimedia Experiments*, pp. 45–59, 2010. 9, 10, 12

[78] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2007, pp. 1–8. 49

[79] T. Pham, N. Maillot, J. Lim, and J. Chevallet, “Latent semantic fusion model for image retrieval and annotation,” in *Proc. ACM International Conference on Information and Knowledge Management*, 2007, pp. 439–444. 12

[80] G. Qi, C. Aggarwal, and T. Huang, “Towards semantic knowledge propagation from text corpus to web images,” in *Proc. ACM International Conference on World Wide Web*, 2011, pp. 297–306. 12, 35, 41, 60

[81] ——, “Towards semantic knowledge propagation from text corpus to web images,” in *Proc. ACM International Conference on World Wide Web*, 2011, pp. 297–306. 56, 66, 68, 87, 88, 89, 90

[82] R. Raina, A. Y. Ng, and D. Koller, “Constructing informative priors using transfer learning,” in *Proc. ACM International Conference on Machine Learning*, 2006, pp. 713–720. 67

[83] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using amazon’s mechanical turk,” in *Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. NAACL HLT, 2010, pp. 139–147. 57, 81, 103

[84] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in *Proc. ACM International Conference on Multimedia*, 2010, pp. 251–260. 33, 81, 101

[85] N. Rasiwasia, P. Moreno, and N. Vasconcelos, “Bridging the gap: Query by semantic example,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007. 11, 18, 29

[86] ——, “Bridging the gap: Query by semantic example,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007. 49, 50, 51, 53, 54, 55, 61, 66, 89, 90, 96

[87] N. Rasiwasia and N. Vasconcelos, “Holistic context models for visual recognition,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 902–917, 2012. 55

[88] ——, “Scene classification with low-dimensional semantic spaces and weak supervision,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2008, pp. 1–6. 61

[89] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006. 80

[90] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000. 67

[91] M. Rohrbach, M. Stark, and B. Schiele, “Evaluating knowledge transfer and zero-shot learning in a large-scale setting,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2011, pp. 1641–1648. 50

[92] M. J. Saberian and N. Vasconcelos, "Multiclass boosting: Theory and algorithms," in *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 2124–2132. 31

[93] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision, Springer*, 2010, pp. 213–226. 66, 88, 89, 90

[94] ——, "Visual domain adaptation using regularized cross-domain transforms," UCB/EECS-2010-106, EECS Department, University of California Berkeley, Tech. Rep., 2010. 66, 88, 89, 90

[95] G. Salton, *The SMART Retrieval System - Experiments in Automatic Document Processing.* Prentice-Hall, Inc., 1971. 9

[96] G. Salton and M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983. 8, 9

[97] S. Sclaroff, M. Cascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Journal of Computer Vision and Image Understanding, Elsevier*, vol. 75, no. 1, pp. 86–98, 1999. 8

[98] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004. 18, 26

[99] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2011, pp. 801–808. 50

[100] M. Slaney, "Semantic-audio retrieval," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 4, 2002, pp. 4108–4111. 12

[101] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Workshop on Multimedia Information Retrieval at Proc. ACM International Conference on Multimedia*, 2006, pp. 321–330. 8, 9, 10, 12

[102] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000. 8, 9, 49

[103] J. Smith and S. Chang, "VisualSEEk: a fully automated content-based image query system," in *Proc. ACM International Conference on Multimedia*, 1997, pp. 87–98. 50

[104] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proc. IEEE International Conference on Multimedia and Expo*, vol. II, 2003, pp. 445–448. 49, 50

[105] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Journal of Multimedia Tools and Applications, Springer*, vol. 25, no. 1, pp. 5–35, 2005. 8, 12

[106] M. A. Stricker and A. Dimai, "Color indexing with weak spatial constraints," in *Electronic Imaging: Science & Technology.* International Society for Optics and Photonics, 1996, pp. 29–40. 48

[107] M. A. Stricker and M. Orengo, "Similarity of color images," in *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology.* International Society for Optics and Photonics, 1995, pp. 381–392. 48

[108] M. Swain and D. Ballard, "Color indexing," *International Journal on Computer Vision, Springer*, vol. 7, no. 1, pp. 11–32, 1991. 38, 48

[109] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recog. using classemes," *European Conference on Computer Vision*, pp. 776–789, 2010. 49, 50

[110] T. Tsikrika and J. Kludas, "Overview of the wikipedia multimedia task at ImageCLEF 2008," *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 539–550, 2009. 8, 10, 12

[111] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008. 15

[112] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002. 15

[113] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009. 98

[114] N. Vasconcelos, "Image indexing with mixture hierarchies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 52

[115] ——, "Exploiting group structure to improve retrieval accuracy and speed in image databases," in *Proc. IEEE International Conference on Image Processing*, vol. 1, 2002, pp. I – 980–983. 52

[116] ——, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004. 9

[117] ——, “Minimum probability of error image retrieval,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004. 50

[118] ——, “From pixels to semantic spaces: Advances in content-based image retrieval,” *IEEE Transactions on Computers*, vol. 40, no. 7, pp. 20–26, 2007. 12

[119] ——, “From pixels to semantic spaces: Advances in content-based image retrieval,” *IEEE Transactions on Computers*, vol. 40, no. 7, pp. 20–26, 2007. 61

[120] N. Vasconcelos and A. Lippman, “A bayesian video modeling framework for shot segmentation and content characterization,” in *Content-Based Access of Image and Video Libraries, 1997. Proceedings. IEEE Workshop on.* IEEE, 1997, pp. 59–66. 49, 50

[121] ——, “Towards semantically meaningful feature spaces for the characterization of video content,” in *Proc. IEEE International Conference on Image Processing*, vol. 1, 1997, pp. 25–28. 49, 50, 61

[122] ——, “Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing,” in *Proc. IEEE International Conference on Image Processing*, 1998, pp. 153–157. 49, 50

[123] ——, “A probabilistic architecture for content-based image retrieval,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1. IEEE, 2000, pp. 216–221. 48

[124] A. Vinokourov, D. Hardoon, and J. Shawe-Taylor, “Learning the semantics of multimedia content with application to web image retrieval and classification,” in *Symposium on Independent Component Analysis and Blind Source Separation*, 2003. 18, 26

[125] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini, “Inferring a semantic representation of text via cross-language correlation analysis,” in *Advances in Neural Information Processing Systems*, 2003, vol. 15, pp. 1473–1480. 13

[126] G. Wang, D. Hoiem, and D. Forsyth, “Building text features for object image classification,” in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2009, pp. 1367–1374. 12

[127] J. Z. Wang and J. Li, “Learning-based linguistic indexing of pictures with 2-D MHMMs,” in *Proc. ACM International Conference on Multimedia*, 2002, pp. 436–445. 10

[128] T. Westerveld, "Image retrieval: Content versus context," in *Content-Based Multimedia Information Access at Recherche d'Information Assistée par Ordinateur*, 2000, pp. 276–284. 12

[129] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010. 50

[130] P. C. Woodland, "Speaker adaptation for continuous density hmms: A review," in *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 2001. 67

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

[132] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM International Conference on Multimedia*, 2009, pp. 175–184. 12, 13

[133] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, 2008. 12, 13

[134] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang, "Weak attributes for large-scale image retrieval," in *Proc. IEEE International Conference on Computer Vision on Pattern Recognition*, 2012, pp. 2949–2956. 50

[135] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *European Conference on Computer Vision, Springer*, 2010, pp. 127–140. 50

[136] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. ACM International Conference on Multimedia*, 2007, pp. 273–276. 12, 13

[137] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical gaussianization for image classification," in *Proc. IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 1971–1977. 67

[138] F. Zhu and L. Shao, "Enhancing action recognition by cross-domain dictionary learning," in *British Machine Vision Conference*, 2013. 67

[139] X. Zhu and A. Goldberg, *Introduction to semi-supervised learning.* Morgan & Claypool, 2009. 67

[140] Y. Zhuang, Y. Yang, F. Wu, and Y. Pan, "Manifold learning based cross-media retrieval: a solution to media object complementary nature," *Journal of VLSI Signal Processing, Springer*, vol. 46, no. 2, pp. 153–164, 2007. 12, 13

[141] Y. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, 2008. 12, 13