

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Semantic Image Representation for Visual Recognition

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

in

Electrical Engineering (Signal and Image Processing)

by

Nikhil Rasiwasia

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Serge Belongie
Professor Kenneth Kreutz-Delgado
Professor David Kriegman
Professor Truong Nguyen

2011

Copyright
Nikhil Rasiwasia, 2011
All rights reserved.

The dissertation of Nikhil Rasiwasia is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2011

DEDICATION

To my nephew Shreyas,
my mother Geeta, my father Niranjan,
my sister Nidhi and my brother-in-law Prashant.

EPIGRAPH



©Bill Watterson

*We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.*

—T. S. Eliot

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	x
List of Tables	xvi
Acknowledgements	xvii
Vita and Publications	xx
Abstract of the Dissertation	xxii
Chapter 1 Introduction	1
1.1 Contributions of the thesis	5
1.1.1 Semantic Image Representation	6
1.1.2 Visual Recognition Systems	6
1.1.3 Holistic Context Modeling	10
1.2 Organization of the thesis	12
Chapter 2 Semantic Image Representation	13
2.1 Preliminaries	14
2.1.1 Notations	14
2.1.2 Image Retrieval Systems	15
2.1.3 Scene Classification Systems	16
2.1.4 Image Representation	18
2.2 Semantic Image Representation	20
2.2.1 The Semantic Multinomial	22
2.2.2 Robust estimation of SMNs	23
2.2.3 SMNs as Posterior Probability Vector	24
2.3 Computing the Semantic Multinomial	25
2.4 Related Work	28
2.5 Acknowledgments	29

Chapter 3	Image Retrieval: Query By Semantic Example	31
	3.1 Introduction	32
	3.2 Related Work	34
	3.3 Query by Semantic Example	36
	3.3.1 Query by Visual Example vs Semantic Retrieval	36
	3.3.2 Query by Semantic Example	38
	3.4 The Proposed Query by Semantic Example System	40
	3.4.1 Similarity Function	41
	3.5 Multiple Image Queries	42
	3.5.1 The Benefits of Query Fusion	43
	3.5.2 Query Combination	44
	3.6 Experimental Evaluation	46
	3.6.1 Evaluation Procedure	46
	3.6.2 Databases	46
	3.6.3 Model Tuning	48
	3.6.4 Performance Within the Semantic Space	51
	3.6.5 Multiple Image Queries	53
	3.6.6 Performance Outside the Semantic Space	54
	3.7 Acknowledgments	59
Chapter 4	Scene Classification with Semantic Representation	60
	4.1 Introduction	61
	4.2 Related Work	63
	4.3 Proposed Approach	64
	4.4 Experimental evaluation	66
	4.4.1 Datasets	66
	4.4.2 Experimental Protocol	66
	4.4.3 Results	67
	4.5 Acknowledgments	76
Chapter 5	Cross Modal Multimedia Retrieval	77
	5.1 Introduction	78
	5.2 Previous Work	80
	5.3 Fundamental Hypotheses	84
	5.3.1 The problem	84
	5.3.2 Multi-modal modeling	85
	5.3.3 The fundamental hypotheses	86
	5.4 Cross-modal Retrieval	88
	5.4.1 Correlation matching (CM)	89
	5.4.2 Semantic matching (SM)	92
	5.4.3 Semantic Correlation Matching (SCM)	95
	5.5 Experimental Setup	95
	5.5.1 Image and text representation	96

	5.6	Parameter selection	99
	5.7	Testing the fundamental hypotheses	104
	5.8	Acknowledgments	108
Chapter 6		Holistic Context Modeling	115
	6.1	Introduction	116
	6.2	Related Work on Context Modeling	118
	6.3	Semantics-based Models and Context Multinomials	120
	6.3.1	Limitations of Semantic Representations	120
	6.3.2	From Semantics to Context	122
	6.3.3	Contextual Concept Models	123
	6.3.4	Contextual Space	125
	6.3.5	Data Augmentation	126
	6.4	Experimental Setup	127
	6.4.1	Datasets	127
	6.4.2	Appearance Features	128
	6.5	Results	129
	6.5.1	Designing the Semantic Space.	129
	6.5.2	Number of Mixture Components	130
	6.5.3	Choice of Appearance Features	132
	6.5.4	Some Examples	134
	6.5.5	Complexity	136
	6.6	Comparison with Previous Work	136
	6.6.1	Scene Classification	139
	6.6.2	Image Retrieval Performance	141
	6.7	Acknowledgments	143
Chapter 7		The Importance of Supervision	145
	7.1	Introduction	146
	7.2	Topic Models	148
	7.2.1	LDA model	148
	7.2.2	Class LDA (cLDA)	149
	7.2.3	Supervised LDA (sLDA)	150
	7.2.4	Geometric Interpretation	151
	7.3	The Importance of Supervision	152
	7.4	Limitations of Existing models	154
	7.4.1	Theoretical Analysis	154
	7.4.2	Experimental Analysis	157
	7.5	Topic supervision	158
	7.5.1	Topics supervision in LDA model	158
	7.5.2	Models and geometric interpretation	159
	7.5.3	Learning and inference with topic-supervision	159
	7.5.4	Experimental analysis	161

7.6	Acknowledgments	162
Chapter 8	Conclusions	164
Appendix A	Datasets.	169
A.1	Datasets	170
A.1.1	Natural Scene Categories (N8, N13, N15)	170
A.1.2	UIUC Sports Dataset (S8)	172
A.1.3	Corel Image Collection (C371, C50, C43, C15)	172
A.1.4	Flickr Images (F18)	177
A.1.5	TVGraz	178
A.1.6	Wikipedia	178
Appendix B	Generalized Expectation maximization (GEM)	180
Appendix C	Computation of Image-SMNs	182
Appendix D	Variational Approximation	184
Appendix E	Parameter Estimation in cLDA	186
Appendix F	Parameter Estimation in topic-supervised LDA models	188
F.1	Learning Topic Conditional Distributions	188
F.2	Learning Class Conditional Distribution	189
Appendix G	Implementation Details of the various systems	190
G.1	Image Representation	190
G.1.1	SIFT Features	190
G.1.2	DCT Features	191
G.1.3	Bag-of-Features	191
G.1.4	Bag-of-Words	192
G.1.5	Semantic Multinomial	192
G.2	Concept/Category Models	193
G.2.1	Appearance Based Models	193
G.2.2	Holistic Context Models	193
G.3	Topic Supervised LDA	194
Bibliography	195

LIST OF FIGURES

Figure 1.1:	Probability of a locomotive image belonging to a number of visual concept classes according to appearance based visual classifiers. Note that, while most of the concepts of largest probability are present in the image, the SMN assigns significant probability to “bridge” and “arch”. This is due to the presence of a geometric structure similar to that of “bridge” and “arch”, shown on the image close-up.	4
Figure 1.2:	An illustration of image representation on the <i>semantic space</i> . An image is represented as a <i>semantic multinomial</i> which is a weight vector obtained using an array of appearance based classifiers.	6
Figure 1.3:	An illustration of “semantic gap” — two images which are similar for humans as they depict the semantic concept of “beach”. However they have different low-level visual properties of color, shape, etc.	7
Figure 2.1:	The generative model underlying image formation at the appearance level. w represents a sample from a vocabulary of scene categories or semantic concepts, and an image \mathcal{I} is composed of N patches, \mathbf{x}_n , sampled independently from $P_{\mathbf{X} W}(\mathbf{x} w)$. Note that, throughout this work, we adopt the standard plate notation of [14] to represent graphical models.	16
Figure 2.2:	Learning the scene category (semantic concept) density from the set \mathcal{D}_w of all training images annotated with the w^{th} caption in $\mathcal{W}(\mathcal{L})$, using hierarchical estimation [21]	17
Figure 2.3:	Image representation in semantic space \mathcal{S} , with a semantic multinomial (SMN) distribution. The SMN is a vector of posterior concept probabilities which encodes the co-occurrence of various concepts in the image, based on visual appearance.	21
Figure 2.4:	SMN for the image shown on the top left computed using (top-right) (2.8), (bottom-left) (2.21) and (bottom-right) (2.23).	25
Figure 2.5:	Alternative generative models for image formation at the appearance level. (a) A concept is sampled per appearance feature vector rather than per image, from $P_{\mathbf{X} W}(\mathbf{x} w)$. (b) Explicit modeling of the contextual variable Π from which a single SMN is drawn per image.	26
Figure 3.1:	An image containing various concepts: ‘train’, ‘smoke’, ‘road’, ‘sky’, ‘railroad’, ‘sign’, ‘trees’, ‘mountain’, ‘shadows’, with variable degrees of presence.	37

Figure 3.2:	Semantic image retrieval. Top: Under QBSE the user provides a query image, probabilities are computed for all concepts, and the image represented by the concept probability distribution. Bottom: Under the traditional SR paradigm, the user specifies a short natural language description, and only a small number of concepts are assigned a non-zero posterior probability.	39
Figure 3.3:	SMN of the <i>train</i> query of 3.6 as a function of the ratio $\frac{L(\alpha-1)}{n}$ adopted for its regularization.	50
Figure 3.4:	Average precision-recall of single-query QBSE and QBVE, Left: Inside the semantic space (<i>Corel371</i>), Right: Outside the semantic space (<i>Flickr18</i>).	51
Figure 3.5:	MAP scores of QBSE and QBVE across the 50 classes of <i>Corel371</i> .	51
Figure 3.6:	Some examples where QBSE performs better than QBVE. The second row of every query shows the images retrieved by QBSE.	52
Figure 3.7:	MAP as a function of query cardinality for multiple image queries. Comparison of QBSE, with various combination strategies, and QBVE. Left: Inside the semantic space (<i>Corel371</i>), Right: Outside the semantic space (<i>Flickr18</i>).	54
Figure 3.8:	Effect of multiple image queries on the MAP score of various classes from <i>Corel371</i> . Left: Classes with highest MAP gains, Right: Classes with lowest MAP gains	55
Figure 3.9:	Best precision-recall curves achieved with QBSE and QBVE on <i>Corel371</i> . Left: Inside the semantic space (<i>Corel371</i>), also shown is the performance with meaningless semantic space. Right: Outside the semantic space (<i>Flickr18</i>).	55
Figure 3.10:	Examples of multiple-image QBSE queries. Two queries (for “Township” and “Helicopter”) are shown, each combining two examples. In each case, two top rows presents the single-image QBSE results, while the third presents the combined query. . .	56
Figure 3.11:	SMN of individual and combined queries from class ‘Township’ of 3.10. Left column shows the first query SMN, center the second and, right the combined query SMN.	57
Figure 3.12:	Performance of QBSE compared to QBVE, based on precision-scope curve for $N = 1$ to 100, Left: Inside the semantic space (<i>Corel371</i>), Right: Outside the semantic space (<i>Flickr18</i>). . . .	58
Figure 4.1:	The proposed scene classification architecture.	65
Figure 4.2:	Theme vectors from each of the scenes of fifteen scene categories.	68
Figure 4.2:	Theme vectors from each of the scenes of fifteen scene categories. (continued)	69
Figure 4.2:	Theme vectors from each of the scenes of fifteen scene categories. (continued)	70

Figure 4.3:	Confusion Table for our method using 100 training image and rest as test examples from each category of Natural15. The average performance is $72.2\% \pm 0.2$	71
Figure 4.4:	Some images from worst performing scene categories in Natural15. (\rightarrow) implies the scene category the image is classified into.	73
Figure 4.5:	Some images from the Corel50 dataset. (\rightarrow) implies the scene category the image is classified into. (a) and (b) show two examples of correctly classified images, (c) and (d) two reasonably misclassified images and (e) and (f) shows two examples of error.	74
Figure 4.6:	The theme vector for the image in Figure 4.5(a).	74
Figure 4.7:	Classification performance as a function of the semantic space dimensions. Also shown, is the growth of the variance of the semantic themes, scaled appropriately.	75
Figure 5.1:	Two examples of image-text pairs: (a) section from the Wikipedia article on the Birmingham campaign (“History” category), (b) part of a Cognitive Science class syllabus from the TVGraz dataset (“Brain” category).	83
Figure 5.2:	Each document (D_i) consists of an <i>image</i> (I_i) and accompanying <i>text</i> (T_i), <i>i.e.</i> , $D_i = (I_i, T_i)$, which are represented as vectors in feature spaces \mathfrak{R}^I and \mathfrak{R}^T , respectively. Documents establish a one-to-one mapping between points in \mathfrak{R}^I and \mathfrak{R}^T	85
Figure 5.3:	Correlation matching (CM) performs joint feature selection in the text and image spaces, projecting them onto two maximally correlated subspaces \mathcal{U}_T and \mathcal{U}_I	87
Figure 5.4:	Cross-modal retrieval using CM. Here, CM is used to find the images that best match a query text.	92
Figure 5.5:	Semantic matching (SM) maps text and images into a semantic space. For each modality, classifiers are used to obtain a semantic representation, <i>i.e.</i> , a weight vector over semantic concepts.	93
Figure 5.6:	Cross-modal retrieval using SM used to find the text that best matches a query image.	94
Figure 5.7:	MAP performance (cross-modal retrieval, validation set) of SCM using two image models: BOW (flat lines) and LDA, for (a) TVGraz and (b) Wikipedia.	101
Figure 5.8:	Cross-modal MAP for CM on TVGraz and Wikipedia (validation sets), as a function of (a) the number of image codewords, (b) the number of text LDA topics, and (c) the number of KCCA components (while keeping the other two parameters fixed at the values reported in 5.5).	104

Figure 5.9:	Confusion matrices on the test set, for both TVGraz (left) and Wikipedia (right). Rows refer to true categories, and columns to category predictions. The more confusion on Wikipedia motivates the lower retrieval performance.	105
Figure 5.10:	top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for TVGraz	107
Figure 5.11:	top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for Wikipedia	108
Figure 5.12:	Per-class MAP for the cross-modal retrieval tasks on TVGraz (left) and Wikipedia (right): text queries (top); image queries (middle); and average performance over both types of queries (bottom).	109
Figure 5.13:	Text query from Biology class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.	110
Figure 5.14:	Text query from 'Warfare' class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.	111
Figure 5.15:	Text query from 'Cactus' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.	112
Figure 5.16:	Text query from 'Butterfly' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.	113
Figure 5.17:	Image-to-text retrieval on TVGraz (first two columns) and Wikipedia (last two columns). Query images are shown on the top row. The four most relevant texts, represented by their ground truth images, are shown in the remaining columns.	114
Figure 6.1:	An image from the "street" class of the N15 dataset (See 6.4.1) along with its SMN. Also highlighted are the two notions of <i>co-occurrence</i> . <i>Ambiguity co-occurrences</i> on the right: image patches compatible with multiple unrelated classes. <i>Contextual co-occurrences</i> on the left: patches of multiple other classes related to "street".	121
Figure 6.2:	Learning the contextual model for the "street" concept, (6.1), on semantic space \mathcal{S} , from the set of all training images annotated with "street".	123

Figure 6.3:	3-component Dirichlet mixture learned for the concept “street”. Also shown, as “*”, are the SMNs associated with each image. The Dirichlet mixture assigns high probability to the concepts “street” and “store”.	125
Figure 6.4:	The Contextual multinomial (CMN) of an image as the vector of co-occurrence probabilities of contextually related concepts. .	126
Figure 6.5:	(a) Classification accuracy as a function of the number of mixture components of the contextual class distributions, for both DCT and SIFT. (b) Dependence of appearance and contextual classification on the accuracy of the appearance modeling for SIFT-GRID features, (c) for DCT features. The performance of contextual classification remains fairly stable across the range of appearance models.	131
Figure 6.6:	Four cluster centers for the class “street” (top) and “forest” (bottom). Note that each class comprises different co-occurrence patterns.	132
Figure 6.7:	top) Two images from the “street” class of N15, and bottom) an image each from the “Ireland” and “Mayan ruins” CD of the Corel collection. Also shown with the images are the SMN and CMN vectors (middle and right column respectively). Notice that the CMN vectors are noise-free and capture the “gist” of the image.	135
Figure 6.8:	Class confusion matrix for classification on the N15 dataset. The average accuracy is 77.20%	141
Figure 6.9:	Precision-recall curves achieved with SMN, CMN, visual matching and chance level image retrieval.	142
Figure 6.10:	Retrieval results for four image queries shown on the left-most column. The first, second, and third row of every query show the five top matches using image matching, SMN, and CMN-based retrieval, respectively.	144
Figure 7.1:	Graphical models for (a) LDA and ts-LDA. (b) cLDA and ts-cLDA. (c) sLDA and ts-sLDA. All models use the standard plate notation [19], with parameters shown in rounded squares. . . .	148
Figure 7.2:	Representation of cLDA and ts-cLDA on a three <i>word simplex</i> . Also shown are sample images from two classes: “o” from class-1 and “x” from class-2. a) cLDA model with two topics. The line segment depicts a one-dimensional <i>topic simplex</i> , whose vertices are topic-conditional word distributions. Each class defines a smooth distribution on the topic simplex, denoted by the contour lines. c) ts-cLDA model. Topic-conditional word distributions are learned with supervision which encapsulate the class attributes.	151

Figure 7.3:	left) Four groups of words with equal word histograms. right) Four groups of edge segments with the equal edge segment histograms. Note that each group can be derived from the others by a displacement of words or edge segments. (This figure is best viewed in color)	152
Figure 7.4:	Classification accuracy as function of the number of topics for sLDA and cLDA, using topics learned with and without class influence and codebooks of size 1024, on (a) N15, (b) N8 and (c) S8. Similar behavior was observed for codebooks of different sizes.	155
Figure 7.5:	Performance of ts-sLDA, ts-cLDA, sLDA, and cLDA as a function of codebook size on (a) N13, (b) N8 and (c) S8. For ts-sLDA and ts-cLDA the number of topics is equal to the number of classes. For sLDA and cLDA, results are presented for the number of topics of best performance.	160
Figure 7.6:	Some example images that were misclassified by cLDA, but correctly classified using ts-cLDA. The expected topic distributions for ts-cLDA and cLDA (using 13 topics) are shown in the middle and bottom rows respectively. For ts-cLDA, topic labels are same as the class labels and the high probability topics are indeed the ones which capture the semantic meaning of the image. For cLDA, the topic labels do not carry any clear semantic meaning.	161

LIST OF TABLES

Table 2.1: SMN Entropy.	26
Table 3.1: Retrieval and Query Database	47
Table 3.2: Effect of SMN regularization on the MAP score of QBSE	49
Table 3.3: Effect of the similarity function on the MAP score of QBSE	50
Table 3.4: MAP of QBVE and QBSE on all datasets considered.	58
Table 4.1: Classification Result for 15 scene categories.	75
Table 4.2: Classification Result for 13 scene category subset.	76
Table 5.1: Taxonomy of the proposed approaches to cross-modal retrieval.	88
Table 5.2: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for TVGraz. μ_p and μ_q are the sample averages for p and q , respectively.	97
Table 5.3: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for Wikipedia. μ_p and μ_q are the sample averages for p and q , respectively.	98
Table 5.4: MAP for CM hypothesis (validation sets).	102
Table 5.5: Best parameter settings for CM, SM and SCM, on both TVGraz and Wikipedia (validation sets).	103
Table 5.6: Cross-modal MAP on TVGraz and Wikipedia (test sets).	106
Table 6.1: Impact of inference model on classification accuracy.	129
Table 6.2: Impact of appearance space on classification accuracy.	133
Table 6.3: Classification Results on Natural Scene Categories.	137
Table 6.4: Classification Results on Natural Scene Categories.	138
Table 6.5: Classification Results on Natural Scene Categories.	138
Table 6.6: Classification Results on Corel Collection.	139
Table 7.1: Classification Results on Natural Scene Categories.	162
Table 7.2: Classification Results on Sports8 and Corel50.	163
Table A.1: Summary of the Natural Scene datasets.	171
Table A.2: Summary of the UIUC Sports dataset.	173
Table A.3: Summary of the C371 dataset.	173
Table A.4: Summary of the TVGraz dataset.	178
Table A.5: Summary of the Wikipedia dataset.	179

ACKNOWLEDGEMENTS

“It takes a village to raise a child” — an old African proverb. So it is with everything else in life. Although, I get the privilege of writing this thesis, it is certain that this thesis would not have been possible, had not the village helped raise it. In this small, yet important section, I take the opportunity to acknowledge many people, who helped shape this thesis. Foremost, my deepest gratitude to life itself, which provides eternal wonder and amazement, motivating me to always learn and know more.

Next, I would like to express my sincerest gratitude to my supervisor, Professor Nuno Vasconcelos. It won't be far from truth, if I said that I did not know the meaning of “research” and it's only through his mentoring and guidance that I slowly understood its true essence. From him I learned the benefits of striving for perfection, of hard work, of hanging on, of just doing it. His contributions to this thesis, definitely equal mine, if not exceeding it. I am also grateful for having an exceptional doctoral committee, and wish to thank Serge Belongie, Kenneth Kreutz-Delgado, David Kriegman, and Truong Nguyen for their valuable input, accessibility and informative courses.

I would like to thank all my colleagues at SVCL, Dr. Dashan Gao, Dr. Antoni Bert Chan, Dr. Hamed Masnadi-Shirazi, Dr. Sunhyoung Han, Vijay Mahadevan, Jose Maria Costa Pereira, Mandar Dixit, Mohammad Saberian, Kritika Muralidharan and Weixin Li for their support, friendship and assistance throughout the years and their collaboration on some of the experiments.

Although the six years, which went in the making of this thesis, were filled with its share of ups and downs, the most cherishable moments — the moments that would flash before my eyes — are the ones I shared with my friends. With the risk of offending the few, who were here with me all along, but to whom, my impaired memory prevents me from doing justice, I would like to thank Ankit Srivastava, Vijay Mahadevan, Himanshu Khatri, Nitin Gupta, Jose Maria Costa Pereira, Gaurav Dhiman, Mayank Kabra, Rathinakumar Appuswamy, Anshuman Gupta, Shikha Misra, Shibin Parameswaran, Adarsh Krishnamurthy, Sethuraman Sankaran, Karthik Sanji, Sravanthi K V, Arun Manohar, Bharath Kumar SV,

Aneesh Subramanian, Nikhil Karamchandani, Vikram Mavalankar, Kowsik Bodi and the many other from whom I again seek forgiveness. A special thanks to my rock climbing friends who showed me what true adventure meant. Also thanks to my friends who walked with me in the years gone by.

In the end, I reserve a special mention for my immediate and extended family. Since any amount of thanks would be incommensurate to their support, I just bow down and seek their blessings.

The text of Chapter 2, in part, is based on the material as it appears in: N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Bridging the Semantic Gap: Query by Semantic Example*’, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007, and N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Query by Semantic Example*’, ACM International Conference on Image and Video Retrieval, LNCS 51-60, Phoenix, 2006. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 3, in part, is based on the material as it appears in: N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Bridging the Semantic Gap: Query by Semantic Example*’, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007, N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Query by Semantic Example*’, ACM International Conference on Image and Video Retrieval, LNCS 51-60, Phoenix, 2006, and N. Rasiwasia and N. Vasconcelos, ‘*A Systematic Study of the role of Context on Image Classification*’, IEEE Conference on Image Processing, 1720-1723, San Diego, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 4, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, ‘*Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision*’, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-6, Anchorage, June 2008. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 5, in part, is based on the material as it appears in: N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy and N. Vasconcelos, ‘*A New Approach to Cross-Modal Multimedia Retrieval*’, ACM

Conference on Multimedia, 251-260, November 2010, and J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R.G. Lanckriet, R. Levy and N. Vasconcelos, ‘*On the role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval*’, submitted to Pattern Analysis and Machine Learning, Sept 2011. The dissertation author was a primary researcher and an author of the cited material. The author would like to thank J. Costa Pereira, E. Coviello and G. Doyle for their helpful comments and contributions to the project.

The text of Chapter 6, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, ‘*Holistic Context Models for Visual Recognition*’, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, N. Rasiwasia and N. Vasconcelos, ‘*Holistic Context Modeling using Semantic Co-occurrences*’, IEEE Conference on Computer Vision and Pattern Recognition, Miami, June 2009, and N. Rasiwasia and N. Vasconcelos, ‘*Image Retrieval using Query by Contextual Example*’, ACM Conference on Multimedia Information Retrieval, pp. 164-171, Vancouver, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter 7, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, ‘*Holistic Context Models for Visual Recognition*’, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, and N. Rasiwasia and N. Vasconcelos, ‘*Generative Models for Image Classification*’, In preparation for IEEE Transactions on Pattern Analysis and Machine Intelligence. The dissertation author was a primary researcher and an author of the cited material.

VITA AND PUBLICATIONS

- 2001-2005 Bachelor of Technology,
Electrical Engineering,
Indian Institute of Technology, Kanpur, India
- 2005–2007 Master of Science
Electrical Engineering (Signal and Image Processing), Uni-
versity of California at San Diego
- 2005–2011 Research Assistant
Statistical and Visual Computing Laboratory
Department of Electrical and Computer Engineering
University of California at San Diego
- 2005-2011 Doctor of Philosophy
Electrical Engineering (Signal and Image Processing),
University of California at San Diego

Journals

N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Bridging the Semantic Gap: Query by Semantic Example*’, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007.

N. Rasiwasia and N. Vasconcelos, ‘*Holistic Context Models for Visual Recognition*’, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence.

J. Costa Pereira, E. Coviello, G. Doyle, **N. Rasiwasia**, G.R.G. Lanckriet, R. Levy and N. Vasconcelos, ‘*On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval*’, Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence.

N. Rasiwasia and N. Vasconcelos, ‘*Generative Models for Image Classification*’, In preparation for IEEE Transactions on Pattern Analysis and Machine Intelligence.

Conferences

A. Kannan, P. Talukdar, **N. Rasiwasia**, and Q. Ke, ‘*Improving Product Classification Using Images*’, To appear in IEEE International Conference on Data Mining, Vancouver, 2011

- R. Kwitt, **N. Rasiwasia** and N. Vasconcelos, '*Learning Pit Pattern Concepts for Gastroenterological Training*', To appear in International Conference on Medical Image Computing and Computer Assisted Intervention, Toronto, September 2011 [**Oral**]
- M. Dixit, **N. Rasiwasia** and N. Vasconcelos, '*Adapted Gaussian Models for Image Classification*', IEEE Conference on Computer Vision and Pattern Recognition, pp. 937-943, Colorado Springs, June 2011
- N. Rasiwasia**, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy and N. Vasconcelos, '*A New Approach to Cross-Modal Multimedia Retrieval*', ACM Conference on Multimedia, 251-260, November 2010 [**Oral**] [**Best student paper**]
- N. Rasiwasia** and N. Vasconcelos, '*Holistic Context Modeling using Semantic Co-occurrences*', IEEE Conference on Computer Vision and Pattern Recognition, Miami, June 2009
- N. Rasiwasia** and N. Vasconcelos, '*Image Retrieval using Query by Contextual Example*', ACM Conference on Multimedia Information Retrieval, pp. 164-171, Vancouver, Oct 2008
- N. Rasiwasia** and N. Vasconcelos, '*A Systematic Study of the role of Context on Image Classification*', IEEE Conference on Image Processing, 1720-1723, San Diego, Oct 2008 [**Oral**]
- N. Rasiwasia** and N. Vasconcelos, '*Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision*', IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-6, Anchorage, June 2008
- N. Rasiwasia** and N. Vasconcelos, '*A study of Query by Semantic Example*', 3rd International Workshop on Semantic Learning and Applications in Multimedia, pp. 1-8, Anchorage, June 2008 [**Oral**]
- N. Rasiwasia**, P. J. Moreno and N. Vasconcelos, '*Query by Semantic Example*', ACM International Conference on Image and Video Retrieval, LNCSPhoenix, 2006 [**Oral**]

ABSTRACT OF THE DISSERTATION

Semantic Image Representation for Visual Recognition

by

Nikhil Rasiwasia

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2011

Professor Nuno Vasconcelos, Chair

A novel image representation, termed semantic image representation, that incorporates contextual information is proposed. In this framework, images are represented by their posterior probabilities with respect to a set of appearance based concept models, built upon the bag-of-features representation. Thus while appearance features are intensity, texture, edge orientations, frequency bases, etc. those of the semantic representation are concept probabilities. Semantic image representation induces a mapping from the space of appearance features to a *semantic space*, where each axis represents a semantic concept. Each concept probability is referred to as a *semantic feature* and the semantic feature vector as the *semantic multinomial* (SMN) distribution. Next, we present design of three different

visual recognition tasks viz. image retrieval, scene classification and cross-modal multimedia retrieval, based on the semantic image representation. First, a novel framework for content based image retrieval, referred to as *query by semantic example* (QBSE) is proposed, which extends the query-by-example paradigm to the semantic space. Current content based image retrieval solutions rely on strict visual similarity, which in most cases, is weakly correlated with the measures of similarity adopted by humans for image comparison. By using the semantic image representation, the retrieval operation is performed at a much higher level of abstraction, which results in retrieval systems that are more accurate than previously possible. QBSE also allows a direct comparison of visual and semantic representations under a common query paradigm, which enables an explicit test of the value of semantic representations for image retrieval. Second, we propose a framework for scene classification based on the semantic image representation. As in previous approaches, we introduce a low dimensional intermediate space, which in the proposed framework is served by the semantic space. However, instead of learning the intermediate “themes” in an unsupervised manner, they are learned with weak supervision, from casual image annotations. When annotations are not available, they are replaced by the scene category labels. A study of the effect of dimensionality on the classification performance is also presented, indicating that the dimensionality of the “theme” space grows sub-linearly with the number of scene categories. Third, the problem of cross-modal retrieval from multimedia repositories is considered. This problem addresses the design of retrieval systems that support queries *across* content modalities, *e.g.*, using text to search for images. A mathematical formulation is proposed, where the design of cross-modal retrieval systems is equated to that of designing isomorphic feature spaces for different content modalities. Three new solutions to the cross-modal retrieval problem are proposed: correlation matching (CM), which models cross-modal correlations between different modalities, semantic matching (SM), which relies on the semantic representation, where different modalities are represented on a common semantic space, and semantic correlation matching (SCM), which combines both. An implementation of the above systems under the minimum probability

of error framework is presented and compared to various existing algorithms in respective visual recognition tasks, on benchmark datasets. It is shown that the proposed semantic image representation is able to achieve superior results. Finally, we discuss the issue of *contextual noise* in semantic representations, due to the inherent ambiguity of the bag-of-features representation. To address this, we propose a novel two-layer framework to context modeling, based on the probability of co-occurrence of objects and scenes. The first layer represents the image in a semantic space, and the second layer introduces distributions of each concept in the semantic space. This facilitates robust inference in the presence of contextual noise. A thorough and systematic experimental evaluation of the proposed context modeling is presented. It is shown that it captures the contextual “gist” of natural images. The effectiveness of the proposed approach to context modeling is further demonstrated through a comparison to existing approaches on scene classification and image retrieval, on benchmark datasets. In all cases, the proposed approach achieves state of the art visual recognition performance.

Chapter 1

Introduction

Humans have an intriguing ability to process visual information amazingly fast and with nearly perfect recognition rates. However, with the proliferation of the Internet, availability of cheap digital cameras, and the ubiquity of cell-phone cameras, the amount of accessible visual information has increased astronomically. Websites such as Flickr alone boast of over 5 billion images, not counting the hundreds of other such websites and countless other images that are not published online. With such enormous collections of available visual content, manual processing becomes prohibitive and it is therefore of great practical importance to build “visual recognition systems”.

Visual recognition is a fundamental problem in computer vision. It subsumes the problems of scene classification [74, 77, 17, 114, 120], image annotation [21, 41, 72, 35, 12], image retrieval [28, 133, 119, 156], object recognition/localization [140, 128, 47], object detection [165, 122, 42] etc. In recent years the application of machine learning technologies — that allow computers to make intelligent decisions based on empirical data — for tackling visual recognition is becoming increasingly popular with advancements being made by both the research communities. While the last decade has produced significant progress towards the solution of the visual recognition problems, the basic strategy has remained the same: 1) identify a number of visual classes of interest, 2) design a set of “appearance” features that are discriminative for those classes, 3) postulate an architecture for their recognition, and 4) rely on sophisticated statistical tools to learn optimal recognizers from training data. We refer to this strategy as *appearance-based* visual recognition, because the associated recognizers rely on image representations which are either image pixels, features, or parts, derived by simple deterministic mappings of those pixels. The main innovations of the last decade have been associated with better appearance-based features e.g. the ubiquitous scale-invariant feature transform (SIFT) descriptor [85], the widespread adoption of statistical modeling e.g. generative graphical models (such as Gaussian Mixture Models (GMM) [157, 21], Latent Dirichlet Allocation [12, 77], etc.), sophisticated families of discriminants (such as support vector machines (SVMs) with various kernels tuned for vision [51, 23, 17, 177, 20] etc.), the application of

powerful machine learning techniques (such as variational learning [14], Markov chain Monte Carlo [50]) etc.) to the design of the recognizers themselves etc.

While there is no question that appearance based classifiers will retain a predominant role in the future of recognition, it is not as clear that they will be *sufficient* to solve the recognition problem. In fact, there is little evidence so far that they can solve all but a small class of problems (such as face detection) with accuracies comparable to those of biological vision. One striking property of the latter, at least in what concerns humans, is that it rarely seems to ground decisions exclusively on low-level visual features. This has been well documented in psychophysics, through unambiguous evidence that scene interpretation depends on *context* [8, 100]. By this, it is usually meant that the detection of an object of interest (e.g. a locomotive) is facilitated by the presence, in the scene, of other objects (e.g. railroad tracks or trains) which may not themselves be of interest. The presence of these *contextual cues* (e.g. that locomotives are usually on tracks and pull trains) increases the detection rate for the object of interest. This is illustrated in Figure 1.1, which shows the posterior probabilities of a locomotive image belonging to a number of visual concept classes, according to a number of appearance based visual detectors trained on those classes. The presence of an ‘arch-like structure in the locomotive’s rooftop makes the weight of the “bridge” concept slightly higher than that of the “locomotive” concept, for the adopted appearance based recognizer. However, by noting that the contextual cues “railroad”, and “train” also have high posterior probability, a context-sensitive recognizer could still assign the image to the “locomotive” class.

Another striking property of human vision, which suggests that raw appearance is not the whole story for recognition, is unveiled by a set of relatively recent findings on the neural structure of the recognition process. In a series of now extensively replicated seminal experiments, Thorpe and collaborators [144] demonstrated an intriguing ability of humans to perform *decent* scene classification with *very small computation*. More precisely, EEG recordings have shown that humans are capable of solving visual recognition problems such as the detection of “food”, “animals”, and so forth, with 90 – 95% accuracy in close to 150 *ms*, i.e. only

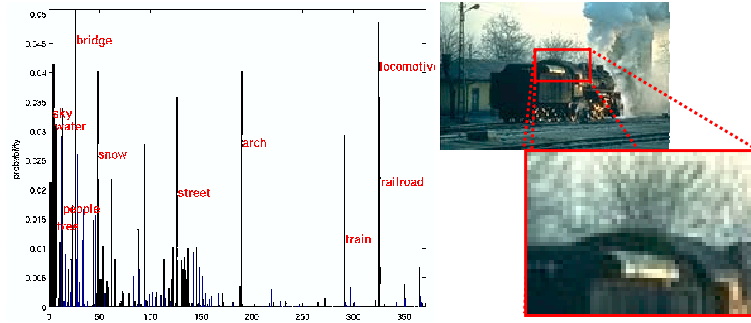


Figure 1.1: Probability of a locomotive image belonging to a number of visual concept classes according to appearance based visual classifiers. Note that, while most of the concepts of largest probability are present in the image, the SMN assigns significant probability to “bridge” and “arch”. This is due to the presence of a geometric structure similar to that of “bridge” and “arch”, shown on the image close-up.

enough time to propagate the visual stimulus (in a feed-forward manner) through a small number of neural layers. Given that 95% is nowhere near the recognition rates that the visual system can achieve for these classes, this raises the question of what these low-grade, but fast, classifiers could be useful for. While we do not profess to know the answer to this question, one possibility is that they could be *contextual classifiers*, whose goal is not to solve the vision problem per se, but detect the contextual cues that could make the solution easier.

In this thesis, image representation and visual recognition form the core body of work, where we address the problem of incorporating contextual cues in the image representation to tackle visual recognition problems. More precisely, the aims of this thesis are twofold. First, the design of a representation that accounts for the contextual cues present in an image. Second, the design of visual recognition systems that build upon the proposed image representation and achieve state of the art visual recognition performance.

1.1 Contributions of the thesis

This thesis provides a novel framework for visual recognition which is based on incorporation of contextual cues. To this end, first a *semantic image representation* is introduced which builds upon recent developments in visual recognition, namely the availability of robust appearance classifiers and image databases annotated with respect to a sizable concept vocabulary. This representation besides being well correlated with the human understanding of the images, is very useful in the design of visual recognition systems that yield state of the art recognition performances. Next, building upon the semantic image representation, we present three frameworks for three different visual recognition tasks, viz. image retrieval, scene classification and cross-modal multimedia retrieval. Under image retrieval, the task is to *retrieve* images from a given image repository in response to a *query* provided by the user. Under scene classification, the task is to assign one of several class labels from a given vocabulary of concepts to a user specified image. Both image retrieval and scene classification are well studied problems in computer vision [133, 28, 119, 77, 74, 105, 120]. Cross-modal multimedia retrieval on the other hand is a relatively recent problem in computer vision, where the retrieval operation is performed across different data modalities e.g. to retrieve text documents in response to an image query.

Next, we show that the semantic image representation, although effective at solving visual recognition problems, suffers from certain drawbacks, in particular the issue of *contextual noise*. In the latter part of this thesis, we introduce the framework of *holistic context modeling*, that addresses these drawbacks. Holistic context models also build upon the semantic image representation and are able to explicitly learn *true* contextual relationship between different concepts directly from the data. Holistic context models are shown to further improve the performance of visual recognition systems. Finally, a formal analysis of the holistic context models in the form of a *generative graphical model* is presented and connections to the existing work in the literature are drawn. In the remainder of this section we briefly discuss the significant contributions of this thesis.

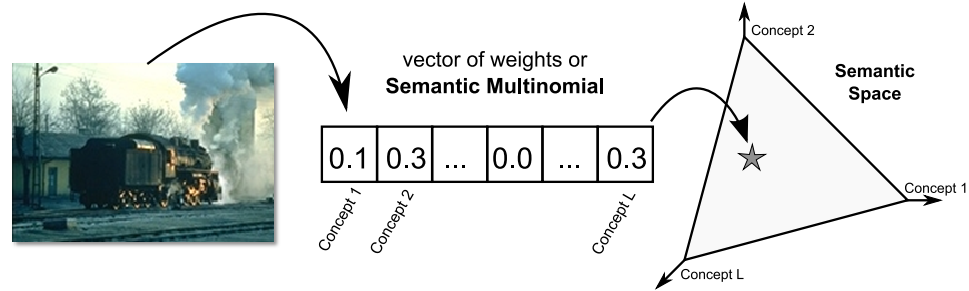


Figure 1.2: An illustration of image representation on the *semantic space*. An image is represented as a *semantic multinomial* which is a weight vector obtained using an array of appearance based classifiers.

1.1.1 Semantic Image Representation

Semantic image representation is a novel image representation, that brings a paradigm shift in the way image are represented. Under semantic image representation, instead representing the images on the space of low level appearance features derived from the image, a *semantic space* — a space where each dimension represents a meaningful visual concept — is introduced, upon which the images are represented and all recognition decisions are performed. To obtain the semantic representation of an image, first a vocabulary of visual concepts is defined and statistical models are learned for all concepts in the vocabulary with existing appearance modeling techniques [21, 74, 77]. Next, the outputs of these appearance classifiers are then interpreted as the dimensions of the semantic space. This is illustrated in 1.2, where an image is represented by the vector of its posterior probabilities under each of the appearance models. This vector is denoted as a *semantic multinomial* (SMN) distribution as the image features themselves define a multinomial distribution over the semantic concepts. An example SMN for a natural image is the probability vector shown in 1.1(left).

1.1.2 Visual Recognition Systems

A significant contribution of this thesis is the design of visual recognition systems based on the proposed semantic image representation. Below we discuss three different visual recognition systems that build upon the semantic image rep-



Figure 1.3: An illustration of “semantic gap” — two images which are similar for humans as they depict the semantic concept of “beach”. However they have different low-level visual properties of color, shape, etc.

resentation.

Image Retrieval: Query by Semantic Example

Current image search engines tend to rely on information extracted from image filenames or neighboring text in the webpage to retrieve the images that best satisfy a given query. This approach is fruitful only if a meticulous and complete textual description of the image is available, but this is rarely the case. It ignores the wealth of information available in the visual information stream itself, i.e. the image content. The image retrieval community studies content-based solutions to the design of retrieval systems. One popular retrieval paradigm is that of *query-by-example* — the user provides a query image, and retrieval consists of finding the closest visual match in an image collection, to this query. However, this paradigm restricts the definition of similarity to a *strict visual form*, declaring images as similar as long as they exhibit identical patterns of color, texture, shape, etc. In most cases, this narrow definition of similarity is weakly correlated with those adopted by humans for image comparison. For example, 1.3 shows two images which are similar for humans as they depict the semantic concept of “beach”, however they have different low-level visual properties of color, shape, etc. This is commonly known as the “semantic gap” between *low-level processing* and the *higher level semantic abstraction* adopted by humans [133, 119].

In this thesis we propose a novel image retrieval framework, Query-by-

semantic-example (QBSE), that addresses the semantic gap. QBSE leverages on the semantic image representation by extending the query-by-example paradigm into the semantic domain, whereby the nearest neighbor retrieval operation is performed directly on the semantic space. This is shown to have two main properties of interest, one mostly practical and the other philosophical. From a practical standpoint, because QBSE has a higher level of abstraction, it enables retrieval systems with higher *generalization ability* that are more accurate than what was previously possible. Philosophically, because it allows a direct comparison of visual and semantic representations under a common query paradigm, QBSE enables the design of experiments that explicitly test the value of semantic representations for image retrieval.

Scene Classification

Scene classification is an important problem for computer vision, and has received considerable attention in the recent past. *Scene* classification differs from *object* classification, in that a scene is composed of several entities often organized in an unpredictable layout[113]. For a given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations. Early efforts at scene classification targeted binary problems, such as distinguishing indoor from outdoor scenes [142], city views from landscape etc. More recently, there has been an effort to solve the problem in greater generality, through design of techniques capable of classifying a relatively large number of scene categories [166, 77, 113, 74, 16, 83], and a dataset of 15 categories has been used to compare the performance of various systems[74, 83]. Several of these approaches aim to provide a compact lower dimensional representation using some intermediate characterization on a latent space, commonly known as the intermediate “theme” or “topic” representation [77]. The rationale for this strategy is that images which share frequently co-occurring visual features have similar representation in the latent space, even if they have no features in common.

In this thesis we propose an alternative solution using the semantic image representation, where the semantic space serves as the intermediary for the low di-

mensional “theme” representation. However instead of the themes being learned in an unsupervised manner, as is the case with existing approaches, they are explicitly defined. The number of semantic classes or themes used, defines the dimensionality of the intermediate semantic space. Experiments show that scene classification based on semantic image representation outperforms the unsupervised latent-space approaches, and achieves performance close to the state of the art, using a much lower dimensional image representation.

Cross Modal Multimedia Retrieval

Classical approaches to information retrieval are of a *uni-modal* nature [125, 133, 84]. Text repositories are searched with text queries, image databases with image queries, and so forth. This paradigm is of limited use in the modern information landscape, where multimedia content is ubiquitous. Recently, there has been a surge of interest in *multi-modal* modeling, representation, and retrieval [106, 148, 132, 138, 28, 60, 31]. Multi-modal retrieval relies on queries combining multiple content modalities (*e.g.* the images and sound of a music video-clip) to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario which limits the applicability of the resulting multimedia models and retrieval systems. For example, these systems are inadequate when the task is to query with objects that do not share the same modality as the retrieval set *e.g.* using images to find similar documents in a text corpus.

In this thesis, a richer interaction paradigm is considered, which is denoted *cross-modal* retrieval. The goal is to build multi-modal content models that enable interactivity with content *across* modalities. Such models can then be used to design *cross-modal retrieval systems*, where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.*, the best accompanying audio-track). The central problem in the design of cross-modal retrieval systems is the inherent inconsistency between the representations of different modalities. To address this, a mathematical formulation is proposed, equating the design of cross-

modal retrieval systems to that of designing isomorphic feature spaces for different content modalities. Semantic image representation naturally lends itself as an effective solution to the design of the isomorphic feature spaces. By extending the semantic image representation to other modalities, all modalities are represented at a higher level of abstraction which establishes a common semantic language between them. This is referred to as the *abstraction hypothesis*. Another solution, based on maximizing correlations between different modalities, denoted as *correlation hypothesis*, is also proposed. By means of extensive experimental evaluation it is concluded that both hypotheses enable design of effective cross-modal retrieval systems and are complementary to each other, although the evidence in favor of the abstraction hypothesis is stronger than that for correlation.

1.1.3 Holistic Context Modeling

While the semantic image representation captures co-occurrences of the semantic concepts present in an image, not all these correspond to *true* contextual relationships. This is usually not due to poor statistical estimation, but due to the inherent *ambiguity* of the underlying features representation. Since appearance based features typically have small spatial support, it is frequently difficult to assign them to a single visual concept e.g. just looking at the close up of the “arch like feature” in 1.1 its is not possible to assert that this feature is from a “locomotive” image and not a “bridge”. Hence, the semantic image representation extracted from an image usually assigns some probability to concepts unrelated to it e.g. “arch” and “bridge” concepts for the “locomotive” image in 1.1. We term this ambiguity as *contextual noise* i.e. casual coincidences due to the ambiguity of the underlying appearance representation (image patches that could belong to either a “locomotive” or an “arch”).

Rather than attempting to eliminate contextual noise by further processing of appearance features, we propose a procedure for *robust* inference of contextual relationships *in the presence of contextual noise*. This is achieved by introducing a second level of representation, that operates on the semantic space. Each visual concept is modeled by the distribution of the posterior probabilities extracted from

all its training images. This *distribution of distributions* is referred as the *contextual model* for the concept. For large enough and diverse enough training sets, these models are dominated by the probabilities of true contextual relationships which can be found by identifying peaks of probability in semantic space. An implementation of contextual modeling is proposed, where concepts are modeled as mixtures of Gaussian distribution on appearance space, and mixtures of Dirichlet distributions on semantic space. It is shown that the contextual descriptions observed in contextual space are substantially less noisy than those characteristic of semantic space, and frequently *remarkably clean*. It is also argued that these probabilities capture the *contextual co-occurrences* of concepts and constitute the global context representation of an image. The effectiveness of the proposed approach to context modeling is further demonstrated through a comparison to existing approaches on scene classification and image retrieval, on benchmark datasets. In all cases, the proposed approach is superior to various contextual modeling procedures in the literature.

We also present a comparison of holistic context models with the existing work on “topic models”, in particular latent Dirichlet allocation (LDA). It is shown that although both these models share a common generative modeling framework, one key property of the holistic context models, that enables it to achieve higher classification accuracy, is that of *supervision*. However, since the holistic context models and LDA use different image representations, its difficult to assess the true gains achieved by supervision in these model. To enable a systematic study of the benefits of supervision, we present a family of topic models, denoted as *topic-supervised* LDA, where supervision is introduced in the LDA framework. All other attributes of the LDA model are kept constant. It is shown that topic-supervised LDA models are able to outperform their unsupervised counterparts, for the task of scene classification.

1.2 Organization of the thesis

The organization of the thesis is as follows. In Chapter 2 we first review two problems of visual recognition viz. retrieval and scene classification, and two popular low-level appearance based image representations viz. discrete cosine transform and scale invariant feature transform. Next we introduce the proposed semantic image representation. In Chapter 3 we highlight the problems of existing image retrieval solutions based on appearance features and introduce query-by-semantic-example retrieval paradigm. Next, in Chapter 4 we present a scene classification system using the semantic features as an intermediate representation. The problem of cross-modal multimedia retrieval is presented in Chapter 5, where we also discuss two novel solutions for retrieval across different content modalities; first based on the semantic image representation and second based on maximizing correlation between different modalities. The issues of contextual noise are discussed in Chapter 6, where we propose holistic context modeling that addresses it. In Chapter 7 we compare the holistic context model to existing the “topic model”. Conclusions are provided in Chapter 8. Finally, a brief discussion on the implementation details of various recognition systems proposed in this work is provided in Appendix G.

Chapter 2

Semantic Image Representation

In this chapter we first present a review of the existing solutions to the problem of image retrieval and scene classification followed by a brief review of low level image representation. We then introduce the semantic image representation for scene classification.

2.1 Preliminaries

We start by briefly reviewing appearance-based modeling and the design of visual recognition systems for image retrieval and scene classification.

2.1.1 Notations

Consider a image database $\mathcal{D} = \{\mathcal{I}_1, \dots, \mathcal{I}_D\}$ where images \mathcal{I}_i are observations from a random variable \mathbf{X} , defined on some feature space \mathcal{X} . For example, \mathcal{X} could be the space of discrete cosine transform (DCT), or SIFT descriptors. Each image is represented as a set of N *low-level feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathcal{X}$, assumed to be sampled independently. This is commonly referred to as the “bag-of-features” (BoF) representation, since the image is represented as an orderless collection of visual features. A popular extension of the BoF representation is the “bag-of-words” (BoW) [27, 74] representation. In BoW representation, the feature space \mathcal{X} is further quantized into $|\mathcal{V}|$ unique bins, defined by a collection of centroids, $\mathcal{V} = \{1, \dots, |\mathcal{V}|\}$, and each feature vector $\mathbf{x}_n, n \in \{1, \dots, N\}$ is mapped to its closest centroid. Each image is then represented as a collection of *visual words*, $\mathcal{I} = \{v_1, \dots, v_N\}, v_n \in \mathcal{V}$, where v_n is the bin that contains the feature vector \mathbf{x}_n . This facilitates the representation of the image as a vector in $\mathbb{R}^{|\mathcal{V}|}$, however it has been argued that feature quantization leads to significant degradation in its discriminative power [15]. In this work, we rely on both BoF and BoW representation, BoF being the default choice of image representation.

2.1.2 Image Retrieval Systems

The starting point for any retrieval system is the image database $\mathcal{D} = \{\mathcal{I}_1, \dots, \mathcal{I}_D\}$. Although several image retrieval formulations are possible, in this work, the framework underlying all query paradigms is that of minimum probability of error retrieval, as introduced in [156]. Under this formulation, each image is considered as an observation from a different class, determined by a random variable Y defined on $\{1, \dots, D\}$. Given a query image \mathcal{I} , the MPE retrieval decision is to assign it to the class of largest posterior probability, i.e.

$$y^* = \arg \max_y P_{Y|\mathbf{X}}(y|\mathcal{I}). \quad (2.1)$$

and image retrieval is based on the mapping $g : \mathcal{X} \rightarrow \{1, \dots, D\}$ of (2.1). Using Bayes rule and under the assumption of independent samples this is equivalent to,

$$y^* = \arg \max_y P_{\mathbf{X}|Y}(\mathcal{I}|y)P_Y(y). \quad (2.2)$$

$$= \arg \max_y \prod_j P_{\mathbf{X}|Y}(\mathbf{x}_j|y)P_Y(y). \quad (2.3)$$

where $P_{\mathbf{X}|Y}(x|y)$ is the class conditional density, which serves as the *appearance model* for the y^{th} image and $P_Y(y)$ the class prior. Although any prior class distribution $P_Y(y)$ can be supported, we assume a uniform distribution in what follows.

To model the appearance distribution, we rely on Gaussian mixture models (GMM). These are popular models for the distribution of visual features [21, 57, 145, 12] and have the form

$$P_{\mathbf{X}|Y}(\mathbf{x}|y; \Gamma_y) = \sum_j \alpha_y^j \mathcal{G}(\mathbf{x}, \mu_y^j, \Sigma_y^j) \quad (2.4)$$

where, α_y is a probability mass function such that $\sum_j \alpha_y^j = 1$, $\mathcal{G}(\mathbf{x}, \mu, \Sigma)$ a Gaussian density of mean μ and covariance Σ , and j an index over the mixture components. Some density estimation [33] procedure can be used to estimate the parameters of this distribution. In this work we use the well known expectation-maximization (EM) algorithm [30]. Henceforth, we refer to the above retrieval paradigm as *query-by-visual-example* (QVBE).

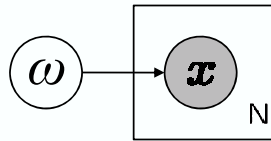


Figure 2.1: The generative model underlying image formation at the appearance level. w represents a sample from a vocabulary of scene categories or semantic concepts, and an image \mathcal{I} is composed of N patches, \mathbf{x}_n , sampled independently from $P_{\mathbf{x}|W}(\mathbf{x}|w)$. Note that, throughout this work, we adopt the standard plate notation of [14] to represent graphical models.

2.1.3 Scene Classification Systems

A scene classification system appends the database \mathcal{D} with a vocabulary of scene category $\mathcal{W} = \{1, \dots, K\}$ and each image with a scene label \mathbf{w}_i , making $\mathcal{D}^{\mathcal{W}} = \{(\mathcal{I}_1, \mathbf{w}_1), \dots, (\mathcal{I}_D, \mathbf{w}_D)\}$. The scene label \mathbf{w}_i is considered to be an observation from a scene category random variable W defined on \mathcal{W} . Note that, for scene classification systems, the label \mathbf{w}_i is an indicator vector such that $\mathbf{w}_{i,j} = 1$ if the i^{th} image is an observation from the j^{th} scene category. Each scene category induces a probability density $\{P_{\mathbf{x}|W}(\mathbf{x}|w)\}_{w=1}^K$ on \mathcal{X} , from which feature vectors are drawn. This is denoted as the *appearance model* for the category w which describes how observations are drawn from the low-level visual feature space \mathcal{X} . As shown in Figure 2.1, the generative model for a feature vector \mathbf{x} thus consists of two steps: first a category label w is selected, with probability $P_W(w) = \pi_w$, and the feature vector then drawn from $P_{\mathbf{x}|W}(\mathbf{x}_n|w)$. Both concepts and feature vectors are drawn independently, with replacement.

Given a new image \mathcal{I} , classification is performed using the minimum probability of error framework, where the optimal decision rule is to assign it to the category of largest posterior probability

$$w^* = \arg \max_w P_{W|\mathbf{x}}(w|\mathcal{I}). \quad (2.5)$$

where $P_{W|\mathbf{x}}(w|\mathcal{I})$ is posterior probability of category w given \mathcal{I} and can be com-

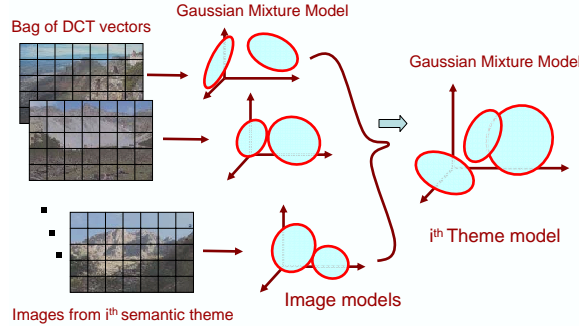


Figure 2.2: Learning the scene category (semantic concept) density from the set \mathcal{D}_w of all training images annotated with the w^{th} caption in $\mathcal{W}(\mathcal{L})$, using hierarchical estimation [21]

puted used Bayes rule under the assumption of independent samples as,

$$P_{W|\mathbf{X}}(w|\mathcal{I}) = \frac{P_{\mathbf{X}|W}(\mathcal{I}|w)P_W(w)}{P_{\mathbf{X}}(\mathcal{I})}. \quad (2.6)$$

$$= \frac{\prod_j P_{\mathbf{X}|W}(\mathbf{x}_j|w)P_W(w)}{\prod_j P_{\mathbf{X}}(\mathbf{x}_j)} \quad (2.7)$$

Although any prior class distribution $P_W(w)$ can be supported, we assume a uniform distribution in what follows. This leads to

$$P_{W|\mathbf{X}}(w|\mathcal{I}) \propto \frac{\prod_j P_{\mathbf{X}|W}(\mathbf{x}_j|w)}{\prod_j P_{\mathbf{X}}(\mathbf{x}_j)} \quad (2.8)$$

The appearance model $P_{\mathbf{X}|W}(x|w)$ is modeled using a GMM, defined by the parameters $\Omega_w = \{\nu_w^j, \Phi_w^j, \beta_w^j\}$,

$$P_{\mathbf{X}|W}(\mathbf{x}|w; \Omega_w) = \sum_j \beta_w^j \mathcal{G}(\mathbf{x}, \nu_w^j, \Phi_w^j) \quad (2.9)$$

where, β_w is a probability mass function such that $\sum_j \beta_w^j = 1$ and j an index over the mixture components. The parameters Ω_w are learned from the set $\mathcal{D}_w^{\mathcal{W}}$ of all training images annotated with the w^{th} category using some density estimation procedure. In this work we rely on a *hierarchical estimation* procedure first proposed in [159], for image indexing. As shown in Figure 2.2, this procedure is itself composed of two steps. First, a Gaussian mixture is learned for each image in $\mathcal{D}_w^{\mathcal{W}}$,

producing a sequence of mixture densities

$$P_{\mathbf{X}|Y,W}(\mathbf{x}|y, w) = \sum_k \alpha_{w,y}^k \mathcal{G}(\mathbf{x}, \mu_{w,y}^k, \Sigma_{w,y}^k), \quad (2.10)$$

where Y is a hidden variable that indicates the index of the image in $\mathcal{D}_w^{\mathcal{W}}$. Note that, if a QBVE has already been implemented, these densities are just replicas of the ones of (2.4). In particular, if the mapping $\mathcal{M} : \{1, \dots, L\} \times \{1, \dots, D\} \rightarrow \{1, \dots, D\}$ translates the index (w, y) of the y^{th} image in $\mathcal{D}_w^{\mathcal{W}}$ into the image's index w on $\mathcal{D}^{\mathcal{W}}$, i.e. $w = \mathcal{M}(w, y)$, then

$$P_{\mathbf{X}|Y,W}(\mathbf{x}|y, w) = P_{\mathbf{X}|W}(\mathbf{x}|\mathcal{M}(w, y)).$$

Omitting, for brevity, the dependence of the mixture parameters on the semantic class w , assuming that each mixture has κ components, and that the cardinality of $\mathcal{D}_w^{\mathcal{W}}$ is D_w , this produces $D_w \kappa$ mixture components of parameters $\{\alpha_y^k, \mu_y^k, \Sigma_y^k\}, y = 1, \dots, D_w, k = 1, \dots, \kappa$. The second step is an extension of the EM algorithm, which clusters the Gaussian components into the mixture distribution of (2.9), using a hierarchical estimation technique (see [21, 159] for details). Because the number of parameters in each image mixture is orders of magnitude smaller than the number of feature vectors extracted from the image, the complexity of estimating concept mixtures is negligible when compared to that of estimating the individual image mixtures.

2.1.4 Image Representation

The literature on image representation is vast and goes back over five decades [1]. Although any type of visual features are acceptable, we only consider *localized features*, i.e., features of limited spatial support [153, 94, 150, 117]. Thus, a localized feature is a representation of a collection of adjoining image pixels, separating it from its immediate neighborhood. Usually image properties — such as intensity, color, texture, edges, edge orientations, frequency spectrum — change across these features. Localized features do not require sophisticated image segmentation procedures, which makes them computationally efficient and robust to scene clutter. Owing to these benefits, in recent years, they have been

quite successful for visual recognition tasks [94]. A large number of localized features have been proposed in the literature, the simplest being a vector of image pixel intensities [77]. Other descriptors emphasize different image properties like color [153, 49], texture [117, 111, 34], shape [49, 7], edges [85, 94], frequency spectrum [46, 62, 118, 156] etc. A comparison of these features for visual recognition tasks was presented in [153, 94]. In this work, since the main aim is to present an image representation that incorporates semantic cues, we do not debate on the choice of low-level feature representation, and rely on two popular localized image representations viz. scale invariant feature transform (SIFT) and discrete cosine transform (DCT). Infact, in Chapter 6 we show that, the semantic image representation improves over low-level visual features and moreover, the choice of low-level feature representation is not critical to the gains achieved. Next we present a brief description of both DCT and SIFT.

Discrete Cosine Transform

The discrete cosine transform (DCT) [62] expresses an image patch in terms of sum of cosine functions oscillating at different frequencies. A DCT of an image patch of size (N_1, N_2) is obtained as,

$$X_{k_1, k_2} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1, n_2} \cos \left[\frac{\pi}{N_1} \left(n_1 + \frac{1}{2} \right) k_1 \right] \cos \left[\frac{\pi}{N_2} \left(n_2 + \frac{1}{2} \right) k_2 \right]. \quad (2.11)$$

The DCT is widely used in image compression, and previous recognition experiments have shown that DCT features can lead to recognition rates comparable to those of many features proposed in the recognition literature [162]. It has also been shown that, for local image neighborhoods, DCT features approximates principal component analysis (PCA). This makes the space of DCT coefficients a natural choice for the feature space, \mathcal{X} , for visual recognition.

In this thesis, DCT features are computed on a dense regular grid, with a step of 8 pixels. 8×8 image patches are extracted around each grid point, and 8×8 DCT coefficients computed per patch and color channel. For monochrome images this results in a feature space of 64 dimensions. For color images the space is 192 dimensional.

Scale Invariant Feature Transform

The scale invariant feature transformation (SIFT), was proposed in [85] as a feature representation invariant to scale, orientation, and affine distortion, and partially invariant to illumination changes. SIFT is a measure of the orientations of the edges pixels in a given image patch. To compute the SIFT, 8-bin orientation histograms are computed in a 4×4 grid. This leads to a SIFT feature vector with $4 \times 4 \times 8 = 128$ dimensions. This vector is normalized to enhance invariance to changes in illumination.

SIFT can be computed for image patches which are selected either 1) by interest point detection, referred to as SIFT-INTR, or 2) on a dense regular grid, referred to as SIFT-GRID. While several interest point detectors are available in the literature, in this thesis SIFT-INTR is computed using interest points obtained with three saliency measures — Harris-Laplace, Laplace-of-Gaussian, and Difference-of-Gaussian — which are merged. These measures also provide scale information, which is used in the computation of SIFT features. For a dense grid, SIFT-GRID, feature points are sampled every 8 pixels. For both the strategies, SIFT features¹ are then computed over a 16×16 neighborhood around each feature point. On average, the two strategies yield similar number of samples per image.

2.2 Semantic Image Representation

While appearance features are intensity, texture, edge orientations, frequency bases, etc. those of the semantic representation are concept probabilities. Semantic image representation differs from appearance based representation in that, images are represented by vectors of *concept counts* $\mathcal{I} = (c_1, \dots, c_L)^T$, rather than being sampled from low-level feature space \mathcal{X} . Each low level feature vector \mathbf{x} for a given image, is assumed to be sampled from the probability distribution of a semantic concept and c_i is the number of low level feature vectors drawn from the i^{th} concept. The count vector for the y^{th} image is drawn from a multinomial

¹Computed using the SIFT implementation made available by LEAR at <http://lear.inrialpes.fr/people/dorko/downloads.html>

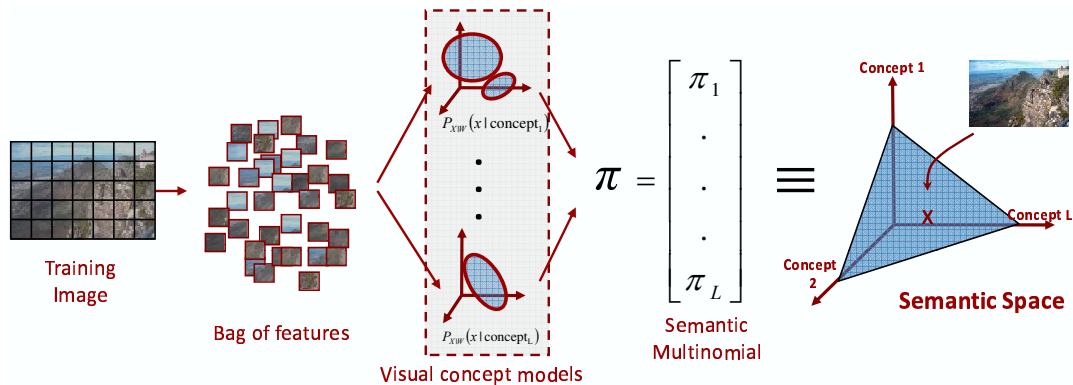


Figure 2.3: Image representation in semantic space \mathcal{S} , with a semantic multinomial (SMN) distribution. The SMN is a vector of posterior concept probabilities which encodes the co-occurrence of various concepts in the image, based on visual appearance.

variable \mathbf{T} of parameters $\boldsymbol{\pi}_y = (\pi_y^1, \dots, \pi_y^L)^T$

$$P_{\mathbf{T}|Y}(\mathcal{I}|y; \boldsymbol{\pi}_y) = \frac{n!}{\prod_{k=1}^L c_k!} \prod_{j=1}^L (\pi_y^j)^{c_j}, \quad (2.12)$$

where π_y^i is the probability that a feature vector is drawn from the i^{th} concept. The random variable \mathbf{T} can be seen as the result of a feature transformation from the space of visual features \mathcal{X} to the L -dimensional probability simplex \mathcal{S}_L . This mapping, $\boldsymbol{\Pi} : \mathcal{X} \rightarrow \mathcal{S}_L$ such that $\boldsymbol{\Pi}(\mathbf{X}) = \mathbf{T}$, maps the image $\mathcal{I} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, thereby the distribution $P_{\mathbf{X}|Y}(\mathcal{I}|y)$, into the multinomials $P_{\mathbf{T}|Y}(\mathcal{I}|y)$, and establishes a correspondence between images and points $\boldsymbol{\pi}_y \in \mathcal{S}_L$, as illustrated by Figure 2.3. We refer to each concept probability $\pi_y^i, i = 1, \dots, L$ a *semantic feature* and the probability vector $\boldsymbol{\pi}_y$ as a *semantic multinomial* (SMN) distribution. The probability simplex \mathcal{S}_L is itself referred to as the *semantic space* [119], which unlike \mathcal{X} has explicit semantics. Semantic features, or concepts, outside the vocabulary simply define directions orthogonal to the learned semantic space. In the example of 1.1, the mapping of the image onto the semantic simplex assigns high probability to (known) concepts such as ‘train’, ‘smoke’, ‘railroad’ etc.

2.2.1 The Semantic Multinomial

Learning the semantic space requires an image database \mathcal{D} and a vocabulary of semantic concepts, $\mathcal{L} = \{1, \dots, L\}$, where each image is labeled with a label vector, \mathbf{c}_d according to \mathcal{L} , making $\mathcal{D}^{\mathcal{L}} = \{(\mathcal{I}_1, \mathbf{c}_1), \dots, (\mathcal{I}_D, \mathbf{c}_D)\}$. \mathbf{c}_d is a binary L -dimensional vector such that $c_{d,i} = 1$ if the d^{th} image was annotated with the i^{th} keyword in \mathcal{L} . The dataset is said to be weakly labeled if absence of a keyword from caption \mathbf{c}_d does not necessarily mean that the associated concept is not present in \mathcal{I}_d . For example, an image containing “sky” may not be explicitly labeled with that keyword. This is usually the case in practical scenarios, since each image is likely to be annotated with a small caption that only identifies the semantics deemed as most relevant to the labeler. We assume weak labeling throughout this work. Note that, the vocabulary of scene categories \mathcal{W} can readily serve as a substitute for the vocabulary of semantic concepts \mathcal{L} . Infact, in absence of datasets annotated with semantic concepts, this is often the modus operandi to learn the semantic space. The only difference between the annotated datasets $\mathcal{D}^{\mathcal{W}}$ and $\mathcal{D}^{\mathcal{L}}$ is that in $\mathcal{D}^{\mathcal{W}}$ an image can be annotated with a single scene category (semantic concept) whereas in $\mathcal{D}^{\mathcal{L}}$ each image can be labeled with multiple concepts.

Given an annotated dataset $\mathcal{D}^{\mathcal{L}}$, appearance based concept models are learned for all the concepts in \mathcal{L} similar to that of learning appearance models for the scene categories. Next, the posterior concept probabilities $P_{W|\mathbf{x}}(w|\mathbf{x}_k)$, $w \in \{1, \dots, L\}$ is computed for *each* feature vector \mathbf{x}_k , $k \in \{1, \dots, N\}$, and \mathbf{x}_k is assigned to the concept of largest probability. Denoting, c_w as the total count of feature vectors assigned to the w^{th} concept in a given image, the maximum likelihood estimate of the semantic feature π_w is then given by [33]

$$\pi_w^{ML} = \arg \max_{\pi_w} \prod_{j=1}^L \pi_j^{c_j} = \frac{c_w}{\sum_j c_j} = \frac{c_w}{N}. \quad (2.13)$$

The vector, $\pi^{ML} = \{\pi_1^{ML}, \dots, \pi_L^{ML}\}$, is the ML estimate of the SMN for a given image.

2.2.2 Robust estimation of SMNs

As is usual in probability estimation, these posterior probabilities can be inaccurate for concepts with a small number of training images. Of particular concern are cases where some of the π_w are very close to zero, and can become ill-conditioned when used for recognition problems, where noisy estimates are amplified by ratios or logs of probabilities. A common solution is to introduce a prior distribution to regularize these parameters. Regularization can then be enforced by adopting a Bayesian parameter estimation viewpoint, where the parameter $\boldsymbol{\pi}$ is considered a random variable, and a prior distribution $P_{\boldsymbol{\Pi}}(\boldsymbol{\pi})$ introduced to favor parameter configurations that are, a priori, more likely.

Conjugate priors are frequently used, in Bayesian statistics [48], to estimate parameters of distributions in the exponential family, as is the case of the multinomial. They lead to a closed-form posterior (which is in the family of the prior), and *maximum a posteriori probability* parameter estimates which are intuitive. The conjugate prior of the multinomial is the Dirichlet distribution

$$\boldsymbol{\pi} \sim \mathbf{Dir}(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_j^L \alpha_j\right)}{\prod_{j=1}^L \Gamma(\alpha_j)} \prod_{j=1}^L \pi_j^{\alpha_j-1}, \quad (2.14)$$

of *hyper-parameters* α_i , and where $\Gamma(\cdot)$ is the Gamma function. Setting² $\alpha_i = \alpha$, the maximum a posteriori probability estimates are

$$\begin{aligned} \pi_w^{posterior} &= \arg \max_{\pi_w} P_{\mathbf{T}|\boldsymbol{\Pi}}(c_1, \dots, c_L | \boldsymbol{\pi}) P_{\boldsymbol{\Pi}}(\boldsymbol{\pi}) \\ &= \arg \max_{\pi_w} \prod_{j=1}^L \pi_j^{c_j} \prod_{j=1}^L \pi_j^{\alpha-1} \\ &= \frac{c_w + \alpha - 1}{\sum_{j=1}^L (c_j + \alpha - 1)}. \end{aligned} \quad (2.15)$$

This is identical to the maximum likelihood estimates obtained from a sample where each count is augmented by $\alpha - 1$, i.e. where each image contains $\alpha - 1$ more feature vectors from each concept. The addition of these vectors prevents zero counts, regularizing $\boldsymbol{\pi}$. As α increases, the multinomial distribution tends to uniform.

²Different hyper-parameters could also be used for the different concepts.

Noting, from (2.13), that $c_w = N\pi_w^{ML}$, the regularized estimates of (2.15) can be written as

$$\pi_w^{posterior} = \frac{\pi_w^{ML} + \pi_0}{\sum_j^L (\pi_j^{ML} + \pi_0)}.$$

with $\pi_0 = \frac{\alpha-1}{N}$.

2.2.3 SMNs as Posterior Probability Vector

The data processing theorem [88] advises against making hard decisions until the very last stages of processing. This suggests that thresholding the individual feature vector posteriors and counting is likely to produce worse probability estimates than those obtained without any thresholding. Motivated by the above argument, it is worth considering an alternative procedure for the estimation of π_w . Instead of (2.13), this consists of equating the semantic features π_w *directly with the posterior probability of the w^{th} semantic concept given the entire image*, i.e.

$$\pi_w^{direct} = P_{W|\mathbf{X}}(w|\mathcal{I}) \quad (2.16)$$

Thus, while in (2.13), posterior probability vector for each feature vector is threshold and aggregated over the entire image, in (2.16) the posterior probability vector is computed directly from the entire collection of the feature vectors. Thus, given an image $\mathcal{I} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ the vector of posterior probabilities

$$\boldsymbol{\pi}^{direct} = (P_{W|\mathbf{X}}(1|\mathcal{I}), \dots, P_{W|\mathbf{X}}(L|\mathcal{I}))^T \quad (2.17)$$

provides a rich description of the image semantics and a robust alternative to the estimation of its SMN. Furthermore, regularized estimates of (2.17) can be obtained with

$$\pi_w^{reg} = \frac{\pi_w^{direct} + \pi_0}{1 + L\pi_0} \quad (2.18)$$

which is equivalent to using maximum a posteriori probability estimates, in the thresholding plus counting paradigm, with the Dirichlet prior of (2.14). In this work we rely on (2.18) to obtain a SMN of a given image.

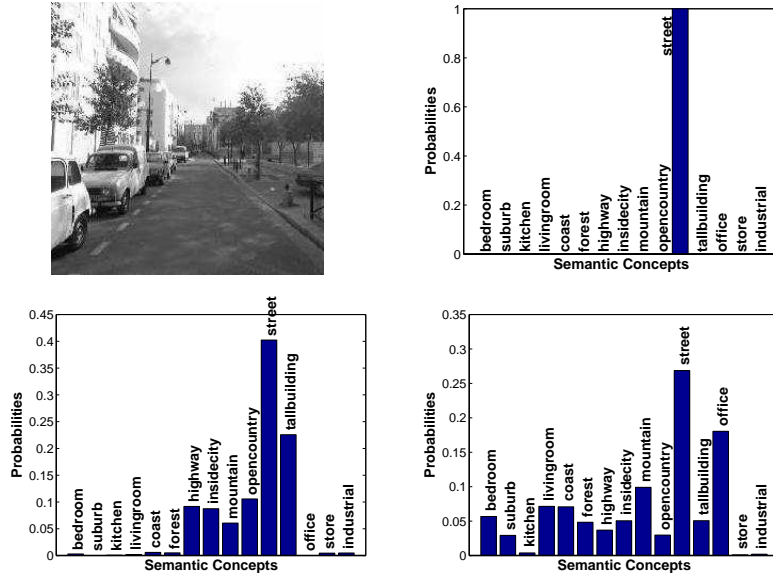


Figure 2.4: SMN for the image shown on the top left computed using (top-right) (2.8), (bottom-left) (2.21) and (bottom-right) (2.23).

2.3 Computing the Semantic Multinomial

It should be noted that the architecture proposed above is generic, in the sense that any appearance recognition system that produces a vector of posterior probabilities $\boldsymbol{\pi}$, can be used to learn the proposed contextual models. In fact, these probabilities can even be produced by systems that do not model appearance explicitly, e.g. multi-class logistic regression, multi-class SVM etc. This is achieved by converting classifier scores to a posterior probability distribution, using probability calibration techniques. For example, the distance from the decision hyperplane learned by an SVM can be converted to a posterior probability using a sigmoidal transform [110]. In practice, however, care must be taken to guarantee that the appearance classifiers are not too strong. If they make very hard decisions, e.g. assign images to a single class, the SMN would simply indicate the presence of a single concept and would not be rich enough to build visual recognition systems. Infact, in Chapter 5 we use multi-class logistic regression to compute the SMNs.

In the MPE implementation above, it is natural to use the posterior probabilities of (2.18) as the SMN of image \mathcal{I} . However, as N tends to be large, there

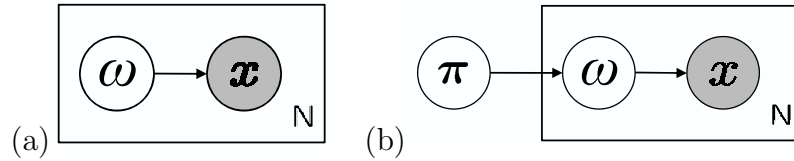


Figure 2.5: Alternative generative models for image formation at the appearance level. (a) A concept is sampled per appearance feature vector rather than per image, from $P_{\mathbf{x}|w}(\mathbf{x}|w)$. (b) Explicit modeling of the contextual variable Π from which a single SMN is drawn per image.

Table 2.1: SMN Entropy.

Model	Entropy
Figure 2.1, Eq (2.8)	0.003 ± 0.044
Figure 2.5(a), Eq (2.21)	2.530 ± 0.435
Figure 2.5(b), Eq (2.23)	2.546 ± 0.593

is usually very strong evidence in favor of one concept, not always that of greatest perceptual significance. For example, if the image has a large region of “sky”, the existence of many sky patches makes the posterior probability of the “sky” concept close to one. This is illustrated in Figure 2.4 (top-right) where the SMN assigns all probability to a single concept. Table 2.1 shows that this happens frequently: the average entropy of the SMNs computed on the N15 Dataset (to be introduced later) is very close to 0. Note that this is the property that enables the learning of the appearance based models from the weakly supervised datasets: when all images containing “sky” are grouped, the overall feature distribution is very close to that of the “sky” concept, despite the fact that the training set contains spurious image patches from other concepts. This is an example of the multiple instance learning paradigm [155], where an image, consisting of some patches from the concept being modeled and some spurious patches from other concepts, serves as the positive bag. Although this dominance of the strongest concept is critical for learning, the data processing theorem advises against it during inference. Or, in other words, while multiple instance learning is required, multiple instance in-

ference is undesirable. In particular, modeling images as bags-of-features from a *single concept*, as in Figure 2.1, does not lend to contextual inference.

One alternative is to perform inference with the much looser model of Figure 2.5(a), where a concept is sampled *per appearance feature vector*, rather than per image. Note that, because labeling information is not available per vector, the models $P_{\mathbf{X}|W}(\mathbf{x}|w)$ are still learned as before, using the multiple instance learning principle. The only difference is the inference procedure. In this case, SMNs are available per image patch denoted as patch-SMN, $\boldsymbol{\pi}^n = P_{W|X}(w_n|x_n), n \in \{1, \dots, N\}$. Determining an SMN, denoted the *Image-SMN*, for the entire image requires computing a representative for this set of patch-SMNs. One possibility is the multinomial of minimum average Kullback-Leibler divergence with all patch-SMNs

$$\boldsymbol{\pi}^* = \arg \min_{\boldsymbol{\pi}} \frac{1}{N} \sum_{n=1}^N KL(\boldsymbol{\pi} || \boldsymbol{\pi}^n) \quad \text{s.t.} \quad \sum_{i=1}^L \pi_i = 1. \quad (2.19)$$

As shown in Appendix C, this is the representative

$$\pi_i^* = \frac{\exp \frac{1}{N} \sum_n \log \pi_i^n}{\sum_i \exp \frac{1}{N} \sum_n \log \pi_i^n}, \quad (2.20)$$

which reduces to

$$\pi_i^* = \frac{\exp \left\{ \frac{1}{n} \sum_n \log P_{X|W}(x_n|i) \right\}}{\sum_j \exp \left\{ \frac{1}{n} \sum_n \log P_{X|W}(x_n|j) \right\}} \quad (2.21)$$

for a uniform prior. This is in contrast to the posterior estimate of (2.8). Note that while (2.8) computes a product of likelihoods, (2.21) computes their geometric mean.

A second possibility is to adopt the generative model of Figure 2.5(b). This explicitly accounts for the contextual variable $\boldsymbol{\Pi}$, from which a single SMN is drawn per image. A concept is then drawn per image patch. In this case, the Image-SMN is

$$\boldsymbol{\pi}^* = \arg \max_{\boldsymbol{\pi}} P_{\boldsymbol{\Pi}|X}(\boldsymbol{\pi}|\mathcal{I}). \quad (2.22)$$

However, this optimization is intractable, and only approximate inference is possible. A number of approximations can be used, including Laplace or variational

approximations, sampling, etc. In Appendix D we show that, for a variational approximation,

$$\pi_i^* = \frac{\gamma_i - 1}{\sum_j \gamma_j - L} \quad (2.23)$$

where, γ_i is computed with the following iteration,

$$\gamma_i^* = \sum_n \phi_{ni} + \alpha_i \quad (2.24)$$

$$\phi_{ni}^* \propto P_{X|W}(x_n|w_n = i) e^{\psi(\gamma_i) - \psi(\sum_j \gamma_j)}. \quad (2.25)$$

Here, α_i is the parameter of the prior $P_{\Pi}(\pi)$ which, for compatibility with the assumption of uniform class priors, we set to 1, $\psi(\cdot)$ the Digamma function, and γ_i , ϕ_{ni} the parameters of the variational distributions. Figure 2.4 shows that the SMNs obtained with (2.21) and (2.23) are rich in contextual information. Table 2.1 shows that the two models lead to approximately the same average SMN entropy on N15, which is much higher than that of (2.8).

Since (2.23) involves an iterative procedure, which is more expensive than the closed form of (2.21), (2.21) is the default choice for computing the SMNs in this work. In Chapter 6 we will show that (2.21) also yield marginally better performance over (2.23), in a scene classification task.

2.4 Related Work

The idea of representing documents as weighted combinations of the words in a pre-defined vocabulary is commonly used in information retrieval. In fact, the classic model for information retrieval is the vector space model of Salton [125, 126]. Under this model, documents are represented as collections of keywords, weighted by importance, and can be interpreted as points in the semantic space spanned by the vocabulary entries. In image retrieval, there have been some proposals to represent images as points in a semantic vector space. The earliest among these efforts [68, 54] were based on semantic information extracted from metadata - viz. origin, filename, image url, keywords from surrounding webpage text, manual annotations, etc.

The closest works, in the literature, to the semantic image representation proposed here, are the systems proposed by Smith *et al.* in [137, 135] and Lu *et al.* in [86]. To the best of our knowledge, [137] pioneered the idea of learning a semantic space by learning a separate statistical model for each concept. The vector of semantic weights, denoted as the ‘model vector’, is learned from the image content. Each image receives a confidence score per semantic concept, based on the proximity of the image to the decision boundary of a support vector machine (SVM) trained to recognize the concept. While laying the foundations for the semantic image representation, [137] does not present any formal definition or systematic analysis of the semantic image representation, as presented in Section 2.2. Moreover in [137], the model vector is used solely for the task of retrieving images known to the system (that were used to learn the SVM classifiers). In Chapter 3 we show that the benefits of semantic representation goes beyond that, and propose image retrieval systems that can generalize well beyond the known vocabulary. Furthermore, we present two novel visual recognition systems, viz. scene classification and cross-modal multimedia retrieval based on the semantic image representation. Infact, the problem of cross-modal multimedia is itself in its nascency and no formal analysis has been presented in the literature, which we do in Chapter 5. Finally, in [137] the model vector is simply used as an alternative image representation, without any analysis of their ability to model semantic “gist” and context of an image. In Chapter 6 we introduce “contextual models” and show that the proposed representation is successful in modeling the “gist” of an image.

2.5 Acknowledgments

The text of Chapter 2, in part, is based on the material as it appears in: N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Bridging the Semantic Gap: Query by Semantic Example*’, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007. and N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Query by Semantic Example*’, ACM International Conference on Image and Video Retrieval, LNCS

51-60, Phoenix, 2006. The dissertation author was a primary researcher and an author of the cited material.

Chapter 3

Image Retrieval: Query By Semantic Example

3.1 Introduction

Content-based image retrieval, the problem of searching for digital images in large image repositories according to their content, has been the subject of significant research in the recent past [133, 101, 107, 160]. Two main retrieval paradigms have evolved over the years: one based on visual queries, here referred to as *query-by-visual-example* (QBVE), and the other based on text, here denoted as *semantic retrieval* (SR). Early retrieval architectures were almost exclusively based on QBVE [61, 134, 90, 101, 107]. Under this paradigm, each image is decomposed into a number of low-level visual features (e.g. a color histogram) and image retrieval is formulated as the search for the best database match to the feature vector extracted from a query image. It was, however, quickly realized that strict visual similarity is, in most cases, weakly correlated with the measures of similarity adopted by humans for image comparison.

This motivated the more ambitious goal of designing retrieval systems with support for semantic queries [109]. The basic idea is to annotate images with semantic keywords, enabling users to specify their queries through a natural language description of the visual concepts of interest. Because manual image labeling is a labor intensive process, SR research turned to the problem of the automatic extraction of semantic descriptors from images, so as to build models of visual appearance of the semantic concepts of interest. This is usually done by the application of machine learning algorithms. Early efforts targeted the extraction of specific semantics [142, 152, 53, 45] under the framework of binary classification. More recently there has been an effort to solve the problem in greater generality, through the design of techniques capable of learning relatively large semantic vocabularies from informally annotated training image collections. This can be done with resort to both unsupervised [5, 35, 12, 41, 72] and weakly supervised learning [70, 22].

In spite of these advances, the fundamental question of whether there is an intrinsic value to building models at a semantic level, remains poorly understood. On one hand, SR has the advantage of evaluating image similarity at a higher

level of abstraction, and therefore better generalization¹ than what is possible with QBVE. On the other hand, the performance of SR systems tends to degrade for semantic classes that they were not trained to recognize. Since it is still difficult to learn appearance models for massive concept vocabularies, this could compromise the generalization gains due to abstraction. This problem is seldom considered in the literature, where most evaluations are performed with query concepts that are known to the retrieval system [5, 12, 35, 41, 72, 22].

In fact, it is not even straightforward to compare the two retrieval paradigms, because they assume different levels of query specification. While a semantic query is usually precise (e.g. ‘the White House’) a visual example (a picture of the ‘White House’) will depict various concepts that are irrelevant to the query (e.g. the street that surrounds the building, cars, people, etc.). It is, therefore, possible that better SR results could be due to a better interface (natural language) rather than an intrinsic advantage of representing images semantically. This may be of little importance when the goal is to build the next generation of (more accurate) retrieval systems. However, given the complexity of the problem, it is unlikely that significant further advances can be achieved without some understanding of the intrinsic value of semantic representations. If, for example, abstraction is indeed valuable, further research on appearance models that account for image taxonomies could lead to exponential gains in retrieval accuracy. Else, if the advantages are simply a reflection of more precise queries, such research is likely to be ineffective.

In this chapter, we introduce a novel image retrieval framework based on semantic image representation, which extends the query-by-example paradigm to the semantic domain. This consists of defining a semantic feature space, where each image is represented by the vector of posterior concept probabilities assigned to it by a semantic labeling system, and performing query-by-example in this space. We refer to the combination of the two paradigms as query-by-semantic-example (QBSE), and present an extensive comparison of its performance with that of QBVE. It is shown that QBSE has significantly better performance for both

¹Here, and throughout this work, we refer to the definition of ‘generalization’ common in machine learning and content-based retrieval: the ability of the retrieval system to achieve low error rates outside of the set of images on which it was trained.

concepts known and unknown to the retrieval system, i.e., it can generalize beyond the vocabulary used for training. It is also shown that, since both QBSE and QBVE share a common framework i.e. that of minimum probability of error retrieval [156], the performance gain of QBSE over QBVE is intrinsic to the semantic nature of image representation.

The chapter is organized as follows. Section 3.2 briefly reviews previous retrieval work related to QBSE. Section 3.3 discusses the limitations of the QBVE and SR paradigms, motivating the adoption of QBSE. Section 3.4 proposes an implementation of QBSE, compatible with the MPE formulation. It is then argued, in Section 3.5, that the generalization ability of QBSE can significantly benefit from the combination of multiple queries, and various strategies are proposed to accomplish this goal. A thorough experimental evaluation of the performance of QBSE is presented in Section 3.6, where the intrinsic gains of semantic image representations (over strict visual matching) are quantified.

3.2 Related Work

Although the task of building semantic image representations for image retrieval, has been on recent interest in the community, few proposals have so far been presented on how to best exploit the semantic space for the design of retrieval systems. A somewhat popular technique to construct content-based semantic spaces, is to resort to active learning based on user’s relevance feedback [161, 87, 56]. The idea is to pool the images relevant to a query, after several rounds of relevance feedback, to build a model for the semantic concept of interest. Assuming that 1) these images do belong to a common semantic class, and 2) the results of various relevance feedback sessions can be aggregated, this is a feasible way to incrementally build a semantic space. An example is given in [75], where the authors propose a retrieval system based on image embeddings. Using relevance feedback, the system gradually clusters images and learns a non-linear embedding which maps these clusters into a hidden space of semantic attributes. Cox *et al.* [26] also focus on the task of learning a predictive model for user selections, by learning a mapping

between 1) the image selection patterns made by users instructed to consider visual similarity and 2) those of users instructed to consider semantic similarity.

These works have focused more on the issue of learning the semantic space than that of its application to retrieval. In fact, it is not always clear how the learned semantic information could be combined with the visual search at the core of the retrieval operation. Furthermore, the use of relevance feedback to train a semantic retrieval system has various limitations. First, it can be quite time consuming, since a sizable number of examples is usually required to learn each semantic model. Second, the assumption that all queries performed in a relevance feedback session are relative to the same semantic concept is usually not realistic, even when users are instructed to do so. For example, a user searching for pictures of ‘cafes in Paris’ is likely to oscillate between searching for pictures of ‘cafes’ and pictures of ‘Paris’.

The closest works in the literature, to the QBSE paradigm adopted here, are those of [137, 135, 86], where retrieval is carried out based on computing L^2 similarity between “model-vectors”, a representation similar to that of semantic image representation. While laying the foundations for QBSE, [137, 135] did not investigate any of the fundamental questions that we now consider. First, because there was no attempt to perform retrieval on databases not used for training, it did not address the problem of generalization to concepts unknown to the retrieval system. As we will see, this is one of the fundamental reasons to adopt QBSE instead of the standard SR query paradigm. Second, although showing that QBSE outperformed a QBVE system, this work did not rely on the same image representation for the two query paradigms. While QBVE was based on either color or edge histogram matching, QBSE relied on a feature space composed of a multitude of visual features, including color and edge histograms, wavelet-based texture features, color correlograms and measures of texture co-occurrence. Because the representations are different, it is impossible to conclude that the improved performance of the QBSE system derives from an *intrinsic* advantage of semantic representations. In what follows, we preempt this caveat by adopting the same image representation and retrieval framework for the design of all systems.

3.3 Query by Semantic Example

Both the QBVE and SR implementations of MPE retrieval have been extensively evaluated in [156] and [21, 22]. Although these evaluations have shown that the two implementations are among the best known techniques for visual and semantic retrieval, the comparison of the two retrieval paradigms is difficult. We next discuss this issue in greater detail, and motivate the adoption of an alternative retrieval paradigm, QBSE, that combines the best properties of the two approaches.

3.3.1 Query by Visual Example vs Semantic Retrieval

Both QBVE and SR have advantages and limitations. Because concepts are learned from collections of images, SR can *generalize* significantly better than QBVE. For example, by using a large training set of images labeled with the concept ‘sky’, containing both images of sky at daytime (when the sky is mostly blue) and sunsets (when the sky is mostly orange), a SR system can learn that ‘sky’ is sometimes blue and others orange. This is a simple consequence of the fact that a large set of ‘sky’ images populate, with high probability, the blue and orange regions of the feature space. It is, however, not easy to accomplish with QBVE, which only has access to two images (the query and that in the database) and can only perform direct matching of visual features. We refer to this type of abstraction, as *generalization inside the semantic space*, i.e., inside the space of concepts that the system has been trained to recognize.

While better generalization is a strong advantage for SR, there are some limitations associated with this paradigm. An obvious difficulty is that most images have multiple semantic interpretations. 3.1 presents an example, identifying various semantic concepts as sensible annotations for the image shown. Note that this list, of relatively salient concepts, is a small portion of the keywords that could be attached to the image. Other examples include colors (e.g. ‘yellow’ train), or objects that are not salient in an abstract sense but could become very relevant in some contexts (e.g. the ‘paint’ of the markings on the street, the ‘letters’ in



Figure 3.1: An image containing various concepts: ‘train’, ‘smoke’, ‘road’, ‘sky’, ‘railroad’, ‘sign’, ‘trees’, ‘mountain’, ‘shadows’, with variable degrees of presence.

the sign, etc.). In general, it is impossible to predict all annotations that may be relevant for a given image. This is likely to compromise the performance of a SR system. Furthermore, because queries are specified as text, a SR system is usually limited by the size of its vocabulary². In summary, SR can generalize poorly *outside the semantic space*.

Since visual retrieval has no notion of semantics, it is not constrained by either vocabulary or semantic interpretations. When compared to SR, QBVE systems can generalize better outside the semantic space. In the example of 3.1, a QBVE would likely return the image shown as a match to a query depicting an industrial chimney engulfed in dark smoke (a more or less obvious query prototype for images of ‘pollution’) despite the fact that the retrieval system knows nothing about ‘smoke’, ‘pollution’, or ‘chimneys’. Obviously, there are numerous examples where QBVE correlates much worse with perceptual similarity than SR. We have already seen that when the latter is feasible, i.e. inside the semantic space, it has better generalization. Overall, it is sensible to expect that SR will perform better

²It is, of course, always possible to rely on text processing ideas based on thesauri and ontologies like WordNet [39] to mitigate this problem. For example, query expansion can be used to replace a query for ‘pollution’ by a query for ‘smoke’, if the latter is in the vocabulary and the former is not. While such techniques are undeniably useful for practical implementation of retrieval systems, they do not reflect an improved ability, by the retrieval system, to model the relationships between visual features and words. They are simply an attempt to fix these limitations a posteriori (i.e. at the language level) and are, therefore, beyond the scope of this work. In practice, it is not always easy to perform text-based query expansion when the vocabulary is small, as is the case for most SR systems, or when the queries report to specific instances (e.g. a person’s name).

inside the semantic space, while QBVE should fare better outside of it. In practice, however, it is not easy to compare the two retrieval paradigms. This is mostly due to the different forms of query specification. While a natural language query is usually precise (e.g. ‘train’ and ‘smoke’), a query image like that of 3.1 always contains a number of concepts that are not necessarily relevant to the query (e.g. ‘mountain’, or even ‘yellow’ for the train color). Hence, the better performance of SR (inside the semantic space) could be simply due to higher query precision. A fair comparison would, therefore, require the optimization of the precision of visual queries (e.g. by allowing the QBVE system to rely on image regions as queries) but this is difficult to formalize.

Overall, both the engineering question of how to design better retrieval systems (with good generalization inside and outside of the semantic space) and the scientific question of whether there is a real benefit to semantic representations, are difficult to answer under the existing query paradigms. To address this problem we propose an alternative paradigm, which is denoted as *query by semantic example* (QBSE).

3.3.2 Query by Semantic Example

A QBSE system operates on a *semantic space* - the space of semantic features introduced in Chapter 2, according to a similarity mapping $f : \mathcal{S}_L \rightarrow \{1, \dots, D\}$ such that

$$f(\boldsymbol{\pi}) = \arg \max_y s(\boldsymbol{\pi}, \boldsymbol{\pi}_y) \quad (3.1)$$

where \mathcal{S}_L is the semantic space, $\boldsymbol{\pi}$ the query SMN and $\boldsymbol{\pi}_y$ the SMN that characterizes the y^{th} database image, and $s(\cdot, \cdot)$ an appropriate similarity function. As shown in 3.2 (top), the user provides a query image, for which a SMN $\boldsymbol{\pi}$ is computed, and compared to all the SMNs $\boldsymbol{\pi}_y$ previously stored for the images in the database. Note that this paradigm differs from SR, as in SR the user specifies a short natural language description which implies only a small number of concepts are assigned non-zero probability. This is illustrated in 3.2 (bottom) where queries in SR are restricted to the edges of the semantic space.

QBSE query paradigm has a number of interesting properties. As discussed

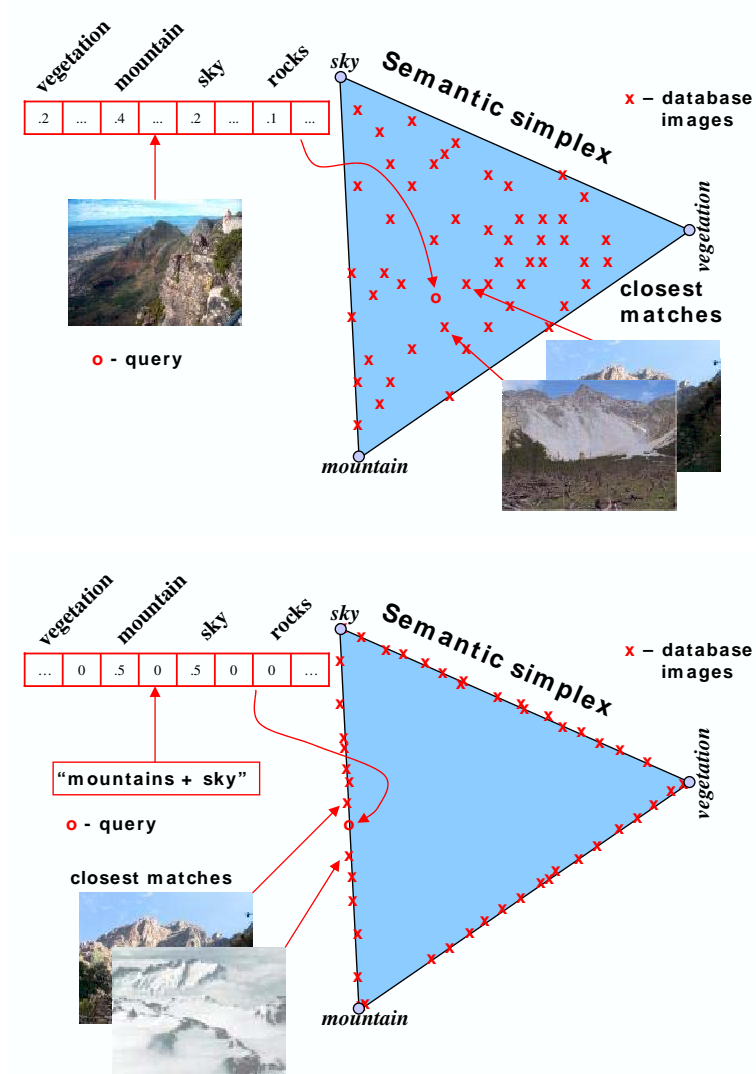


Figure 3.2: Semantic image retrieval. Top: Under QBSE the user provides a query image, probabilities are computed for all concepts, and the image represented by the concept probability distribution. Bottom: Under the traditional SR paradigm, the user specifies a short natural language description, and only a small number of concepts are assigned a non-zero posterior probability.

in Chapter 2, the semantic space \mathcal{S}_L is defined by the concepts in the vocabulary known to the system. The semantic features, or concepts, outside the vocabulary simply define directions orthogonal to the learned semantic space. This implies that, by projecting these dimensions onto the simplex, the QBSE system can gen-

eralize beyond the known semantic concepts. In the example of 3.1, the mapping of the image onto the semantic simplex assigns high probability to (known) concepts such as ‘train’, ‘smoke’, ‘railroad’ etc. This makes the image a good match for other images containing large amounts of ‘smoke’, such as those depicting industrial chimneys or ‘pollution’ in general. The system can therefore establish a link between the image of 3.1 and ‘pollution’, despite the fact that it has no *explicit* knowledge of the ‘pollution’ concept³. Second, when compared to QBVE, QBSE complements all the advantages of query by example with the advantages of a semantic representation. Moreover, since in both cases queries are specified by the same examples, any differences in their performance can be directly attributed to the semantic vs. visual nature of the associated image representations⁴. This enables the objective comparison of QBVE and QBSE.

3.4 The Proposed Query by Semantic Example System

QBSE is a generic retrieval paradigm and, as such, can be implemented in many different ways. Any implementation must specify a method to estimate the SMN that describes each image, and a similarity function between SMNs. In the implementation presented herein, the SMN vectors π_i are learned with a semantic labeling system described in 2.2, which implements the mapping $\mathbf{\Pi}$, by computing an estimate of posterior concept probabilities given the observed feature vectors

$$\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I}). \quad (3.2)$$

In the rest of this section, we describe the various similarity functions.

³Note that this is different from text-based query expansion, where the link between ‘smoke’ and ‘pollution’ must be *explicitly* defined. In QBSE, the relationship is instead inferred automatically, from the fact that both concepts have commonalities of visual appearance.

⁴This assumes, of course, that a common framework, such as MPE, is used to implement both the QBSE and QBVE systems.

3.4.1 Similarity Function

There are many known methods to measure the distance between two probability distributions, all of which can be used to measure the similarity of two SMNs. Furthermore, because the latter can also be interpreted as normalized vectors of counts, this set can be augmented with all measures of similarity between histograms. We have compared various similarity functions for the purpose of QBSE.

Kullback-Leibler (KL) Divergence

The KL divergence between two distributions $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ is

$$s_{KL}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\boldsymbol{\pi}') = \sum_{i=1}^L \pi_i \log \frac{\pi_i}{\pi'_i}. \quad (3.3)$$

It is non-negative, and equal to zero when $\boldsymbol{\pi} = \boldsymbol{\pi}'$. For retrieval, it also has an intuitive interpretation as the asymptotic limit of (2.1) when Y is uniformly distributed [158]. However, it is not symmetric, i.e. $KL(\boldsymbol{\pi}||\boldsymbol{\pi}') \neq KL(\boldsymbol{\pi}'||\boldsymbol{\pi})$. A symmetric version can be defined as

$$s_{symmKL}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\boldsymbol{\pi}') + KL(\boldsymbol{\pi}'||\boldsymbol{\pi}) \quad (3.4)$$

$$= \sum_{i=1}^L \pi_i \log \frac{\pi_i}{\pi'_i} + \sum_{i=1}^L \pi'_i \log \frac{\pi'_i}{\pi_i}. \quad (3.5)$$

Jensen-Shannon Divergence

The Jensen-Shannon divergence (JS) is a measure of whether two samples, as defined by their empirical distributions, are drawn from the same source distribution [25]. It is defined as

$$s_{JS}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}) + KL(\boldsymbol{\pi}'||\hat{\boldsymbol{\pi}}) \quad (3.6)$$

where $\hat{\boldsymbol{\pi}} = \frac{1}{2}\boldsymbol{\pi} + \frac{1}{2}\boldsymbol{\pi}'$. This divergence can be interpreted as the average distance (in the KL sense) between each distribution and the average of all distributions.

Correlation

The correlation between two SMNs is defined as

$$s_{CO}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \boldsymbol{\pi}^T \boldsymbol{\pi}' = \sum_i^L \pi_i \times \pi'_i. \quad (3.7)$$

Unlike the KL or JS divergence, which attain their minimum value (zero) for equal distributions, correlation is maximum in this case. The maximum value is, however, a function of the distributions under consideration. This limitation can be avoided by the adoption of the *normalized correlation*,

$$s_{NC}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \frac{\boldsymbol{\pi}^T \boldsymbol{\pi}'}{\|\boldsymbol{\pi}\| \|\boldsymbol{\pi}'\|} = \frac{\sum_i^L \pi_i \times \pi'_i}{\sqrt{\sum_j \pi_j^2} \sqrt{\sum_j \pi_j'^2}}. \quad (3.8)$$

Other Similarity Measures

A popular set of image similarity metrics is that of L^p distances

$$s_{L^p}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \left(\sum_{i=1}^L |\pi_i - \pi'_i|^p \right)^{\frac{1}{p}}. \quad (3.9)$$

These distances are particularly common in color-based retrieval, where they are used as metrics of similarity between color histograms. Another popular metric is the histogram intersection (HI) [141],

$$s_{HI}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \sum_{i=1}^L \min(\pi_i, \pi'_i), \quad (3.10)$$

the maximization of which is equivalent minimizing the L^1 norm.

3.5 Multiple Image Queries

A QBSE system can theoretically benefit from the specification of queries through multiple examples. We next give some reasons for this and discuss various alternatives for query combination.

3.5.1 The Benefits of Query Fusion

Semantic image labeling is, almost by definition, a noisy endeavor. This is a consequence of the fact that various interpretations are usually possible for a given arrangement of image intensities. An example is given in 1.1 where we show an image and the associated SMN. While most of the probability mass is assigned to concepts that are present in the image (‘railroad’, ‘locomotive’, ‘train’, ‘street’, or ‘sky’), two of the concepts of largest probability do not seem related to it: ‘bridge’, and ‘arch’. Close inspection of the image (see close-up presented in the figure), provides an explanation for these labels: when analyzed locally, the locomotive’s roof actually resembles the arch of a bridge. This visual feature seems to be highly discriminant, since when used as a query in a QBVE system, most of the top matches are images with arch-like structures, not trains (see 3.6). While these types of errors are difficult to avoid, they are *accidental*. In particular, the arch-like structure of 1.1 is the result of viewing a particular type of train, at a particular viewing angle, and a particular distance. It is unlikely that similar structures will emerge consistently over a set of train images. There are obviously other sources of error, such as classification mistakes for which it is not possible to encounter a plausible explanation. But these are usually even less consistent, across a set of images, than those due to accidental visual resemblances. A pressing question is then whether it is possible to exploit the lack of consistency of these errors to obtain a better characterization of the query image set?

We approach this question from a *multiple instance* learning perspective [92, 2], formulating the problem as one of learning from *bags of examples*. In QBSE, each image is modeled as a bag of feature vectors, which are drawn from the different concepts according to the probabilities π_i . When the query consists of multiple images, or bags, the negative examples that appear across those bags are inconsistent (e.g. the feature vectors associated with the arch-like structure which is prominent in 1.1 but does not appear consistently in all train images), and tend to be spread over the feature space (because they also depict background concepts, such as roads, trees, mountains, etc., which vary from image to image). On the other hand, feature vectors corresponding to positive examples are likely

to be concentrated within a small region of the space. It follows that, although the distribution of positive examples may not be dominant in any individual bag, the consistent appearance in all bags makes it dominant over the entire query ensemble. This suggests that a better estimate of the query SMN should be possible by considering a set of multiple query images.

In addition to higher accuracy, a set of multiple queries is also likely to have better generalization, since a single image does not usually exhibit all possible visual manifestations of a given semantic class. For example, images depicting ‘bikes on roads’ and ‘cars in garage’ can be combined to retrieve images from the more general class of ‘vehicles’. A combination of the two query image sets enables the retrieval system to have a more complete representation of the vehicle class, by simultaneously assigning higher weights to the concepts ‘bike’, ‘cars’, ‘road’, and ‘garage’. This enables the retrieval of images of ‘bikes in garage’ and ‘cars on roads’, matches that would not be possible if the queries were used individually.

3.5.2 Query Combination

Under MPE retrieval, query combination is relatively straightforward to implement by QBVE systems. Given two query images $\mathcal{I}_q^1 = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \dots, \mathbf{x}_n^1\}$ and $\mathcal{I}_q^2 = \{\mathbf{x}_1^2, \mathbf{x}_2^2, \dots, \mathbf{x}_n^2\}$, the probability of the composite query $\mathcal{I}_q^C = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \dots, \mathbf{x}_n^1, \mathbf{x}_1^2, \mathbf{x}_2^2, \dots, \mathbf{x}_n^2\}$ given class $Y = y$ is

$$\begin{aligned} P_{\mathbf{X}|Y}(\mathcal{I}_q^C|y) &= \prod_{k=1}^n P_{\mathbf{X}|Y}(\mathbf{x}_k^1|y) \prod_{l=1}^n P_{\mathbf{X}|Y}(\mathbf{x}_l^2|y) \\ &= P_{\mathbf{X}|Y}(\mathcal{I}_q^1|y) P_{\mathbf{X}|Y}(\mathcal{I}_q^2|y). \end{aligned} \quad (3.11)$$

The MPE decision of (2.1) for the composite query is obtained by combining (3.11) with (2.4) and Bayes rule.

In the context of QBSE, there are at least three possibilities for query combination. The first is equivalent to (3.11), but based on the probability of the

composite query \mathcal{I}_q^C given semantic class $W = w$,

$$\begin{aligned} P_{\mathbf{X}|W}(\mathcal{I}_q^C|w) &= \prod_{k=1}^n P_{\mathbf{X}|W}(\mathbf{x}_k^1|w) \prod_{l=1}^n P_{\mathbf{X}|W}(\mathbf{x}_l^2|w) \\ &= P_{\mathbf{X}|W}(\mathcal{I}_q^1|w) P_{\mathbf{X}|W}(\mathcal{I}_q^2|w), \end{aligned} \quad (3.12)$$

which is combined with (2.9) and Bayes rule to compute the posterior concept probabilities of (3.2). We refer to (3.12) as the ‘LKLD combination’ strategy for query combination. It is equivalent to taking a geometric mean of the probabilities of the individual images given the class.

A second possibility is to represent the query as a mixture of SMNs. This relies on a different generative model than that of (3.12): the i^{th} query is first selected with probability λ_i and a count vector is then sampled from the associated multinomial distribution. It can be formalized as

$$P_{\mathbf{T}}(\mathcal{I}_q^C; \boldsymbol{\pi}_q) = \frac{n!}{\prod_{k=1}^L c_k!} \prod_{j=1}^L (\lambda_1 \pi_1^j + \lambda_2 \pi_2^j)^{c_j}, \quad (3.13)$$

where $P_{\mathbf{T}}(\mathcal{I}_q^C; \boldsymbol{\pi}_q)$ is the multinomial distribution for the query combination, of parameter $\boldsymbol{\pi}_q = \lambda_1 \boldsymbol{\pi}_1 + \lambda_2 \boldsymbol{\pi}_2$. $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are the parameters of the individual multinomial distribution, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ the vector of query selection probabilities. If $\lambda_1 = \lambda_2$, the two SMNs are simply averaged. We adopt the uniform query selection prior, and refer to this strategy as ‘SMN combination’. Geometrically, it sets the combined SMN to the centroid of the simplex that has the SMNs of the query images as vertices. This ranks highest the database SMN which is closest to this centroid.

The third possibility, henceforth referred to as ‘KL combination’, is to execute the multiple queries separately, and combine the resulting image rankings. For example, when similarity is measured with the KL divergence, the divergence between the combined image SMN, $\boldsymbol{\pi}_q$, and database SMNs $\boldsymbol{\pi}_y$ is,

$$s_{KL}(\boldsymbol{\pi}_q, \boldsymbol{\pi}_y) = \frac{1}{2} KL(\boldsymbol{\pi}_1 || \boldsymbol{\pi}_y) + \frac{1}{2} KL(\boldsymbol{\pi}_2 || \boldsymbol{\pi}_y). \quad (3.14)$$

It is worth noting that this combination strategy is closely related to that used in QBVE. Note that the use of (3.11) is equivalent to using the arithmetic average (mean) of log-probabilities which, in turn, is identical to combining image rankings, as in (3.14). For QBVE the two combination approaches are identical.

3.6 Experimental Evaluation

In this section we report on an extensive evaluation of QBSE. We start by describing the evaluation procedure and the various databases used. This is followed by some preliminary tuning of the parameters of the QBSE system and the analysis of a number of retrieval experiments, that can be broadly divided into two classes. Both compare the performance of QBSE and QBVE, but while the first is performed inside the semantic space, the second studies retrieval performance outside of the latter.

3.6.1 Evaluation Procedure

In all cases, performance is measured with *precision* and *recall*, a classical measure of information retrieval performance [125], which is also widely used by the image retrieval community [136], and one of the metrics adopted by the TRECVID evaluation benchmark. Given a query and the top ‘ N ’ database matches, also called as *scope*, if ‘ r ’ of the retrieved objects are relevant (where relevant means belonging to the class of the query), and the total number of relevant objects in the database is ‘ R ’, then precision is defined as ‘ r/N ’, i.e. the percentage of N which are relevant and recall as ‘ r/R ’, which is the percentage of all relevant images contained in the retrieved set. Precision-recall is commonly summarized by the mean average precision (MAP)[41]. This consists of averaging the precision at the ranks where recall changes, and taking the mean over a set of queries. Because some authors [123] consider the characterization of retrieval performance by curves of *precision-scope* more expressive for image retrieval, we also present results with this measure.

3.6.2 Databases

The evaluation of a QBSE system requires three different databases. The first is a *training database*, used by the semantic labeling system to learn concept probabilities. The second is a *retrieval database* from which images are to be retrieved. The third is a database of *query images*, which do not belong to either

Table 3.1: Retrieval and Query Database

Database	<i>Corel371</i>	<i>Corel15</i>	<i>Flickr18</i>
Semantic Space	Inside	Outside	Outside
Source	Corel CDs	Corel CDs	flickr.com
# Retrieval Images	4500	1200	1440
# Query Images	500	300	360
# Classes	50	15	18

the training or retrieval databases. In the first set of experiments, the training and retrieval databases are identical, and the query images are inside the semantic space. This is the usual evaluation scenario for semantic image retrieval [35, 72, 41]. In the second, designed to evaluate generalization, both query and retrieval databases are outside the semantic space.

Training Database

We relied on *Corel371* dataset as the training database for all experiments. A detailed description of *Corel371* dataset is provided in Appendix A.1.3. 4,500 training images are used to learn the semantic space. Since overall there are 371 concepts, this leads to a 371-dimensional semantic space. With respect to image representation, all images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space. Image observations were derived from 8×8 patches obtained with a sliding window, moved in a raster-scan fashion. A feature transformation was applied to this space by computing the 8×8 discrete cosine transform (DCT) of the three color components of each patch. The parameters of the semantic class mixture hierarchies were learned in the subspace of the resulting 192-dimension feature space composed of the first 21 DCT coefficients from each channel. In all experiments, the SMN associated with each image was computed with these semantic class-conditional distributions.

Retrieval and Query Database

To evaluate retrieval performance we carried out tests on three databases *Corel371*, *Corel15* and *Flickr18*, the details of which are provided in Appendix A.1.3 and Appendix A.1.4.

Inside the Semantic Space Retrieval performance inside the semantic space was evaluated by using *Corel371* as both retrieval and query database. More precisely, the 4500 training images served as the *retrieval database* and the remaining 500 as the *query database*. This experiment relied on clear ground truth regarding the relevance of the retrieved images, based on the theme of the CD to which the query belonged.

Outside the Semantic Space To test performance outside the semantic space, we relied on two additional databases viz *Corel15* and *Flickr18*. For both databases, 20% of randomly selected images served as *query images* and the remaining 80% as the *retrieval database*. 3.1 summarizes the composition of the databases used.

A QBVE system only requires a query and a retrieval database. In all experiments, these were made identical to the query and retrieval databases used by the QBSE system. Since the performance of QBVE does not depend on whether queries are inside or outside the semantic space, this establishes a benchmark for evaluating the generalization of QBSE.

3.6.3 Model Tuning

All parameters of our QBVE system have been previously optimized, as reported in [156]. Here, we concentrate on the QBSE system, reporting on the impact of 1) SMN regularization, and 2) choice of similarity function on the retrieval performance. The parameters resulting from this optimization were used in all subsequent experiments.

Table 3.2: Effect of SMN regularization on the MAP score of QBSE

Ratio	MAP score		
	<i>Corel371</i>	<i>Corel15</i>	<i>Flickr18</i>
100	0.1544	0.1878	0.1447
10	0.1744	0.2030	0.1557
1	0.1833	0.2156	0.1625
0.1	0.1768	0.2175	0.1615
0.01	0.1709	0.2160	0.1594
0.001	0.1683	0.2150	0.1578
0.0001	0.1672	0.2144	0.1569
0.00001	0.1667	0.2141	0.1564

Effect of regularization on QBSE

3.2 presents the MAP obtained with values of $L\pi_0$ (2.18), ranging from 10^{-5} to 100. 3.3 presents the SMN of the *train* query of 3.6, for some of the values of $L\pi_0$. It can be seen that very large values of α force the SMN towards a uniform distribution, e.g. 3.3c, and almost all semantic information is lost. 3.3b shows the SMN regularized with the optimal value of $\pi_0 = 1/L$, where exceedingly low concept probabilities are lower-bounded by the value of 0.001. This regularization is instrumental in avoiding very noisy distance estimates during retrieval.

Effect of the Similarity Function on QBSE

3.3 presents a comparison of the seven similarity functions discussed in the text. It is clear that L^2 distance and histogram intersection do not perform well. All information theoretic measures, KL divergence, symmetric KL divergence and Jensen-Shanon divergence, have superior performance, with an average improvement of 15%. Among these the KL divergence performs the best. The closest competitors to KL divergence are the correlation and normalized correlation metrics. Although, they outperform KL divergence inside the semantic space (*Corel371*), their performance is inferior for databases outside the semantic space (*Flickr18*,

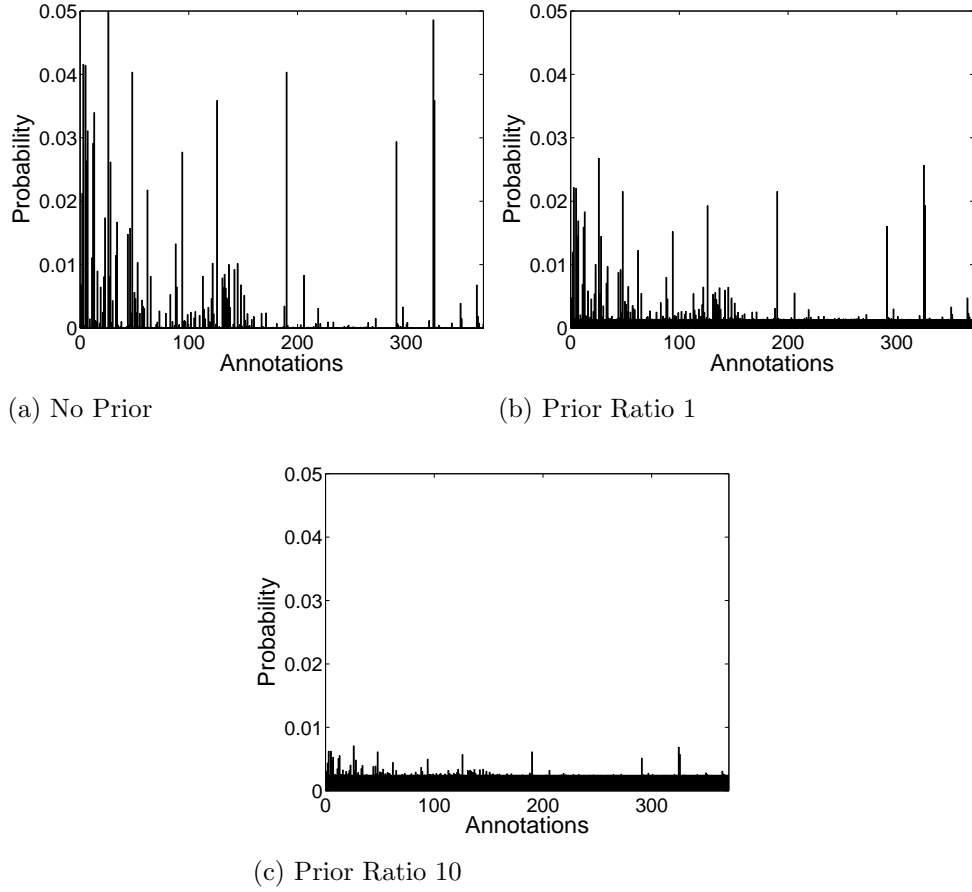


Figure 3.3: SMN of the *train* query of 3.6 as a function of the ratio $\frac{L(\alpha-1)}{n}$ adopted for its regularization.

Table 3.3: Effect of the similarity function on the MAP score of QBSE

Similarity Function	MAP score		
	<i>Corel371</i>	<i>Corel15</i>	<i>Flickr18</i>
KL divergence	0.1768	0.2175	0.1615
Symmetric KLD	0.1733	0.2164	0.1602
Jensen-Shanon	0.1740	0.2158	0.1611
Correlation	0.2108	0.1727	0.1392
Normalized Correlation	0.1938	0.2041	0.1595
L2 distance	0.1461	0.1830	0.1408
Histogram Intersection	0.1692	0.2119	0.1600

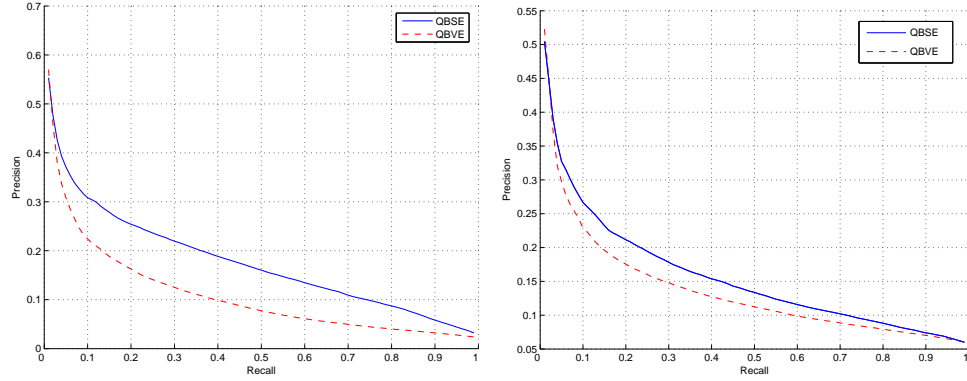


Figure 3.4: Average precision-recall of single-query QBSE and QBVE, Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

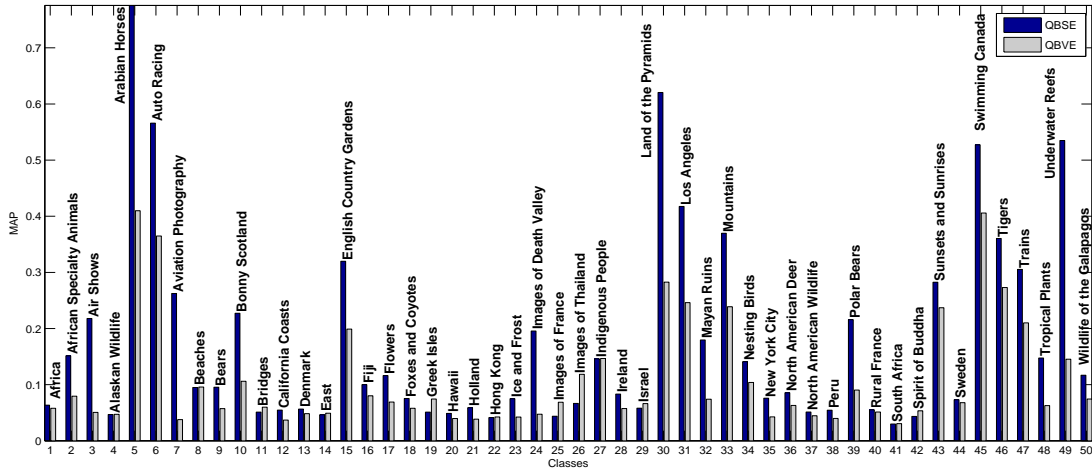


Figure 3.5: MAP scores of QBSE and QBVE across the 50 classes of *Corel371*.

Corel15). This indicates that the KL divergence is likely to have better generalization. While further experiments will be required to reach definitive conclusions, this has led us to adopt the KL divergence in the remaining experiments.

3.6.4 Performance Within the Semantic Space

3.4 (left) presents the precision-recall curves obtained on *Corel371* with QBVE and QBSE. It can be seen that the precision of QBSE is significantly higher than that of QBVE, at most levels of recall. The competitive performance of QBVE at low recall can be explained by the fact that there are always some database

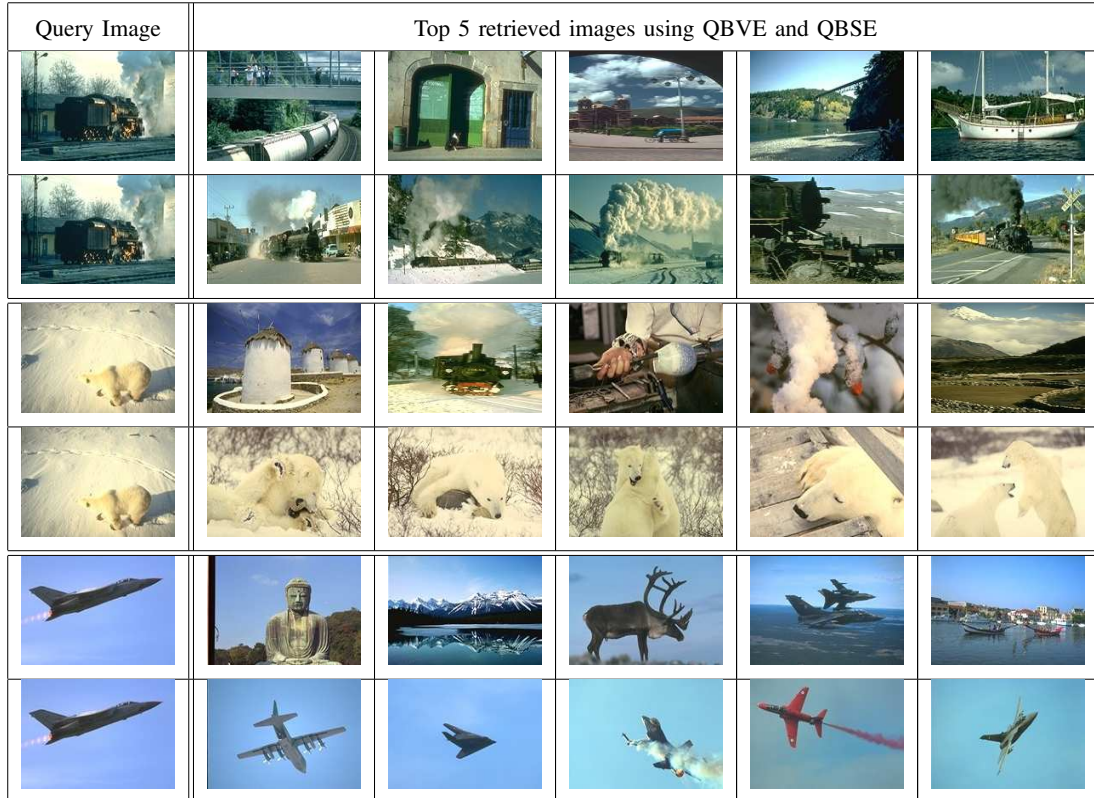


Figure 3.6: Some examples where QBSE performs better than QBVE. The second row of every query shows the images retrieved by QBSE.

images which are visually similar to the query. However, performance decreases much more dramatically than that of QBSE, as recall increases, confirming the better generalization ability of the latter. The MAP scores for QBSE and QBVE are 0.1665 and 0.1094 respectively and the chance MAP performance is 0.0200. 3.5 presents a comparison of the performance on individual classes, showing that QBSE outperforms QBVE in almost all cases.

The advantages of QBSE are also illustrated by 3.6, where we present the results of some queries, under both QBVE and QBSE. Note, for example, that for the query containing *white smoke* and a large area of *dark train*, QBVE tends to retrieve images with *whitish* components, mixed with *dark* components, that have little connection to the *train* theme. Furthermore, the arch-like structure highlighted in 1.1 seems to play a prominent role in visual similarity, since three of the five top matches contain arches. Due to its higher level of abstraction, QBSE

is successfully able to generalize the main semantic concepts of *train*, *smoke* and *sky*, realizing that the white color is an irrelevant attribute to this query (as can be seen in the last column, where an image of *train with black smoke* is successfully retrieved).

3.6.5 Multiple Image Queries

Since the test set of *Corel371* contains 9 to 11 images from each class, it is possible to use anywhere from 1 to 9 images per query. When the number of combinations was prohibitively large (for example, there are close to 13,000 combinations of 5 queries), we randomly sampled a suitable number of queries from the set. 3.7 (left) shows the MAP values for multiple image queries, as a function of query cardinality, under both QBVE and QBSE for *Corel371*. In the case of QBSE, we also compare the three possible query combination strategies: ‘*LKLD*’, ‘*SMN*’, and ‘*KL Combination*’. It is clear that, inside the semantic space, the gains achieved with multiple QBSE queries are unparalleled on the visual domain. In [143], the authors have experimented with multiple query images on a QBVE system. They show that, using two examples, precision increases by around 15% at 10% recall (over single example queries) but no further improvements are observed for three or more images. We have found that, while the MAP of QBSE increases with the number of images, no gain is observed under QBVE. For QBSE, among the various combination methods, combining SMNs yields best results, with a gain of 29.8% over single image queries. ‘*LKLD*’ and ‘*KL Combination*’ exhibit a gain of 17.3% and 26.4% respectively. For QBSE, the increase of precision with query cardinality is experienced at all levels of recall.

3.8 shows the performance of 1-9 image queries for the best and the worst ten classes, sorted according to the gain in MAP score. It is interesting to note that in all of the best 10 classes, single image query performs well above chance, while the opposite holds for the worst 10. This means that moderate performance of a QBSE system can be considerably enhanced by using multiple query images, but this is not a cure for fundamental failures. Overall, the MAP score increases with the number of queries for 76% of the classes. For the classes with unsatisfactory

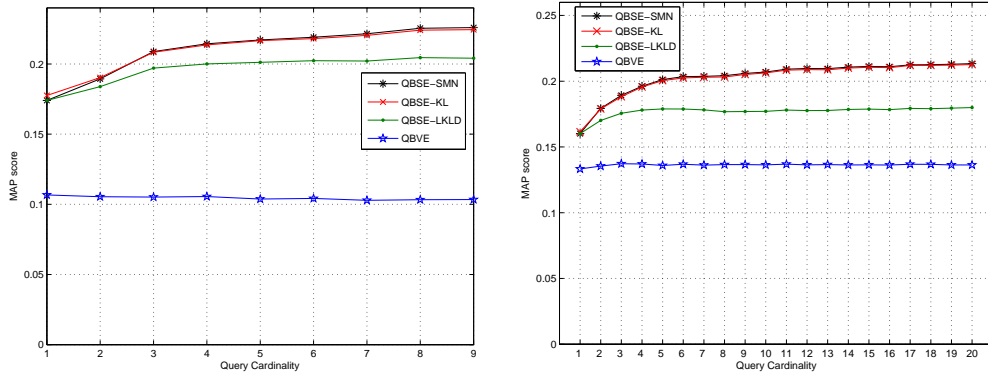


Figure 3.7: MAP as a function of query cardinality for multiple image queries. Comparison of QBSE, with various combination strategies, and QBVE. Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

MAP score, poor performance can be explained by 1) significant inter-concept overlap (e.g., ‘Air Shows’ vs. ‘Aviation Photography’), 2) incongruous concepts that would be difficult even for a human labeler (e.g. ‘Holland’ and ‘Denmark’), or 3) failure to learn semantic homogeneity among the images, e.g. ‘Spirit of Buddha’. Nevertheless, for 86% of the classes QBSE outperforms QBVE by an average MAP score of 0.136. On the remaining QBVE is only marginally better than QBSE, by an average MAP score of 0.016. 3.9 (Left) presents the average precision-recall curves, obtained with the number of image queries that performed best, for QBSE and QBVE on *Corel371*. It is clear that QBSE significantly outperforms QBVE at all levels of recall, the average MAP gain being of 111.73%.

3.6.6 Performance Outside the Semantic Space

3.4 (Right) presents precision-recall curves obtained on *Flickr18*⁵, showing that outside the semantic space single-query QBSE is marginally better than QBVE. When combined with 3.4 (Left), it confirms that, overall, single-query QBSE has better generalization than visual similarity: it is substantially better inside the semantic space, and has slightly better performance outside of it.

⁵For brevity, we only document the results obtained with *Flickr18*, those of *Corel15* were qualitatively similar

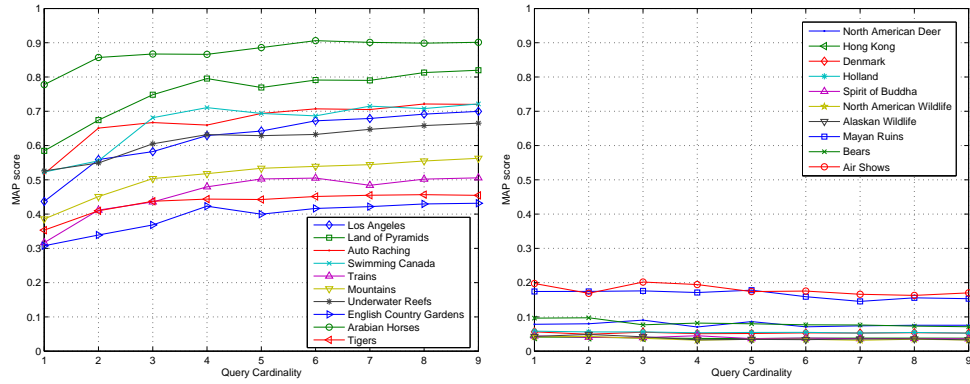


Figure 3.8: Effect of multiple image queries on the MAP score of various classes from *Corel371*. Left: Classes with highest MAP gains, Right: Classes with lowest MAP gains

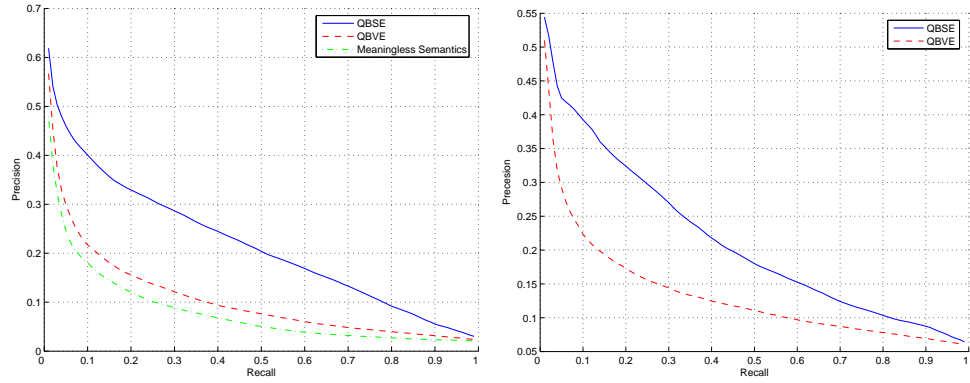


Figure 3.9: Best precision-recall curves achieved with QBSE and QBVE on *Corel371*. Left: Inside the semantic space (*Corel371*), also shown is the performance with meaningless semantic space. Right: Outside the semantic space (*Flickr18*).
































Query Image	Multiple Image Query					
						
						
township						
						
						
Helicopter						

Figure 3.10: Examples of multiple-image QBSE queries. Two queries (for “Township” and “Helicopter”) are shown, each combining two examples. In each case, two top rows presents the single-image QBSE results, while the third presents the combined query.

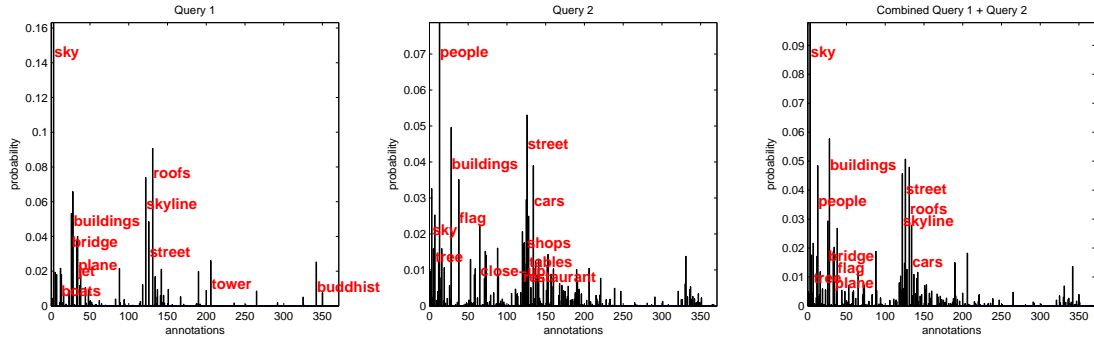


Figure 3.11: SMN of individual and combined queries from class ‘Township’ of 3.10. Left column shows the first query SMN, center the second and, right the combined query SMN.

For multiple image queries we performed experiments with up to 20 images per query (both databases contain 20 test images per class). As was the case for *Corel371*, multiple image queries benefit QBSE substantially but have no advantage for QBVE. This is shown in 3.7 (Right), where we present the MAP score as a function of query cardinality. With respect to the combination strategy, ‘SMN’ once again outperforms ‘KL’(slightly) and ‘LKLD Combination’ (significantly).

An illustration of the benefits of multiple image queries is given in 3.10. The two top rows present query images from the class ‘Township’(Flickr18) and single-query QBSE retrieval results. The third row presents the result of combining the two queries by ‘SMN combination’. It illustrates the wide variability of visual appearance of the images in the ‘Township’ class. While single-image queries fail to express the semantic richness of the class, the combination of the two images allows the QBSE system to expand ‘indoor market scene’ and ‘buildings in open air’ to an ‘open market street’ or even a ‘railway platform’. This is revealed, by the SMN of the combined query, presented in 3.11 (right), which is a semantically richer description of the visual concept ‘Township’, containing concepts (like ‘sky’, ‘people’, ‘street’, ‘skyline’) from both individual query SMNs. The remaining three rows of 3.10 present a similar result for the class ‘Helicopter’ (*Corel15*).

Finally, 3.9 presents the best results obtained with multiple queries under both the QBSE and QBVE paradigms. A similar comparison, using the precision-

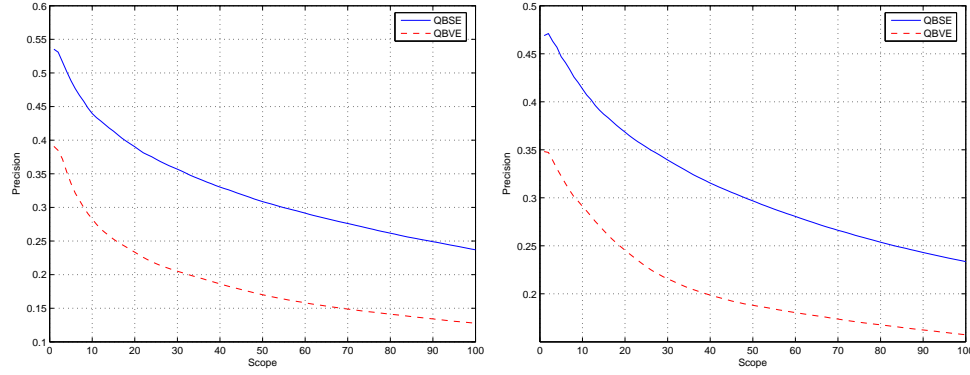


Figure 3.12: Performance of QBSE compared to QBVE, based on precision-scope curve for $N = 1$ to 100, Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

Table 3.4: MAP of QBVE and QBSE on all datasets considered.

Database	Chance	QBVE	QBSE	% increase
<i>Corel371</i>	0.0200	0.1067	0.2259	111.73
<i>Corel15</i>	0.0667	0.2176	0.2980	36.95
<i>Flickr18</i>	0.0556	0.1373	0.2134	55.47

scope curve is shown in 3.12. It is clear that, when multiple image queries are adopted, QBSE significantly outperforms QBVE, even outside the semantic space. 3.4 summarizes the MAP gains of QBSE, over QBVE, for all datasets considered. In the case of *Flickr18* the gain is of 55.47%. Overall, the table emphatically points out that QBSE significantly outperforms QBVE, both inside and outside the semantic space. Since the basic visual representation (DCT features and Gaussian mixtures) is shared by the two approaches, this is strong indication that *there is a benefit* to the use of semantic representations in image retrieval. To further investigate this hypothesis we performed a final experiment, based on QBSE with a semantically meaningless space. Building on the fact that all semantic models are learned by grouping images with a common semantic concept, this was achieved by replicating the QBSE experiments with random image groupings. That is, instead of a semantic space composed of concepts like ‘sky’ (learned from images

containing sky), we created a ‘semantic space’ of nameless concepts learned from random collections of images. 3.9 (left) compares (on *Corel371*) the precision-recall obtained with QBSE on this ‘meaningless semantic space’, with the previous results of QBVE and QBSE. It is clear that, in the absence of semantic structure, QBSE has *very poor* performance, and is *clearly inferior* to QBVE.

3.7 Acknowledgments

The author would like to thank Gustavo Carneiro and Pedro Moreno for helpful discussions and comments.

The text of Chapter 3, in part, is based on the material as it appears in: N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Bridging the Semantic Gap: Query by Semantic Example*’, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007, N. Rasiwasia, P. J. Moreno and N. Vasconcelos, ‘*Query by Semantic Example*’, ACM International Conference on Image and Video Retrieval, LNCS 51-60, Phoenix, 2006, and N. Rasiwasia and N. Vasconcelos, ‘*A Systematic Study of the role of Context on Image Classification*’, IEEE Conference on Image Processing, 1720-1723, San Diego, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.

Chapter 4

Scene Classification with Semantic Representation

In this chapter we introduce the problem of scene classification and present a novel solution based on semantic image representation.

4.1 Introduction

Scene classification is an important problem for computer vision, and has received considerable attention in the recent past. It differs from the conventional object detection/classification, to the extent that a scene is composed of several entities often organized in an unpredictable layout[113]. Images of scenes also differ from images of objects with respect to the distance between the camera and the elements in the image [104]. For a given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations. Frequently, images from two different scene categories are visually similar, e.g., it can be difficult to distinguish between scenes such as “open country” and “mountain” (see Sec. 4.4).

Early efforts at scene classification targeted binary problems, such as distinguishing indoor from outdoor scenes [142], city views from landscape etc. Subsequent research was inspired by the literature on human perception. In [9], it was shown that humans can recognize scenes by considering them in a “holistic” manner, without recognizing individual objects. Recently, it was also found that humans can perform high-level categorization tasks extremely rapidly [144] in the near absence of attention [78]. Drawing inspiration from the perceptual literature, [104] proposes a low dimensional representation of scenes, based on several global properties such as “naturalness”, “openness”, etc. More recently, there has been an effort to solve the problem in greater generality, through design of techniques capable of classifying relatively large number of scene categories [166, 77, 113, 74, 16, 83], and a dataset of 15 categories has been used to compare the performance of various systems[74, 83]. These methods tend to rely on *local region descriptors*, modeling an image as a bag-of-features (BoF, see Section 2.1.1. The space of local region descriptors is then quantized, based on some clustering mechanism, and the mean vectors of these clusters, commonly known

as “visual-words”¹ are chosen as their representatives, thereby yielding the bag-of-words (BoW) representation. This representation is motivated by the time-tested BoW model, widely used in text-retrieval [125]. The analogy between visual-words and text-words is also explored in [130].

Lately, various extensions of this basic BoW model have been proposed [77, 113, 16, 83]. All such methods aim to provide a compact lower dimensional representation using some intermediate characterization on a latent space, commonly known as the intermediate “theme” or “topic” representation [77]. The rationale is that images which share frequently co-occurring visual-words have similar representation in the latent space, even if they have no visual-words in common. This leads to representations robust to the problems of polysemy - a single visual-words may represent different scene content, and synonymy - different visual-words may represent the same content [113]. It also helps to remove the redundancy that may be present in the basic BoW model, and provides a semantically more meaningful image representation. Moreover, a lower dimensional latent space speeds up computation: for example, the time complexity of a Support Vector Machine (SVM) is linear in the dimension of the feature space. Finally, it is unclear that the success of the basic BoW model would scale to very large problems, containing both large image corpuses and a large number of scene categories. In fact, this has been shown not to be the case in text-retrieval, where it is now well established that a flat representation is insufficient for large scale systems, and the use of intermediate latent spaces leads to more robust solutions [58, 14]. However, a direct translation of these methods to computer vision has always incurred a loss in performance, and latent models have not yet been shown to be competitive with the flat BoW representation [83, 74].

In this chapter we propose an alternative solution, based on semantic image representation. Like the latent model approaches we introduce an intermediate space - the semantic space, however, instead of learning the themes in an unsupervised manner from the BoW representation as is done in existing works, the

¹In the literature the terms “textons”, “keypoints”, “visterms”, “visual-terms” or “visterms” have been used with approximately the same meaning, i.e. mean vectors of the clusters in a high-dimensional space.

semantic themes are explicitly defined and the images are casually annotated with respect to their presence. This can *always* be done since, in the absence of “thematic” annotations, the “themes” can be made equal to the class labels, which are always available. The number of semantic themes used defines the dimensionality of the intermediate theme space, henceforth referred to as “semantic space”. Each theme induces a probability density on the space of low-level features, and the image is represented as the vector of posterior theme probabilities. An implementation of this approach is presented and compared to existing algorithms on benchmark datasets. It is shown that the proposed low dimensional representation outperforms the unsupervised latent-space approaches, and achieves performance close to the state of the art, previously only accessible with the flat BoW representation using a much higher dimensional image representation.

The paper is organized as follows. Section 4.2 discusses related work. Section 4.3 presents the approach now proposed, and Section 4.4 an empirical evaluation on benchmark datasets, allowing comparison to previous results.

4.2 Related Work

Low dimensional representations for scene classification have been studied in [77, 113, 16, 83]. On one hand, it is noticed that increasing the size of the codebook improves classification performance [102]. Csurka et al. [27] compare different codebook sizes ranging from 100 to 2500 visual-words, showing that performance degrades monotonically as size decreases. They choose a size of 1000, based on a trade-off between accuracy and speed. Quelhas et al. [113] also experience a monotonic degradation of performance for 3-class classification, and use a codebook of 1000 visual-words. In [74], Lazebnik et al. show that performance increases when codebook size is increased from 200 to 400 visual-words.

On the other hand, there is a strong desire for low dimensional representations, for the benefits elucidated in Section 4.1. This is achieved by resorting to techniques from the text-processing literature, such as Latent Dirichlet Allocation (LDA) [14], Probabilistic Latent Semantic Analysis (pLSA) [58] etc., which

produce an intermediate latent “theme” representation. Fei-Fei et al. [77] motivate the use of intermediate representations, citing the use of “textons” in texture retrieval. They then propose two variations of LDA to generate the intermediate theme representation. In [113], Quelhas et al. use pLSA, to generate the compact representation. They argue that pLSA has the dual ability to generate a robust, low dimensional scene representation, and to automatically capture meaningful scene aspects or themes. pLSA is also used by Bosch et al. in [16]. Another approach to two-level representation based on the Maximization of Mutual Information (MMI) is presented in [83]. However, a steep drop in classification performance is often experienced as a result of dimensionality reduction [83, 74].

4.3 Proposed Approach

A scene classification system can be broadly divided into two modules. The first defines the image representation, while the second delineates the classifier used for decision making. Since the main goal of this thesis is to present a low-dimensional semantic theme representation, we do not dwell on the choice of classifier, simply using an SVM. This is the standard choice in the scene classification literature [166, 102, 27].

Semantic Theme Representation

Under the proposed classification framework, an image is represented by its semantic multinomial (SMN). This is similar in principle to the two level image representations of [77, 113, 16], where an intermediate “theme” space is learned in an unsupervised fashion. In the proposed formulation the semantic space serves as the surrogate for the intermediate “theme” space. As discussed in Chapter 2, learning a semantic space requires a vocabulary of semantic concepts \mathcal{L} and a dataset annotated with respect to \mathcal{L} . These semantic concepts serve the same role as the intermediate “themes” in the existing work [77, 113, 16]. In general, semantic concepts or “themes” are different from image classes. For example, images in the “Street” class of Figure 4.2i contain themes such as “Road”, “Sky”, “People”, or

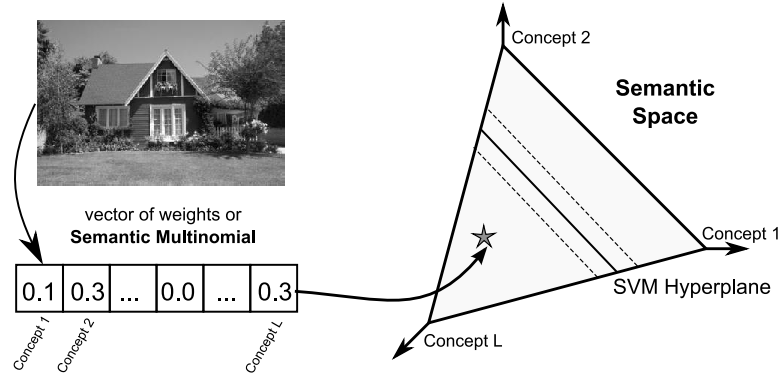


Figure 4.1: The proposed scene classification architecture.

“Cars”. However, current popular scene classification datasets lack such semantic theme annotations and in the absence of these, the set of scene categories $\mathcal{W} = \{1, \dots, K\}$, e.g. “Street”, can serve as a proxy for \mathcal{L} . In this case, each image is only explicitly annotated with one “theme”, even though it may depict multiple: e.g. most images in the “Street” class of Figure 4.2i also depict “Buildings”. We refer to this limited type of scene labeling as *casual annotation*. This is the annotation mode for all results reported in this paper, to enable comparison to previous scene classification work. We will see that supervised learning of the intermediate theme space with casual annotations can be far superior to unsupervised learning of a latent theme space, as previously proposed [77].

Scene Classification

Due to the limited information contained in casual annotations, images cannot be simply represented by the caption vectors \mathbf{c}_i . In fact, \mathbf{c}_i is only available for training images, and $\mathbf{c}_{i,j} = 0$ does not mean that the i^{th} image does not contain the j^{th} theme, simply that it was not annotated with it. Instead, the proposed classification system represents images by vectors of theme frequency, or counts. In this way, an image can be associated with multiple themes, even when there are no multiple associations in the labels used for training. As shown in Figure 4.1, the scene classifier (e.g. SVM) then operates on this feature space.

4.4 Experimental evaluation

We now present an empirical evaluation of the model as a low dimensional semantic theme representation for two publicly available datasets, comparing performance with [83, 16, 77, 74]. We also present a study of classification accuracy as a function of semantic space dimensions.

4.4.1 Datasets

Scene classification results are presented on two public datasets: 1) Natural15 [74] and 2) Corel50 photos, used in [21] for image annotation comprising of 50 scene categories. The details of these datasets are discussed in Appendix A.1.1 and Appendix A.1.3 respectively. The use of the Natural15 dataset allow us to directly compare with the existing results on scene classification. In particular, we show a comparison of our results using low-dimensional representation with those of [83, 74, 77, 16]. The Corel50 dataset has 100 high resolution images per category, which we resize to an average of 180×120 pixels. To the best of our knowledge, this is the database with maximum number of scene categories so far studied in the literature (viz. 50). Since the dimension of our semantic theme representation directly depends on the number of scene categories (see Sec. 4.3), this dataset enables the study of the effects of dimensionality as the number of categories grows.

4.4.2 Experimental Protocol

At the low level, images are represented as bags of 8×8 vectors of discrete cosine transform (DCT) coefficients sampled on a uniform grid. The Corel50 dataset consists of color images which are converted from RGB to YcrCb colorspace². The Natural15, consist of grayscale images hence no such conversion is required. Semantic theme densities are learned on a 36(out of 64) / 64(out of 192) dimensional subspace of the DCT coefficients for Natural15 and Corel50 dataset respectively,

²We also conducted experiments with the CIE lab colorspace and the results are almost similar.

with each theme modeled as a mixture of 128 Gaussian components. The images at the semantic theme level are represented by 15 (50) dimensional theme vectors for Natural15 (Corel50). Later on, we also show that not all 50 themes are equally informative on Corel50. 100 (90) images per scene are used to learn the theme density for Natural15 (Corel50), and the rest of the images are used as the test set. All experiments on Natural15 are repeated 5 times with different randomly selected train and test images. For Corel50 dataset, we use the same training and test images as used in [21, 35]. A multi-class SVM using one-vs-all strategy with Gaussian kernel is used for classification, with the parameters obtained by 3-fold cross validation.

4.4.3 Results

We start by studying scene classification accuracy.

Scene classification

Figure 4.2 shows an example from each of the fifteen scene categories of Natural15, along with their semantic theme representation. All images shown are actually classified correctly by the classifier. Two interesting observations can be made: 1) semantic theme vectors *do capture* the different semantic meanings of the images, hence correlating well with human perception. For example, the theme vector shown for the scene from the category “Forest” in Figure 4.2n, has large weights for themes such as “*forest*”, “*mountain*” and “*open-country*”, which are suitable themes for the scene, and 2) in many examples (viz. Figure 4.2(d)-(f),(h),(i)), even though the semantic theme corresponding to the same semantic scene category does not have the highest probability, the scene is still classified correctly. For example in Figure 4.2i, in spite of the “*street*” theme having much lower probability than “*tall-building*”, “*inside-city*”, “*highway*”, the image is classified as belonging to the “Street” category. This is a direct consequence of the classifier learning *associations* between themes, despite the casual nature of the annotations. Figure 4.4 presents some of the misclassified images from the worst performing scene categories, along with the scene category they are classified into.

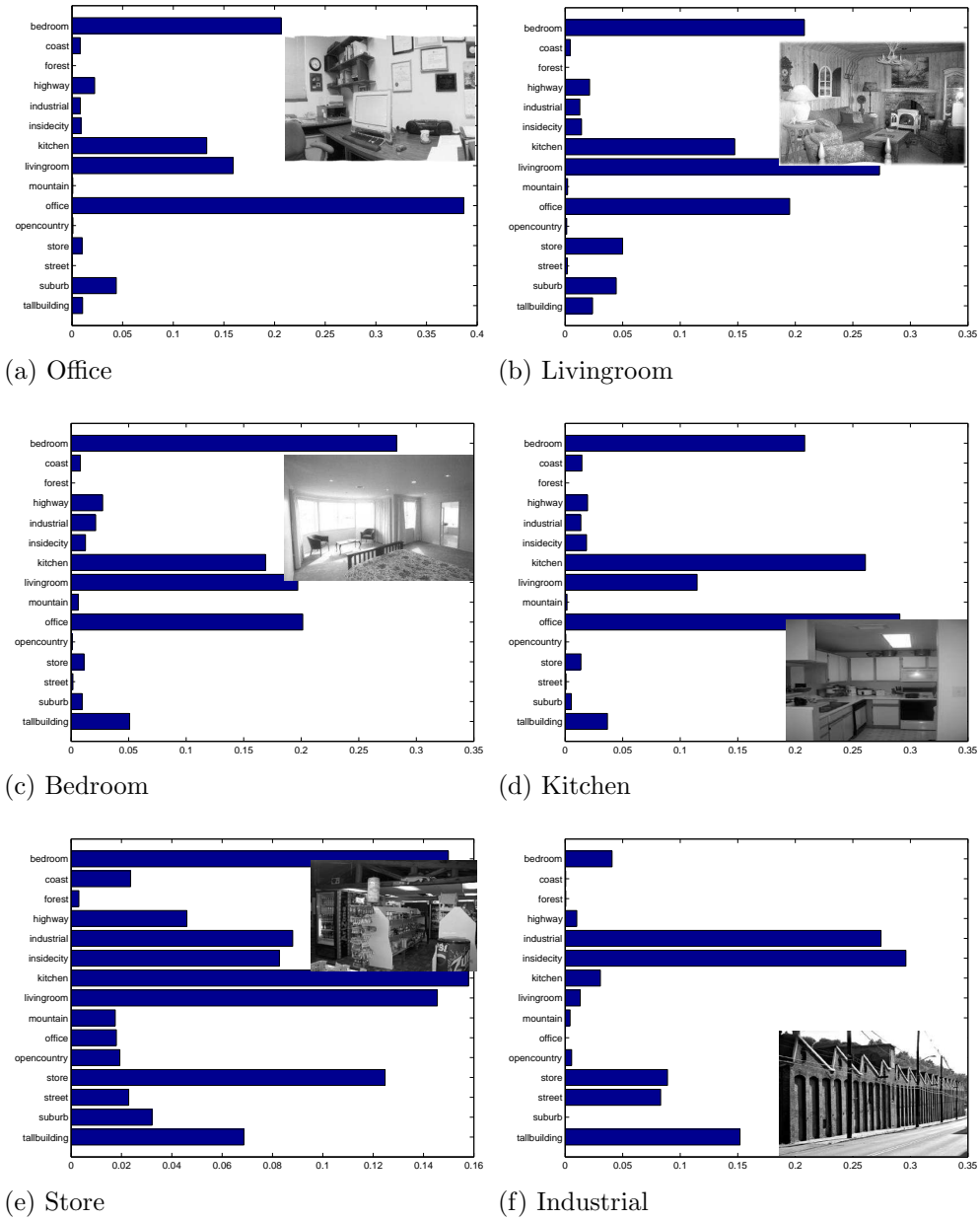


Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories.

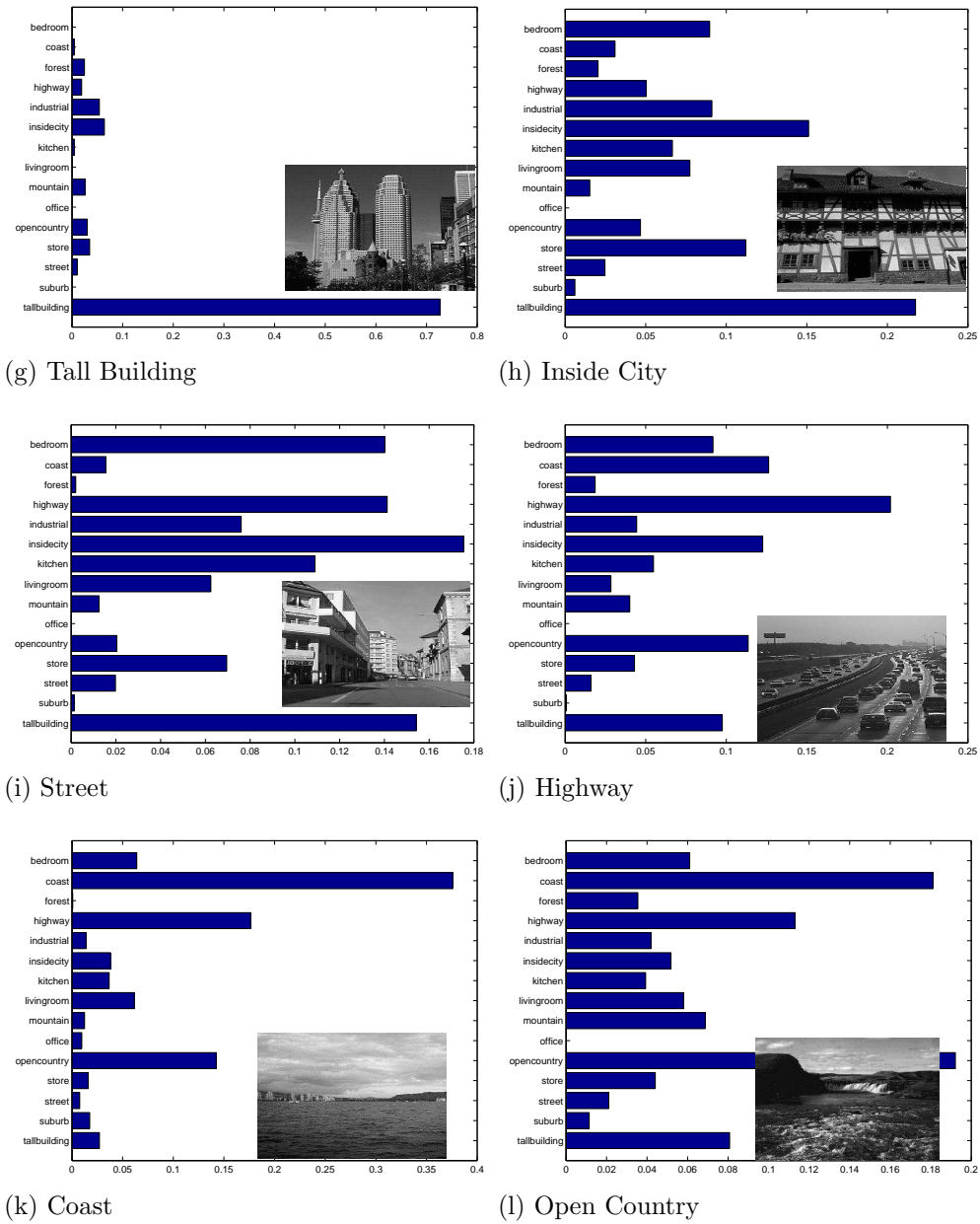


Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories.
(continued)

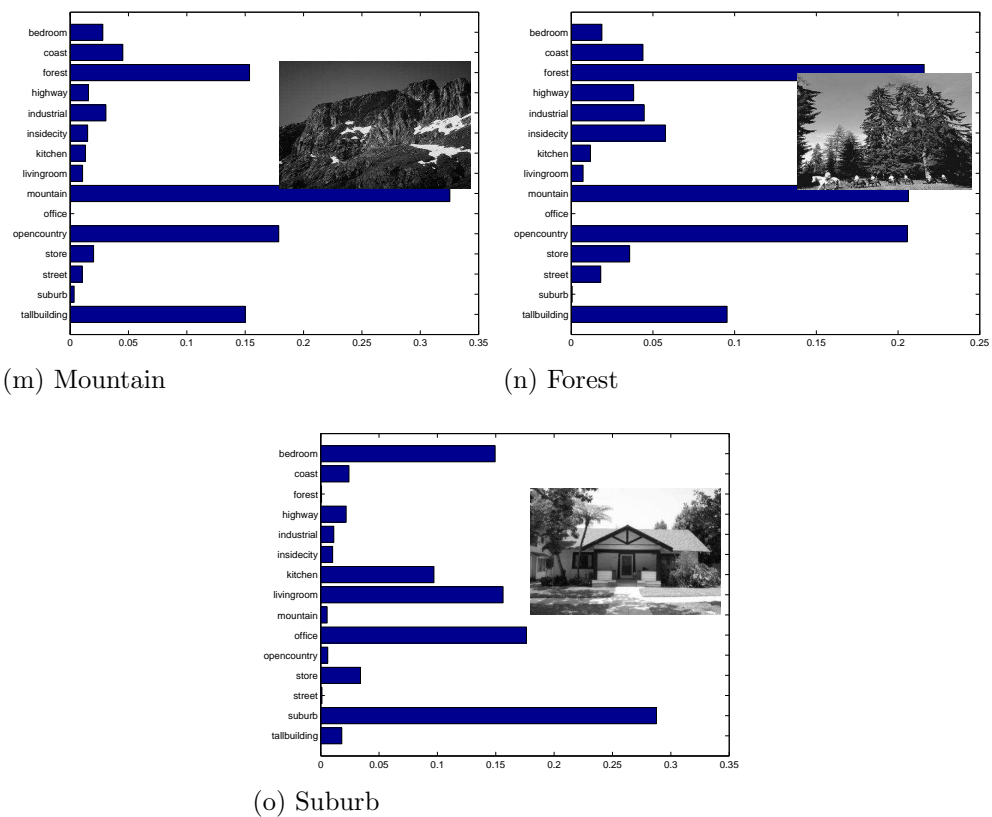


Figure 4.2: Theme vectors from each of the scenes of fifteen scene categories.
(continued)

	Office	Livingroom	Bedroom	Kitchen	Store	Industrial	TallBuilding	InsideCity	Street	Highway	Coast	Opencountry	Mountain	Forest	Suburb
Office	.96	.01	.02	.01	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00
Livingroom	.05	.55	.19	.05	.01	.00	.02	.03	.08	.00	.01	.01	.00	.00	.01
Bedroom	.09	.25	.36	.14	.03	.03	.01	.02	.04	.01	.00	.00	.02	.00	.00
Kitchen	.07	.07	.04	.66	.08	.05	.00	.02	.01	.00	.00	.00	.00	.00	.00
Store	.00	.02	.01	.07	.80	.08	.00	.00	.01	.00	.00	.00	.00	.00	.00
Industrial	.00	.00	.05	.04	.12	.69	.03	.01	.02	.00	.00	.01	.00	.01	.00
TallBuilding	.00	.02	.02	.00	.01	.04	.71	.05	.09	.02	.00	.00	.00	.01	.03
InsideCity	.00	.03	.01	.02	.01	.02	.06	.74	.07	.01	.00	.00	.00	.00	.01
Street	.00	.03	.03	.01	.02	.02	.10	.06	.66	.06	.00	.01	.00	.00	.02
Highway	.00	.01	.01	.00	.01	.02	.01	.02	.02	.78	.05	.06	.02	.00	.00
Coast	.00	.00	.00	.00	.00	.02	.00	.01	.00	.04	.77	.13	.02	.00	.00
Opencountry	.00	.00	.01	.00	.01	.00	.00	.00	.01	.04	.10	.66	.08	.08	.00
Mountain	.01	.00	.01	.00	.01	.01	.01	.00	.01	.01	.04	.10	.71	.07	.01
Forest	.00	.00	.00	.00	.04	.00	.01	.00	.01	.01	.00	.05	.06	.82	.01
Suburb	.00	.00	.00	.00	.00	.00	.02	.01	.00	.00	.00	.00	.00	.01	.96

Figure 4.3: Confusion Table for our method using 100 training image and rest as test examples from each category of Natural15. The average performance is $72.2\% \pm 0.2$

The confusion table for Natural15 is shown in Figure 4.3. The average classification accuracy, over all categories is $72.2 \pm 0.2\%$. As was experienced by [74], there is confusion between indoor categories such as “Bedroom”, “Livingroom” and “Kitchen” and outdoor categories like “Opencountry” and “Mountain”. In fact close to 25% of images from the category “Bedroom” were classified as “Livingroom”. On Corel50, the classification accuracy stands at 56.8%, the chance classification accuracy being 2%. Figure 4.5 shows some of the images from various scene categories of Corel50 dataset. Also shown in Figure 4.6 is the theme vector for the image of Figure 4.5(a).

Comparison with existing work

4.1 compares the classification accuracy of the proposed method on Natural15, using 15 dimensional theme vectors, with the existing results in the literature. It is evident that when compared to the MMI based dimensionality reduction of Liu et al. [83], which achieves a rate of 63.32% using a 20 dimensional space, the method performs substantially better, achieving a rate of 72.2% on an even lower dimensional space of 15 themes. The performance is equal to that of Lazebnik et al. [74]³, who represent images as the basic BoW model, using 200 visual-words. A similar comparison on the thirteen subcategories of the dataset used in [77, 16], is presented in 4.1. Again, the proposed low-dimensional theme vector based representation performs close to the best results in the literature, with a much lower dimensional space. This dataset also shows that the proposed method substantially outperforms the latent-space method of Fei-Fei et al. [77], and achieves equivalent performance the latent-space method of Bosch et al. [16] with roughly half of its dimensionality.

Informative semantic themes

In all the experiments conducted above, scene categories served as a proxy for the intermediate themes. This is a practical approach to scene classification where the images are devoid of other annotations. However, it might seem that the extension of the current framework to very large-scale problems involving thousands of categories, will annul the benefits gained by the proposed representation, as the dimension of the semantic space would grow with the number of categories. The effects of varying the dimensions of the semantic space on the classification accuracy is studied, on Corel50 dataset. Semantic spaces of k dimensions were produced by ordering the semantic themes by the variance of their posterior probabilities, and selecting the k of largest variance (for k ranging from 2 to 50). Clas-

³Note that the best results on this dataset, are obtained by incorporating spatial information, and representing images as histograms at different spatial resolution, with Spatial Pyramid Matching [74]. The accuracy is 81.1%, with a 4200 dimensional feature space. Multi-resolution semantic representations would also be possible with the proposed method, as well as the incorporation of spatial information, but these extensions are beyond the scope of the current discussion.

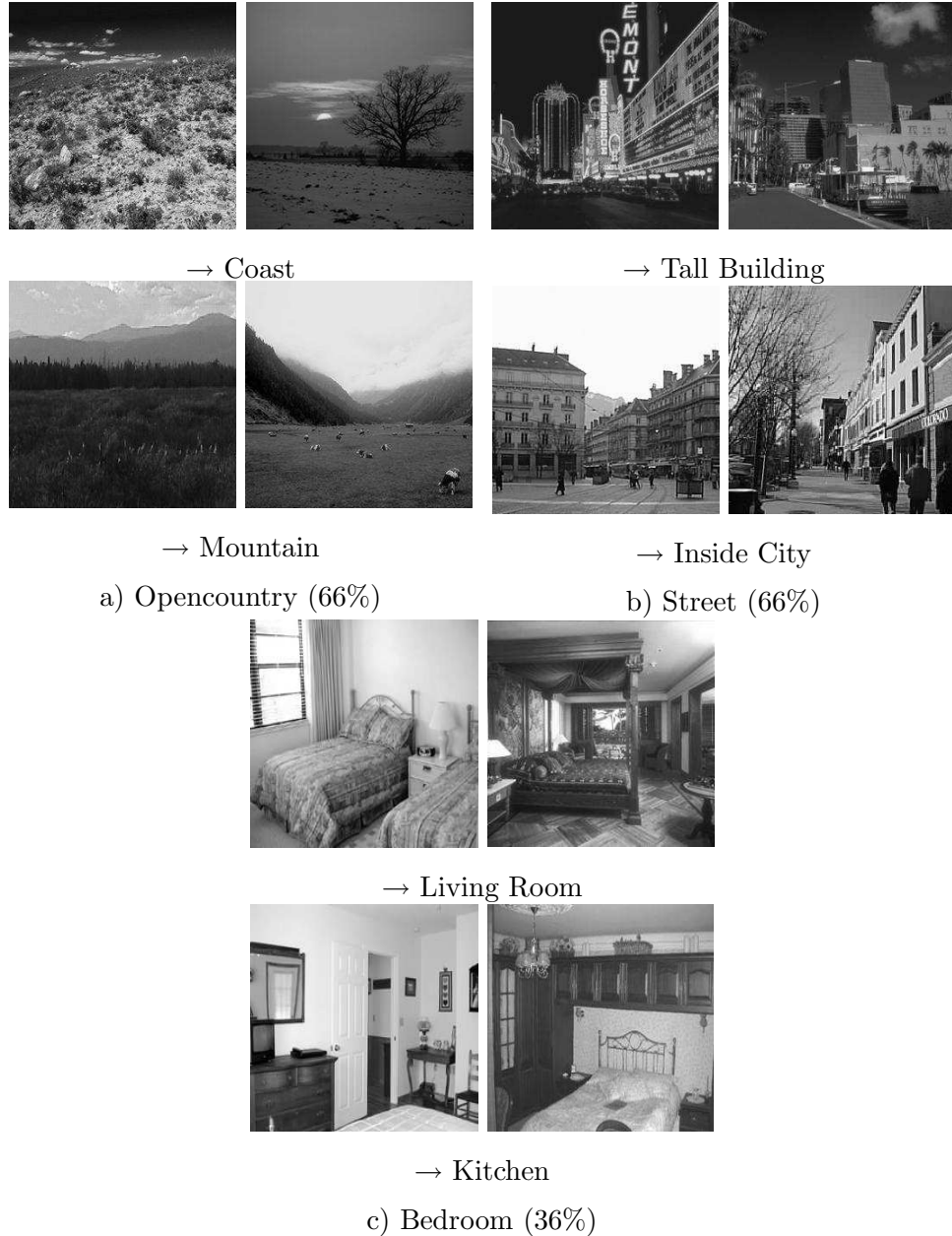


Figure 4.4: Some images from worst performing scene categories in Natural15. (→) implies the scene category the image is classified into.

sification was performed on each of these resulting spaces and Figure 4.7 presents the performance as a function of the dimension. It can be observed that not all of the 50 dimensions are equally informative, as moving from 40 to 50 dimensions increases performance by only 3.8% (a relative gain of 6.7%). This can be explained



Figure 4.5: Some images from the Core50 dataset. (\rightarrow) implies the scene category the image is classified into. (a) and (b) show two examples of correctly classified images, (c) and (d) two reasonably misclassified images and (e) and (f) shows two examples of error.

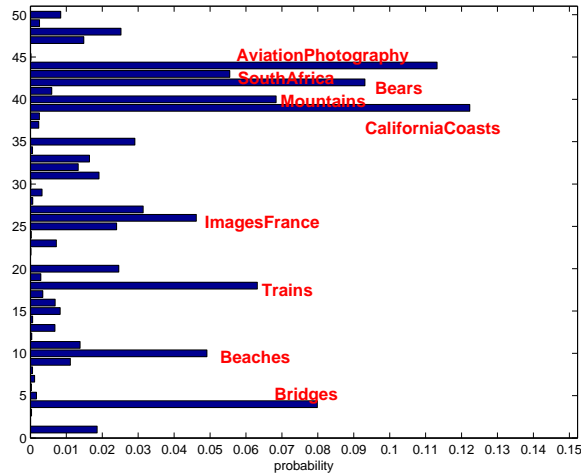


Figure 4.6: The theme vector for the image in Figure 4.5(a).

by the plot of variance of the posterior probabilities for the 50 themes (in the same figure). For very large scale problems, where most of the variance is expected to be captured by a subset of the features, the correlation of classification performance with the variance of the themes indicates that the number of informative themes

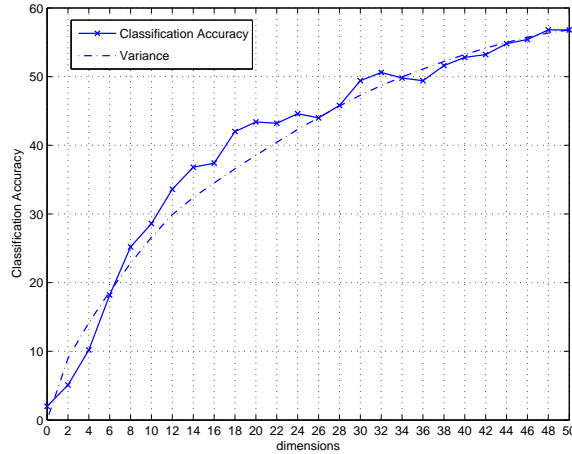


Figure 4.7: Classification performance as a function of the semantic space dimensions. Also shown, is the growth of the variance of the semantic themes, scaled appropriately.

Table 4.1: Classification Result for 15 scene categories.

Method	Dimensions	Classification Accuracy
<i>Our method</i>	15	72.2 ± 0.2
<i>Liu et al. [83]</i>	20	63.32
<i>Liu et al. [83]</i>	200	75.16
<i>Lazebnik et al. [74]</i>	200	72.2 ± 0.6

would grow sub-linearly as the number of scene categories is increased. It is unclear that this type of behavior will hold for the flat BoW representations. In the works previously presented in the literature, the codebook has *linear* size on the number of classes.

The results presented above allow a number of conclusions. While low dimensional semantic representations are desirable for the reasons discussed in Section 4.1, previous approaches based on latent-space models have failed to match the performance of the flat BoW model, which has high dimensionality. We have shown that this is indeed possible, with methods that have much lower complexity than the latent-space approaches previously proposed, but make better use of

Table 4.2: Classification Result for 13 scene category subset.

Method	Dimensions	Classification Accuracy
<i>Our method</i>	13	72.7 ± 0.3
<i>Bosch et al. [16]</i>	25	73.4
<i>Fei-Fei et al. [77]</i>	40	65.2
<i>Lazebnik et al. [74]</i>	200	74.7

the available labeling information. We have also shown that the proposed method extracts meaningful semantic image descriptors, despite the casual nature of the training annotations, and is able to learn co-occurrences of semantic themes without explicit training for these. Finally a study of the effect of dimensionality on the classification performance was presented, and indicated that the dimensionality would grow sub-linearly with the number of scene categories. This could be a significant advantage over the flat BoW model which, although successful for the limited datasets in current use, will likely not scale well when the class vocabulary increases.

4.5 Acknowledgments

The text of Chapter 4, in part, is based on the material as it appears in: Rasiwasia, N., Vasconcelos, N. “Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, June 2008 The dissertation author was a primary researcher and an author of the cited material.

Chapter 5

Cross Modal Multimedia Retrieval

5.1 Introduction

Over the last decade there has been a massive explosion of multimedia content on the web. This explosion has not been matched by an equivalent increase in the sophistication of multimedia content modeling technology. Today, the prevailing tools for searching multimedia repositories are still *uni-modal* in nature. Text repositories are searched with text queries, image databases with image queries, and so forth. To address this problem, the academic community has devoted itself to the design of models that can account for *multi-modal* data, i.e. data with multiple content modalities. Recently, there has been a surge of interest in multi-modal modeling, representation, and retrieval [106, 148, 132, 138, 28, 60, 31]. Multi-modal retrieval relies on queries combining multiple content modalities (*e.g.* the images and sound of a music video-clip) to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). These efforts have, in part, been spurred by a variety of large-scale research and evaluation experiments, such as TRECVID [132] and ImageCLEF [106, 148], involving datasets that span multiple data modalities. However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario. Typically, the different modalities are fused into a representation that does not allow individual access to any of them, *e.g.* some form of dimensionality reduction of a large feature vector that concatenates measurements from images and text. Classical uni-modal techniques are then applied to the low-dimensional representation. This limits the applicability of the resulting multimedia models and retrieval systems.

An important requirement for further progress in these areas is the development of sophisticated joint models for multiple content modalities. In this chapter, we consider a richer interaction paradigm, which is denoted *cross-modal* retrieval. The goal is to build multi-modal content models that enable interactivity with content *across* modalities. Such models can then be used to design *cross-modal retrieval systems*, where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.*, the best accompanying audio-track). This form of retrieval can be seen as a generalization of current content labeling systems, where one dominant modality is augmented with simple information from

another, which can be subsequently searched. Examples include keyword-based image [4, 97, 21] and song [151, 149, 89, 36] retrieval systems. One property of cross-modal retrieval is that, by definition, it requires *representations that generalize across content modalities*. This implies the ability to establish cross-modal links between the attributes (of different modalities) characteristic of each document, or document class. Detecting these links requires much deeper content understanding than the classical matching of 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 *abstract* that the word “rose” matches the visual attribute “red blob”. This is much closer to what humans do than simple color or word matching. Hence, cross-modal retrieval is a better context than uni-modal retrieval for the study of fundamental hypotheses on multimedia modeling.

We exploit this property to study two hypotheses on the joint modeling of images and text. The first, denoted the *correlation hypothesis*, is that explicit modeling of low-level correlations between the different modalities is of importance for the success of the joint models. The second, denoted the *abstraction hypothesis*, is that the modeling benefits from semantic abstraction, *i.e.*, the representation of images and text in terms of semantic (rather than low-level) descriptors. These hypotheses are partly motivated by previous evidence that correlation, *e.g.*, correlation analysis on fMRI [55], and abstraction, *e.g.*, hierarchical topic models for text clustering [14] or semantic representations for image retrieval (see Chapter 3), 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. Both semantic matching and semantic correlation matching build upon the proposed semantic image representation (see Chapter 2).

The hypotheses are tested by measuring the retrieval performance of these models on two reciprocal cross-modal retrieval tasks: 1) the retrieval of text documents in response to a query image, and 2) the retrieval of images in response

to a query text. These are basic cross-modal retrieval problems, central to many applications of practical interest, such as finding pictures that effectively illustrate a given text (*e.g.*, to illustrate a page of a story book), finding the texts that best match a given picture (*e.g.*, a set of vacation accounts about a given landmark), or searching using a combination of text and images. Model performance on these tasks is evaluated with two datasets: TVGraz [66] and a novel dataset based on Wikipedia’s featured articles. These experiments show independent benefits to both correlation modeling and abstraction. In particular, best results are obtained by a model that accounts for both low-level correlations — by performing a kernel canonical correlation analysis (KCCA) [127, 163] — and semantic abstraction — by projecting images and texts into a common semantic space (see Chapter 2) designed with logistic regression. This suggests that the abstraction and correlation hypotheses are complementary, each improving the modeling in a different manner. Individually, the gains of abstraction are larger than those of correlation modeling.

This chapter is organized as follows. Section 5.2 discusses previous work in multi-modal and cross-modal multimedia modeling. Section 5.3 presents a mathematical formulation for cross-modal modeling and discusses the two fundamental hypotheses analyzed in this work. Section 5.4 introduces the models underlying correlation, semantic, and semantic correlation matching. Section 5.5 discusses the experimental setup used to evaluate the hypotheses. Model validation and parameter tuning are detailed in Section 5.6. The hypotheses are finally tested in Section 5.7.

5.2 Previous Work

The problems of image and text retrieval have been the subject of extensive research in the fields of information retrieval, computer vision, and multimedia [28, 133, 132, 106, 93]. In all these areas, the emphasis has been on *uni-modal* approaches, where query and retrieved documents share a single modality [125, 124, 156, 28, 133]. For example, in [124], a query text and in [156], 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. For example, the existence of a well known *semantic gap*, (see Chapter 1) between current image representations and those adopted by humans, severely limits the performance of uni-modal image retrieval systems [133](see Chapter 3).

In general, successful retrieval from large-scale image collections requires that the latter be augmented with text metadata provided by human annotators. These manual annotations are typically in the form of a few keywords, a small caption, or a brief image description [106, 148, 132]. When this metadata is available, the retrieval operation tends to be uni-modal and ignore the images — the text metadata of the query image is simply matched to the text metadata available for images in the database. Because manual image labeling is labor-intensive, recent research has addressed the problem of automatic image labeling¹ [21, 63, 41, 73, 96, 4]. As we saw in Chapter 2, rather than labeling images with a small set of most relevant semantic concepts, images can be represented as a weighted combination of all concepts in the vocabulary, by projecting them into a *semantic space*, where each dimension is a semantic concept. Semantic space was used for uni-modal image retrieval in Chapter 3, which enabled retrieval of images using *semantic similarity* — by combining the semantic space with a suitable similarity function.

In parallel, advances have been reported in the area of *multi-modal* retrieval systems [106, 148, 132, 138, 28, 60, 31]. 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 [171, 108, 37], or by learning different models for different modalities and fusing their predictions [168, 69]. One popular approach is to concatenate features from different modalities into a common vector and rely on unsupervised structure discovery algorithms, such as latent semantic analysis (LSA), to find statistical patterns that span the different modalities. A good overview of these methods is given in [37], which also discusses the combination of uni-modal and

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

multi-modal retrieval systems. Multi-modal integration has also been applied to retrieval tasks including audio-visual content [99, 44]. In general, the inability to access each data modality individually (after the fusion of modalities) limits the applicability of these systems to cross-modal retrieval.

Recently, there has been progress towards multi-modal systems that do not suffer from this limitation. These include retrieval methods for corpora of images and text [31], images and audio [178, 76], text and audio [131], or images, text, and audio [175, 178, 182, 181, 176]. One popular approach is to rely on graph-based manifold learning techniques [175, 178, 182, 181, 176]. 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 methods in this class is the lack of out-of-sample generalization. Since there is no computationally efficient way to project the query into the manifold, queries are restricted to the training set used to learn the latter. Hence, all unseen queries must be mapped to their nearest neighbors in this training set, defeating the purpose of manifold learning. An alternative solution is to learn correlations between different modalities [76, 178, 164]. For example, [76] 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 [164] 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 [55].

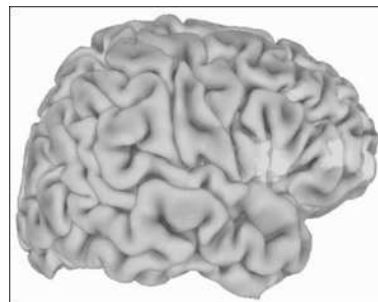
Despite these advances in multi-modal modeling, current approaches tend to rely on a limited textual representation, in the form of keywords, captions, or small text snippets. We refer to all of these as forms of *light annotation*. This is at odds with the ongoing explosion of multimedia content on the web, where it is now possible to collect large sets of extensively annotated data. Examples



(a)

Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in "Time" magazine as saying, "The new administration should have been given a chance to confer with the various groups interested in change." Black hotel owner A. G. Gaston stated, "I regret the absence of continued communication between white and Negro leadership in our city." A white Jesuit priest assisting in desegregation negotiations attested, "These demonstrations are poorly timed and misdirected." Protest organizers knew they would meet with violence from the Birmingham Police Department but chose a confrontational approach to get the attention of the federal government. Reverend Wyatt Tee Walker, one of the SCLC founders and the executive director from 1960-1964, 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 to attract the media, and in turn induce national sympathy and attention to the everyday segregated circumstance of a person living in the Deep South," Walker said. He headed the planning of what he called Project C, which stood for "confrontation". According to this historians Isserman and Kazin, the demands on the city authorities were straightforward: desegregate the economic life of Birmingham its restaurants, hotels, public toilets, and the unwritten policy of hiring blacks for menial jobs only Maurice Isserman and Michael Kazin, *America Divided: The Civil War of the 1960s*, (Oxford, 2008), p.90. (...)

Home - Courses - Brain and Cognitive Sciences - A Clinical Approach to the Human Brain 9.22J / HST.422J A Clinical Approach to the Human Brain Fall 2006 Activity in the highlighted areas in the prefrontal cortex may affect the level of dopamine in the mid-brain, in a finding that has implications for schizophrenia. (Image courtesy of the National Institutes of Mental Health.) Course Highlights This course features summaries of each class in the lecture notes section, as well as an extensive set of readings. Course Description This course is designed to provide an understanding of how the human brain works in health and disease, and is intended for both the Brain and Cognitive Sciences major and the non-Brain and Cognitive Sciences major. Knowledge of how the human brain works is important for all citizens, and the lessons to be learned have enormous implications for public policy makers and educators. The course will cover the regional anatomy of the brain and provide an introduction to the cellular function of neurons, synapses and neurotransmitters. Commonly used drugs that alter brain function can be understood through a knowledge of neurotransmitters. Along similar lines, common diseases that illustrate normal brain function will be discussed. Experimental animal studies that reveal how the brain works will be reviewed. Throughout the seminar we will discuss clinical cases from Dr. Byrne's experience that illustrate brain function; in addition, articles from the scientific literature will be discussed in each class. (...)



(b)

Figure 5.1: Two examples of image-text pairs: (a) section from the Wikipedia article on the Birmingham campaign ("History" category), (b) part of a Cognitive Science class syllabus from the TVGraz dataset ("Brain" category).

include news archives, blog posts, or Wikipedia pages, where pictures are related to complete text articles, not just a few keywords. We refer to these datasets as *richly annotated*. While potentially more informative, rich annotation establishes a much more nuanced connection between images and text than that of *light annotation*. Indeed, keywords usually are explicit image labels and, therefore, clearly relate to it, while many of the words in rich text may be unrelated to the image used to illustrate it. For example, Figure 5.1a shows a section of the Wikipedia article on the “Birmingham campaign”, along with the associated image. Notice that, although related to the text, the image is clearly not representative of all the words in the article. The same is true for the web-page in Figure 5.1b, from the TVGraz dataset [66] (see Appendix A for more details on both Wikipedia and TVGraz datasets). This is a course syllabus that, beyond the pictured brain, includes course information and other unrelated matters. A major long-term goal of modeling richly annotated data is to recover this *latent* relationship between the text and image components of a document, and exploit it in benefit of practical applications.

5.3 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.

5.3.1 The problem

We consider the problem of information retrieval from a database $\mathcal{B} = \{D_1, \dots, D_{|\mathcal{B}|}\}$ of *documents* comprising *image* and *text* components. In practice, these documents can be quite diverse: from documents where a single text is complemented by one or more images (*e.g.*, a newspaper article) to documents containing multiple pictures and text sections (*e.g.*, a Wikipedia page). For simplicity, we consider the case where each document consists of a single *image* and its

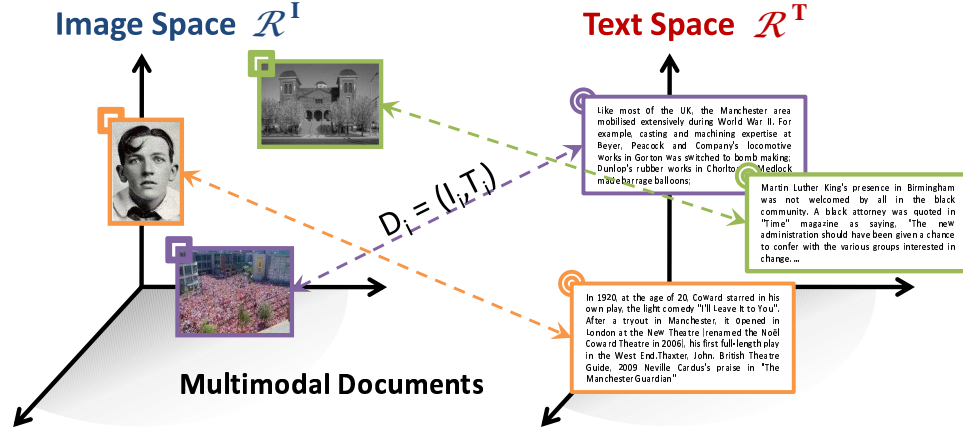


Figure 5.2: Each document (D_i) consists of an *image* (I_i) and accompanying *text* (T_i), *i.e.*, $D_i = (I_i, T_i)$, which are represented as vectors in feature spaces \mathcal{R}^I and \mathcal{R}^T , respectively. Documents establish a one-to-one mapping between points in \mathcal{R}^I and \mathcal{R}^T .

accompanying *text*, *i.e.*, $D_i = (I_i, T_i)$. Images and text are represented as vectors in feature spaces \mathcal{R}^I and \mathcal{R}^T respectively², as illustrated in Figure 5.2, documents establish a one-to-one mapping between points in \mathcal{R}^I and \mathcal{R}^T . Given a text (image) query $T_q \in \mathcal{R}^T$ ($I_q \in \mathcal{R}^I$), the goal of *cross-modal retrieval* is to return the closest match in the image (text) space \mathcal{R}^I (\mathcal{R}^T).

5.3.2 Multi-modal modeling

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

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

be an invertible mapping between the two spaces. Given a query T_q in \mathcal{R}^T , it suffices to find the nearest neighbor to $\mathcal{M}(T_q)$ in \mathcal{R}^I . Similarly, given a query I_q in \mathcal{R}^I , it suffices to find the nearest neighbor to $\mathcal{M}^{-1}(I_q)$ in \mathcal{R}^T . In this case,

²Note that, in this chapter we deviate from the standard representation of an image (adopted in this work) as a bag of N feature vectors, $\mathcal{I} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathcal{X}$, to one where an image is represented as a vector in \mathcal{R}^I . The motivation is to maintain a simple and consistent representation across all different modalities. See Section 2.1.1 for a brief description on mapping images from \mathcal{X}^N to \mathcal{R}^I

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 \mathfrak{R}^I and \mathfrak{R}^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. First, consider learning invertible mappings

$$\mathcal{M}_I : \mathfrak{R}^I \rightarrow \mathcal{V}^I \quad \mathcal{M}_T : \mathfrak{R}^T \rightarrow \mathcal{V}^T$$

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

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

between these two spaces. In this case, given a text query T_q in \mathfrak{R}^T , cross-modal retrieval reduces to finding the nearest neighbor of

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

in \mathfrak{R}^I . Similarly, given an image query I_q in \mathfrak{R}^I , the goal is to find the nearest neighbor of

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

in \mathfrak{R}^T . This formulation can be generalized to learning non-invertible mappings \mathcal{M}_I and \mathcal{M}_T by seeking the nearest neighbors of $\mathcal{M} \circ \mathcal{M}_T(T_q)$ and $\mathcal{M}^{-1} \circ \mathcal{M}_I(I_q)$ in the intermediate spaces \mathcal{V}^I and \mathcal{V}^T , respectively, and matching them up with the corresponding image and text, in \mathfrak{R}^I and \mathfrak{R}^T . Under this formulation, followed in this work, 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).

5.3.3 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

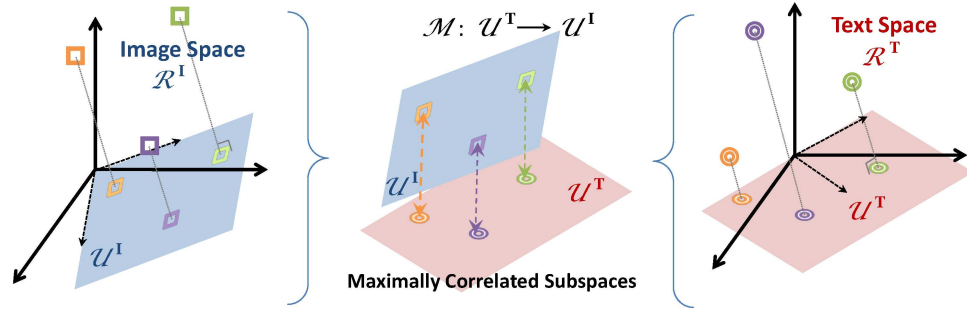


Figure 5.3: Correlation matching (CM) performs joint feature selection in the text and image spaces, projecting them onto two maximally correlated subspaces \mathcal{U}_T and \mathcal{U}_I .

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. Considering a query for “swan” 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 just match “white” to “white” and “swan” to “swan”, a cross-modal retrieval system cannot solve the task without *abstracting* that “white is a visual attribute of swan”. Hence, cross-modal retrieval is a more effective paradigm for testing fundamental hypotheses in multimedia representation than uni-modal or multi-modal retrieval. In this work, we exploit the cross-modal retrieval problem to test two such hypotheses regarding the joint modeling of images and text.

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

The hypotheses are tested by comparing three possibilities for the design of the intermediate spaces \mathcal{V}^I and \mathcal{V}^T of cross-modal retrieval. In the first case, two

Table 5.1: Taxonomy of the proposed approaches to cross-modal retrieval.

	correlation hypothesis	abstraction hypothesis
CM	✓	
SM		✓
SCM	✓	✓

feature transformations map \mathfrak{R}^I and \mathfrak{R}^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).

5.1 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.

5.4 Cross-modal Retrieval

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

5.4.1 Correlation matching (CM)

The design of a mapping from \mathfrak{R}^T and \mathfrak{R}^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) [29] and principal component analysis (PCA) [64]. 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 the introduction of a number of multi-modal subspace methods in the literature. In this work, we consider *cross-modal factor analysis* (CFA), *canonical correlation analysis* (CCA), and *kernel canonical correlation analysis* (KCCA). All these methods include a training stage, where the subspaces \mathcal{U}^I and \mathcal{U}^T are learned, followed by a projection stage, where images and text are projected into these spaces. Figure 5.3 illustrates this process. Cross-modal retrieval is finally performed within the low-dimensional subspaces.

Linear subspace learning

CFA seeks transformations that best represent coupled patterns between different subsets of features (e.g., different modalities) describing the same objects [76]. It finds the orthonormal transformations Ω_I and Ω_T that project the two modalities onto a shared space, $\mathcal{U}^I = \mathcal{U}^T = \mathcal{U}$, where the projections have minimum distance

$$\|X_I\Omega_I - X_T\Omega_T\|_F^2. \quad (5.1)$$

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

$$\text{trace}(X_I\Omega_I\Omega_T'X_T'), \quad (5.2)$$

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.*,

$$X_I'X_T = \Omega_I \Lambda \Omega_T, \quad (5.3)$$

where Λ is the matrix of singular values of $X_I'X_T$ [76].

CCA [59] learns the d -dimensional subspaces $\mathcal{U}^I \subset \mathfrak{R}^I$ (image) and $\mathcal{U}^T \subset \mathfrak{R}^T$ (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 $w_i \in \mathfrak{R}^I$ and $w_t \in \mathfrak{R}^T$, but seeks directions along which the data is maximally correlated

$$\max_{w_i \neq 0, w_t \neq 0} \frac{w_i' \Sigma_{IT} w_t}{\sqrt{w_i' \Sigma_I w_i} \sqrt{w_t' \Sigma_T w_t}} \quad (5.4)$$

where Σ_I and Σ_T are the empirical covariance matrices for images $\{I_1, \dots, I_{|D|}\}$ and text $\{T_1, \dots, T_{|D|}\}$ respectively, and $\Sigma_{IT} = \Sigma_{TI}'$ the cross-covariance between them. Repetitively solving (5.4), 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 $\phi_T : \mathfrak{R}^T \rightarrow \mathcal{F}^T$ and $\phi_I : \mathfrak{R}^I \rightarrow \mathcal{F}^I$. Application of CFA or CCA in these spaces can then recover complex patterns of dependency in the original feature space. As is common in machine learning, the transformations $\phi_T(\cdot)$ and $\phi_I(\cdot)$ are computed only implicitly, by the introduction of two kernel functions $\mathcal{K}_T(\cdot, \cdot)$ and $\mathcal{K}_I(\cdot, \cdot)$, specifying the inner products in \mathcal{F}^T and \mathcal{F}^I , *i.e.*, $\mathcal{K}_T(T_m, T_n) = \langle \phi_T(T_m), \phi_T(T_n) \rangle$ and $\mathcal{K}_I(I_m, I_n) = \langle \phi_I(I_m), \phi_I(I_n) \rangle$, respectively.

KCCA [127, 163] 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

$$\max_{\alpha_i \neq 0, \alpha_t \neq 0} \frac{\alpha_i' K_I K_T \alpha_t}{V(\alpha_i, K_I) V(\alpha_t, K_T)}, \quad (5.5)$$

where $V(\alpha, K) = \sqrt{(1 - \kappa)\alpha' K^2 \alpha + \kappa \alpha' K \alpha}$, $\kappa \in [0, 1]$ is a regularization parameter, and K_I and K_T are the kernel matrices of the image and text representations, *e.g.*, $(K_I)_{mn} = \mathcal{K}_I(I_m, I_n)$. Given optimal α_i and α_t for (5.5), w_i and w_t are obtained as linear combinations of the training examples $\{\phi_I(I_k)\}_{k=1}^{|\mathcal{B}|}$, and $\{\phi_T(T_k)\}_{k=1}^{|\mathcal{B}|}$, with α_i and α_t as weight vectors, *i.e.*, $w_i = \Phi_I(X_I)^T \alpha_i$ and $w_t = \Phi_T(X_T)^T \alpha_t$, where $\Phi_I(X_I)$ ($\Phi_T(X_T)$) is the matrix whose rows contain the high-dimensional representation of the image (text) features. To optimize (5.5), we solve a generalized eigenvalue problem using the software package of [163]. The first d generalized eigenvectors provide us with d weight vectors $\{\alpha_{i,k}\}_{k=1}^d$ and $\{\alpha_{t,k}\}_{k=1}^d$, from which 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 $\mathcal{U}^T \subset \mathcal{F}^T$ can be derived, with $1 \leq d \leq |\mathcal{B}|$.

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) $I \in \mathfrak{R}^I$ ($T \in \mathfrak{R}^T$) 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 $I \in \mathfrak{R}^I$ 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

$$\begin{aligned} p_{I,k} &= \langle \phi_I(I), w_{i,k} \rangle \\ &= \langle \phi_I(I), [\phi_I(I_1), \dots, \phi_I(I_{|\mathcal{B}|})] \alpha_{i,k} \rangle \\ &= [\mathcal{K}_I(I, I_1), \dots, \mathcal{K}_I(I, I_{|\mathcal{B}|})] \alpha_{i,k}, \end{aligned} \quad (5.6)$$

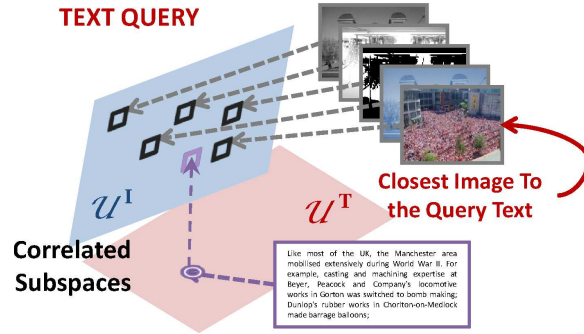


Figure 5.4: Cross-modal retrieval using CM. Here, CM is used to find the images that best match a query text.

where $k = 1, \dots, d$. Analogously, a text $T \in \mathfrak{R}^T$ 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 $\mathcal{K}_T(\cdot, \cdot)$.

Correlation matching

For all methods, a natural invertible mapping between the projections onto \mathcal{U}^I and \mathcal{U}^T follows from the correspondence between the d -dimensional bases of the subspaces, as $w_{i,1} \leftrightarrow w_{t,1}, \dots, w_{i,d} \leftrightarrow w_{t,d}$. 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 5.3. Given an image query I with projection p_I , the text $T \in \mathfrak{R}^T$ that most closely matches it is that for which p_T minimizes

$$D(I, T) = d(p_I, p_T), \quad (5.7)$$

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 $I \in \mathfrak{R}^I$ is that for which p_I minimizes $d(p_I, p_T)$. An illustration of cross-modal retrieval using CM is given in Figure 5.4.

5.4.2 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{B} with a vocabulary

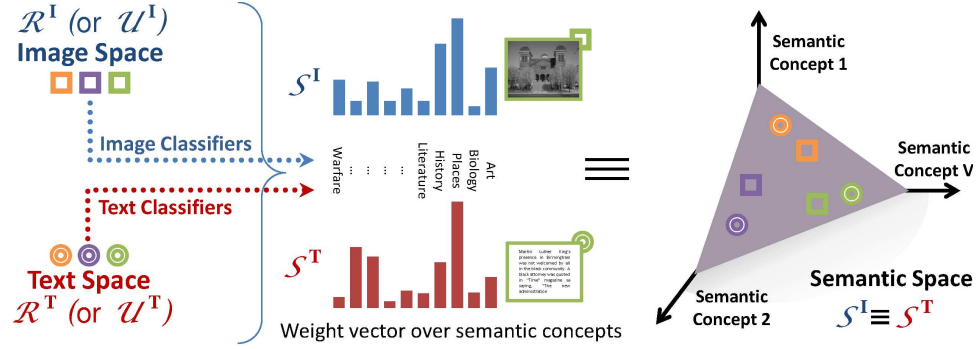


Figure 5.5: Semantic matching (SM) maps text and images into a semantic space. For each modality, classifiers are used to obtain a semantic representation, *i.e.*, a weight vector over semantic concepts.

$\mathcal{L} = \{1, \dots, L\}$ of semantic concepts, such as “History” or “Biology”. Individual documents are grouped into these classes. Two mappings $\mathbf{\Pi}_T$ and $\mathbf{\Pi}_I$ are then implemented using classifiers of text and images, respectively. $\mathbf{\Pi}_T$ maps a text $T \in \mathbb{R}^T$ into a vector π_T of posterior probabilities $P_{W|T}(w|T), w \in \{1, \dots, L\}$ with respect to each of the classes in \mathcal{L} . The space \mathcal{S}^T of these vectors is referred to as the *semantic space for text*, and the probabilities $P_{W|T}(w|T)$ as *semantic text features*. Similarly, $\mathbf{\Pi}_I$ maps an image I into a vector π_I of *semantic image features* $P_{W|I}(w|I), w \in \{1, \dots, L\}$ 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 standard features in \mathbb{R}^T and \mathbb{R}^I are the result of unsupervised learning, and frequently have no obvious interpretation (*e.g.*, image features tend to be edges, edge orientations or frequency bases), the features in \mathcal{S}^T and \mathcal{S}^I are semantic concept probabilities (*e.g.*, the probability that the image belongs to the “History” or “Biology” document classes). In Chapter 3, it was shown that this increased semantic abstraction can lead to substantially better generalization for tasks such as image retrieval. 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* document classes. Hence, the spaces can be treated as being the same, *i.e.*, $\mathcal{S}^T = \mathcal{S}^I$, leading to the schematic representation in Figure 5.5.

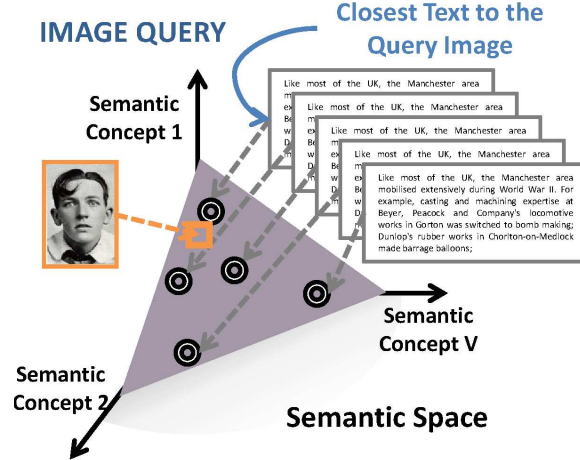


Figure 5.6: Cross-modal retrieval using SM used to find the text that best matches a query image.

In Chapter 2, it was highlighted that it is not necessary to model each class explicitly and any system that computes posterior probabilities can be employed to obtain the semantic representation. For the evaluation of cross-modal retrieval systems, the posterior probability distributions are computed through multi-class logistic regression which produces linear classifiers with a probabilistic interpretation. Logistic regression based classification is chosen due to its simplicity. Under this, the posterior probability of class w is computed, by fitting the image (text) features to a logistic function,

$$P_{W|X}(w|x; \boldsymbol{\beta}) = \frac{1}{Z(x, \boldsymbol{\beta})} \exp(\beta_w^T x), \quad (5.8)$$

where $Z(x, \boldsymbol{\beta}) = \sum_w \exp(\beta_w^T x)$ is a normalization constant, W the class label, X the feature vector in the input space, and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_L\}$ with β_w a vector of parameters for class w . A multi-class logistic regression is learned for the image and text modality, by making X the image and text representation, $I \in \mathfrak{R}^I$ and $T \in \mathfrak{R}^T$, respectively. In our implementation we use the software package Liblinear [38]. Given a query image I (text T), represented by $\pi_I \in \mathcal{S}^I$ ($\pi_T \in \mathcal{S}^T$), cross-modal retrieval will find the text T (image I), represented by $\pi_T \in \mathcal{S}^T$ ($\pi_I \in \mathcal{S}^I$), that minimizes

$$D(I, T) = d(\pi_I, \pi_T), \quad (5.9)$$

for some suitable distance measure d between probability distributions. An illustration of cross-modal retrieval using SM is given in Figure 5.6.

5.4.3 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. 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 with correlation modeling. Logistic regressors $\mathbf{\Pi}_I$ and $\mathbf{\Pi}_T$ 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 (5.9), based on the semantic mappings $\pi_I = \mathbf{\Pi}_I(p_I)$ and $\pi_T = \mathbf{\Pi}_T(p_T)$ after projecting them onto \mathcal{U}^I and \mathcal{U}^T , respectively.

5.5 Experimental Setup

In this section, we describe an extensive experimental evaluation of the proposed framework. Two tasks were considered: text retrieval from an image query, and image retrieval from a text query. The cross-modal retrieval performance is measured with *precision-recall* (PR) curves and *mean average precision* (MAP) scores. The standard 11-point interpolated PR curves [91] are used. The MAP score is the average precision at the ranks where recall changes. Both metrics are evaluated at the level of *in-* or *out-of-category*, which is a popular choice in the information retrieval literature [119].

Dataset

For the evaluation of the cross-modal retrieval system we use two different datasets, viz. TVGraz and Wikipedia. The TVGraz dataset is a collection of web-

pages compiled by Khan *et al.* [66] and contains 2,058 image-text pairs divided into 10 categories (see Appendix A.1.5 for more details). Wikipedia is novel dataset assembled from the “Wikipedia featured articles”, a continually updated collection of Wikipedia articles, and contains a total of 2,866 image-text pairs again divided into 10 categories (see Appendix A.1.6 for more details)

The two datasets have important differences. TVGraz images are archetypal members of the categories, due to the collection procedure [66]. The dataset is eminently visual, since its categories (*e.g.*, “Harp”, “Dolphin”) are specific objects or animals, and the classes are semantically well-separated, with little or no semantic overlap. For example, the syllabus of a Neuroscience class can be attached to a picture of a brain. However, the texts are small and can be less representative of the categories to which they are associated. In Wikipedia, on the other hand, the category membership is assessed based on text content. Hence, texts are mostly of good quality and representative of the category, while the image categorization is more ambiguous. For example, a portrait of 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. Frequently, documents could be classified into one or more categories. Individually, the images can be difficult to classify, even for a human. Together, the two datasets represent an important subset of the diversity of practical cross-modal retrieval scenarios: applications where there is more uniformity of text than images, and vice-versa.

5.5.1 Image and text representation

For both modalities, the base representation is a bag-of-words (BOW) representation. Text words were obtained by stemming the text with the Python Natural Language Toolkit³. Direct word histograms were not suitable for text because the large lexicon made the correlation analysis intractable. Instead, a latent Dirichlet allocation (LDA) [14] model was learned from the text features, using the implementation of [32]. LDA summarizes a text as a mixture of topics. More precisely, a text is modeled as a multinomial distribution over topics, each of which

³<http://www.nltk.org/>

Table 5.2: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for TVGraz. μ_p and μ_q are the sample averages for p and q , respectively.

			TVGraz		
Experiment	measure	$d(p, q)$	img query	txt query	avg
CM	ℓ_1	$\sum_i p_i - q_i $	0.376	0.418	0.397
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.391	0.444	0.417
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.498	0.476	0.487
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.486	0.462	0.474
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.296	0.546	0.421
	ℓ_1	$\sum_i p_i - q_i $	0.412	0.548	0.480
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.380	0.550	0.465
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.533	0.560	0.546
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.579	0.556	0.568
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.576	0.636	0.606
	ℓ_1	$\sum_i p_i - q_i $	0.637	0.645	0.641
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.614	0.63	0.622
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.669	0.646	0.658
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.678	0.641	0.660

Table 5.3: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for Wikipedia. μ_p and μ_q are the sample averages for p and q , respectively.

			Wikipedia		
Experiment	measure	$d(p, q)$	img query	txt query	avg
CM	ℓ_1	$\sum_i p_i - q_i $	0.193	0.234	0.214
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.199	0.243	0.221
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.288	0.239	0.263
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.287	0.239	0.263
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.188	0.276	0.232
	ℓ_1	$\sum_i p_i - q_i $	0.232	0.276	0.254
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.211	0.278	0.245
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.315	0.278	0.296
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.354	0.272	0.313
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.287	0.282	0.285
	ℓ_1	$\sum_i p_i - q_i $	0.329	0.286	0.308
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.307	0.286	0.296
	NC	$\frac{p^T q}{\ p\ \ q\ }$	0.375	0.288	0.330
	NC_c	$\frac{(p - \mu_p)^T (q - \mu_q)}{\ p - \mu_p\ \ q - \mu_q\ }$	0.388	0.285	0.337

is in turn modeled as a multinomial distribution over words. Each word in a text is generated by first sampling a topic from the text-specific topic distribution, and then sampling a word from that topic’s multinomial. This serves two purposes: it reduces dimensionality and increases feature abstraction, by representing text as a distribution over topics instead of a distribution over words. In text modeling the number of topics in LDA ranged from 5 to 800.

Image words were learned with the scale invariant feature transformation (SIFT-GRID) [85] computed on a grid of image patches. A bag of SIFT descriptors was first extracted from each image in the training set, using the SIFT implementation of LEAR⁴. A codebook, or dictionary of visual words was then learned with the K-means clustering algorithm. The SIFT descriptors extracted from each image were vector quantized with this codebook, producing a vector of visual word counts per image. Besides this BOW representation, we also use a lower-dimensional representation for images, similar to that for text, by fitting an LDA model to visual word histograms and representing images as a distribution over topics. Preliminary experiments indicated that this outperformed an image representation of reduced dimensionality through principal component analysis (PCA). In image modeling for LDA representation the number of topics ranged from 5 to 4,000, for BOW the number of visual words ranged from 128 to 8,192.

5.6 Parameter selection

The combination of three retrieval modes (CM, SM, and SCM), three correlation matching approaches (CFA, CCA, KCCA), two image representations (BOW, LDA), and various distance measures d generates a large number of possibilities for the implementation of cross-modal retrieval. Since each configuration has a number of parameters to tune, it is difficult to perform an exhaustive comparison of all possibilities. Instead, we pursued a sequence of preliminary comparisons to prune the configuration space, using a random 80/20 split of the training set, for training and validation respectively (splitting TVGraz’ training set into 1,245

⁴<https://lear.inrialpes.fr/people/dorko/downloads.html>

training and 313 validation examples, and Wikipedia’s into 1,738 training and 435 validation documents). This suggested a cross-modal retrieval architecture that combines i) the centered normalized correlation (for distances d), ii) a BOW (rather than LDA) representation for images, and iii) KCCA to learn correlation subspaces. Supporting experiments are presented below. For each retrieval mode – CM, SM, SCM for image queries or text queries – and each dataset – TVGraz, Wikipedia –, the codebook size (for image representation), the number of topics (for text representation) and/or the number of KCCA components were determined, where applicable, by performing a grid search and adopting the settings with maximum retrieval performance on the validation set, unless indicated otherwise. In the following section, the top performing approaches are compared on the test set.

Distance Measures

We started by comparing a number of distance measures d , for the evaluation of (5.7) and (5.9), in CM, SM, and SCM retrieval experiments (using KCCA to produce the subspaces for CM and SCM, and BOW to represent images). The distance measures are listed in 5.2 and 5.3 for TVGraz and Wikipedia respectively, and include the Kullback-Leibler divergence (KL), ℓ_1 and ℓ_2 norms, normalized correlation (NC), and centered normalized correlation (NC_c). The KL divergence was not used with CM because this technique does not produce a probability simplex. 5.2 and 5.3 present the MAP scores achieved with each measure, on the validation set. NC_c achieved the best average performance in all experiments other than CM-based retrieval on TVGraz, where it was outperformed by NC . Since the difference was small even in this case, NC_c was adopted as distance measure in all remaining experiments.

Text and image representation

Due to the intractability of word counts, we considered only the LDA representation for text. In the image domain, we compared the performance of the BOW and LDA representations, using an SCM system based on KCCA subspaces

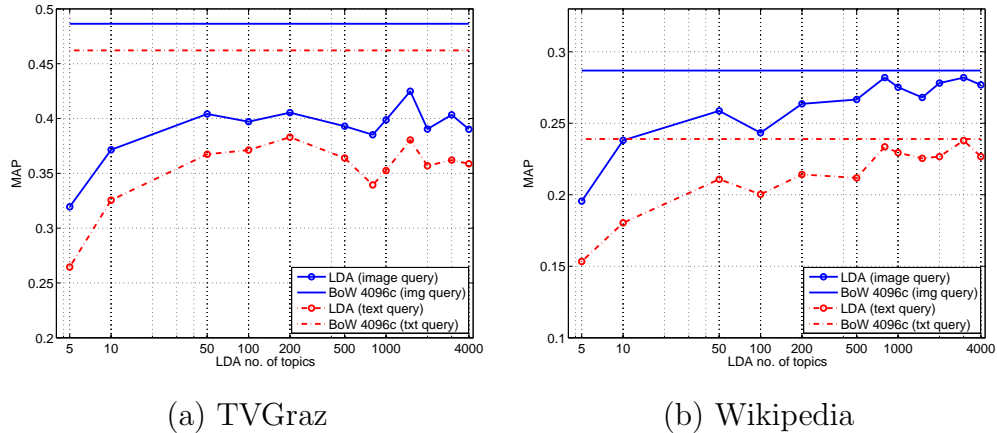


Figure 5.7: MAP performance (cross-modal retrieval, validation set) of SCM using two image models: BOW (flat lines) and LDA, for (a) TVGraz and (b) Wikipedia.

and 4,096 codewords for BOW (an optimal setting, as evidenced in Section 5.6). Figure 5.7 presents the results for both text and image queries. Since the retrieval performance of LDA was inferior to that of BOW, for all topic cardinalities, BOW was adopted as the image representation for all remaining experiments.

Correlation matching

The next set of experiments was designed to compare the different CM methods. These methods have different degrees of freedom and thus require different amounts of parameter tuning. The most flexible representation is KCCA, whose performance varies with the choice of kernel and regularization parameter κ of (5.5). We started by comparing various combinations of text and image kernels. Best results were achieved for a *chi-square radial basis function* kernel⁵ for images combined with a *histogram intersection* kernel [141, 18] for text. Combinations involving other kernels (*e.g.*, linear, Gaussian, exponential) achieved inferior validation set performance. Regarding regularization, best results were obtained with $\kappa = 10\%$ on TVGraz and $\kappa = 50\%$ on Wikipedia. The need for a stronger regular-

⁵ $\mathcal{K}(x, y) = \exp\left(\frac{d_{\chi^2}(x, y)}{\gamma}\right)$ where $d_{\chi^2}(x, y)$ is the chi-square distance between x and y and γ the average chi-square distance among training points.

Table 5.4: MAP for CM hypothesis (validation sets).

Experiment	Image Query	Text Query	Average	Average Gain	Dataset
KCCA	0.486	0.462	0.474	-	TVGraz
CCA	0.284	0.254	0.269	76%	
CFA	0.195	0.179	0.187	153%	
KCCA	0.287	0.239	0.263	-	Wiki.
CCA	0.210	0.174	0.192	37%	
CFA	0.195	0.156	0.176	50%	

izer in Wikipedia suggests that there are more spurious correlations on this dataset, which could lead to over-fitting. This is sensible, given the greater diversity and abstraction of the concepts in this dataset.

For CCA (CFA), the only free parameter is the number of canonical components (dimensionality of the shared space) used for both image and text representation. This parameter also remains to be tuned for KCCA. For each experiment and data set, a grid search was performed and the parameter of best retrieval performance was adapted under each method (CFA, CCA, KCCA). 5.4 presents best CM performances achieved with each method. In all cases, KCCA yields top performance. On TVGraz, the average gain (for text and image queries) is 153% over CFA and 76% over CCA. On Wikipedia, the gain over CFA is 50% and over CCA 37%. KCCA was chosen to implement the correlation hypothesis in the remaining experiments.

Parameter Tuning

For a cross-modal retrieval architecture combining the best of the above, i.e., KCCA (to learn correlation subspaces), NC_c (as distance measure), and the BOW representation for images, we take a closer look at the codebook size for image (BOW) representation, the number of topics for text (LDA) representation and the number of KCCA components. Figure 5.5 summarizes the optimal parameter

Table 5.5: Best parameter settings for CM, SM and SCM, on both TVGraz and Wikipedia (validation sets).

	CM	SM	SCM	
MAP image / text query	0.49 / 0.46	0.59 / 0.56	0.68 / 0.64	TVGraz
BOW codewords	4096			
LDA topics	200	100	400	
KCCA components	8	-	1125	
MAP image / text query	0.29 / 0.24	0.35 / 0.27	0.39 / 0.29	Wikipedia
BOW codewords	4096			
LDA topics	20	600	200	
KCCA components	10	-	38	

settings (after performing a grid search with cross-validation) and corresponding retrieval performance on the validation set, for CM, SM and SCM experiments. 5.8 provides more detail on how varying each parameter individually affects the performance, for CM. Note that the best MAP scores are obtained with a small number of KCCA components (< 10). For the image representation, best performance was achieved with codebooks of 4,096 visual words, on both datasets. For text, 200 topics performed best on TVGraz and 20 on Wikipedia. Note that in the test set experiments of Section 5.7, the number of KCCA components of 5.5 is scaled by the ratio of the number of training points of the test experiments and that of the validation experiments (see A.4 and A.5 in Appendix A), so that a comparable fraction of correlation is preserved after dimensionality reduction⁶.

⁶KCCA seeks directions of maximum correlation in $\text{Span}\{\phi_I(I_1), \dots, \phi_I(I_{|\mathcal{B}|})\}$ and $\text{Span}\{\phi_T(T_1), \dots, \phi_T(T_{|\mathcal{B}|})\}$, where $|\mathcal{B}|$ is the training set size. This is larger for test than for validation experiments (2,173 v.s. 1,738 on Wikipedia and 1,558 v.s. 1,245 on TVGraz). Hence, on average, a KCCA component will explain less correlation in the test than in the validation experiments. It follows that a larger number of KCCA components are needed to capture the same fraction of the total correlation.

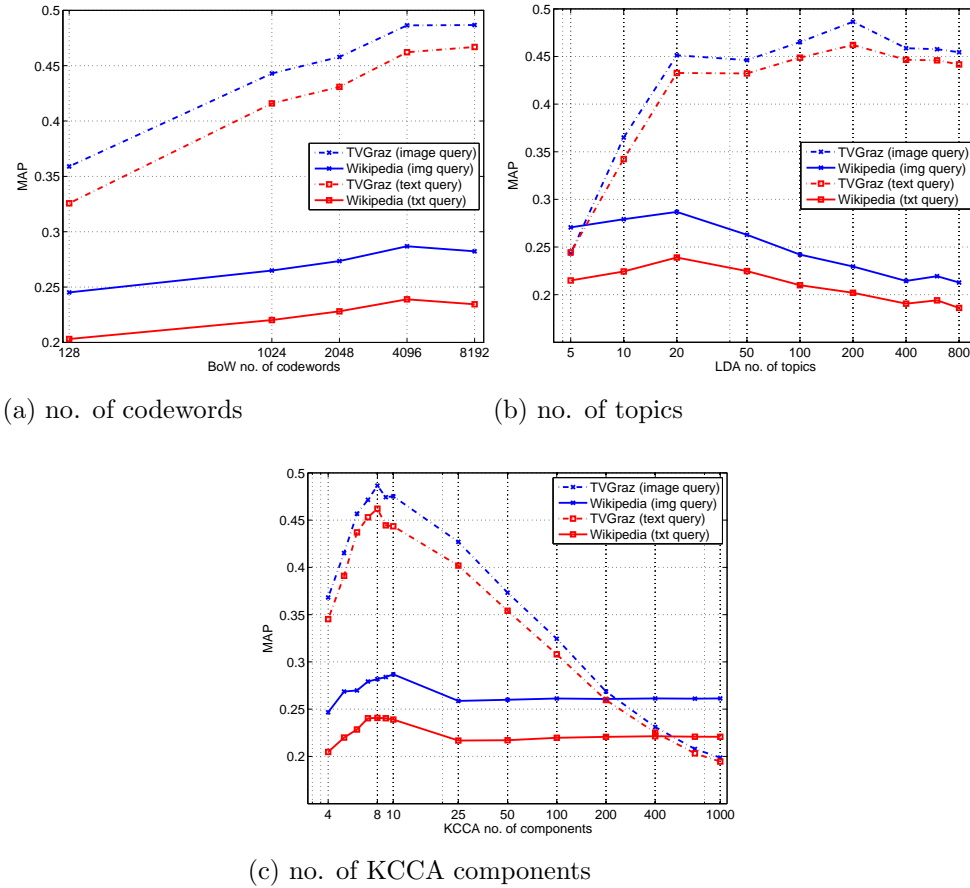


Figure 5.8: Cross-modal MAP for CM on TVGraz and Wikipedia (validation sets), as a function of (a) the number of image codewords, (b) the number of text LDA topics, and (c) the number of KCCA components (while keeping the other two parameters fixed at the values reported in 5.5).

5.7 Testing the fundamental hypotheses

In this section, we compare the performance of CM, SM, and SCM on the test set. In all cases, the parameter configurations are those that achieved best cross-validation performance in the previous section. 5.6 compares the MAP scores of cross-modal retrieval — text-to-image, image-to-text, and their average — using CM, SM and SCM, to chance-level performance⁷. Two distinct observations can be made from this table with regards to TVGraz. First, it provides evidence in

⁷Random images (text) returned in response to a text (image) query.

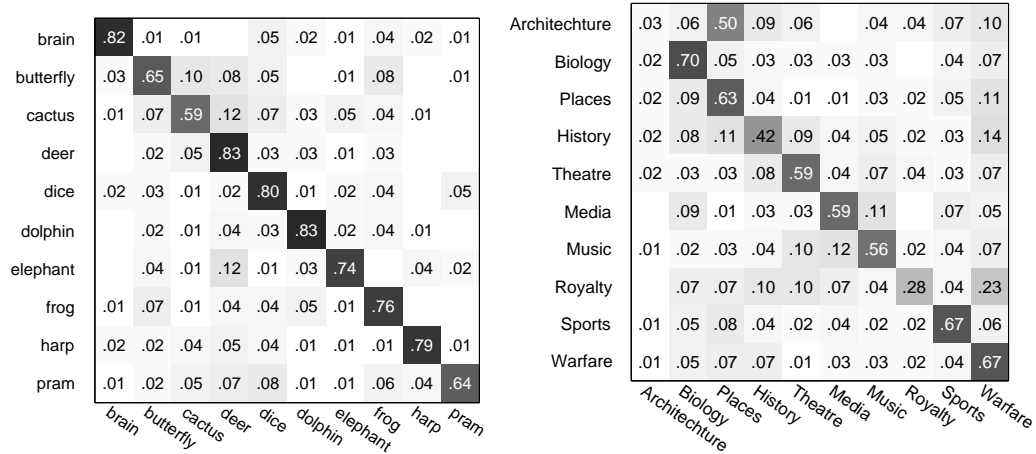


Figure 5.9: Confusion matrices on the test set, for both TVGraz (left) and Wikipedia (right). Rows refer to true categories, and columns to category predictions. The more confusion on Wikipedia motivates the lower retrieval performance.

support of the two hypotheses of Section 5.3.3. Both joint dimensionality reduction (CM) and semantic abstraction (SM) are beneficial for multi-modal modeling, leading to a non-trivial improvement over chance-level performance. For example, in TVGraz, CM achieves an average MAP score of 0.497, over four times the random retrieval performance of 0.114. SM yields an even greater improvement, attaining a MAP score of 0.622. Second, combining correlation modeling with semantic abstraction (SCM) is desirable, leading to higher MAP scores. On TVGraz, SCM improves about 12% over SM and 40% over CM, achieving an average MAP score of 0.694. This suggests that the contributions of cross-modal correlation and semantic abstraction are *complementary*: not only is there an independent benefit to both correlation modeling and abstraction, but the *best performance is achieved when the approaches underlying the two hypotheses are combined*. The gains hold for both cross-modal retrieval tasks, *i.e.*, image and text queries.

Similar conclusions can be drawn for Wikipedia. However, the improvement of SCM over SM is less substantial than in TVGraz. In fact, the retrieval performances on Wikipedia are generally lower than those on TVGraz. As discussed in Section 5.5, this is likely due to the broader scope of the Wikipedia categories. In

Table 5.6: Cross-modal MAP on TVGraz and Wikipedia (test sets).

Experiment	Image Query	Text Query	Average	Average Gain	
SCM	0.693	0.696	0.694	-	TVGraz
SM	0.625	0.618	0.622	11.6%	
CM	0.507	0.486	0.497	39.6%	
Random	0.114	0.114	0.114	509%	
SCM	0.372	0.268	0.320	-	Wiki.
SM	0.362	0.252	0.307	4.2%	
CM	0.282	0.225	0.253	26.5%	
Random	0.119	0.119	0.119	170%	

this dataset, a significant fraction of documents could be classified into multiple categories, making the data harder to model. This explanation is supported by the confusion matrices of Figure 5.9. These were built by assigning each text and image query to the class of highest MAP in the ranking produced by SCM⁸. Note, for example, the significant confusion between the categories “Architecture” and “Places”, or “Royalty” and “Warfare”. Figure 5.10 and 5.11 presents PR curves and precision at N curves, of cross-modal retrieval with CM, SM and SCM for TVGraz and Wikipedia respectively. All methods yield non-trivial precision improvements, at all levels of recall, when compared to the random baseline. On TVGraz, SM has higher precision than CM, and SCM has higher precision than SM, at all levels of recall. On Wikipedia, SCM improves over CM, at all levels of recall, but the improvement over SM is small. Figure 5.12 shows the MAP scores achieved per category by all approaches. SCM has a significantly higher MAP than CM and SM on all classes of TVGraz, and is either comparable or better than CM and SM on the majority of classes of Wikipedia.

Few examples of text queries and corresponding retrieval results, using the SCM methodology, are shown in Figure 5.13, 5.14, Figure 5.15, and 5.16. The text

⁸Note that this is not ideal for classification, since the MAP is computed over a ranking of the test set.

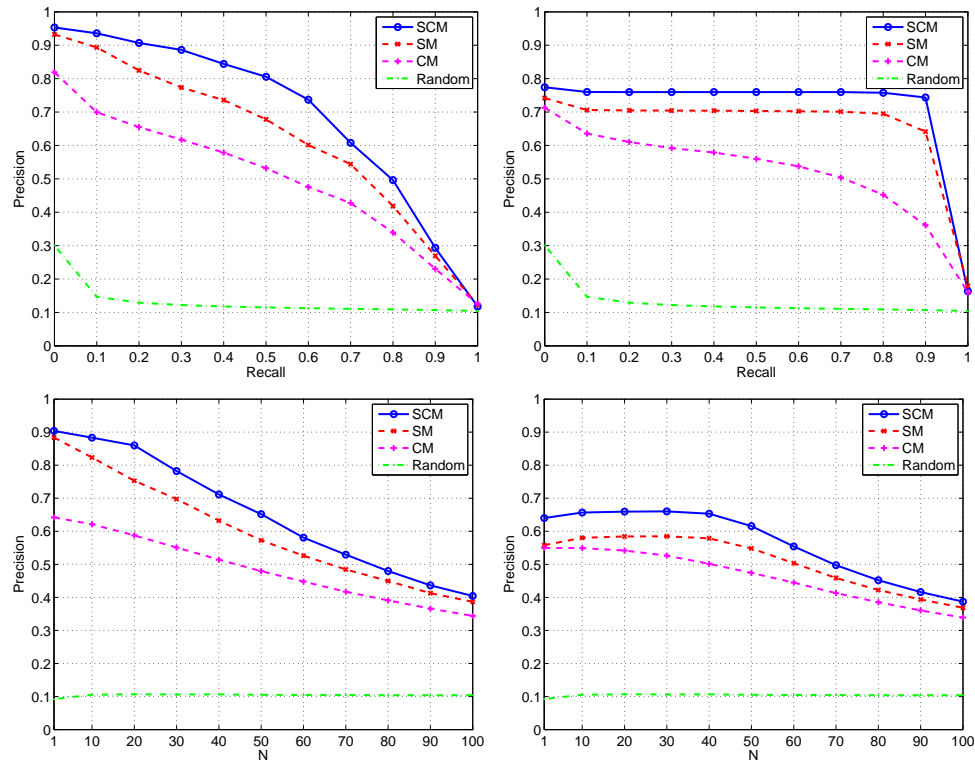


Figure 5.10: top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for TVGraz

query is presented along with its probability vector π_T and the ground truth image. The top five image matches are shown below the text, along with their probability vectors π_I . Note that SCM assigns these images the highest ranks in the retrieved list because their semantic vectors (π_I) most closely match that of the text (π_T). For the TVGraz example (Figure 5.16) this can be verified by noting the common concentration of probability mass around the “Butterfly” bin. In the Wikipedia example (Figure 5.14) the probability is concentrated around the “Warfare” bin. Finally, Figure 5.17 shows some examples of image-to-text retrieval. The query images are shown on the top row, and the images associated with the four best text matches are shown on the bottom.

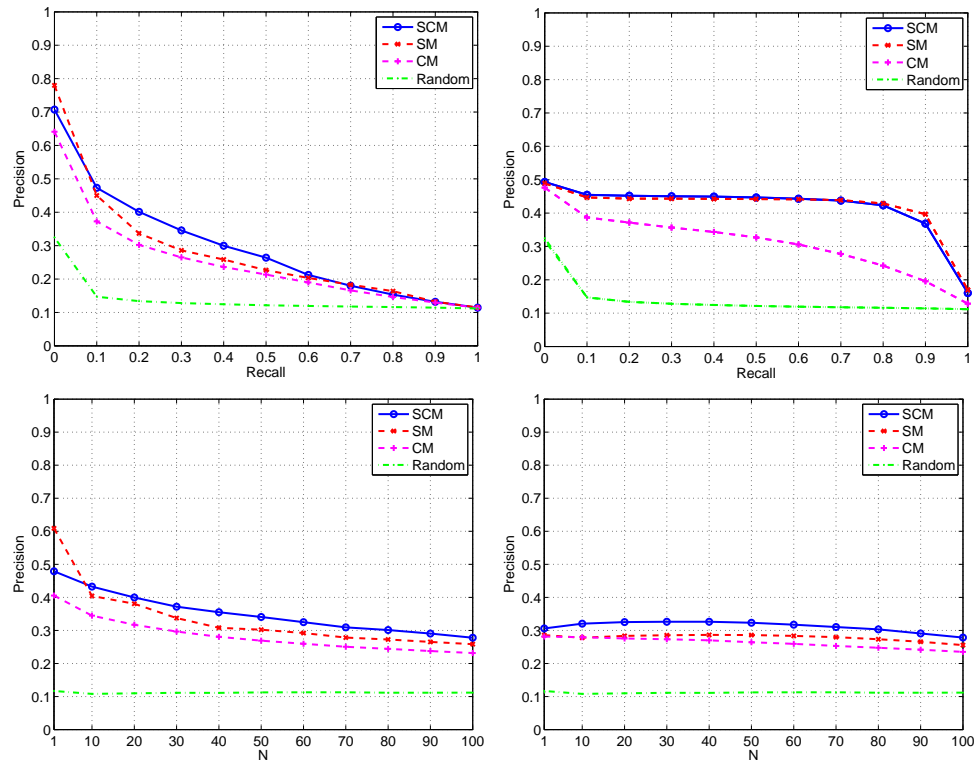


Figure 5.11: top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for Wikipedia

5.8 Acknowledgments

The author would like to thank Jose Costa Pereira, Emanuele Coviello, Gabe Doyle, Gert Lanckriet and Roger Levy, for their help and contribution in developing the cross-model multimedia system.

The text of Chapter 5, in part, is based on the material as it appears in: N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos “A New Approach to Cross-Modal Multimedia Retrieval”, ACM Proceedings of the 15th international conference on Multimedia, Florence, Italy, Oct 2010. The dissertation author was a primary researcher and an author of the cited material.

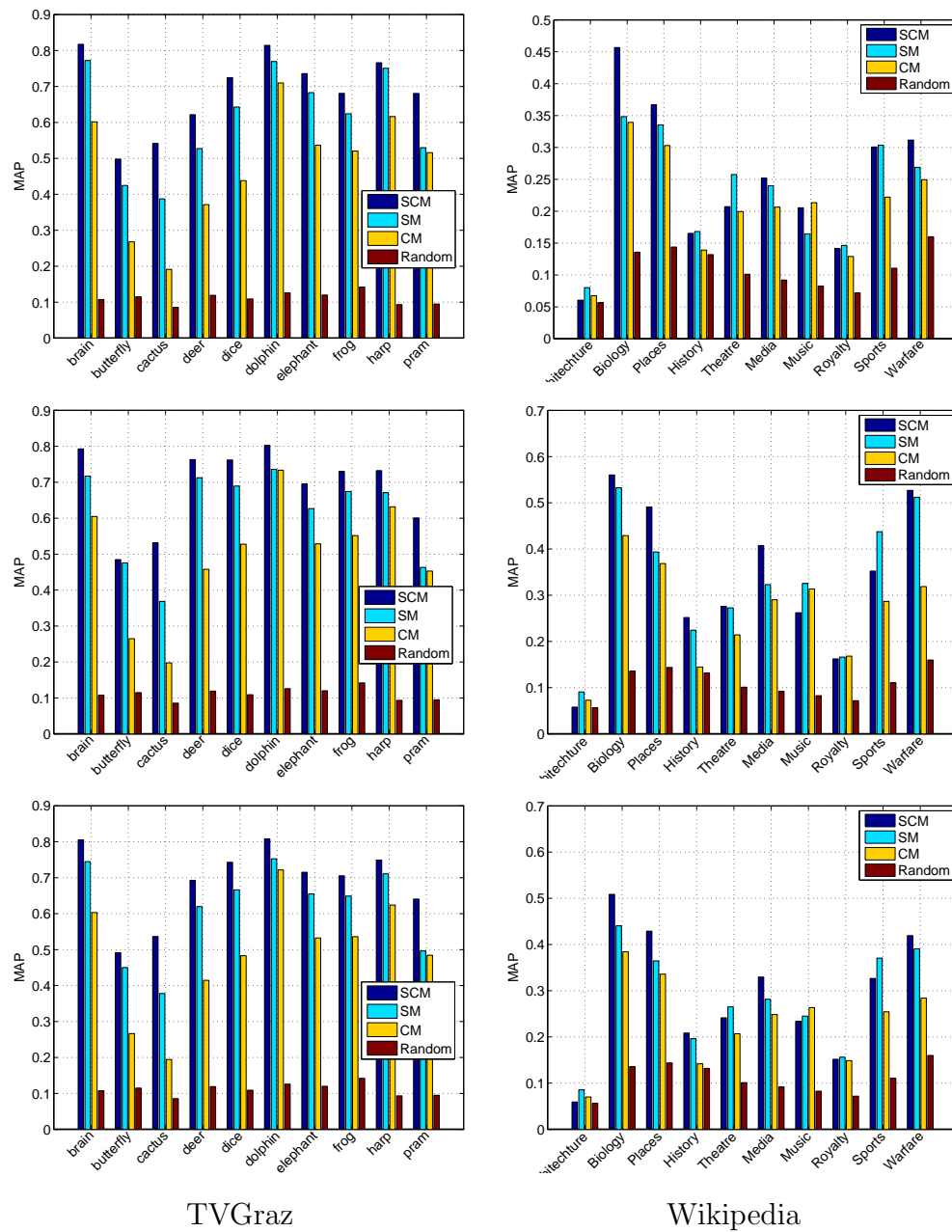


Figure 5.12: Per-class MAP for the cross-modal retrieval tasks on TVGraz (left) and Wikipedia (right): text queries (top); image queries (middle); and average performance over both types of queries (bottom).

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

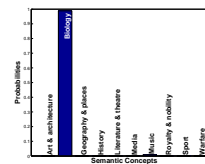
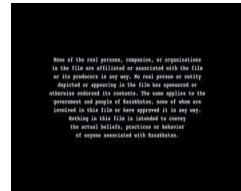
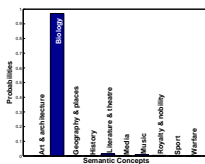
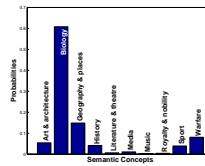
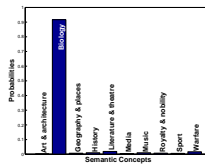
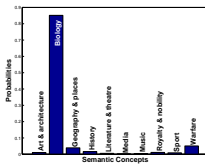
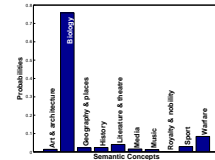


Figure 5.13: Text query from Biology class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

Between October 1 and October 17, the Japanese delivered 15,000 troops to Guadalcanal, giving Hyakutake 20,000 total troops to employ for his planned offensive. Because of the loss of their positions on the east side of the Matanikau, the Japanese decided that an attack on the U.S. defenses along the coast would be prohibitively difficult. Therefore, Hyakutake decided that the main thrust of his planned attack would be from south of Henderson Field. His 2nd Division (augmented by troops from the 38th Infantry Division), under Lieutenant General Masao Maruyama and comprising 7,000 soldiers in three infantry regiments of three battalions each was ordered to march through the jungle and attack the American defences from the south near the east bank of the Lunga River. Shaw, "First Offensive", p. 34, and Rottman, "Japanese Army", p. 63. (...)

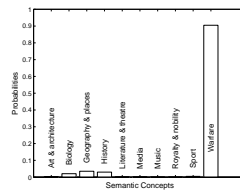
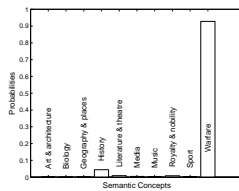
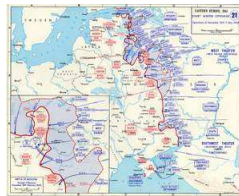
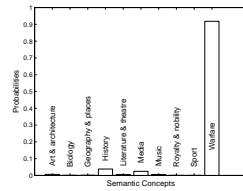
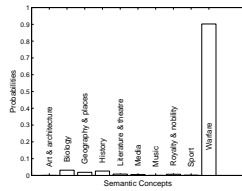
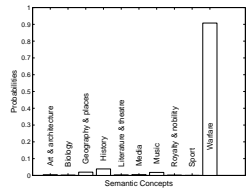
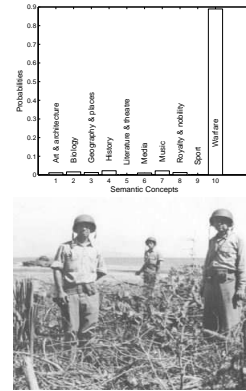


Figure 5.14: Text query from 'Warfare' class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

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

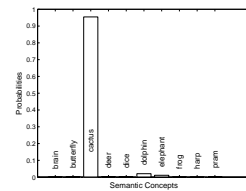
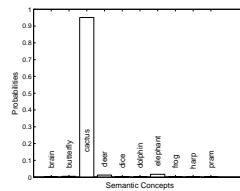
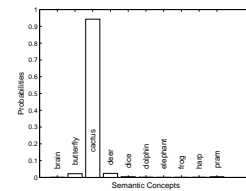
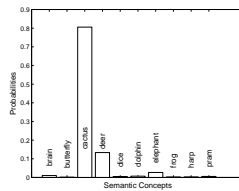
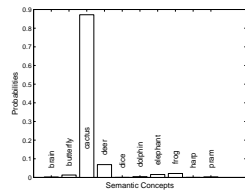
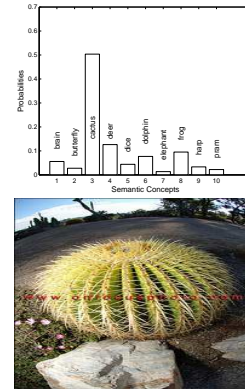


Figure 5.15: Text query from 'Cactus' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

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.

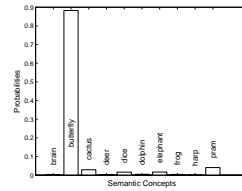
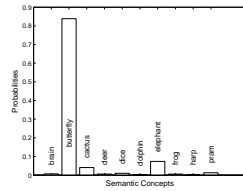
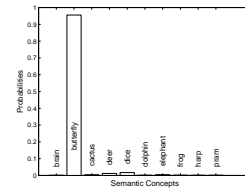
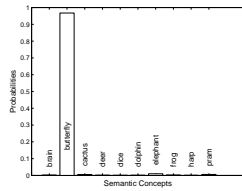
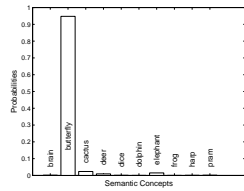
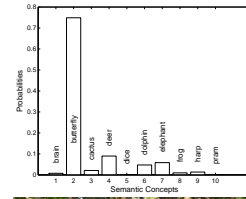


Figure 5.16: Text query from 'Butterfly' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

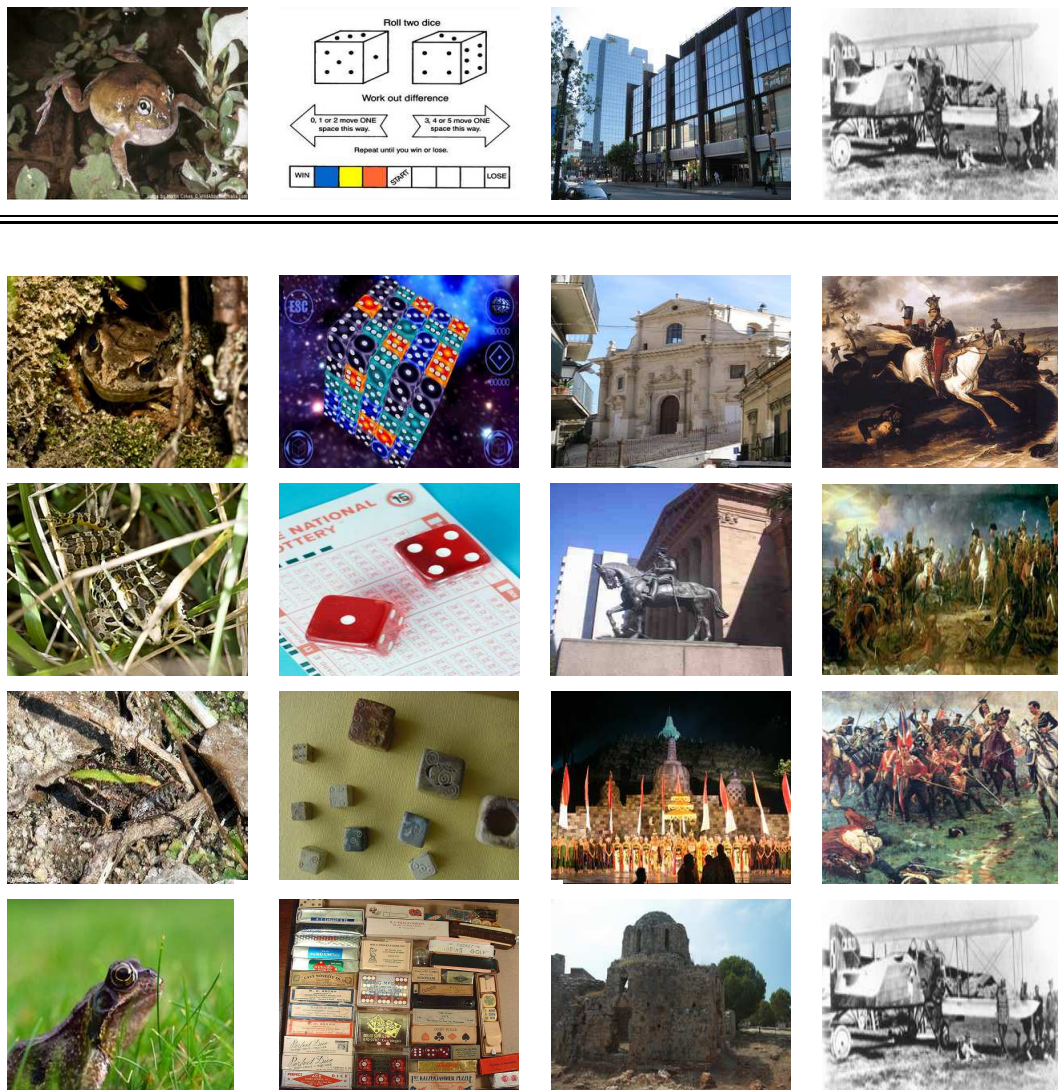


Figure 5.17: Image-to-text retrieval on TVGraz (first two columns) and Wikipedia (last two columns). Query images are shown on the top row. The four most relevant texts, represented by their ground truth images, are shown in the remaining columns.

Chapter 6

Holistic Context Modeling

In this chapter we discuss some of the drawbacks of the proposed semantic image representation and introduce the framework of “holistic context modeling” that, while addressing these drawbacks, yields robust visual recognition systems.

6.1 Introduction

Recent psychophysics studies have shown that humans rarely guide recognition *exclusively* by the appearance of the concepts to recognize. Most frequently, appearance is complemented by the *analysis of contextual relationships* with other visual concepts in the field of view [10]. In general, the detection of a concept of interest (e.g. buildings) is facilitated by the presence, in the scene, of other concepts (e.g. street, city) which *may not* themselves be of interest. Psychophysical studies have shown that context can depend on multiple clues. For example, object recognition is known to be affected by properties such as support (objects do not float in the air), interposition (objects occupy different volumes), probability (objects appear in different scenes with different probabilities), position (objects appear in typical locations), and size (objects have typical relative sizes) [10].

In this chapter, we investigate an approach to context modeling based on the probability of co-occurrence of objects and scenes. This modeling is quite simple, and builds upon the semantic representation of the images introduced in Chapter 2. Semantic image representation itself builds upon the *bag-of-features* (BoF) representation (see Chapter 2 for details regarding BoF representation), thereby inheriting several of its benefits. Most notably, it is strongly invariant to scene configurations, an essential attribute for robust scene classification and object recognition, and has low complexity, a property that enables large training sets and good generalization. Its main advantage over BoF is a higher level of *abstraction*, which can lead to substantially better generalization — as established in Chapter 3, by comparing the performance of nearest-neighbors classification in an image retrieval context. However, the semantic representation also has some limitations that can be traced back to the BoF representation itself. Most notable among these is a certain amount of *contextual noise*, i.e., noise in the probabilities

that compose the SMN. This is usually not due to poor statistical estimation, but due to the intrinsic *ambiguity* of the underlying BoF representation. Since appearance based features have small spatial support, it is frequently difficult to assign them to a single visual concept. Hence, the SMN extracted from an image usually assigns some probability to concepts unrelated to it (e.g. the concepts “bedroom” and “kitchen” for the “street” image of Figure 6.1).

Thus, while the SMN representation captures co-occurrences of the semantic concepts present in an image, not all these correspond to *true* contextual relationships. In fact, we argue that many (e.g. “bedroom” and “kitchen” in Figure 6.1) are *accidental*, i.e., casual coincidences due to the ambiguity of the underlying appearance representation (image patches that could belong to either a bed or a kitchen counter). Rather than attempting to eliminate contextual noise by further processing of appearance features, we propose a procedure for *robust* inference of contextual relationships *in the presence of accidental co-occurrences*. The idea is to keep the robustness of the appearance representation, but perform the classification at a higher level of *abstraction*, where ambiguity can be more easily detected.

This is achieved by introducing a second level of representation, that operates in the space of semantic features. The intuition is that, in this space, accidental co-occurrences are events of much smaller probability than true contextual co-occurrences: while “street” co-occurs with “buildings” in most images, it accidentally co-occurs with “bedroom” or “kitchen” in a much smaller set. True contextual relationships can thus be found by identifying peaks of probability in semantic space. Each visual concept is modeled by the distribution of the posterior probabilities extracted from all its training images. This *distribution of distributions* is referred as the *contextual model* for the concept. For large enough and diverse enough training sets, these models are dominated by the probabilities of true contextual relationships. Minimum probability of error (MPE) contextual classification can thus be implemented by simple application of Bayes’ rule. This suggests representing images as vectors of posterior probabilities under the contextual concept models, which we denote by *contextual multinomials* (CMN). These

are shown much less noisier than the SMNs learned at the appearance level.

An implementation of contextual modeling is proposed, where concepts are modeled as mixtures of Gaussian distribution on appearance space, and mixtures of Dirichlet distributions on semantic space. It is shown that 1) the contextual representation outperforms the appearance based representation, and 2) this holds irrespectively of the choice and accuracy of the underlying appearance models. An extensive experimental evaluation, involving the problems of scene classification and image retrieval shows that, despite its simplicity, the proposed approach is superior to various contextual modeling procedures in the literature.

The chapter is organized as follows. Section 6.2 briefly reviews the literature on context modeling. Section 6.3 then discusses the limitations of semantic image representation built upon appearance classifiers and introduces contextual models. An extensive experimental evaluation of contextual modeling is then presented in Section 6.4, Section 6.5, and Section 6.6.

6.2 Related Work on Context Modeling

Recent efforts towards context based recognition can be broadly grouped in two classes. The first, an *object-centric* approach, consists of methods that model contextual relationships between sub-image entities, such as objects. Examples range from simply accounting for the co-occurrence of different objects in a scene [115, 43], to explicit learning of the spatial relationships between objects [47, 174], or an object and its neighboring image regions [57]. Methods in the second class adopt a *scene-centric* representation, whereby context models are learned from entire images, generating a holistic description of the scene or its “gist” [104, 166, 77, 105, 74]. Various recent works have shown that semantic descriptions of natural images can be obtained with these representations, without explicit image segmentation [104]. This is consistent with evidence from the psychology [103] and cognitive neuroscience [3] literatures.

The scene-centric representation has itself been explored in two ways. One approach is to equate context to a vector of statistics of low-level visual measure-

ments taken over the entire image. For example, [104] models scenes according to the differential regularities of their second order statistics. A second approach is to rely on the BoF/BoW representation. Here, low-level features are computed locally and aggregated across the image, to form a holistic context model [166, 77, 121]. Although these methods usually ignore spatial information, some extensions have been proposed to weakly encode the latter. These consist of dividing the image into a coarse grid of spatial regions, and modeling context within each [104, 74].

The proposed context modeling combines aspects of both the object-centric and scene-centric strategies. Like the object-centric methods, we exploit relationships between co-occurring semantic concepts in natural scenes to derive contextual information. This is, however, accomplished without demarcating individual concepts or regions in the image. Instead, all conceptual relations are learned through global scene representations. Moreover, these relationships are learned in a purely data-driven fashion, i.e. no external guidance about the statistics of high-level contextual relationships is required, and the representation consists of full probability distributions, not just statistics. The proposed representation can be thought as modeling the “gist” of the scene by the co-occurrences of semantic visual concepts that it contains.

The representation closest to that now proposed is probably the family of latent topic models, recently popular in vision [77, 114, 17]. These models were originally proposed in the text literature, to address the ambiguity of BoW. It was realized that word histograms cannot account for polysemy (the same word may represent different meanings) and synonymy (different words may represent same meaning) [14, 58]. This led to the introduction of intermediate latent representations, commonly known as “themes” or “topics”. Borrowing from the text literature, several authors applied the idea of latent spaces to visual recognition [12, 4, 129, 140, 77, 114, 17]. The rationale is that images which share frequently co-occurring features have a similar representation in the latent space. Although successful for text, the benefits of topic discovery have not been conclusively established for visual recognition. In fact, a drop in classification performance is often experienced when unsupervised latent representations are introduced [83, 114, 74].

This issue is discussed in detail in the next chapter, where we argue that unsupervised topic discovery is not a good idea for recognition and show that the architecture now proposed can be interpreted as a modified topic model, where the topics are pre-specified and learned in a weakly supervised manner. This is shown to increase the recognition performance.

The use of appearance based classifier outputs as feature vectors has also been proposed in [120, 169, 147]. In these works a classifier is first learned for a given keyword vocabulary — [169, 147] learn discriminative classifiers from flickr/bing images, [120] learns a generative model using a labeled image set — and the outputs of these classifiers are then used as feature vectors for a second layer of classification. In these works, classifier outputs are simply used as an alternative low dimensional image representation, without any analysis of their ability to model context. We discuss the limitations of using appearance models for context modeling and introduce “contextual models” that address these limitations. We also present extensive experimental evidence supporting the benefits of these higher level models, and show that they achieve higher classification accuracies on benchmark datasets.

6.3 Semantics-based Models and Context Multinomials

6.3.1 Limitations of Semantic Representations

One major source of difficulties is that semantic models built upon the BoF representation of appearance inherit the ambiguities of the latter. There are two main types of ambiguity. The first is that contextually unrelated concepts (for example smoke and clouds) can have similar appearance representation under BoF. The second is that the resulting semantic descriptors can account for contextual frequencies of co-occurrence, but not true contextual dependencies. These two problems are illustrated in Figure 6.1. First, image patches frequently have ambiguous interpretation. When considered in isolation, they can be compatible with

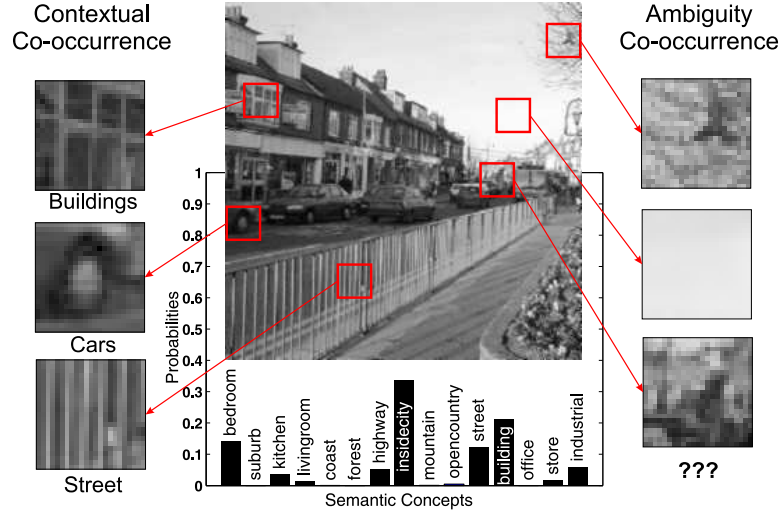


Figure 6.1: An image from the “street” class of the N15 dataset (See 6.4.1) along with its SMN. Also highlighted are the two notions of *co-occurrence*. *Ambiguity co-occurrences* on the right: image patches compatible with multiple unrelated classes. *Contextual co-occurrences* on the left: patches of multiple other classes related to “street”.

many concepts. For example it is unclear that even a human could confidently assign the patches shown on the right of Figure 6.1 to the “street” concept, with which the image is labeled. Second, appearance-based models lack information about the interdependence of the semantics of the patches which compose the images in a class. For example, the fact that, as shown on the left, images of street scenes typically contain patches of street, car wheels, and building texture.

We refer to these two observations as *co-occurrences*. In the first case, a patch can accidentally co-occur with multiple concepts (the equivalent to *polysemy* in text analysis). In the second, patches from multiple concepts typically co-occur in scenes of a given class (the equivalent to *synonymy* for text). While only the co-occurrences of the second type are indicative of *true* contextual relationships, SMNs learned from appearance models capture *both* types of co-occurrences. This is again illustrated by the example of Figure 6.1. On one hand, the displayed SMN reflects the *ambiguity* that sometimes exists between patches of “street scenes” and “bedrooms”, “kitchens” or “living rooms”. These are all man-made structures

which, for example, contain elongated edges due to buildings, beds, furniture, etc. Note that all classes that typically do not have such structures (e.g. natural scenes such as “mountain”, “forest”, “coast”, or “open country”) receive close to zero probability. On the other, the SMN reflects the likely co-occurrence, in “street scenes”, of patches of “inside city”, “street”, “buildings”, and “highway”. In summary, while SMN probabilities can be interpreted as semantic features, which account for co-occurrences due to both ambiguity and context, they are not purely *contextual features*.

One possibility to deal with the ambiguity of the semantic representation is to explicitly model contextual dependencies. This can be done by introducing *constraints* on the appearance representation, by modeling constellations of parts [42, 40] or object relationships [146, 47]. However, the introduction of such constraints increases complexity, and reduces the invariance of the representation, sacrificing generalization. A more robust alternative is to keep BoF, but represent images at a higher level of *abstraction*, where ambiguity can be more easily detected. This is the strategy pursued in this work, where we exploit the fact that the two types of SMN co-occurrences have different *stability*, to extract *more reliable* contextual features.

6.3.2 From Semantics to Context

The basic idea is that, while images from the same concept are expected to exhibit similar contextual co-occurrences, this is not likely for ambiguity co-occurrences. Although the “street scene” of Figure 6.1 contains some patches that could also be attributed to the “bedroom” concept, it is unlikely that this will hold for most images of street scenes. By definition, ambiguity co-occurrences are *accidental*, otherwise they would reflect common semantics of the two concepts, and would be contextual co-occurrences. Thus, while impossible to detect from a single image, stable contextual co-occurrences should be detectable by joint inspection of *all* SMNs derived from the images of a concept.

This is accomplished by extending concept modeling by one further layer of semantic representation. As illustrated in Figure 6.2, each concept k is modeled

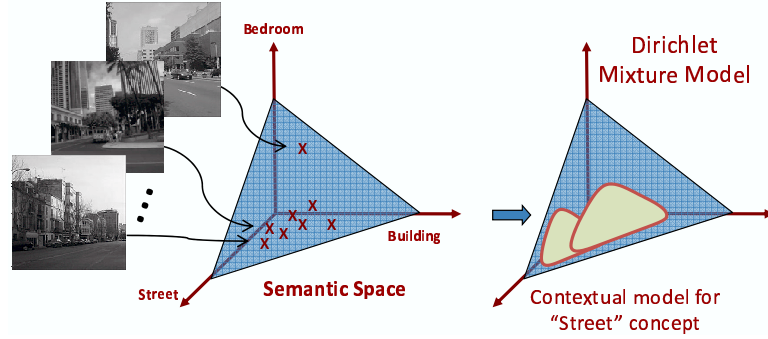


Figure 6.2: Learning the contextual model for the “street” concept, (6.1), on semantic space \mathcal{S} , from the set of all training images annotated with “street”.

by the probability distribution of the SMNs derived from all training images in its training set, \mathcal{D}_k . We refer to this SMN distribution as the *contextual model* for k . If \mathcal{D}_k is large and diverse, this model is dominated by the stable properties of the features drawn from concept k . In this case, the features are SMNs and their stable properties are the true contextual relationships of k . Hence, concept models assign high probability to regions of the semantic space occupied by contextual co-occurrences, and small probability to those of ambiguity co-occurrences.

For example, since streets typically co-occur with buildings, the contextual model for “street” assigns high probability to SMNs that include both concepts. On the other hand, because “street” only co-occurs accidentally with “bedroom”, SMNs including this concept receive low-probability. Hence, representing images by their posterior distribution under contextual models emphasizes contextual co-occurrences, while suppressing accidental coincidences due to ambiguity. As a parallel to the nomenclature of Chapter 2, we refer to the posterior probabilities at this higher level of abstraction as *contextual features*, the probability vector associated with each image as a *contextual multinomial* distribution, and the space of such vectors as the *contextual space*.

6.3.3 Contextual Concept Models

Contextual concept models are learned in the semantic space \mathcal{S} . Under the most general formulation, concepts are drawn from a random variable K defined

on the index set $k \in \{1, \dots, K\}$ of a concept vocabulary \mathcal{K} . In this work, we assume that this vocabulary is the concept vocabulary \mathcal{L} used in visual space \mathcal{X} , i.e. $\mathcal{K} = \mathcal{L}$. Note that this assumption implies that if \mathcal{L} is composed of scenes (objects), then the contextual models account for relationships between scenes (objects). A trivial extension would be to make concepts on semantic space \mathcal{S} different from those on visual space \mathcal{X} , promoting a concept hierarchy. For example, K could be defined on the vocabulary of scenes, $\mathcal{K} = \{\textit{‘desert’}, \textit{‘beach’}, \textit{‘forest’}\}$ and W on objects, $\mathcal{L} = \{\textit{‘sand’}, \textit{‘water’}, \textit{‘sky’}, \textit{‘trees’}\}$. In this way, scenes in \mathcal{K} would be naturally composed of objects in \mathcal{L} , enabling the contextual models to account for relationships between scenes and objects. This would, however, require training images (weakly) labeled with respect to both \mathcal{L} and \mathcal{K} . We do not pursue such hierarchical concept taxonomies in what follows.

Since \mathcal{S} is itself a probability simplex, one natural model for a concept k in \mathcal{S} is the mixture of Dirichlet distributions

$$P_{\Pi|K}(\boldsymbol{\pi}|k; \Lambda^k) = \sum_m \beta_m^k \mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}_m^k). \quad (6.1)$$

This model has parameters $\Lambda^k = \{\beta_m^k, \boldsymbol{\alpha}_m^k\}$, where β_m is a probability mass function ($\sum_m \beta_m^k = 1$). $\mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha})$ a Dirichlet distribution of parameter $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_L\}$,

$$\mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^L \alpha_i)}{\prod_{i=1}^L \Gamma(\alpha_i)} \prod_{i=1}^L (\pi_i)^{\alpha_i - 1} \quad (6.2)$$

and $\Gamma(\cdot)$ the Gamma function. As illustrated in Figure 6.2, the parameters Λ^k are learned from the SMNs $\boldsymbol{\pi}$ of all images in \mathcal{D}_k , i.e. the images annotated with the k^{th} concept in \mathcal{L} . Learning is implemented by maximum likelihood estimation, using the generalized expectation-maximization (GEM) algorithm discussed in Appendix B.

Figure 6.3 shows an example of a 3-component Dirichlet mixture learned for the semantic concept “street”, on a three-concept semantic space. This model is estimated from 100 images (shown as data points on the figure). Note that, although some of the image SMNs exhibit ambiguity co-occurrences with the “forest” concept, the Dirichlet mixture is strongly dominated by the true contextual

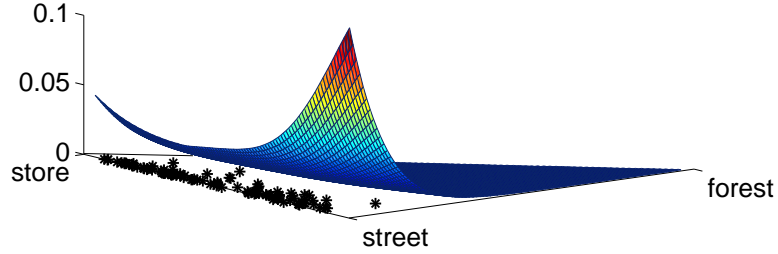


Figure 6.3: 3-component Dirichlet mixture learned for the concept “street”. Also shown, as “*”, are the SMNs associated with each image. The Dirichlet mixture assigns high probability to the concepts “street” and “store”.

co-occurrences between the concepts “street” and “store”. This is an illustration of the ability of the model to lock onto the true contextual relationships.

6.3.4 Contextual Space

The contextual models $P_{\Pi|K}(\boldsymbol{\pi}|k)$ play, in semantic space \mathcal{S} , a similar role to that of the appearance models $P_{\mathbf{X}|W}(\mathbf{x}|w)$ in visual space \mathcal{X} . It follows that MPE concept detection, on a test image \mathcal{I} of SMN $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_L\}$, can be implemented with a Bayes decision rule based on the posterior concept probabilities

$$P_{K|\Pi}(k|\boldsymbol{\pi}) = \frac{P_{\Pi|K}(\boldsymbol{\pi}|k)P_K(k)}{P_{\Pi}(\boldsymbol{\pi})}. \quad (6.3)$$

This is the semantic space equivalent of (2.8) and, once again, we assume a uniform concept prior $P_K(k)$.

As in Chapter 2, it is also possible to design a new semantic space, by retaining all posterior contextual concept probabilities $\theta_k = P_{K|\Pi}(k|\boldsymbol{\pi})$. We denote the vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^T$ as the *contextual multinomial* (CMN) distribution of image \mathcal{I} . As illustrated in Figure 6.4, CMN vectors lie on a new probability simplex \mathcal{C} , here referred to as the *contextual space*. In this way, the contextual representation establishes a mapping from images in \mathcal{X} to CMNs $\boldsymbol{\theta}$ in \mathcal{C} . In 6.4 we show that CMNs are much more reliable contextual descriptors than SMNs.

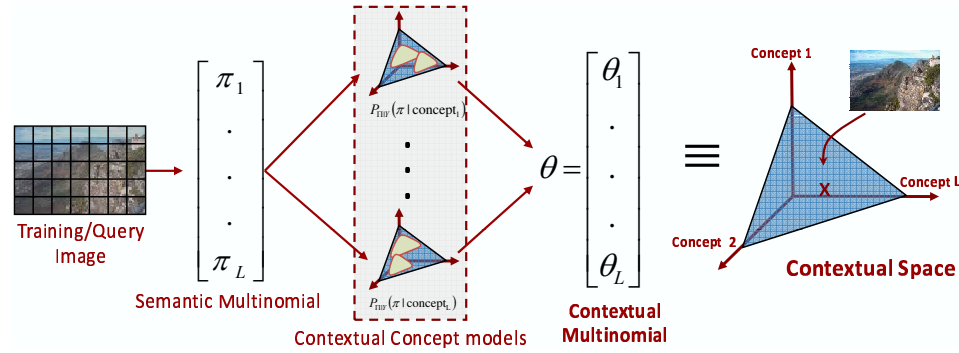


Figure 6.4: The Contextual multinomial (CMN) of an image as the vector of co-occurrence probabilities of contextually related concepts.

6.3.5 Data Augmentation

It should be noted that, similar to learning the semantic representation, this architecture is generic, in the sense that any appearance recognition system that produces a vector of posterior probabilities π , can be used to learn the proposed contextual models. However, when as above, an SMN is computed per image, the number of training images upper bounds the cardinality of the training set for contextual models. Since there is usually a limited number of labeled images per concept, this can lead to over fitting. For example, the 100 images available per concept on N15 are sufficient to learn appearance models (each image contains thousands of patches), but 100 SMNs do not suffice to learn Dirichlet mixtures in a 15 dimensional space. One possibility is to use the patch-SMNs, $\pi^{(n)}$ (see Section 2.3), which are abundant. These, however, tend to be too noisy, due to the ambiguities discussed above. To overcome this problem we resort to a middle ground between patch-SMNs and image-SMNs: multiple SMNs are estimated per image, from random patch subsets. More precisely, a set of patches is first selected, randomly, from the image. An SMN is then estimated from this set, as would be done if the image consisted of these patches alone. The process is repeated with different patch subsets, generating a number of SMNs per image. By controlling the number of random sets, it is possible to control the cardinality of the training set for each contextual model. The use of random patch subsets simultaneously alleviates the problems of data scarcity (many subsets can be drawn per image),

and estimation noise (each SMN pools information from multiple patches). Moreover, similar to the learning of appearance models, learning contextual models with data augmentation also relies on the multiple instance learning paradigm where each image, being a collection of SMNs, serves as the positive bag, with some SMNs depicting true contextual co-occurrences and some others ambiguity co-occurrences. In 6.5.1, we show that this data augmentation strategy leads to significant improvements in classification accuracy.

6.4 Experimental Setup

In this section, we describe the experimental setup used to evaluate performance of the proposed contextual modeling. The evaluation consists of two vision tasks, viz. scene classification and image retrieval.

6.4.1 Datasets

To test the proposed contextual modeling framework, we adopt datasets previously used in the scene classification and image retrieval literatures.

Scene Classification

Scene classification results are presented for two publicly available datasets viz. “Natural Scene Categories” and “Corel Image Collection”.

Natural Scene Categories (N15, N13, N8) We present results on all three subsets of the “Natural Scene Categories” dataset, viz. Natural15 (N15), Natural13 (N13) and Natural8 (N8). These dataset allows direct comparison with published results on scene classification. To learn the concept models, 100 images per scene are used, the remaining being used as test set. All experiments are repeated six times, with random train/test splits. A detailed description of these datasets are provided in Appendix. A.1.1.

Corel Image Collection (C50, C43) We also present results of the “Corel Image Collection” which has much higher number of classes as compared to the “Natural Scene Categories” dataset. We construct two different datasets from this

collection, viz. Corel50 (C50) and Corel43(C43) with 50 and 43 classes respectively. For C50, 90 images from each CD are used to learn class models and the remaining for testing. For C43, 90 images per label are used to learn the class models and the remainder are used for testing. All images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space. A detailed description of these datasets are provided in Appendix. A.1.3.

Image Retrieval

To evaluate retrieval performance, we use two datasets introduced in [119]. **Corel Image Collection (C15)** consists of 1,500 images from another 15 Corel Stock Photo CDs, divided into a retrieval set of 1,200 images and a query set of 300 images. CD themes are used as the ground truth image concepts, creating a 15-dimensional semantic and contextual space. A detailed description of C15 is provided in Appendix. A.1.3.

Flickr Images (F18) consists of 1,800 images from `www.flickr.com` divided into 18 classes resulting in an 18 dimensional semantic and contextual space. A set of 1,440 images serves as the retrieval dataset, and the remaining 360 as the query set. A detailed description of F18 is provided in Appendix. A.1.4.

Note that, for all datasets except C43, each image is explicitly annotated with just one concept, even though it may depict multiple. Thus, the co-occurrence information learned from these datasets is purely data driven. In C43, although multiple annotations are available per image, their co-occurrences are not explicitly used to learn context. In summary, no high level co-occurrence information is used to train the contextual models.

6.4.2 Appearance Features

Both SIFT and DCT features are used for appearance representation. SIFT features are computed either by interest point detection, SIFT-INTR, or on a dense regular grid SIFT-GRID. The two strategies yield about 1000 samples per image. DCT features are computed on a dense regular grid, with a step of 8 pixels. 8×8 image patches are extracted around each grid point, and 8×8

Table 6.1: Impact of inference model on classification accuracy.

Model	Classification Accuracy (%)		
	Appearance	Contextual	
		Image	RandomPatch
Figure 2.1, Eq (2.8)	71.67 ± 1.17	71.67 ± 1.17	-
Figure 2.5(a), Eq (2.21)	71.67 ± 1.17	73.33 ± 0.69	77.20 ± 0.39
Figure 2.5(b), Eq (2.23)	54.97 ± 0.58	73.43 ± 0.99	75.14 ± 0.75

DCT coefficients computed per patch and color channel. For monochrome images this results in a feature space of 64 dimensions. For color images the space is 192 dimensional. In this case, appearance distributions are learned in the 129 dimensional subspace composed of the first 43 DCT coefficients from each channel. For datasets exclusively comprised of color images, only the DCT features are used.

6.5 Results

A number of classification experiments were performed (N15 dataset) to evaluate the impact of the various parameters of the proposed contextual representation on recognition performance.

6.5.1 Designing the Semantic Space.

In Section 2.3, we discussed three strategies to compute Image-SMNs. 6.1 reports their classification accuracy, for both appearance and contextual modeling with SIFT-GRID. Contextual models learned from SMNs computed with (2.8) fail to improve upon the (already high performing) appearance classifiers. This is not totally surprising, since these SMNs lack co-occurrence information (see discussion of Section 2.3). In comparison, SMNs computed with (2.21) or (2.23) are rich in such information, enabling contextual models to outperform their appearance counterparts.

Note that, although the LDA-like inference algorithm of (2.23) yields significantly lower classification performance at the appearance level than that of (2.21), both strategies attain a classification accuracy of $\sim 73.3\%$ at the contextual level. Note also that, despite much weaker performance at appearance-level than (2.8), (2.23) performs substantially better at the contextual level. Together, these results suggest that the recognition performance at the appearance level is not necessarily a good predictor of performance at the contextual level. In particular, the relative performances of the three inference procedures advise against inference procedures that make hard decisions at the lower levels of recognition.

To increase the cardinality of the training sets used for contextual modeling, 800 random sets of 30 patches are sampled per image, yielding 800 patch-SMNs per image. Image-SMNs are then computed from these, with (2.21) or (2.23). 6.1 reports the benefits of this data augmentation, showing that performance improves in both cases. For (2.21) classification accuracy improves from 73.33% to 77.20%, for (2.23) from 73.43% to 75.14%. Since (2.23) involves an iterative procedure, which is more expensive than the closed form of (2.21), and has weaker performance, we use (2.21) in the remaining experiments.

6.5.2 Number of Mixture Components

Figure 6.5(a) presents the classification performance as a function of the number of contextual mixture components, for SIFT-GRID, SIFT-INTR and DCT features. In all cases, a single Dirichlet distribution is insufficient to model the semantic co-occurrences of N15. As the number of mixture components increases from 1 to 8, performance rises substantially for SIFT (e.g. from 72.58% to 76.13% for SIFT-GRID), and dramatically (from 55.93% to 70.48%) for the DCT. Above 8 components, the gain is moderate in all cases, with a maximum accuracy of 77.20% for SIFT-GRID and 73.05% for the DCT. Figure 6.6 shows the cluster centers learned with a four-component Dirichlet mixture using DCT features, for the “street” and “forest” classes. These cluster centers can be interpreted as the SMNs of the dominant co-occurrence patterns learned for these classes. Two interesting observations can be made. First, the class mixtures indeed account for different

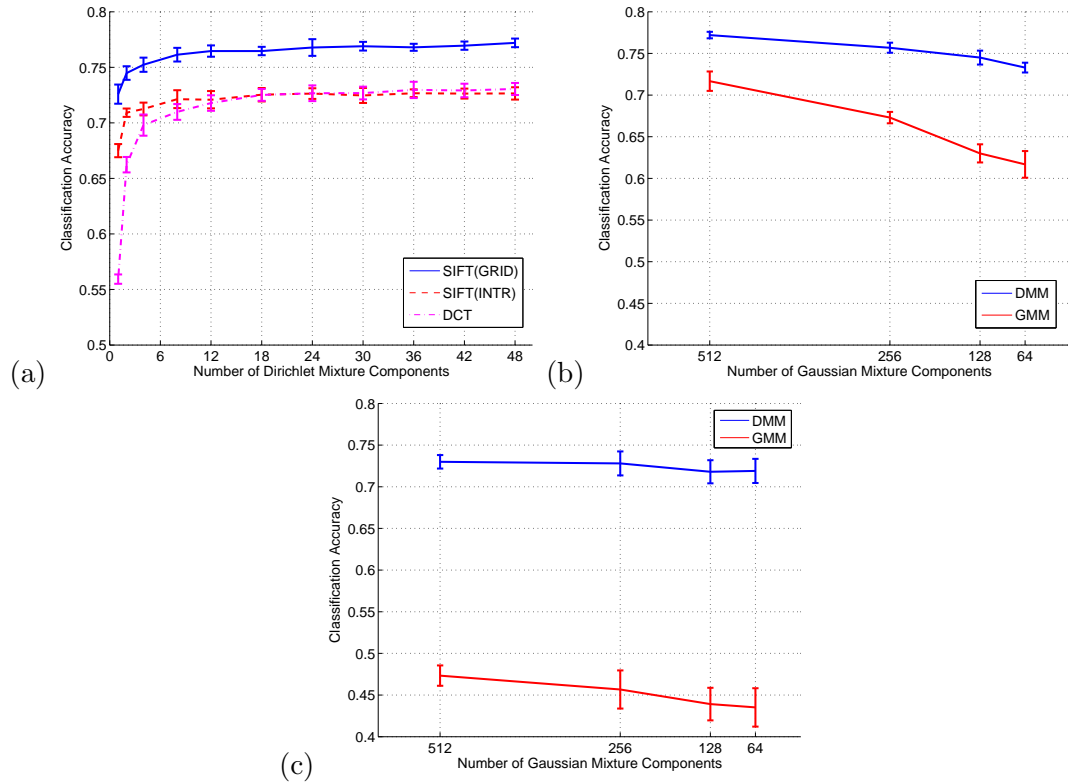


Figure 6.5: (a) Classification accuracy as a function of the number of mixture components of the contextual class distributions, for both DCT and SIFT. (b) Dependence of appearance and contextual classification on the accuracy of the appearance modeling for SIFT-GRID features, (c) for DCT features. The performance of contextual classification remains fairly stable across the range of appearance models.

co-occurrence patterns: in both cases the four cluster centers are quite distinct. Second, not all cluster centers assign high probability to the feature vector which is namesake of the class. In the “street” example, although one of the centers assigns high probability to the “street” concept, the remaining ones assign higher probability to alternative concepts, e.g. “tall building”, “inside city”, “highway” etc. than to “street” itself.

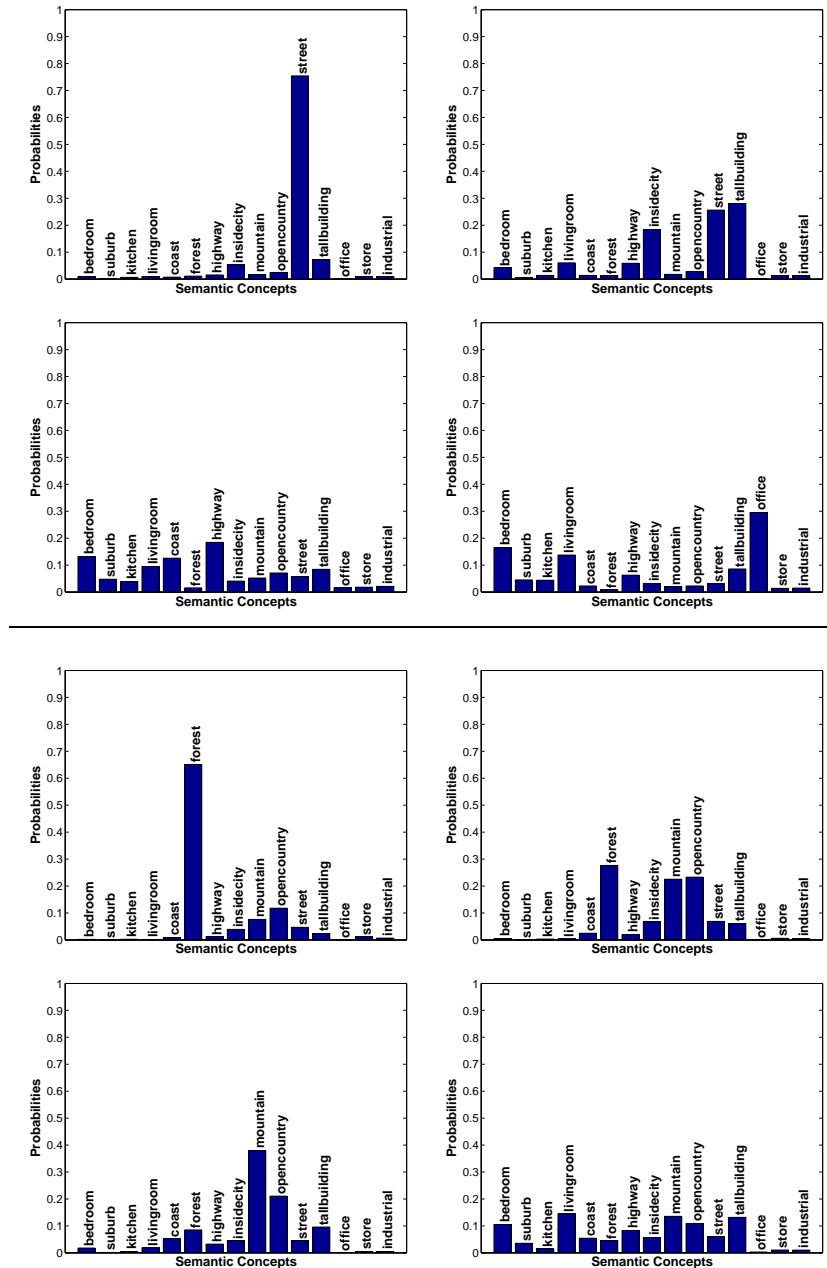


Figure 6.6: Four cluster centers for the class “street” (top) and “forest” (bottom). Note that each class comprises different co-occurrence patterns.

6.5.3 Choice of Appearance Features

6.2 compares the classification performance of the three appearance representations. In all cases, the contextual models yield improved performance, with a

Table 6.2: Impact of appearance space on classification accuracy.

Feature	Classification Accuracy (%)		Gain
	Appearance Models	Contextual Models	
SIFT-GRID using (2.21)	71.67 ± 1.17	77.20 ± 0.39	7.7%
SIFT-GRID using (2.23)	54.97 ± 0.58	75.14 ± 0.75	36.7%
SIFT-INTR	68.58 ± 0.41	72.65 ± 0.56	5.9%
DCT	47.33 ± 1.22	73.05 ± 0.54	54.3%

gain of 7.7%, 5.9% and over 54% for SIFT-GRID, SIFT-INTR and DCT, respectively. Note that the contextual models achieve high performance (over 72%) for *all* appearance features. More interestingly, this performance is almost unaffected by that of the underlying appearance classification, in the sense that very large variations in the latter lead to relatively small differences in the former.

This hypothesis was studied in greater detail, by measuring how contextual-level performance depends on the “quality” of the appearance classification. The number of Gaussian components in the appearance models was the parameter adopted to control this “quality”. Figure 6.5(b) and (c) shows that decreasing this parameter leads to a *substantial* degradation of appearance-level recognition, for both SIFT and DCT. Nevertheless, the performance of the contextual classifiers, built with these appearance classifiers, *does not change substantially*. On the contrary, the contextual classifiers assure a classification gain that *compensates* for the losses in appearance classification. For SIFT-GRID, this gain ranges from about 20% at 64 Gaussian mixture components, to about 8% at 512. For the DCT, corresponding gains are of 65% and 54% respectively. In result, while the appearance classifier experiences a drop of 17% (21%) for DCT (SIFT-GRID) as the number of components is reduced from 512 to 64, the performance of contextual classification drops by only a small margin of 2% (5%).

Overall, the performance of the contextual classifier is not even strongly

affected by the feature transformation adopted. While, at the appearance level, the performance of the DCT is not comparable to that of SIFT, the choice of transform is much less critical when contextual modeling is included: the two transforms lead to similar performance at the contextual level. This suggests that 1) any reasonable architecture could, in principle, be adopted for appearance classification, and 2) there is no need for extensive optimization at this level. This is an interesting conclusion, given that accurate appearance classification has been a central theme in the recognition literature over the last decades.

6.5.4 Some Examples

The ability of contextual modeling to compensate for classification noise at the appearance level can be observed by simple inspection of the posterior distributions at the two levels. Figure 6.7 shows two images from the “street” class of N15, and an image each from the “Ireland” and “Mayan ruins” CD of the Corel Collection. The SMN and CMN vectors computed from each image are shown in the second and third column, respectively. Two observations can be made. First, as discussed in 6.3.1, the SMN vectors can include substantial *contextual noise*, reflecting *both* types of concept co-occurrences. For example, patches from the first image (“street” class) have high probability under concepts such as “bedroom”, “livingroom”, “kitchen”, “inside city”, “tall building”. Some of these co-occurrences (“bedroom”, “livingroom”, “kitchen”) are due to patch ambiguities. Others (“inside city”, “tall building”) are consistent with the fact that the concepts are contextually dependent. The SMN representation has no power to disambiguate between the two types of co-occurrences. This is more pronounced for larger semantic spaces: the SMNs of Corel images (43 dimensional space) exhibit much denser co-occurrence patterns than those of N15.

Second, CMNs are remarkably noise-free for all semantic spaces considered. They capture the “gist” of the underlying scenes, assigning high probability only to truly contextual concepts. This increased robustness follows from the fact that contextual models learn the statistical structure of the contextual co-occurrences that characterize *all* SMNs associated with each class. This makes class models

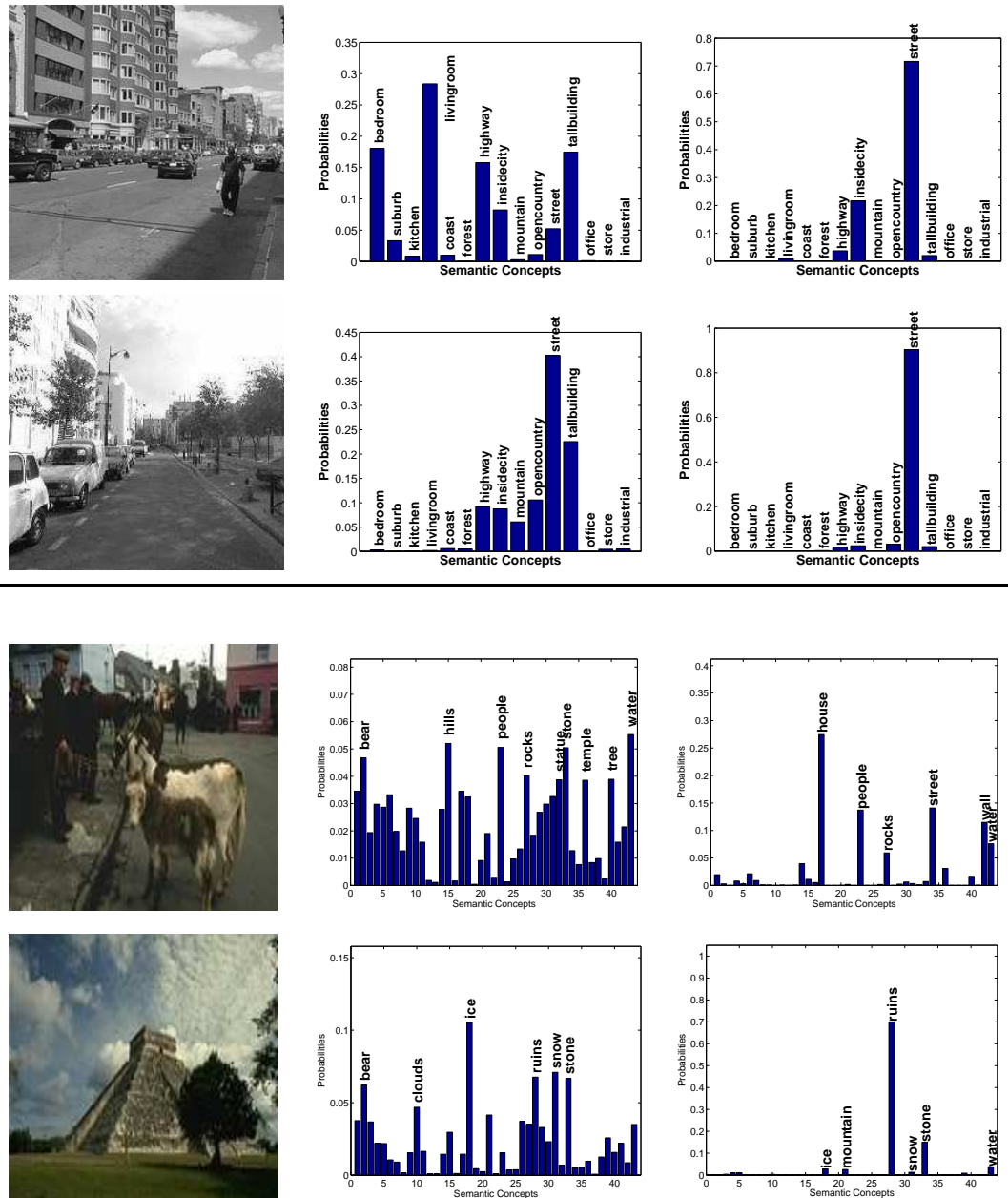


Figure 6.7: top) Two images from the “street” class of N15, and bottom) an image each from the “Ireland” and “Mayan ruins” CD of the Corel collection. Also shown with the images are the SMN and CMN vectors (middle and right column respectively). Notice that the CMN vectors are noise-free and capture the “gist” of the image.

at contextual level mitigate ambiguity co-occurrences, which tend to be spurious, while accentuating true contextual co-occurrences, which are stable. Consider, for example, the image in the third row. Its SMN is a frequently occurring training example for contextual models of “street”, “house”, “people” (this is true even though the image has low probability of “street” and “house” under appearance modeling), etc. On the other hand, it is an unlikely training pattern for contextual models of “bear” and “hills”, which only accidentally co-occur with “street” or “house”. Hence, this SMN has large posterior probability under contextual models for “house” and “street”, but not for “bear” or “hills”.

6.5.5 Complexity

In this section we report approximate running times for training and testing, under both the appearance and contextual class models. All experiments are conducted on an 2x Intel Xeon E5504 Quad-core 2.00GHz processor, with average image size of 270×250 pixels. Learning of appearance models requires computing SIFT/DCT features, which takes about 800/20ms per image respectively. Given these features, 512 component Gaussian mixture models are learned from 100 training images in about 3 minutes per class, using the hierarchical approach of [159]. For testing, computing the likelihood of a given image requires about 50ms per class. These likelihoods serve as features for the contextual models. A 42 component Dirichlet mixture model, learned from 100 training images, with 800 SMNs per image, requires about 2 minutes to learn. During testing, it takes about 30ms to compute the likelihood of an image under each contextual class model.

6.6 Comparison with Previous Work

In this section we compare the proposed contextual recognition with existing solutions to scene classification and image retrieval.

Table 6.3: Classification Results on Natural Scene Categories.

Method	Classif.	Dims. ^a	Accuracy (%)
	N15 Dataset		
Contextual Models	Bayes	15	77.20 ± 0.39
pLSA [17] ^b	SVM	40	72.7
pLSA [74]	SVM	60	63.3
LDA [77] ^e	Bayesian	40	59.0
“gist” like [74]	SVM	16	45.3 ± 0.5
BoW [74]	SVM	400	74.8 ± 0.3
BoW [74]	SVM	200	72.2 ± 0.6
Bag of Concepts [83] ^c	SVM	100	73.01
Kernel Codebook [154]	SVM	3200	~75 ^d
Diffusion Distance [82]	SVM	2000	74.9
SIS [24]	SVM	200	74.94
Semantic Space [120]	SVM	15	73.95 ± 0.74

^a Dimensionality of the space on which classification is performed

^b Uses half of the dataset for training

^c Uses a subset of test images per concept

^d Accuracy estimated from figure

^e Our implementation of the algorithm

Table 6.4: Classification Results on Natural Scene Categories.

Method	Classif.	Dims. ^a	Accuracy (%)
N13 Dataset			
Contextual Models	Bayes	13	80.86 ± 0.50
LDA [77]	Bayesian	40	65.2
pLSA [17] ^b	SVM	35	74.3
pLSA [114]	SVM	40	60.8
pLSA [74]	SVM	60	65.9
BoW [74]	SVM	200	74.7
Taxonomy [6]	Bayesian	40	68
“gist” features [65]	SVM	512	~55 ^c
Semantic Space [120]	SVM	13	77.57 ± 1.12

^a Dimensionality of the space on which classification is performed

^b Uses half of the dataset for training

^c Accuracy estimated from figure

Table 6.5: Classification Results on Natural Scene Categories.

Method	Classif.	Dims. ^a	Accuracy (%)
N8 Dataset			
Contextual Models	Bayes	8	85.60 ± 0.70
Context Ancestry [80]	Logistic	484	82
pLSA [17] ^b	SVM	25	82.5
HDP-HMT [67]	Bayesian	200	84.5
“gist” [104] ^c	SVM	512	83.7
Semantic Space [120]	SVM	8	84.24 ± 0.71

^a Dimensionality of the space on which classification is performed

^b Uses half of the dataset for training

^c Gist features implicitly uses weak spatial information

Table 6.6: Classification Results on Corel Collection.

Method ^a	Classif.	Dims.	Accuracy (%)
C50 Dataset			
Contextual Models	Bayes	50	57.8
Appearance Models	Bayes	129	53.6
Bag of Words [74]	SVM	512	48.4
pLSA [17]	SVM	50	40.2
LDA [77]	Bayes	50	31.0
C43 Dataset			
Contextual Models	Bayes	43	42.9
Appearance Models	Bayes	129	39.9
Bag of Words [74]	SVM	512	36.3
pLSA [17]	SVM	50	33.0
LDA [77]	Bayes	50	24.6

^a Our implementation of the algorithms

6.6.1 Scene Classification

Given the posterior probabilities of (6.3), MPE scene classification can be implemented by application of Bayes rule. This consists of assigning image \mathcal{I} , of SMN π , to the scene class k of largest posterior $P_{K|\Pi}(k|\pi)$. 6.3, 6.4 and 6.5 compare the resulting classification accuracies for N15, N13, and N8 respectively, with those of many methods in the literature. A number of observations can be made from the table. First, contextual modeling achieves the best results on all three datasets. Its performance is quite superior to that of topic discovery models (LDA [77], pLSA [17, 114]), of which only [17] is remotely competitive. Even so, the classification rates of the latter (72.7% on N15, 74.7% on N13, and 82.5% on N8) are well below those of the former (77.2%, 80.86%, and 85.6%). Somewhat closer to this (74.8% on N15, 74.7% on N13) is the performance of SVMs with the BoW

representation¹. Note, however, that these require much higher dimensional spaces, e.g. a 400 visual-word vocabulary [74], and storage of a number of support vectors that grows with the number of classes and training examples. Contextual modeling has lower dimensionality, lower complexity, and achieves a higher classification accuracy². Also reported is a baseline with discriminative learning [120] where an SVM classifier is applied to the vector of outputs of the appearance classifiers. Again, the proposed context models achieve superior classification performance on all datasets.

Within the area of context modeling, e.g. comparing to the methods of [104, 80], the proposed approach is again more effective. For the N8 (N13, N15) dataset, [104] ([65], [74]) report a classification accuracy of 83.7% (55%, 45.3%³), respectively, using the “gist” features of [104]. The corresponding figures for the proposed contextual models are 85.6% (80.86%, 77.2%). The scene confusion matrix for N15 is also shown in Figure 6.8. Note that most errors are due to confusion between “coast” and “open country,” “living room” and “bed room,” or “living room” and “kitchen.” These are very tolerable errors, given the similarity of scenes in these classes. In fact, their images are sometimes difficult to discriminate even for a human.

Finally, 6.6 presents classification results for the C50 and C43 datasets. Contextual modeling again improves on the classification accuracy achievable with appearance classifiers. For C50 the absolute gain is of 4.2%, for C43 of 3%. When compared to the top performing published methods on the natural scene dataset [74, 17] the proposed contextual modeling again achieves significantly

¹Note that BoW representation is obtained by vector quantizing the space of descriptors and representing an image with a visual word histogram.

²We note that better results have been reported for an extension of the BoW representation that includes a weak encoding of spatial information [74, 179]. These results are the current state-of-the-art for N15: 81.4% [74] using a SVM classifier on an 8400 dimensional space; 85.2% [179] using a nearest neighbor classifier on an 8192 dimensional space. Note that the performance of these approaches without the additional spatial encoding is 74.8% and 75.8%, respectively, which is well below the 77.2% achieved by the proposed contextual models. Although contextual classification could also be augmented with weak encoding of spatial information — one possibility is to learn contextual class models for different image sub-regions and model the overall contextual class model as a mixture of these sub-region models — it remains to be determined if the gains would be as large as for the BoW representation. We leave this as a topic for future work.

³Using a 16 dimensional “gist” like feature instead of the commonly used 512 dimensions.

	office	kitchen	livingroom	bedroom	store	industrial	tallbuilding	insidecity	street	highway	coast	opencountry	mountain	forest	suburb
office	.83	.07	.03	.06	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
kitchen	.06	.71	.11	.08	.02	.02	.00	.00	.00	.00	.00	.00	.00	.00	.00
livingroom	.05	.09	.60	.11	.07	.08	.00	.00	.00	.00	.00	.00	.00	.00	.01
bedroom	.03	.06	.19	.55	.03	.08	.00	.03	.00	.00	.00	.00	.01	.00	.01
store	.01	.03	.07	.00	.75	.10	.00	.03	.00	.00	.00	.00	.00	.00	.00
industrial	.00	.02	.05	.05	.12	.63	.04	.03	.00	.01	.01	.00	.01	.00	.02
tallbuilding	.00	.01	.01	.00	.01	.06	.82	.05	.01	.00	.00	.00	.01	.01	.00
insidecity	.00	.00	.00	.00	.01	.03	.05	.77	.06	.05	.00	.00	.00	.00	.00
street	.00	.00	.00	.00	.02	.04	.03	.87	.03	.00	.00	.02	.00	.00	.00
highway	.00	.00	.00	.01	.01	.03	.00	.01	.03	.86	.03	.02	.01	.00	.00
coast	.00	.00	.00	.00	.00	.00	.00	.00	.02	.83	.11	.03	.00	.00	.00
opencountry	.00	.00	.00	.00	.00	.00	.00	.00	.03	.13	.71	.08	.04	.00	.00
mountain	.00	.00	.00	.00	.00	.00	.01	.00	.01	.01	.07	.88	.01	.00	.00
forest	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.07	.05	.88	.00	.00
suburb	.01	.00	.02	.00	.03	.01	.00	.00	.00	.00	.00	.01	.00	.93	.00

Figure 6.8: Class confusion matrix for classification on the N15 dataset. The average accuracy is 77.20%

higher accuracy. On C50, its accuracy is 57.8% while [74] and [17] achieve classification rates of 48.4% and 40.2%, respectively. On C43, the corresponding numbers are 42.9%, 36.3%, and 33.0%. Overall, it can be concluded that the proposed contextual modeling consistently outperforms existing context-based scene classification methods in the literature.

6.6.2 Image Retrieval Performance

Finally, the benefits of holistic context modeling were evaluated on the task of content based image retrieval, using the query-by-example paradigm. This is a nearest-neighbor classifier, where a vector of global image features extracted from a query image is used to retrieve the images of closest feature vector in an image database. In Chapter 3, we have shown that state-of-the-art results for this type of operation are obtained by using appearance-level posterior distributions (SMNs) as feature vectors. In this work, we compare results of using the distributions obtained at the contextual (CMN) and appearance (SMN) levels. The similarity between the distributions of the query and database images is measured with the Kullback-Leibler divergence [119].

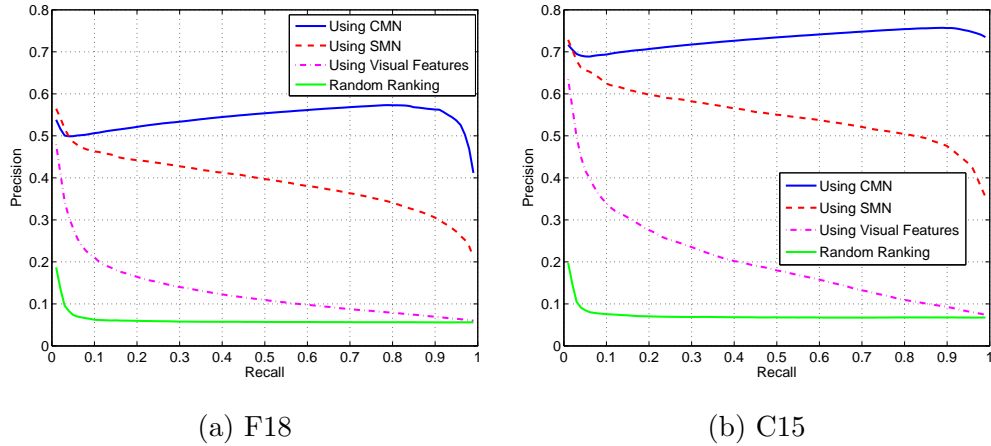


Figure 6.9: Precision-recall curves achieved with SMN, CMN, visual matching and chance level image retrieval.

Figure 6.9, presents precision-recall (PR) curves on C15 and F18. Also shown are the performance of the image matching system of [156], which is based on the MPE retrieval principle now used but does not rely on semantic modeling, and chance-level retrieval. Note how the precision of contextual modeling is *significantly* superior to those of the other methods at *all* levels of recall. For example, on C15, the mean-average precision (area under PR curve) of CMN (0.73) is 32% higher than that of SMN (0.55). The respective figures for F18 are 0.54 and 0.39, a gain of over 38%. Overall, the PR curves of CMN are remarkably flat, attaining high precision at high levels of recall. This is unlike any other retrieval method that we are aware of. It indicates very good generalization: while most retrieval approaches (even image matching) can usually find a few images in the class of the query, it is much more difficult to generalize to images in the class that *are not* visually similar to the query.

Figure 6.10 illustrates the improved generalization of contextual modeling. It presents retrieval results for the three systems (top three rows of every query show the top retrieved images using visual matching, SMN, and CMN respectively). The first column shows the queries while the remaining columns show the top five retrieved images. Note how visual matching has no ability to bridge the semantic gap, simply matching semantically unrelated images of similar color and texture.

This is unlike the semantic representations (SMN and CMN) which are much more effective at bridging the gap, leading to a much smaller number of semantically irrelevant matches. In particular, the ability of the CMN-based system to retrieve images in the query's class is quite impressive, given the high variability of visual appearance.

6.7 Acknowledgments

The text of Chapter 6, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, '*Holistic Context Models for Visual Recognition*', Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, N. Rasiwasia and N. Vasconcelos, '*Holistic Context Modeling using Semantic Co-occurrences*', IEEE Conference on Computer Vision and Pattern Recognition, Miami, June 2009, and N. Rasiwasia and N. Vasconcelos, '*Image Retrieval using Query by Contextual Example*', ACM Conference on Multimedia Information Retrieval, pp. 164-171, Vancouver, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.

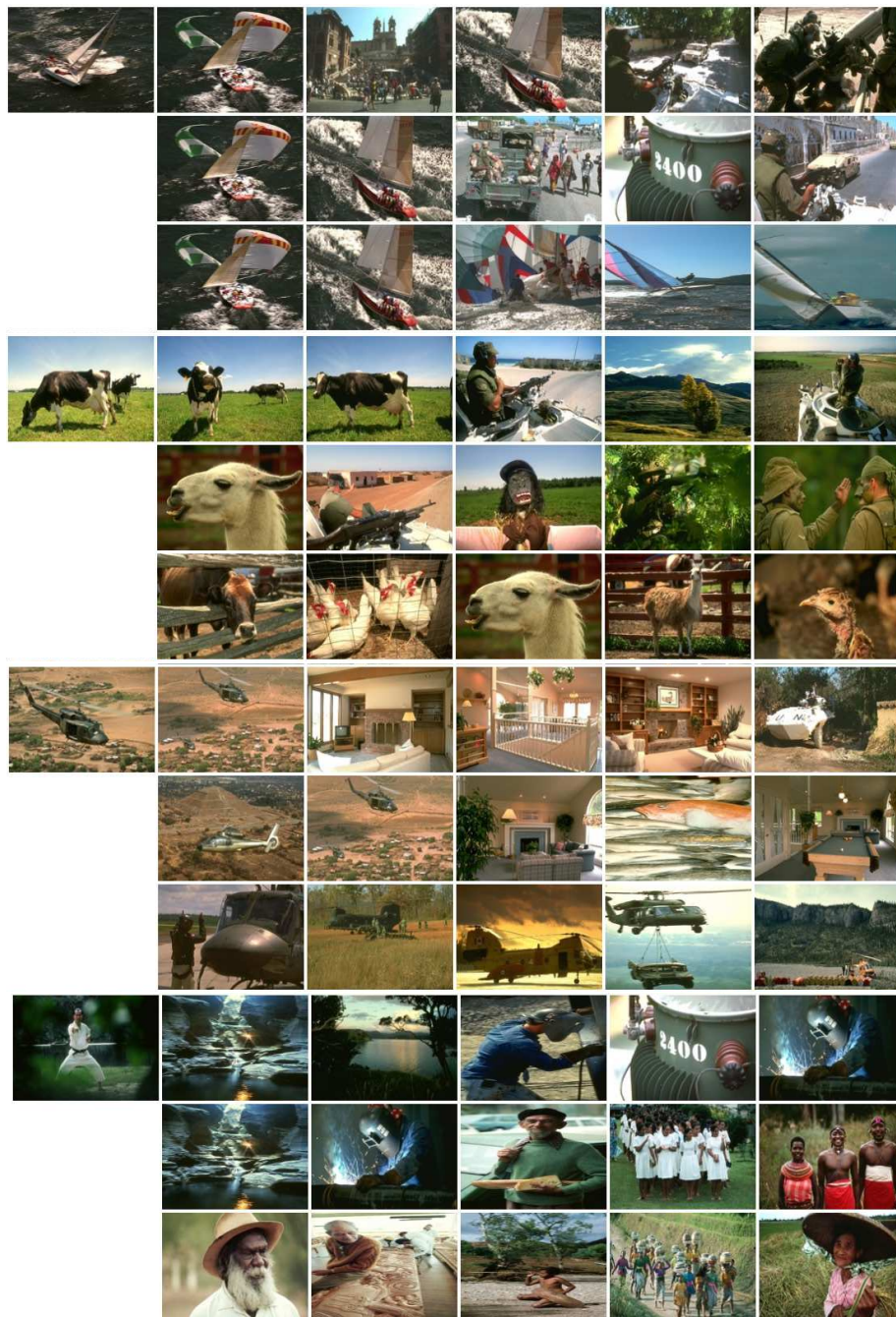


Figure 6.10: Retrieval results for four image queries shown on the left-most column. The first, second, and third row of every query show the five top matches using image matching, SMN, and CMN-based retrieval, respectively.

Chapter 7

The Importance of Supervision

7.1 Introduction

The architecture proposed in Chapter 6 has several properties in common with the family of *theme* or *topic models*, [14, 58]. Topic models were introduced to facilitate the *discovery* of hidden structure in a corpus of data in the text processing literature. Popular examples include latent Dirichlet allocation (LDA) [14] and probabilistic latent semantic analysis (pLSA) [58]. In these models, each entry in a corpus is represented as a finite mixture over an intermediate set of *topics* discovered in an *unsupervised* fashion. However, in their original formulations, topic models do not incorporate supervised information and can not be directly employed for classification.

Several extensions of the LDA model have been proposed to address this limitation in both the text and vision literatures¹. One popular extension is to apply a classifier, such as a SVM, to the topic representation learned by these models [14, 17, 114]. A second approach is to incorporate a class label variable in the generative model [77, 13, 167, 71, 180, 112]. These are denoted generative extensions. Two popular extensions in this family, for scene classification, are that of [77], here referred to as classLDA (cLDA), and [167], commonly known as supervisedLDA (sLDA). The latter was first proposed for supervised text prediction in [13]. Thus, like the representation of holistic context models, topic models for supervised tasks have two layers. Appearance features are used to compute topic probabilities (that correspond to the proposed SMNs), which are hierarchically propagated to a more abstract layer that computes class probabilities (correspondent to the proposed CMNs).

In this chapter, we discuss the generative extensions of the LDA model in context of the proposed holistic context models (see Chapter 6). We start by highlighting the similarities and differences between the cLDA model and the holistic context model. Although the Bayesian network for both these models are very similar, there are fundamental differences, the most important being the level of supervision. Existing generative extensions of LDA such as cLDA and sLDA

¹Note that some of these models were discussed in Chapter 4, however for clarity of the presentation these models are reviewed again in this chapter

rely on unsupervised discovery of topic. This fundamentally restricts their efficacy for the task of visual recognition. This is shown by 1) a theoretical analysis of the learning algorithms, and 2) experimental evaluation on classification problems. Theoretically, it is shown that the impact of class information on the topics discovered by cLDA and sLDA is *very weak* in general, and vanishes for large samples. Experiments show that the classification accuracies of cLDA and sLDA *are not superior* to those of unsupervised topic discovery. Although the holistic context models are effective at addressing this limitation, they have a different learning and inference procedure, which prevent a systematic study of the benefits of supervision in these models. Infact, existing approaches rely on the bag-of-words representation whereas bag-of-features was the choice of image representation in holistic context model (see 2.1.1 for details). In this chapter, to test the benefits of supervision in LDA models, we propose a family of LDA models which we denote as *topic supervised (ts)*. Instead of relying on *discovered* topics, topic-supervised LDA *equates topics to the classes of interest* for scene classification, establishing a one-to-one mapping between topics and class labels. This *forces LDA to pursue semantic regularities in the data*.

Note that the only, subtle yet significant, difference between the existing generative extensions and the proposed topic supervised extensions, is that the topics are no longer *discovered*, but *specified*. Both these systems rely on the same image representation, that of bag-of-words, and the same learning/inference procedures (although as we shall see in 7.5.3, learning in topic supervised models is much more simplified). This enables us to attribute any difference in their performance, to the difference in the level of supervision. It is shown that, topic supervision significantly improves on the classification accuracy of existing supervised LDA extensions. This is demonstrated by the introduction of *topic supervised* versions of LDA, cLDA and sLDA, denoted *ts-LDA*, *ts-cLDA* and *ts-sLDA* respectively. In all cases, the performance of topic supervised models is superior to that of the corresponding LDA models learned without topic-supervision.

The chapter is organized as follows. Section 7.2 briefly reviews the literature on generative models for scene classification. Topic models, in particular cLDA

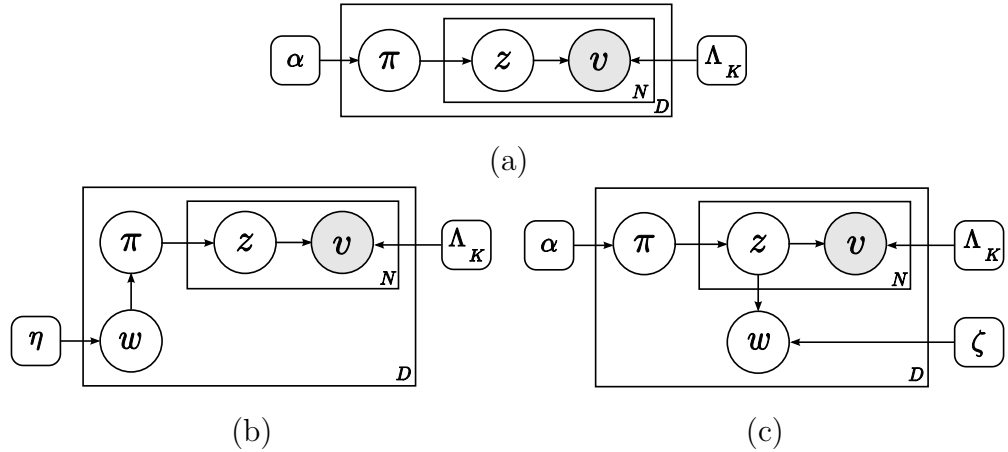


Figure 7.1: Graphical models for (a) LDA and ts-LDA. (b) cLDA and ts-cLDA. (c) sLDA and ts-sLDA. All models use the standard plate notation [19], with parameters shown in rounded squares.

model is compared to the holistic context models in Section 7.3. The limitations of existing models are highlighted in Section 7.4. Next, in Section 7.5 we introduce the topic-supervised model. An extensive experimental evaluation of the proposed frameworks is presented in Sections 7.5.4.

7.2 Topic Models

We start by reviewing LDA and its various generative extensions for classification.

7.2.1 LDA model

LDA is the generative model of Figure 7.1(a). Under it, images are sampled as follows.

for each image **do**

sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.

for $i \in \{1, \dots, N\}$ **do**

sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{L} = \{1, \dots, K\}$, where \mathcal{L} is the set of topics.

```

    sample a visual word  $v_i \sim P_{V|Z}(v_i|z_i; \Lambda_{z_i})$ .
  end for
end for

```

where $P_{\Pi}()$ and $P_{V|Z}()$ are the prior and topic-conditional distributions respectively. $P_{\Pi}()$ is a Dirichlet distribution on \mathcal{L} with parameter $\boldsymbol{\alpha}$, and $P_{V|Z}()$ a categorical distribution on \mathcal{V} with parameters $\Lambda_{1:K}$. Although the parameters of the model can be learned with the well known expectation maximization (EM) algorithm, the E-step yields an intractable inference problem. To address this, a wide range of approximate inference methods have been proposed [11], such as Laplace or variational approximations, sampling methods, etc. In this work, we adopt variational inference for all models where exact inference is intractable. Variational inference for the LDA model is briefly discussed in Appendix D². In its original formulation, LDA does not incorporate class information and cannot be used for classification. We next discuss two models proposed to address this limitation.

7.2.2 Class LDA (cLDA)

ClassLDA (cLDA) was introduced in [77] for image classification. In this model, shown in Figure 7.1(b), a class variable W is introduced as the parent of the topic prior Π . In this way, each class defines a prior distribution in topic space, conditioned on which the topic probability vector $\boldsymbol{\pi}$ is sampled. Images are sampled as follows

```

for each image do
  sample a class label  $w \sim P_W(w; \boldsymbol{\eta})$ ,  $w \in \mathcal{W}$ 
  sample  $\boldsymbol{\pi} \sim P_{\Pi|W}(\boldsymbol{\pi}|w; \boldsymbol{\alpha}_w)$ .
  for  $i \in \{1, \dots, N\}$  do
    sample a topic,  $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$ ,  $z_i \in \mathcal{L} = \{1, \dots, K\}$ .
    sample a visual word  $v_i \sim P_{V|Z}(v_i|z_i; \Lambda_{z_i})$ 
  end for
end for

```

²Note that the variational inference procedure is detailed for the LDA model of Figure 2.5(b), which has notational differences with Figure 7.1(a), but the variational inference procedure is identical.

end for
end for

where, $\boldsymbol{\alpha}_w = \{\alpha_{w1}, \dots, \alpha_{wK}\}$. Parameter learning for cLDA is similar to that of LDA [77] and detailed in Appendix E.

Given image \mathcal{I}_q , classification is performed by MPE decision rule, where the posterior $P_{W|V}(w|\mathcal{I}_q)$ can be approximated using a variational approximation [77].

7.2.3 Supervised LDA (sLDA)

The sLDA model was proposed in [13]. As shown in Figure 7.1(c), the class variable W is conditioned by the topics Z . The original formulation uses unconstrained real-valued response variables W and is not suitable for classification. An extension to discrete responses, using a softmax function, was introduced in [167]. An alternative extension to binary image annotation was proposed in [112], using a multi-variate Bernoulli variable for W . In [180], the max-margin principle is used to train sLDA, which is denoted maximum entropy discrimination LDA (medLDA). In this work, sLDA refers to the formulation of [167], since this was the one previously used for scene classification. Images are sampled as follows

for each image **do**
 sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.
 for $i \in \{1, \dots, N\}$ **do**
 sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{L} = \{1, \dots, K\}$
 sample a visual word $v_i \sim P_{V|Z}(v_i|z_i; \Lambda_{z_i})$.
 end for
 sample a class label $w \sim P_{W|Z}(w|\bar{\mathbf{z}}; \boldsymbol{\zeta}_{1:C})$, $w \in \mathcal{W}$
end for

where, $\bar{\mathbf{z}}$ is the mean topic assignment vector $\bar{\mathbf{z}}_k = \frac{1}{N} \sum_{n=1}^N \delta(z_n, k)$, and

$$P_{W|Z}(w|\bar{\mathbf{z}}; \boldsymbol{\zeta}) = \frac{\exp(\boldsymbol{\zeta}_w^T \bar{\mathbf{z}})}{\sum_{l=1}^C \exp(\boldsymbol{\zeta}_l^T \bar{\mathbf{z}})} \quad (7.1)$$

a softmax activation function with parameter $\boldsymbol{\zeta}_c \in \mathbb{R}^K$. The parameters of this model can be learned with variational inference, as described in [167].

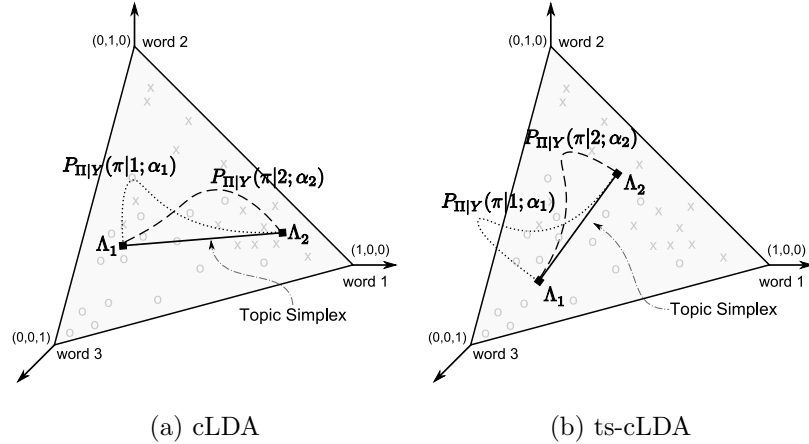


Figure 7.2: Representation of cLDA and ts-cLDA on a three *word simplex*. Also shown are sample images from two classes: “o” from class-1 and “x” from class-2. a) cLDA model with two topics. The line segment depicts a one-dimensional *topic simplex*, whose vertices are topic-conditional word distributions. Each class defines a smooth distribution on the topic simplex, denoted by the contour lines. c) ts-cLDA model. Topic-conditional word distributions are learned with supervision which encapsulate the class attributes.

7.2.4 Geometric Interpretation

The models discussed above have an elegant geometric interpretation [14, 139]. Given a vocabulary of $|\mathcal{V}|$ distinct words, a $|\mathcal{V}|$ dimensional space can be constructed where each axis represents the occurrence of a particular word. A standard $|\mathcal{V}| - 1$ -simplex in this space, here referred to as *word simplex*, represents all probability distributions over words. Each image (when represented as a word histogram) is a point on this space. Figure 7.2(a) illustrates the two dimensional simplex of all probability distributions over three words. Also shown are some sample images from two classes, “o” from class-1 and “x” from class-2.

Figure 7.2(a) shows a schematic of cLDA with two topics. Each topic in an LDA model defines a probability distribution over words and is represented as a point on the word simplex. Since topic probabilities add to one, a set of K topics defines a $K - 1$ simplex, here denoted the *topic simplex*. When the number of topics K is smaller than the number of words $|\mathcal{V}|$, the topics span a low-dimensional

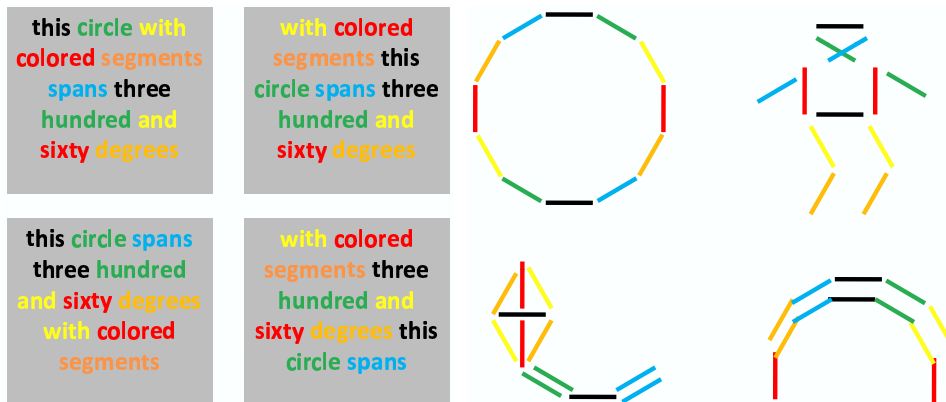


Figure 7.3: left) Four groups of words with equal word histograms. right) Four groups of edge segments with the equal edge segment histograms. Note that each group can be derived from the others by a displacement of words or edge segments. (This figure is best viewed in color)

sub-simplex of the word simplex. The projection of images on the topic simplex can be thought of as *dimensionality reduction*. In Figure 7.2(a), the two topics are represented by Λ_1 and Λ_2 , and span a one-dimensional simplex, shown as a connecting line segment. In cLDA, each class defines a distribution (parameterized by α_w) on the topic simplex. The distributions of class-1 and class-2 are depicted in the figure as dotted and dashed lines, respectively. Similar to cLDA, sLDA can also be represented on the topic simplex, where each class defines a softmax function³.

7.3 The Importance of Supervision

The architecture of holistic context models bear close resemblance to that of cLDA. In Section 2.3, it was shown that SMNs can be computed using the graphical model of Figure 2.5(b). In fact, the graphical model of Figure 2.5(b) is that of LDA. Holistic context models introduce a second layer of modeling, using

³Strictly speaking, the softmax function is defined on the average of the sampled topic assignment labels \bar{z} . However, when the number of features N is sufficiently large, \bar{z} is proportional to the topic distribution π . Thus, the softmax function can be thought of as defined on the topic simplex.

multi-modal Dirichlet distribution, on top of the SMNs obtained using the LDA framework. This is similar in principle to the cLDA model where a uni-modal Dirichlet distribution is introduced. Figure 7.1(b) presents the complete version of this model, including the concept variable W at the semantic level. Given the equivalence of the graphical models, it is worth discussing in detail the differences between the two approaches. The fundamental difference is the *level of abstraction* of the intermediate stage of the representation (topics vs. SMNs). While topics are learned in an unsupervised manner, SMN features have *explicit* semantics.

Recall the *semantic gap* between appearance features and visual classes. While text features (words) are intrinsically semantic, this is *not* the case for vision, where localized appearance features (e.g. edge segments) *have no semantic interpretation*. This is illustrated in Figure 7.3, where we present four groups of text (words) and appearance (edge segments) features *with identical distributions*. Because the word features are semantic, it is very difficult to construct a group (sentence) with the same words that is semantically far from the others. This is absolutely not the case for vision, where equivalence of feature distributions places almost no constraint on the group semantics. As the figure shows, the exact same segments can very easily be used to construct groups that depict completely unrelated concepts. The fact that *equivalence of feature distributions does not translate into semantic equivalence* is denoted a semantic gap.

While the semantic gap is small for text (semantic features), it is large for images. Thus, the success of a representation for text classification is an unreliable predictor of its success for scene classification. In particular, the observation that unsupervised topic discovery produces semantic topics for text [14, 58], is very weak evidence that it will be successful for visual recognition. In fact, Figure 7.3 shows that it cannot. In the absence of explicit supervision for topic semantics, it is impossible to learn that the four edge groupings of (c) belong to different topics. On the contrary, the four groups form a perfect appearance cluster, since their segment histograms *are identical*. Unfortunately, due to the semantic gap, this cluster has no well defined semantics *as a whole*. Hence, unsupervised topic learning has no ability to bridge the semantic gap between local appearance and

visual classes. This is unlike the proposed architecture, where SMN features are learned with explicit supervision, and it does make sense to talk about a *semantic space*.

It should be emphasized that in this toy example, although explicit topic supervision results in four classes of *identical* distribution (a highly suboptimal clustering under any unsupervised learning criteria), it produces the *semantically correct* statistical description of the data under the chosen image representation. Note that, under this model, all images of Figure 7.3(right) have an equal chance of being assigned to any of the classes. This is a classifier of higher probability of error than that learned without supervision. In fact, it is the weakest possible classifier. On the other hand, unsupervised topic modeling produces a much stronger classifier: all images assigned to one class with high probability, other classes mostly noise. In summary, the supervised model reflects *both* the true semantics of the data and the ambiguity of the image representation. It attempts to perform the *right* classification but can only do so with high uncertainty. The unsupervised model *invents* an alternative classification problem, which has nothing to do with the image semantics but *can be* solved very accurately. In addition to producing a semantically useless image description, it is also confident on its accuracy.

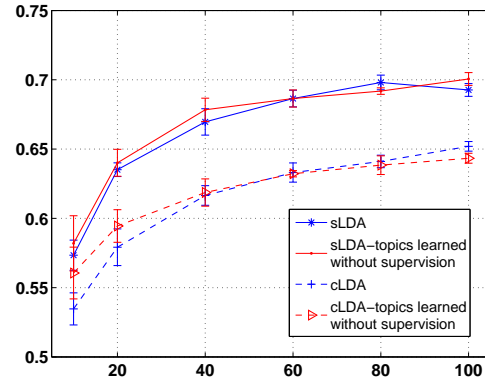
7.4 Limitations of Existing models

In this section we present theoretical and experimental evidence that, contrary to popular belief, topics discovered by sLDA and cLDA are not more suitable for discrimination than those of standard LDA.

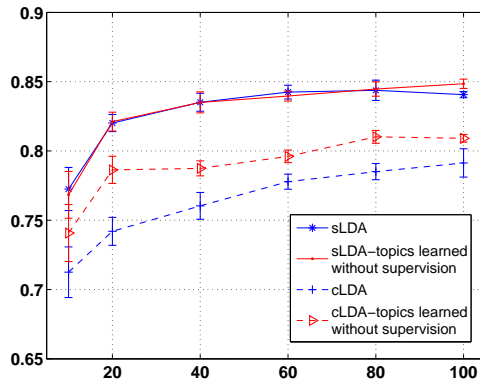
7.4.1 Theoretical Analysis

We start by showing that, in both cLDA and sLDA, the class label has a very weak influence in the learning of topic distributions. This is accomplished by an analysis of the learning equations for both cLDA and sLDA, using the variational approximation framework.

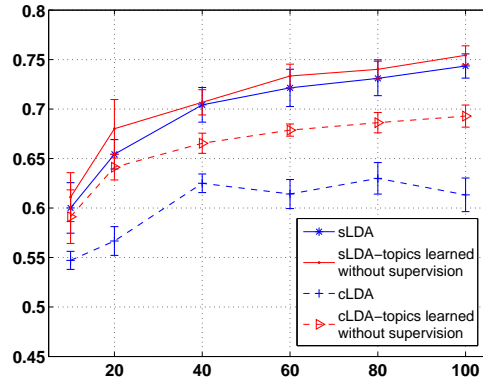
In both sLDA and cLDA the parameters $\Lambda_{1:K}$ of the topic distributions are



(a) N15



(b) N8



(c) S8

Figure 7.4: Classification accuracy as function of the number of topics for sLDA and cLDA, using topics learned with and without class influence and codebooks of size 1024, on (a) N15, (b) N8 and (c) S8. Similar behavior was observed for codebooks of different sizes.

obtained via the variational M-step as:

$$\Lambda_{kv} \propto \sum_d \sum_n \delta(v_n^d, v) \phi_{nk}^d \quad (7.2)$$

where d indexes the images, $\sum_v \Lambda_{kv} = 1$, $\delta()$ is a Kronecker delta function and ϕ_{nk} is the parameter of the variational distribution $q(z)$. This parameter is computed

in the E-step with

$$\text{For cLDA: } \quad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{w^d k} \quad (7.3)$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \exp[\psi(\gamma_k^d)] \quad (7.4)$$

$$\text{For sLDA: } \quad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_k \quad (7.5)$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \exp \left[\psi(\gamma_k^d) + \frac{\zeta_{w^d k}}{N} - \frac{\sum_c \exp \frac{\zeta_{ck}}{N} \prod_{m \neq n} \sum_j \phi_{mj}^d \exp \frac{\zeta_{cj}}{N}}{\sum_c \prod_m \sum_j \phi_{mj}^d \exp \frac{\zeta_{cj}}{N}} \right] \quad (7.6)$$

where, γ is the parameter of the variational distribution $q(\boldsymbol{\pi})$ (see [14] for the details of variational inference in LDA). The important point to note is that the class label w^d only influences the topic distributions through (7.3) for cLDA (where $\boldsymbol{\alpha}_{w^d}$ is used to compute the parameter $\boldsymbol{\gamma}^d$) and (7.6) for sLDA (where the variational parameter ϕ_{nk}^d depends on the class label w^d through $\zeta_{w^d k}/N$).

We next consider the case of cLDA. Given that $q(\boldsymbol{\pi})$ is a posterior Dirichlet distribution (and omitting the dependence on d for simplicity), the estimate of γ_k has two components: $\hat{l}_k = \sum_n \phi_{nk}$, which acts as a vector of counts, and α_{wk} which is the parameter from the prior distribution. As the number of samples increases, the amplitude of the count vector, $\hat{\mathbf{l}}$, increases proportionally, while the prior $\boldsymbol{\alpha}_w$ remains constant. Hence, for a sufficiently large sample size N , the prior $\boldsymbol{\alpha}_w$ has a very weak influence on the estimate of $\boldsymbol{\gamma}$. This is a hallmark of Bayesian parameter estimation, where the prior only has impact on the posterior estimates for small sample sizes. It follows that the connection between class label W and the learned topics Z_i is *extremely weak*. This is not a fallacy of the variational approximation. In cLDA (Figure 7.1(b)), the class label distribution is simply a prior for the remaining random variables. This prior is *easily overwhelmed* by the evidence collected at the feature-level, whenever the sample is large.

A similar effect holds for sLDA, where the only dependence of the parameter estimates on the class label is through the term $\zeta_{w^d k}/N$. This clearly diminishes as the sample size N increases. In summary, topics learned with either cLDA or

sLDA are very *unlikely* to be informative of semantic regularities of interest for classification, and much more likely to capture generic regularities, common to all classes.

7.4.2 Experimental Analysis

To confirm the observations above, we performed experiments with topics learned under two approaches. In the first, we used the original learning equations, i.e. (7.3) and (7.4) for cLDA and (7.5) and (7.6) for sLDA. In the second we severed all connections with the class label variable *during learning* (of the topics), by reducing the variational E-step (for both cLDA and sLDA) to,

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha \quad (7.7)$$

$$\phi_{nk}^{d*} \propto \Lambda_{kv_n^d} \exp [\psi(\gamma_k^d)] \quad (7.8)$$

with $\alpha = 1$. This guarantees that the topic-conditional distributions are learned without any class influence. The remaining parameters (α_w for cLDA, ζ_w for sLDA) are still learned using the original equations. The rationale for these experiments is that, if supervision makes any difference, models learned with the original algorithms should perform better.

Figure 7.4 shows the scene classification performance of cLDA and sLDA, under the two learning approaches, on the N15, N8, and S8 datasets (see Appendix A for details on the experimental setup). The plots were obtained with a 1024 words codebook, and between 10 and 100 topics. Clearly, the classification performance of the original models *is not* superior to that of the ones learned without class supervision. The sLDA model has almost identical performance under the two approaches, on the three datasets. For cLDA, unsupervised topic discovery is in fact *superior* on the N8 and S8 dataset. This can be explained by poor regularization of the original cLDA algorithm. We have observed small values of α_{wk} , which probably led to poor estimates of the topic distributions in (7.3). For example, the maximum, median and minimum values of α_{wk} learned with 10 topics on S8 were 0.61, 0.12, 0.04 respectively. In contrast, the corresponding values for

unsupervised topic discovery were 7.09, 1.09, 0.55. Similar effects were observed in experiments with codebooks of different size. These results are clear evidence that the performance of cLDA and sLDA is similar (if not inferior) to that of topic learning without class supervision. In both cases, the class variable has very weak impact on the learning of topic distributions.

7.5 Topic supervision

In this section introduce topic supervision for LDA models, and its impact in learning and inference.

7.5.1 Topics supervision in LDA model

The simplest solution to the limitations discussed in the last section, is to *force* topics to reflect the semantic regularities of interest. This consists of equating topics to class labels, and is denoted *topic supervised LDA*. Topic supervision was previously proposed in semi-LDA [170] and labeled-LDA [116], for action and text classification respectively. However, its impact on classification performance is difficult to ascertain from these works, for several reasons. First, none of them performed a systematic comparison to existing LDA methods. Second, both are topic-supervised versions of LDA. Intuitively, topic supervised versions of classification models, namely cLDA and sLDA, should achieve better performance. Third, semi-LDA adopts an unconventional inference process, which assumes that $p(z_n|v_1, v_2, \dots, v_n) \propto p(z_n|\boldsymbol{\pi})p(z_n|v_n)$. It is unclear how this affects the performance of the topic-supervised model. Finally, the goal of labeled-LDA is to assign multiple labels per document. This is somewhat different from scene classification, although labeled-LDA reduces to a topic-supervised model for classification if there is a single label per item.

7.5.2 Models and geometric interpretation

To analyze the impact of topic-supervision on the various LDA models, we start by noting that the graphical model of the topic supervised extension of any LDA model is *exactly* the same as that of the model without topic supervision. The only, subtle yet significant, difference is that the topics are no longer *discovered*, but *specified*. It is thus possible to introduce topic-supervised versions of all models in the literature. In this work, we consider three such versions, viz. “topic supervised LDA (ts-LDA)”, “topic-supervised class LDA (ts-cLDA)”, and “topic-supervised supervised LDA (ts-sLDA)”. These are the topic-supervised versions of LDA, cLDA and sLDA, respectively, with the following three distinguishing properties,

- the set of topics \mathcal{L} is the set of class labels \mathcal{W} .
- the samples from the topic variables Z_i are class labels.
- the topic conditional distributions $P_{V|Z}()$ are learned in a supervised manner.

We will shortly see that this has the added advantage of substantially simpler learning.

Figure 7.2(b) shows the schematic of ts-cLDA for a two class problem on a three word simplex. As with cLDA, Figure 7.2(a), Λ_1 and Λ_2 represent two topic-distributions. There is, however, a significant difference. For cLDA, topic distributions are learned in a bottom up manner and can be positioned anywhere on the word simplex, by the topic discovery algorithm. For ts-cLDA, the topics are specified: each topic is an image class.

7.5.3 Learning and inference with topic-supervision

The introduction of topic-level supervision decouples the learning of the topic-conditional distribution $P_{V|Z}()$ from that of the other model parameters, substantially reducing learning complexity. In general, learning topic distributions would require a strongly supervised training set, however in absence of these labels, all patch labels in an image are made equal to its class label, i.e. $z_n^d = w^d \forall n, d$.

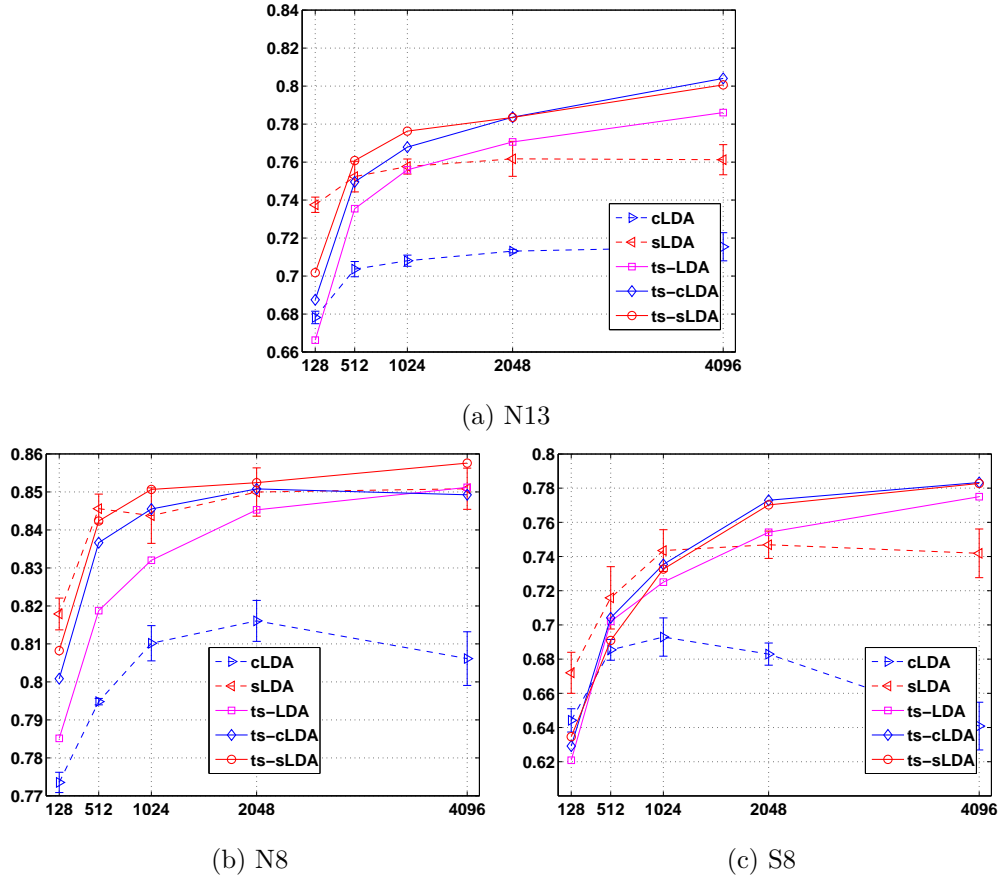


Figure 7.5: Performance of ts-sLDA, ts-cLDA, sLDA, and cLDA as a function of codebook size on (a) N13, (b) N8 and (c) S8. For ts-sLDA and ts-cLDA the number of topics is equal to the number of classes. For sLDA and cLDA, results are presented for the number of topics of best performance.

This type of learning has shown to be effective, both through the design of image labeling systems [21] and theoretical connections to multiple instance learning [155]. The ML estimate of Λ_k is

$$\Lambda_{kv}^* = \arg \max_{\Lambda_k} \sum_d \sum_n \delta(w^d, k) \delta(v_n^d, v) \log \Lambda_{kv} \quad (7.9)$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. The solution to this optimization problem is

$$\Lambda_{kv} = \frac{\sum_d \sum_n \delta(w^d, k) \delta(v_n^d, v)}{\sum_j \sum_d \sum_n \delta(w^d, j) \delta(v_n^d, v)}. \quad (7.10)$$

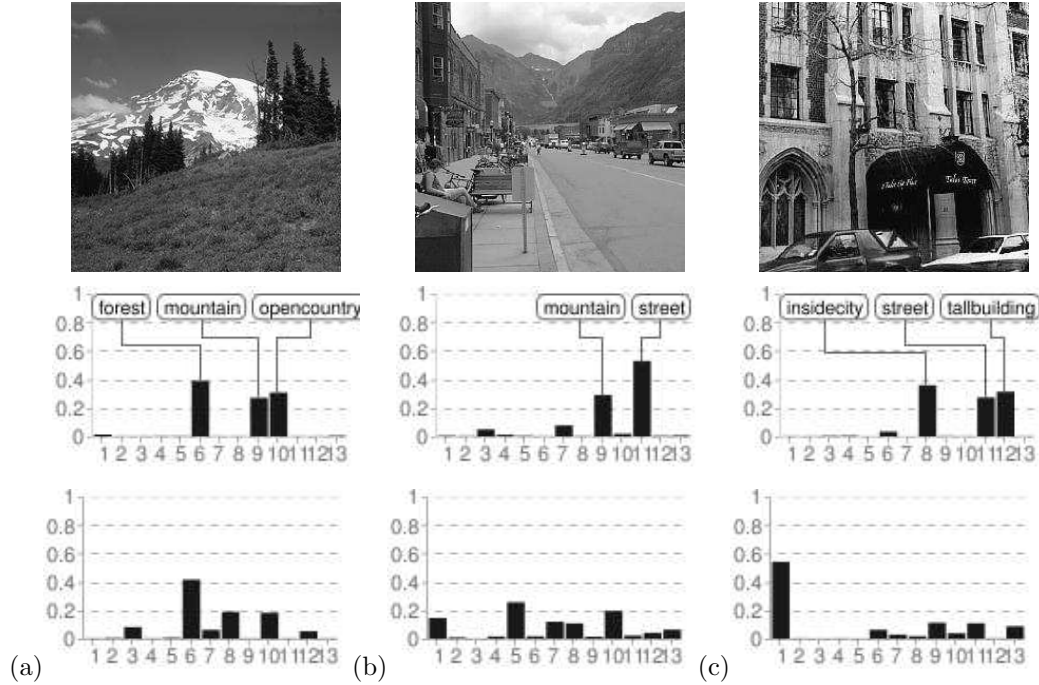


Figure 7.6: Some example images that were misclassified by cLDA, but correctly classified using ts-cLDA. The expected topic distributions for ts-cLDA and cLDA (using 13 topics) are shown in the middle and bottom rows respectively. For ts-cLDA, topic labels are same as the class labels and the high probability topics are indeed the ones which capture the semantic meaning of the image. For cLDA, the topic labels do not carry any clear semantic meaning.

Given the topic-conditional distributions, all other parameters can be learned as in the original models. Parameter estimation for ts-cLDA is detailed in Appendix F.

7.5.4 Experimental analysis

Figure 7.5 presents classification results of ts-LDA, ts-cLDA and ts-sLDA, as a function of codebook size, under the experimental conditions of Figure 7.4. Compared to sLDA and cLDA, all three topic supervised approaches achieve superior classification performance. This is true for all datasets across different codebook size when compared to cLDA, and for all datasets and codebooks with over 1024 codewords when compared to sLDA. The best performance across dif-

Table 7.1: Classification Results on Natural Scene Categories.

model	Dataset		
	N15	N13	N8
ts-sLDA	74.82 ± 0.68	79.70 ± 0.48	86.33 ± 0.69
ts-cLDA	74.38 ± 0.78	78.92 ± 0.68	86.25 ± 1.23
ts-LDA	72.60 ± 0.51	78.10 ± 0.31	85.53 ± 0.41
sLDA	70.87 ± 0.48	76.17 ± 0.92	84.95 ± 0.51
cLDA	65.50 ± 0.32	72.02 ± 0.58	81.30 ± 0.55

ferent codebooks and topics cardinality is reported in 7.1 and 7.2. On average, across datasets, topic-supervision improves the classification accuracies of cLDA and sLDA by 12% and 5% respectively. Among the three topic-supervised models, ts-cLDA and ts-sLDA achieve comparable performance, which is superior to that of the simpler ts-LDA model.

Figure 7.6 shows some images incorrectly classified by cLDA but correctly classified by ts-cLDA, on the N15 dataset. Also shown are the topic histograms obtained in each case, with ts-cLDA in the middle and cLDA in the bottom row. The figures illustrate the effectiveness of ts-sLDA at capturing semantic regularities — topics with high probability are indeed representative of the image semantics. Note that such an interpretation is only possible as the topic labels in ts-cLDA have a one-to-one correspondence with the class labels. For cLDA, topic histograms merely represent visual clusters.

7.6 Acknowledgments

The text of Chapter 7, in full, is based on the material as it appears in: N. Rasiwasia and N. Vasconcelos, ‘*Holistic Context Models for Visual Recognition*’, Accepted to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, and N. Rasiwasia and N. Vasconcelos, ‘*Generative Models for Image Classification*’, In preparation for IEEE Transactions on Pattern Analysis and Machine Intelligence. The dissertation author was a primary researcher and an author

Table 7.2: Classification Results on Sports8 and Corel50.

	Dataset	
model	S8	C50
ts-sLDA	78.37 \pm 0.80	42.33
ts-cLDA	77.43 \pm 0.97	40.80
ts-LDA	77.77 \pm 1.02	39.20
sLDA	74.95 \pm 1.03	39.22
cLDA	70.33 \pm 0.86	34.33

of the cited material.

Chapter 8

Conclusions

In this thesis, we proposed a novel semantic image representation based on co-occurrence of semantic concepts. The proposed modeling is quite simple, and builds upon the bag-of-features appearance representation and the availability of robust appearance classifiers. Images are represented by their posterior probabilities with respect to a set of semantic concepts. This results in mapping of the images from the space of appearance feature to that of semantic features. Denoted as the semantic space, each dimension of this space encodes an appearance-based posterior probability with respect to a semantic concept. Semantic image representation is shown to have a higher level of abstraction than bag-of-features appearance representation. Three novel visual recognition systems; for the task of image retrieval, scene classification and cross-modal multimedia retrieval, were proposed. All three recognition systems build upon the proposed semantic image representation.

First, the design of a content based retrieval system, query-by-semantic-example (QBSE), was introduced, where the retrieval operation was carried on the semantic space. QBSE system, apart from yielding state of the art retrieval performance, was instrumental in evaluating the intrinsic benefit of semantic representation for image retrieval. The results above provide *strong* support in favor of the argument, that is *semantic representations have an intrinsic benefit for image retrieval*. While this could be dismissed as a trivial conclusion, we believe that doing so would be unwise, for two main reasons. First, it had not been previously shown that query-by-text systems can generalize beyond the restricted vocabulary on which they are trained. This is certainly not the case for the current standard text based query paradigm. Second, the results above suggest some interesting hypotheses for future research, which could lead to long-term gains that are more significant than simple out-of-vocabulary generalization. For example, given that the higher abstraction of the semantic representation enables better performance than visual matching, it appears likely that semantic spaces constructed with better abstraction or that exploits the structure of natural language, can lead to better retrieval systems. The QBSE paradigm now proposed could be easily extended to the multi-resolution semantic spaces that are likely to result from a hierarchical

concept representation. Furthermore, it would allow an objective characterization of the gains achievable at the different levels of the taxonomy. We intend to explore these questions in future work.

Second, the design of a scene classification system based on semantic image representation was presented. Inspired from the recent works on scene classification, where a low-level intermediate “theme” space is introduced, a framework based on semantic space – which serves as a proxy for the intermediate space, was proposed. All classification decisions were performed on this space. An implementation of the proposed framework was presented and compared to various existing algorithms, on benchmark datasets. The results allow a number of conclusions. First, while low dimensional semantic representations are desirable for the reasons discussed in Section 4.1, previous approaches based on latent-space models have failed to match the performance of the flat bag-of-words model, which has high dimensionality. We have shown that this is indeed possible, with methods that have much lower complexity than the latent-space approaches previously proposed, but make better use of the available labeling information. Next, a study of the effect of dimensionality on the classification performance was presented, which indicates that the dimensionality would grow sub-linearly with the number of scene categories. This could be a significant advantage over the flat bag-of-words models which, although successful for the limited datasets in current use, will likely not scale well when the class vocabulary increases.

Third, the design of cross-modal multimedia retrieval system was proposed. This entails the retrieval of database entries from one content modality in response to queries from another. While the emphasis was on cross-modal retrieval of images and text, the proposed models support many other content modalities. By requiring representations that can generalize across modalities, cross-modal retrieval establishes a suitable context for the objective investigation of fundamental hypotheses in multimedia modeling. We have considered two such hypotheses, regarding the importance of low-level cross-modal correlations and semantic abstraction in multi-modal content modeling. The hypotheses were objectively tested by comparing the performance of three new approaches to cross-modal retrieval:

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.*, text and images) to a pair of isomorphic spaces, where a natural correspondence can be established for cross-modal retrieval purposes. The retrieval performance of the three solutions was extensively tested on two datasets, “Wikipedia” and “TVGraz”, containing documents that combine images and text. While the two fundamental hypotheses were shown to hold for the two datasets, where both CM and SM achieved significant improvements over chance retrieval, SM achieved overall better performance than CM. This implies stronger evidence for the abstraction than for the correlation hypothesis. The two hypotheses were also found to be complementary, with SCM achieving the best results of all methods considered.

Finally, the design of a two-layer holistic context modeling system based on the probability of co-occurrence of objects and scenes was proposed. The first layer represents the images in a semantic space, which has a higher level of abstraction, but suffers from a certain amount of contextual noise, due to the inherent ambiguity of classifying image patches. The second layer enables robust inference in the presence of this noise, by modeling the distribution of each concept in the semantic space. An image is then represented by its posterior probabilities with respect to these *contextual* distributions. This was shown to produce posterior distributions that emphasize concept co-occurrences due to true contextual relationships and inhibit accidental co-occurrences due to ambiguity. Interestingly, we found a weak correlation between the quality of the appearance classification and the corresponding quality at the contextual level. In fact, some variations of the representation with weak appearance-level performance were top-performers at the contextual level. It appears that, while supervision is critical to bridging the semantic gap during learning, soft appearance-level decisions are more effective during inference. This is an interesting finding, given the emphasis on highly accurate appearance classification in the literature. Recognition systems that operates on the clean contextual representation were shown to outperform both noisy semantic representation and the appearance representation in the tasks of scene

classification and image retrieval. In both cases, it was also shown that, despite its simplicity, the proposed contextual models are superior to various previous proposals in the literature. The gains with respect to appearance modeling were shown to hold irrespectively of the choice and accuracy of the underlying appearance models.

The overall representation is similar to a topic model, but where topics are learned in a supervised manner. Supervised learning is a necessary condition for overcoming the semantic gap between the low-level patch representation and the higher-level contextual relationships. While multiple instance learning is required to cope with the uncertainty of the appearance representation, multiple instance inference was shown ineffective. Best results are obtained with weaker, patch-based, inference that leads to an appearance representation of higher entropy. This prevents a greedy commitment to premature image explanations that, while consistent with appearance statistics, do not take context into account. The latter goal is better served by inference procedures that assign non-zero probability to multiple plausible classes, at the appearance level. We proposed topic supervised topic models that address the limitations of the existing topic models, enabling them to achieve better classification accuracies.

It should be noted that our current implementation does not incorporate spatial information of any form. Current evidence [74, 83] suggests that integration of weak spatial information, by dividing an image in a 2×2 or 4×4 grid of spatial bins, can improve the accuracy of visual recognition systems. Furthermore, in this thesis the proposed semantic and contextual image representations, were tested on datasets composed of ten to a few hundred concepts. The benefits of the proposed representation in recognition tasks with much higher number of semantic concepts, remains to be tested. We intend to explore these issues as a part of future work.

Appendix A

Datasets.

A.1 Datasets

In this work we adopt several datasets previously used in visual recognition task such as image annotation, image retrieval, scene classification etc. In addition to the existing datasets, we introduce three new datasets — two datasets for the task of image retrieval and one for cross-modal retrieval. Next, we briefly discuss the salient properties of these datasets.

A.1.1 Natural Scene Categories (N8, N13, N15)

The Natural Scene Categories, is a collection of three datasets, viz. “LabelMe Natural Scenes”, “Thirteen Natural Scenes” and “Fifteen Natural Scenes”, where “LabelMe Natural Scenes” is a subset of “Thirteen Natural Scenes” which itself is a subset of the “Fifteen Natural Scenes” dataset.

LabelMe Natural Scenes (N8)

“LabelMe Natural Scenes” dataset, henceforth referred to as “Natural8” (N8), consists of 2688 images classified into eight classes viz “Coast”, “Forest”, “Highway”, “Inside City”, “Mountain”, “Open Country”, “Street”, “Tall Building”. This dataset was first proposed in [104] and has been later used in several scene classification literatures [114, 17, 80, 67] etc. Although the images are available with color, in this work as is commonly done we convert all the images to gray scale. The average size of each image is 250×250 pixels. N8 dataset is primarily used for scene classification task, where 100 images per class serve as the training set and the rest of the images as the test set. A.1 provides a detailed description of various classes.

Thirteen Natural Scenes (N13)

“Thirteen Natural Scenes” dataset here referred to as “Natural13 (N13)”, was first proposed in [77] where five more scene categories, viz. “Bedroom”, “Suburb”, “Kitchen”, “Livingroom”, “Office”, were added to the N8 dataset. N13 dataset has been used by several authors to evaluate scene classification systems

Table A.1: Summary of the Natural Scene datasets.

Natural8 (N8)			
Category	Training set	Test set	Total
Coast	100	260	360
Forest	100	228	328
Highway	100	160	260
Inside City	100	208	308
Mountain	100	274	374
Open Country	100	310	410
Street	100	192	292
Tall Building	100	256	356
total	800	1888	2688
Natural13 (N13) Additional Classes			
Bedroom	100	116	216
Suburb	100	141	241
Kitchen	100	110	210
Livingroom	100	189	289
Office	100	115	215
total	1300	2559	3859
Natural15 (N15) Additional Classes			
Store	100	215	315
Industrial	100	211	311
total	1500	2985	4485

[17, 114, 74, 6, 65]. A.1 provides a detailed description of various additional classes of the N13 dataset.

Fifteen Natural Scenes (N15)

“Fifteen Natural Scenes” dataset, here referred to as “Natural15” (N15) is currently one of the most popular dataset used for the evaluation of scene recognition systems. N15 dataset was first proposed in [74], where two more scene categories, viz. “Store”, “Industrial” were added to the N13 dataset. Thus, N15 dataset consists of fifteen classes of natural scenes where each class contains 200 to 400 images, of average size 270×250 pixels. In all the experiments using N15 dataset, 100 images per scene are used to learn the model, the remaining being used as test set. A.1 provides a detailed description of the additional classes in the N15 dataset.

A.1.2 UIUC Sports Dataset (S8)

UIUC Sports dataset, henceforth referred to as “Sports8” (S8), consists of 1579 images classified into eight sports categories, viz. {“badminton”, “bocce”, “croquet”, “polo”, “rock climbing”, “rowing”, “sailing”, “snowboarding”}. It was first proposed in [79] for Latent Dirichlet Allocation based (LDA) based classification, and subsequently used by [167] to evaluate supervised-LDA. Each category has 137 to 250 images with an average size of over 1000×1000 pixels. For our experiments, the images were resized to a maximum of 256 pixels along the larger border. In all, there are 1579 images. In this work S8 dataset is used to evaluate scene classification systems. As in [79], 70 images per scene are used to learn the model, and 60 images are used as test set. A.2 provides a detailed description of all the classes in the S8 dataset.

A.1.3 Corel Image Collection (C371, C50, C43, C15)

The Corel Image Collection consists of the Corel Stock Photo CDs. Each CD includes 100 images of a common topic. We construct four different datasets from this collection.

Table A.2: Summary of the UIUC Sports dataset.

Category	Training set	Test set	Total
Coast	70	60	200
Forest	70	60	137
Highway	70	60	236
Inside City	70	60	182
Mountain	70	60	194
Open Country	70	60	250
Street	70	60	190
Tall Building	70	60	190
total	560	480	1579

Corel371 (C371)

The first dataset is “Corel371” (C371) which was first proposed in [35] for the task of automatic image annotation. C371 consists of 5,000 images from 50 Corel Stock Photo CDs. Each image is further labeled with 1-5 semantic concepts. Overall there are 371 concepts in the vocabulary. C371 has since then been used to evaluate several other image annotation systems [41, 72, 21, 22] etc where 4500 images are used to train the system and the rest 500 for evaluation. A.3 provides a list of the annotation available for the C371 dataset along with the number of training and testing images per concept (in brackets). All images in this collection are available with color information. In this work, all the images from the Corel Collection are normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space.

Table A.3: Summary of the C371 dataset.

water (1005,116); sky (883,105); tree (854,93); people (670,74); grass (446,51);
 buildings (408,54); mountain (307,38); flowers (269,27); snow (267,31); clouds (254,

Table A.3: (continued)

26); rocks (228,22); stone (212,20); street (203,26); plane (199,25); bear (198,22); field (198,17); sand (184,19); birds (179,17); beach (177,18); boats (155,15); jet (147,19); leaf (136,12); cars (134,17); plants (129,15); house (124,19); bridge (123,15); polar (122,13); valley (122,11); garden (117,10); hills (113,18); close-up (112,10); ruins (107,12); statue (106,11); horses (103,12); tracks (103,11); sun (101,10); ice (99,12); wall (98,14); ocean (96,9); cat (96,11); temple (94,10); train (94,11); tiger (91,10); coral (89,9); scotland (89,11); swimmers (85,8); coast (84,5); window (79,8); branch (78,2); pool (77,11); foals (77,9); sunset (76,7); sculpture (76,10); frost (74,7); head (71,2); forest (71,11); fox (71,9); nest (71,7); mare (69,9); city (67,10); railroad (63,8); ground (60,4); horizon (59,4); shops (59,4); petals (59,4); arch (57,4); reefs (56,5); palace (56,4); reflection (55,9); park (55,2); desert (55,11); skyline (53,6); locomotive (53,9); shore (51,8); castle (49,6); pillar (49,9); river (48,4); town (48,9); road (47,4); deer (47,4); waves (45,4); smoke (44,10); sea (43,2); church (42,6); market (40,2); tower (40,7); coyote (37,2); light (37,6); courtyard (37,2); sign (37,2); zebra (37,4); bush (36,1); fence (35,2); village (35,7); door (35,2); landscape (35,4); pyramid (35,3); black (34,2); roofs (34,2); tundra (33,9); display (32,1); shadows (32,3); elk (32,6); island (31,2); flight (30,1); grizzly (30,7); harbor (30,4); rodent (30,4); runway (29,1); stems (29,2); palm (28,3); tulip (28,3); antlers (28,4); dunes (28,1); man (28,1); woman (28,1); turn (28,3); fish (27,6); restaurant (27,4); formula (27,4); buddha (26,1); white-tailed (26,2); kauai (26,4); hut (25,6); herd (25,4); formation (24,2); wood (24,4); food (24,2); museum (23,4); indian (22,3); oahu (22,1); ships (21,3); flag (21,2); prop (21,1); hillside (21,3); farms (21,2); bengal (21,6); cliff (21,0); hats (21,2); lizard (21,1); prototype (21,4); gate (20,2); shrine (20,0); frozen (20,4); face (19,2); log (18,2); arctic (18,3); bulls (18,5); caribou (18,4); moose (18,1); canyon (18,3); baby (18,1); buddhist (18,3); straightaway (18,0); tables (17,2); costume (17,3); hotel (17,2); fountain (17,1); night (17,2); tortoise (17,0); path (16,1); stairs (16,2); figures (16,0); lawn (16,2); giant (16,0);

Table A.3: (continued)

giraffe (16,1); steel (16,0); hawaii (16,3); land (15,1); meadow (15,3);
cubs (15,1); autumn (15,0); umbrella (15,0); crystals (15,1); booby (15,5);
seals (15,0); maui (15,2); lake (14,1); windmills (14,2); monastery (14,2);
facade (14,0); mule (14,2); tusks (14,1); sphinx (14,1); anemone (13,1);
clothes (13,1); writing (13,0); ceremony (13,1); cottage (13,3); elephant (13,3);
monks (13,3); iguana (13,3); marine (13,3); reptile (13,1); f-16 (12,1);
tails (12,1); pagoda (12,0); fruit (12,2); poppies (12,0); pots (12,3);
albatross (12,1); girl (12,3); cow (11,4); guard (11,0); athlete (11,3);
steps (11,0); horns (11,1); fly (11,1); prayer (11,0); shrubs (10,3);
post (10,2); crab (10,1); entrance (10,1); column (10,2); relief (10,1);
penguin (10,0); row (10,0); antelope (10,2); bay (9,0); fan (9,1);
sunrise (9,1); vegetation (9,1); sailboats (9,0); chapel (9,0); paintings (9,0);
plaza (9,1); pond (9,0); vines (9,1); bench (9,0); waterfalls (9,0);
slope (9,1); goat (9,2); wolf (9,0); dog (8,0); stream (8,0);
lion (8,3); barn (8,2); glass (8,1); architecture (8,1); fog (8,0);
stick (8,0); wings (8,0); blooms (8,1); mosque (8,1); squirrel (8,2);
rainbow (7,0); dress (7,1); run (7,0); sheep (7,2); detail (7,1);
room (7,0); cathedral (7,2); monument (7,3); canal (7,1); interior (7,3);
mist (7,2); vineyard (7,1); lynx (7,1); african (7,1); pups (7,0);
carvings (6,0); kit (6,1); den (6,1); balcony (6,1); art (6,2);
decoration (6,2); chairs (6,0); crowd (6,0); cheese (6,0); silhouette (6,1);
terrace (6,1); cactus (6,2); outside (6,1); basket (5,1); drum (5,0);
winter (5,0); rockface (5,0); pair (5,0); nets (5,1); pattern (5,0);
blossoms (5,0); store (5,1); needles (5,1); designs (5,0); lily (5,0);
lighthouse (5,2); truck (5,1); marsh (5,1); porcupine (5,1); range (5,0);
pole (5,0); dance (5,1); plain (4,0); peaks (4,1); helicopter (4,0);
fall (4,0); sponges (4,0); star (4,0); cave (4,2); vegetables (4,0);
rose (4,0); dock (4,1); pottery (4,0); fawn (4,0); chrysanthemums (4,0);
trunk (4,2); eagle (4,0); whales (4,1); rabbit (4,0); animals (4,0);
shell (3,0); storm (3,0); crafts (3,1); festival (3,1); mural (3,0);
butterfly (3,1); carpet (3,0); floor (3,0); vendor (3,1); parade (3,0);

Table A.3: (continued)

doorway (3,1); texture (3,0); dust (3,0); pack (3,0); dall (3,0);
trail (3,0); shirt (3,0); pebbles (3,0); snake (3,1); moon (2,0);
cafe (2,1); angelfish (2,0); perch (2,0); sidewalk (2,2); spider (2,0);
tent (2,0); clearing (2,0); hands (2,0); crops (2,0); vehicle (2,1);
rice (2,0); tomb (2,0); calf (2,1); school (2,0); boeing (1,0);
diver (1,0); sails (1,1); model (1,0); railing (1,0); ladder (1,0);
rapids (1,0); military (1,0); mushrooms (1,0); hawk (1,0); orchid (1,1);
saguaro (1,0); mast (1,0); pepper (1,0); insect (1,0); glacier (1,0);
harvest (1,0); shade (1,0); ceiling (1,0); furniture (1,0); lichen (1,0);
remains (1,0); leopard (1,0); jeep (1,0); cougar (1,1); canoe (1,0);
race (1,0); grouper (0,1); moss (0,1); aerial (0,1);

Corel50,Corel43 (C50,C43)

We also use the image from C371 dataset to construct two more dataset, “Corel50” (C50) and “Corel43” (C43) for the task of scene classification task, relying on the CD labels for groundtruth instead of the image annotations. C50 contains 50 scene classes, each corresponding to one CD in the collection. For each CD, 90 images are used to learn class models and the remaining for testing. It has been argued that CD labels lead to an easy classification problem [173] as there is high variability between images from different CDs and high similarity among those from the same CD. To address these concerns, we construct another dataset from this collection, C43 that uses a set of manual annotations (disjoint from the CD labels) as ground truth. 43 semantic concepts are chosen from the set of annotations of [35] (those with a minimum of 100 annotated images) and 100 images are randomly selected per concept. Since an image can be labeled with more than one concept, this results in a total of 3102 images. Of these, 2766 are randomly selected to create a test set with approximately 90 images per label, and the remainder are used for testing. A correct classification is declared whenever the top predicted label matches any of the groundtruth labels.

Corel15 (C15)

Corel15 (C15) consists of 1,500 images from another fifteen previously unused Corel Stock Photo CDs, viz. “Adventure Sailing”, “Autumn”, “Barnyard Animals”, “Caves”, “Cities of Italy”, “Commercial Construction”, “Food”, “Greece”, “Helicopters”, “Military Vehicles”, “New Zealand”, “People of World”, “Residential Interiors”, “Sacred Places”, “Soldier”. Once again, the CD themes (non-overlapping with those of C50) served as the ground truth. This dataset is used for the evaluation of image retrieval systems where 1,200 images serve as the retrieval set and the remaining 300 images as the query set.

A.1.4 Flickr Images (F18)

To address some criticism that ‘Corel is easy’ [98, 172], we collected a second database from the online photo sharing website www.flickr.com. The images in this database were extracted by placing queries on the flickr search engine, and manually pruning images that appeared irrelevant to the specified queries. Note that the judgments of relevance did not take into account how well a content-based retrieval system would perform on the images, simply whether they appeared to be search errors (by flickr) or not. The images are shot by flickr users, and hence differ from the Corel Stock photos, which have been shot professionally. This database, “Flickr18” (F18), contains 1800 images divided into 18 classes viz. “Automobiles”, “Building and Landscapes”, “FacialCloseUp”, “Flora”, “FlowersCloseup”, “Food and Fruits”, “Frozen”, “Hills and Valley”, “Horsesl and Foal”, “JetPlanes”, “Sand”, “Sculpture and Statues”, “SeaandWaves”, “Solar”, “Township”, “Train”, “Underwater”, “Waterfun”, according to the manual annotations provided by the online users. F18 is again used for evaluating image retrieval systems where 20% of randomly selected images served as the query set and the remaining 80% as the retrieval set.

Table A.4: Summary of the TVGraz dataset.

Category	Training set	Test set	Total
Brain	109	47	156
Butterfly	195	51	246
Cactus	137	37	174
Deer	223	51	274
Dice	169	50	219
Dolphin	163	59	222
Elephant	120	54	174
Frog	215	67	282
Harp	131	42	173
Pram	96	42	138
total	1558	500	2058

A.1.5 TVGraz

The TVGraz dataset is a collection of web-pages compiled by Khan *et al.* [66]. The Google Image search engine was used to retrieve 1,000 web-pages for each of ten categories from the Caltech-256 [52] dataset. The results were filtered into a set of 2,592 positive web-pages, containing both text and image data, for which the image belonged to the query category. Due to copyright issues, the TVGraz database is stored as a list of URLs, and must be recompiled by each new user. We collected 2,058 image-text pairs, since some URLs were defunct and we discarded web-pages that did not contain at least 10 words and one image. The median text length, per web-page, is 289 words. A random split was used to produce 1,558 training and 500 test documents, as summarized in A.4.

A.1.6 Wikipedia

A novel dataset was assembled from the “Wikipedia featured articles”, a continually updated collection of Wikipedia articles, which contained 2,669 entries

Table A.5: Summary of the Wikipedia dataset.

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

when the data was collected, in October 2009. These articles, which are selected and reviewed for style and quality by Wikipedia’s editors, are often accompanied by one or more pictures from the Wikimedia Commons, supplying a text-image pairing. The Wikipedia featured articles are divided into 29 categories, but some contain very few entries. We considered only articles from the 10 most populated categories, which were used as a semantic vocabulary. Since the featured articles tend to have multiple images and span multiple topics, each article was split into sections, based on its section headings. Each image was assigned to the section in which it was placed by the author(s). This produced a total of 7,114 sections, which are internally more coherent and usually contain a single picture. The dataset was then pruned, by keeping only sections with exactly one image and at least 70 words. The final corpus contains a total of 2,866 documents. The median text length is 200 words. A random split was used to produce a training set of 2,173 documents and a test set of 693 documents, as summarized in A.5.

Appendix B

Generalized Expectation maximization (GEM)

The parameters $\Lambda^w = \{\beta_k^w, \boldsymbol{\alpha}_k^w\}$ of the contextual class models of (6.1) are learned using GEM. This is an extension of the well known EM algorithm, applicable when the M-step of the latter is intractable. It consists of two steps. The E-Step is identical to that of EM, computing the expected values of the component probability mass β_k . The generalized M-step estimates the parameters $\boldsymbol{\alpha}_k$. Rather than solving for the parameters of maximum likelihood, it simply produces an estimate of higher likelihood than that available in the previous iteration. This is known to suffice for convergence of the overall EM procedure [30]. We resort to the Newton-Raphson algorithm to obtain these improved parameter estimates, as suggested in [95] for single component Dirichlet distributions. Omitting the dependence on the concept index w for brevity, the algorithm iterates between two steps,

E-step: compute

$$h_{dk} = \frac{\mathcal{D}ir(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_k)\beta_k}{\sum_l \beta_l \mathcal{D}ir(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_l)} \quad (\text{B.1})$$

M-step: set

$$(\beta_k)^{new} = \frac{N_k}{N}, \quad \text{where } N = \sum_{dk} h_{dk}, N_k = \sum_d h_{dk} \quad (\text{B.2})$$

$$(\alpha_k)^{new} = (\alpha_k)^{old} + \mathcal{H}^{k-1} \mathbf{g}^k \quad (\text{B.3})$$

$$\text{where } \mathbf{g}_i^k = N_k (\Psi(\sum_{p=1}^L \alpha_p) - \Psi(\alpha_i)) + \sum_d h_{dk} \log \pi_{id} \quad (\text{B.4})$$

$$\text{and } \mathcal{H}_{ii}^k = N_k (\Psi'(\sum_{p=1}^L \alpha_p) - \Psi'(\alpha_i)) \quad (\text{B.5})$$

$$\mathcal{H}_{ij}^k = N_k (\Psi'(\sum_{p=1}^L \alpha_p)), \quad (\text{B.6})$$

Ψ and Ψ' are the Digamma and Trigamma functions [95].

Appendix C

Computation of Image-SMN

Given N patch-based SMNs, $\boldsymbol{\pi}^{(n)}$, the Image-SMN $\boldsymbol{\pi}^*$ is

$$\begin{aligned}\boldsymbol{\pi}^* &= \arg \min_{\boldsymbol{\pi}} \frac{1}{N} \sum_{n=1}^N KL(\boldsymbol{\pi} || \boldsymbol{\pi}^{(n)}) \\ &= \arg \min_{\boldsymbol{\pi}} \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^L \pi_i \log \frac{\pi_i}{\pi_i^{(n)}} \\ &= \arg \min_{\boldsymbol{\pi}} \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^L \left[\pi_i \log \pi_i - \pi_i \log \pi_i^{(n)} \right]\end{aligned}$$

subject to $\sum_{i=1}^L \pi_i = 1$. This has Lagrangian

$$\mathcal{L}(\boldsymbol{\pi}, \lambda) = \sum_{i=1}^L \pi_i \log \pi_i - \frac{1}{N} \sum_{i=1}^L \pi_i \sum_{n=1}^N \log \pi_i^{(n)} + \frac{\lambda}{N} (1 - \sum_{i=1}^L \pi_i).$$

Setting derivatives with respect to π_i to zero leads to

$$1 + \log \pi_i - \frac{1}{N} \sum_{n=1}^N \log \pi_i^{(n)} - \frac{\lambda}{N} = 0, \quad (\text{C.1})$$

$$\text{or} \quad \pi_i = \exp \left(\hat{\lambda} + \langle \log \pi_i \rangle \right) \quad (\text{C.2})$$

where $\langle \log \pi_i \rangle = \frac{1}{N} \sum_{n=1}^N \log \pi_i^{(n)}$ and $\hat{\lambda} = \frac{\lambda}{N} - 1$. Summing over i and using the constraint $\sum_i \pi_i = 1$,

$$1 = \exp(\hat{\lambda}) \sum_{i=1}^L \exp \langle \log \pi_i \rangle \quad (\text{C.3})$$

$$\exp(\hat{\lambda}) = \frac{1}{\sum_{i=1}^L \exp \langle \log \pi_i \rangle}. \quad (\text{C.4})$$

Substituting (C.4) in (C.2),

$$\pi_i^* = \frac{\exp \langle \log \pi_i \rangle}{\sum_{i=1}^L \exp \langle \log \pi_i \rangle} \quad (\text{C.5})$$

$$= \frac{\exp \frac{1}{N} \sum_n \log \pi_i^{(n)}}{\sum_i \exp \frac{1}{N} \sum_n \log \pi_i^{(n)}}. \quad (\text{C.6})$$

Appendix D

Variational Approximation

Variational methods approximate the posterior $P(\boldsymbol{\pi}, w_{1:N}|x_{1:N})$ by a mean-field variational distribution $q(\boldsymbol{\pi}, w_{1:N})$, indexed by free variational parameters, within some class of tractable probability distributions \mathcal{F} . These distributions usually assume independent factors,

$$q(\boldsymbol{\pi}, w_{1:N}) = q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \prod_n q(w_n; \phi_n) \quad (\text{D.1})$$

where $q(y)$ and $q(z_i)$ are categorical models, and $q(\boldsymbol{\pi})$ a Dirichlet distribution. Given an observation $x_{1:N}$, the optimal variational approximation minimizes the Kullback-Leibler (KL) divergence between the two posteriors

$$q^* = \arg \min_{q \in \mathcal{F}} KL(q(\boldsymbol{\pi}, w_{1:N}) || P(\boldsymbol{\pi}, w_{1:N}|x_{1:N})) \quad (\text{D.2})$$

$$= \mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})) + \log P(x_{1:N}) \quad (\text{D.3})$$

where,

$$\mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})) = E_q[\log q(\boldsymbol{\pi}, w_{1:N})] - E_q[\log P(\boldsymbol{\pi}, w_{1:N}, x_{1:N})]. \quad (\text{D.4})$$

Since the data likelihood $P(x_{1:N})$ is constant for an observed image, the optimization problem is identical to

$$q^*(\boldsymbol{\pi}, w_{1:N}) = \arg \min_{q \in \mathcal{F}} \mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})), \quad (\text{D.5})$$

From Appendix A.3 of [14], the update rule for coordinate descent of the variational parameters is

$$\gamma_i^* = \sum_n \phi_{ni} + \alpha_i \quad (\text{D.6})$$

$$\phi_{ni}^* \propto P_{X|W}(x_n|w_n = i) e^{\psi(\gamma_i) - \psi(\sum_j \gamma_j)} \quad (\text{D.7})$$

such that $\sum_i \phi_{ni} = 1$ and, where α_i are the parameters of the prior class distribution $P(\boldsymbol{\pi}; \boldsymbol{\alpha})$ and ψ is the Digamma function [95]. Once the parameters of the variational distribution are obtained, the SMN for an image can be computed as,

$$\boldsymbol{\pi}^* = \arg \max_{\boldsymbol{\pi}} q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \quad (\text{D.8})$$

$$= \arg \max_{\boldsymbol{\pi}} \log q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \quad (\text{D.9})$$

$$= \arg \max_{\boldsymbol{\pi}} \sum_j^L (\gamma_j - 1) \log \pi_j \quad (\text{D.10})$$

$$\text{such that, } \sum_j \pi_j = 1 \quad (\text{D.11})$$

Using the Lagrange multiplier, λ , we get

$$J(\boldsymbol{\pi}, \lambda) = \sum_j^L (\gamma_j - 1) \log \pi_j + \lambda(1 - \sum_j^L \pi_j) \quad (\text{D.12})$$

Taking partial derivatives with respect to, π_j and λ and setting them to zero we get,

$$\frac{\partial J}{\partial \pi_j} = \frac{(\gamma_j - 1)}{\pi_j} - \lambda = 0, \forall j \quad (\text{D.13})$$

$$\frac{\partial J}{\partial \lambda} = 1 - \sum_j^L \pi_j = 0 \quad (\text{D.14})$$

From (D.13) and (D.14) we get,

$$\pi_j = \frac{\gamma_j - 1}{\sum_j \gamma_j - L} \quad (\text{D.15})$$

Appendix E

Parameter Estimation in cLDA

The parameters $(\boldsymbol{\eta}, \boldsymbol{\alpha}_{1:C}, \Lambda_{1:K})$ of cLDA are learned using variational Expectation Maximization (EM) algorithm. This iterates between:

Variational E-Step consists of approximating the posterior $P(\boldsymbol{\pi}^d, z_{1:N}^d | \mathcal{I}^d, y^d)$ for an image $\mathcal{I}^d = \{w_1^d, \dots, w_N^d\}$ using the variational distribution,

$$q(\boldsymbol{\pi}^d, z_{1:N}^d) = q(\boldsymbol{\pi}^d; \boldsymbol{\gamma}^d) \prod_n q(z_n^d; \phi_n^d) \quad (\text{E.1})$$

Similar to the variational inference of LDA (see Appendix D), the variational parameters can be computed using the update rules,

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k} \quad (\text{E.2})$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \exp[\psi(\gamma_k^d)] \quad (\text{E.3})$$

where, $\sum_k \phi_{nk}^d = 1$. Note that in cLDA, since each class is associated with a separate prior over the topic simplex, (E.2) differs from (D.6), in that $\boldsymbol{\alpha}$ parameters are class specific.

M-Step consists of computing the values of the parameters $(\boldsymbol{\alpha}_{1:C}, \Lambda_{1:K})$, where $\boldsymbol{\alpha}_y$ is obtained by maximizing,

$$\boldsymbol{\alpha}_y^* = \arg \max_{\boldsymbol{\alpha}_y} - \sum_d \delta(y^d, y) \log \mathcal{B}(\boldsymbol{\alpha}_y) + \sum_d \sum_k \delta(y^d, y) (\alpha_{y^d k} - 1) E_q[\log \pi_k^d] \quad (\text{E.4})$$

where,

$$E_q[\log \pi_k^d] = \psi\left(\sum_l \gamma_l^d\right) - \psi(\gamma_k^d) \quad (\text{E.5})$$

$$\mathcal{B}(\boldsymbol{\alpha}_y) = \frac{\prod_k (\Gamma(\alpha_{yk}))}{\Gamma(\sum_k \alpha_{yk})} \quad (\text{E.6})$$

and $\Gamma(\cdot)$ is the Gamma function. The above optimization can be carried out using the method of Newton-Raphson gradient ascent as detailed in [95].

Λ_k is obtained by maximizing,

$$\Lambda_{kv}^* = \arg \max_{\Lambda_k} \sum_d \sum_n \delta(w_n^d, v) \phi_{nk}^d \log \Lambda_{kv} \quad (\text{E.7})$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$, using the method of Lagrange multipliers which results in the closed form update,

$$\Lambda_{kv} \propto \sum_d \sum_n \delta(w_n^d, v) \phi_{nk}^d \quad (\text{E.8})$$

where, proportionality symbols means that Λ_k is normalized to sum to 1. Note that its common to assume a uniform class prior and we assume $\boldsymbol{\eta}_y = \frac{1}{C}, \forall y \in \mathcal{Y}$.

Appendix F

Parameter Estimation in topic-supervised LDA models

In this section, we discuss the parameter estimation for ts-cLDA. The parameter for other topic-supervised models can be computed using a similar approach. Topic supervision decouples cLDA learning into two steps: 1) learning of the parameters $\Lambda_{1:K}$ of the topic-conditional distributions, and 2) learning of the parameters $\alpha_{1:C}$ of the class-conditional distributions¹.

F.1 Learning Topic Conditional Distributions

As discussed in Section 7.5, since the topics are defined over the class vocabulary $\mathcal{T} = \mathcal{V}$, in absence of the individual topic labels z_n^d for the visual words w_n^d during learning, we assume all topic labels are equal to the image class y^d , i.e. $z_n^d = y^d \forall n, d$. Although, this is not true in reality, such an approximation has been shown to be effective both through the design of image labeling systems [21] and through theoretical connections to multiple instance learning. Infact, this is an implicit assumption in learning the parameters of the flat model. Thus, the ML estimates of Λ_k can be obtained from

$$\Lambda_{kv}^* = \arg \max_{\Lambda_k} \sum_d \sum_n \delta(y^d, k) \delta(w_n^d, v) \log \Lambda_{kv} \quad (\text{F.1})$$

¹Note that η is again assumed to follow a uniform distribution

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. Using the method of Lagrange multipliers, the solution to the optimization problem is given by,

$$\Lambda_{kv} = \frac{\sum_d \sum_n \delta(y^d, k) \delta(w_n^d, v)}{\sum_j \sum_d \sum_n \delta(y^d, j) \delta(w_n^d, v)} \quad (\text{F.2})$$

F.2 Learning Class Conditional Distribution

Once the topic-conditional distributions are learned, the parameters α_c of the class-conditional distributions can be learned by the maximizing the likelihood of the data, $P(y^d, w_{1:N}^d)$ using the standard variational EM algorithm, this approach iterates between two steps:

Variational E-Step consists of computing,

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k} \quad (\text{F.3})$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \exp[\psi(\gamma_k^d)] \quad (\text{F.4})$$

where, proportionality symbols means that ϕ_n^d is normalized to sum to 1.

M-Step consists of computing the values of the parameters $\alpha_{1:C}$ (note that $\Lambda_{1:K}$ is already computed) similar to (E.4).

Appendix G

Implementation Details of the various systems

We conclude this thesis by discussing some implementation details of various recognition systems proposed in this work. This discussion is intended mostly for those interested in replication portions or the entirety of the work described in the thesis. Although, many of the details have been mentioned in the previous chapters, we believe that it is useful to present a cohesive summary of the most important points.

G.1 Image Representation

Given a database of images, images are represented either using DCT or SIFT descriptors, where both BoF and BoW models are employed.

G.1.1 SIFT Features

To compute SIFT, image patches are selected either 1) by interest point detection, referred to as SIFT-INTR, or 2) on a dense regular grid, referred to as SIFT-GRID. For SIFT-INTR, interest points computed using three operators — Harris-Laplace, Laplace-of-Gaussian, and Difference-of-Gaussian — which are then merged. These measures also provide scale information, which is used in the com-

putation of SIFT features. For SIFT-GRID, feature points are sampled every 8 pixels and the descriptor is computed over a 16×16 neighborhood around each feature point. Both interest points and SIFT features are computed with the implementation of LEAR — <http://lear.inrialpes.fr/people/dorko/downloads.html>. On average, the two strategies yield similar number of samples per image. The SIFT descriptors are scaled by a factor of 100 to prevent numerical instabilities during learning of the Gaussian mixture models.

G.1.2 DCT Features

DCT features are computed on a dense regular grid, with a step of 1-to-8 pixels (usually improved performance is obtained with lower step size, but at the cost of computation). 8×8 image patches are extracted around each grid point, and 8×8 DCT coefficients computed per patch and color channel. The DCT coefficients are vectorized into a row vector using the coefficient scanning mechanism defined by the MPEG standard. For monochrome images this results in a feature space of 64 dimensions. For color images the space is 192 dimensional, where the vectors for corresponding channels are interleaved. We currently use the YBR color-space defined by MPEG, but this selection has not been subject to detailed scrutiny.

G.1.3 Bag-of-Features

Using the bag of features representation, each image is modeled as a Gaussian Mixture Model (GMM) with a fixed number C of mixture components. The default value is $C = 8$ for DCT and $C = 16$ for SIFT, but can be modified when the database is initialized. In general using more mixture components result in improved performance, however the gains diminish over 16 mixture components. All Gaussian mixture parameters are estimated using the EM algorithm. The implementation is fairly standard, the only details worth mentioning are the following.

1. All Gaussians have diagonal covariances.

2. In order to avoid singularities, a variance limiting constrain is applied. This constrain places a minimum value of 10(0.01) on the variance along each dimension of the DCT(SIFT) feature space. Note, if new features are being introduced, a good estimate of minimum covariance using cross-validation techniques should be obtained.
3. Initialization is performed with a vector quantizer designed by the Linde-Buzo-Gray (LBG) algorithm, using a variation of the cell splitting method described in [81]. For more details please see [162].
4. The EM iterations for DCT(SIFT) features are restricted to 5(15) iterations. More iterations are required for SIFT features as 1) there are more mixture components, 2) unlike DCT, every dimension of SIFT has high variance.

G.1.4 Bag-of-Words

To obtain the bag-of-words representation, the space of image features is quantized using the LBG algorithm with a fixed number B of clusters. The default value is $B = 256$ codeword for both DCT and SIFT, although several experiments are performed with codewords as high as 4096 (e.g. topic-supervised LDA models). In general increasing the size of codebook leads to improved performance. Note that the initialization of GMM uses the codebooks learned with LBG algorithm.

G.1.5 Semantic Multinomial

To compute the Semantic Multinomial(SMN), the posterior probability of the concepts given an image, is computed using (2.21) for BoF and using (2.23) for BoW. (2.21) yields SMN which are almost uniform for BoW. SMN are regularized using $\pi_0 = 1$ for QBSE and $\pi_0 = 0.0001$ for holistic context models. Unless otherwise mentioned, similarity between two SMNs is computed using KL divergence.

G.2 Concept/Category Models

G.2.1 Appearance Based Models

GMM is the choice of probability distribution for the appearance based concept models. Given a training set of images, along with their GMM learned using the approach described above, appearance models are learned using hierarchical estimation technique proposed in [162]. For DCT(SIFT), 128(512) mixture components are used, although more mixture components leads to better performance.

G.2.2 Holistic Context Models

Contextual class models are learned using the outputs of appearance based models. Dirichlet Mixture Model (DMM) is the choice of probability distribution. Given a training set of images, DMM can be learned using image-SMNs, however since most datasets used in this work has only ~ 100 images per concept available for training, data augmentation techniques of Section 6.3.5 is employed. To increase the cardinality of the training sets used for contextual modeling, 800 random sets of 30 patches are sampled per image, yielding 800 patch-SMNs per image. Image-SMNs are then computed from these, with (2.21) or (2.23).

The parameters of the contextual class models are learned using GEM as described in Appendix B. The implementation is fairly standard, the only details worth mentioning are the following.

1. The number of mixture components is set to 42. Given sufficient training data, in general higher number of mixture components yield better recognition accuracies, however the benefits are limited over 40 mixture components.
2. In order to avoid singularities, a variance limiting constrain is applied. Since both high and low values of α parameter can lead to low variance, α values are constrained to a minimum and a maximum value of 0.01 and 100 respectively. The performance of the system is not very sensitive to the choice of these values, however a sanity check must be performed to ensure that they are not beyond reasonable limits.

3. Initialization is performed with a vector quantizer designed by the Linde-Buzo-Gray (LBG) algorithm, using a variation of the cell splitting method described as described above. The similarity between different SMN is computed using KL divergence.
4. The GEM iterations are restricted to 15 iterations.

G.3 Topic Supervised LDA

Topic supervised LDA models are built using BoW representation with the vocabulary size ranging from 128 to 4096. For each dataset, codebooks are generated from a random collection of 300 examples per training image. For experiments using LDA and sLDA we use the code available online¹. This code was modified for cLDA and topic-supervised LDA. The number of topics is varied from 10 to 100 for topic discovery approaches. For topic-supervised models, the number of topics is equal to the number of classes. The α_k parameter is set to 1 in all experiments except for cLDA and ts-cLDA, where an asymmetric α_y parameter is learned per class. Although not explicitly shown in Figure Figure 7.1, we use the “smoothed” version of various LDA models with a Dirichlet prior on the topic-distributions [14], using a symmetric hyper-parameter of 0.001. The performance of various models is not very sensitive to the choice of both α_k and the smoothing parameter.

¹<http://www.cs.princeton.edu/~blei/lda-c/> and <http://www.cs.princeton.edu/~chongw/slda/> respectively.

Bibliography

- [1] F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.
- [2] P. Auer. On learning from multi-instance examples: Empirical evaluation of a theoretical approach. In *Proceedings of 14th International Conference on Machine Learning*, pages 21–29, 1997.
- [3] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.
- [4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.
- [5] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *ICCV*, volume 2, pages 408–415, Vancouver, 2001.
- [6] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, pages 1–8, 2008.
- [7] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 509–522, 2002.
- [8] I. Biederman. On the semantics of a glance at a scene. *Perceptual organization*, pages 213–263, 1981.
- [9] I. Biederman. Aspects and extension of a theory of human image understanding. *Computational processes in human vision: An interdisciplinary perspective*, pages Ablex Publishing Corporation, New Jersey, 1988.
- [10] I. Biederman, RJ Mezzanotte, and JC Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. In *Cognitive Psychology*, volume 14, pages 143–77, 1982.
- [11] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

- [12] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 2003.
- [13] D.M. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.
- [14] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [15] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. *Proc. CVPR*, 2008.
- [16] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *9th European Conference on Computer Vision*, pages 517 – 30, Graz, Austria, 2006.
- [17] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 30(4):712, 2008.
- [18] S. Boughorbel, J.P. Tarel, and N. Boujemaa. Generalized histogram intersection kernel for image recognition. In *IEEE International Conference on Image Processing*, volume 3. IEEE, 2005.
- [19] W.L. Buntine. Operations for learning with graphical models. *Arxiv preprint cs/9412102*, 1994.
- [20] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. *Proc. Int. Conf. on Data Mining*, pages 427–432, 2007.
- [21] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29(3):394–410, March, 2007.
- [22] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego*, 2005.
- [23] AB Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *IEEE CVPR*, volume 1, 2005.
- [24] H. Cheng, Z. Liu, and J. Yang. Sparsity induced similarity measure for label propagation. *ICCV*, 2009.
- [25] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

- [26] Ingemar J. Cox, Joumana Ghosn, Thomas V. Papatomas, and Peter N. Yianilos. Hidden annotation in content based image retrieval. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.
- [27] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
- [28] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 39:65, 2008.
- [29] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- [30] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *J. of the Royal Statistical Society*, B-39, 1977.
- [31] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. *Information Processing & Management*, 40(5):807–827, 2004.
- [32] G. Doyle and C. Elkan. Accounting for word burstiness in topic models. In *Proceedings 26th International Conference on Machine Learning*, 2009.
- [33] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001.
- [34] D. Dunn and W. Higgins. Optimal gabor filters for texture segmentation. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 7(4), July 1995.
- [35] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.
- [36] 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*, 2007.
- [37] H.J. Escalante, C.A. Hérnadez, L.E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings 1st ACM International Conference on Multimedia Information Retrieval*, pages 172–179. ACM, 2008.
- [38] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

- [39] C. Fellbaum. *Wordnet: an electronic lexical database*. MIT Press, 1998.
- [40] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 2009.
- [41] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.
- [42] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, June 2003.
- [43] M. Fink and P. Perona. Mutual boosting for contextual inference. *Neural Information Processing Systems*, 2004.
- [44] J.W. Fisher, T. Darrell, W.T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. *Advances in Neural Information Processing Systems*, pages 772–778, 2001.
- [45] D. Forsyth and M. Fleck. Body plans. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 678–683, 1997.
- [46] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, 1946.
- [47] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. *IEEE Conference in Computer Vision and Pattern Recognition (CVPR) 2008, Anchorage, USA.*, 2008.
- [48] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman Hall, 1995.
- [49] T. Gevers and A. Smeulders. Picktoseek: Combining color and shape invariant features for image retrieval. *IEEE Trans. on Image Processing*, 9(1):102–119, January 2000.
- [50] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.
- [51] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

- [52] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical report, Caltech, 2006.
- [53] N. Haering, Z. Myles, and N. Lobo. Locating deducuous trees. In *Workshop in Content-based Access to Image and Video Libraries*, pages 18–25, 1997, San Juan, Puerto Rico.
- [54] Junwei Han and Lei Guo. A new image retrieval system supporting query by semantics and example. In *ICIP (3)*, 2002.
- [55] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
- [56] X. He, O. King, W.Y. Ma, M. Li, and H.J. Zhang. Learning a semantic space from user’s relevance feedback for image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(1):39–48, 2003.
- [57] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. *10th European Conference on Computer Vision, Marseille, France*, page 30, 2008.
- [58] T. Hofmann. Probabilistic latent semantic indexing. *ACM SIGIR*, pages 50–57, 1999.
- [59] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
- [60] J. Iria, F. Ciravegna, and J. Magalhães. Web news categorization using a cross-media document graph. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8. ACM, 2009.
- [61] A. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition Journal*, 29, August 1996.
- [62] N. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, 1984.
- [63] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.
- [64] IT Jolliffe. *Principal Component Analysis*. Springer, 2002.
- [65] AJ Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR. IEEE*, 2009.

- [66] I. Khan, A. Saffari, and H. Bischof. Tvgraz: Multi-modal learning of object categories by combining textual and visual features. In *Proceedings 33rd Workshop of the Austrian Association for Pattern Recognition*, 2009.
- [67] J.J. Kivinen, EB Sudderth, and MI Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *ICCV*. Citeseer, 2007.
- [68] Yasushi Kiyoki, Takashi Kitagawa, and Takanari Hayama. A metadatabase system for semantic image search by a mathematical model of meaning. *SIGMOD Rec.*, 23(4):34–41, 1994.
- [69] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. Combining image captions and visual analysis for image concept classification. In *Proceedings 9th International Workshop on Multimedia Data Mining at ACM SIG Knowledge Discovery and Data Mining*, pages 8–17. ACM New York, NY, USA, 2008.
- [70] H. Kueck, P. Carbonetto, and N. Freitas. A constrained semi-supervised learning approach to data association. In *European Conference on Computer Vision*, Prague, Czech Republic, 2004.
- [71] S. Lacoste-Julien, F. Sha, and M.I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems 21 (NIPS08)*, 2008.
- [72] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS, Vancouver*, 2003.
- [73] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Neural Information Processing Systems, Vancouver, Canada*, 2003.
- [74] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. IEEE Conf. Comp. Vision Patt. Recog*, 2005.
- [75] C. S. Lee, W.-Y. Ma, and H. Zhang. Information embedding based on user’s relevance feedback for image retrieval. In *Proc. SPIE Vol. 3846, p. 294-304, Multimedia Storage and Archiving Systems IV*, pages 294–304, 1999.
- [76] D. Li, N. Dimitrova, M. Li, and I.K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings 11th ACM International Conference on Multimedia*, pages 604–611. ACM, 2003.
- [77] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, pages 524–531, 2005.

- [78] F.F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci US A*, 99(14):9596–601, 2002.
- [79] L.J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. *Proc. ICCV*, 2007.
- [80] J.J. Lim, P. Arbeláez, C. Gu, and J. Malik. Context by region ancestry. In *ICCV*. Citeseer, 2010.
- [81] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, 1980.
- [82] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *CVPR*. IEEE, 2009.
- [83] Jingen Liu and Mubarak Shah. Scene modeling using co-clustering. *International Conference on Computer Vision*, 2007.
- [84] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *IEEE International Conference on Multimedia and Expo*, 2001.
- [85] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [86] Jing Lu, Shao ping Ma, and Min Zhang. Automatic image annotation based-on model space. In *IEEE NLP-KE*, 2005.
- [87] Ye Lu, HongJiang Zhang, Liu Wenyin, and Chunhui Hu. Joint semantics and feature based image retrieval using relevance feedback. *IEEE Transactions on Multimedia*, 5(3):339–347, 2003.
- [88] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge Univ Pr, 2003.
- [89] M.I. Mandel and D.P.W. Ellis. Multiple-instance learning for music information retrieval. In *Proceedings of International Symposium of Music Information Retrieval*, 2008.
- [90] R. Manmatha and S. Ravela. A syntatic characterization of appearance and its application to image retrieval. In *SPIE Conference on Human Vision and Electronic Imaging II*, volume 3016, 1997, San Jose, California.
- [91] C.D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [92] O. Maron and T. Lozano-Perez. A framework for multiple instance learning. In *Neural Information Processing Systems, Denver, Colorado*, 1998.

- [93] C.T. Meadow, B.R. Boyce, D.H. Kraft, and C.L. Barry. *Text Information Retrieval Systems*. Emerald Group Pub Ltd, 2007.
- [94] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, pages 1615–1630, 2005.
- [95] T.P. Minka. Estimating a dirichlet distribution. <http://research.microsoft.com/~minka/papers/dirichlet/>, 1:3, 2000.
- [96] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007.
- [97] Y. Mori, H. Takahashi, and R. Oka. Automatic word assignment to images based on image division and vector quantization. In *Proceedings of Recherche d’Information Assistée par Ordinateur (RIAO)*. Citeseer, 2000.
- [98] Henning Muller, Stephane Marchand-Maillet, and Thierry Pun. The truth about corel - evaluation in image retrieval. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, 2002.
- [99] S. Nakamura. Statistical multimodal integration for audio-visual speech processing. *IEEE Transactions on Neural Networks*, 13(4):854–866, 2002.
- [100] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.
- [101] W. Niblack and et al. The qbic project: Querying images by content using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, pages 173–181, SPIE, Feb. 1993, San Jose, California.
- [102] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. *Proc. ECCV*, 4:490–503, 2006.
- [103] A. Oliva and P.G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41(2):176–210, 2000.
- [104] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [105] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception*, 2006.
- [106] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the ImageCLEF 2009 photo task. *CLEF working notes*, 2009.

- [107] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int. Journal of Computer Vision*, Vol. 18(3):233–254, June 1996.
- [108] T.T. Pham, N.E. Maillot, J.H. Lim, and J.P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings 16th ACM International Conference on Information and Knowledge Management*, pages 439–444. ACM, 2007.
- [109] R. Picard. Digital libraries: Meeting place for high-level and low-level vision. In *Proc. Asian Conf. on Computer Vision*, December 1995, Singapore, USA.
- [110] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 6174, 1999.
- [111] M. Porat and Y. Zeevi. Localized texture processing in vision: Analysis and synthesis in the gaborian space. *IEEE Trans. on Biomedical Engineering*, 36(1):115–129, January 1989.
- [112] D. Putthividhya, HT Attias, and SS Nagarajan. Supervised topic model for automatic image annotation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.
- [113] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. *Proceedings. Tenth IEEE International Conference on Computer Vision*, Vol. 1:883 – 90, 2005.
- [114] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1575–1589, 2007.
- [115] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. *Computer Vision, IEEE 11th International Conference on*, pages 1–8, 2007.
- [116] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.
- [117] T. Randen and J.H. Husoy. Filtering for texture classification: A comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291–310, 1999.

- [118] K. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, 1990.
- [119] N. Rasiwasia, PL Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on*, 9(5):923–938, 2007.
- [120] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [121] L.W. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, 2004.
- [122] HA Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–208, 1996.
- [123] Y. Rui and T. Huang. Optimizing learning in image retrieval. *IEEE CVPR*, 2000.
- [124] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
- [125] G. Salton and J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
- [126] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.
- [127] J. Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [128] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *International Journal of Computer Vision*, pages 1–22, 2007.
- [129] J. Sivic, B.C. Russell, A. Efros, A. Zisserman, and W.T. Freeman. Discovering object categories in image collections. *Proc. ICCV*, 1:65, 2005.
- [130] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477, 2003.
- [131] M. Slaney. Semantic-audio retrieval. In *IEEE International Conference on Acoustics Speech and Signal Processing*, volume 4, pages 4108–4111. IEEE, 2002.

- [132] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proceedings 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
- [133] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [134] J. Smith and S. Chang. Visualeek: a fully automated content-based image query system. In *ACM Multimedia, Boston, Massachusetts*, pages 87–98, 1996.
- [135] J. R. Smith, C.-Y. Lin, M. R. Naphade, A. Natsev, and B. L. Tseng. Validity-weighted model vector-based retrieval of video. In *Proceedings of the SPIE, Volume 5307, pp. 271-279 (2003)*., pages 271–279, 2003.
- [136] JR Smith. Image retrieval evaluation. *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.
- [137] J.R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. *ICME*, pages 445–448, 2003.
- [138] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.
- [139] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427, 2007.
- [140] EB Sudderth, A. Torralba, WT Freeman, and AS Willsky. Learning hierarchical models of scenes, objects, and parts. *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, 2, 2005.
- [141] M. Swain and D. Ballard. Color indexing. *Int. Journal of Computer Vision*, Vol. 7(1):11–32, 1991.
- [142] Martin Szummer and Rosalind Picard. Indoor-outdoor image classification. In *Workshop in Content-based Access to Image and Video Databases*, 1998, Bombay, India.
- [143] S. M. M. Tahaghoghi, James A. Thom, and Hugh E. Williams. Are two pictures better than one? In *ADC '01: Proceedings of the 12th Australasian database conference*, pages 138–144, Washington, DC, USA, 2001. IEEE Computer Society.
- [144] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

- [145] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, pages 169–191, 2003.
- [146] A. Torralba, K.P. Murphy, and W.T. Freeman. Contextual models for object detection using boosted random fields. *Advances in Neural Information Processing Systems*, 2004.
- [147] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. *ECCV*, pages 776–789, 2010.
- [148] T. Tsirikia and J. Kludas. Overview of the wikipedia multimedia task at ImageCLEF 2009. In *Working Notes for the CLEF Workshop*, 2009.
- [149] 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*, 16(2):467–476, February 2008.
- [150] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.
- [151] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.
- [152] A. Vailaya, A. Jain, and H. Zhang. On image classification: City vs. landscape. *Pattern Recognition*, 31:1921–1936, December 1998.
- [153] K.E.A. Van De Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1582–1596, 2009.
- [154] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, and A.W.M. Smeulders. Kernel codebooks for scene categorization. *Proc. ECCV*, pages 696–709, 2008.
- [155] M. Vasconcelos, N. Vasconcelos, and G. Carneiro. Weakly supervised top-down image segmentation. *CVPR*, pages 1001–1006, 2006.
- [156] N. Vasconcelos. Minimum probability of error image retrieval. *IEEE Trans. on Signal Processing*, August 2004.
- [157] N. Vasconcelos. Minimum probability of error image retrieval. *IEEE Trans. on Signal Processing*, 52(8), August 2004.
- [158] N. Vasconcelos. A unified view of image similarity. In *Proc. Int. Conf. Pattern Recognition*, Barcelona, Spain, 2000.

- [159] N. Vasconcelos. Image indexing with mixture hierarchies. In *Proc. IEEE CVPR*, Kawai, Hawaii, 2001.
- [160] N. Vasconcelos and M. Kunt. Content-based retrieval from image databases: Current solutions and future directions. In *Proc. Int. Conf. Image Processing*, Thessaloniki, Greece, 2001.
- [161] N. Vasconcelos and A. Lippman. Learning over multiple temporal scales in image databases. In *Proc. European Conference on Computer Vision, Dublin, Ireland*, 2000.
- [162] Nuno Vasconcelos. *Bayesian models for visual information retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.
- [163] A. Vinokourov, D.R. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *4th International Symposium on Independent Component Analysis and Blind Source Separation*, 2003.
- [164] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. *Advances in Neural Information Processing Systems*, pages 1497–1504, 2003.
- [165] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 1(2), 2002.
- [166] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. *DAGM.04 Annual Pattern Recognition Symposium*, 2004.
- [167] C. Wang, D. Blei, and F.F. Li. Simultaneous image classification and annotation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.
- [168] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *Proceedings of 19th International Conference on Pattern Recognition*, 2009.
- [169] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *IEEE International Conference on Computer Vision*, pages 428–435, 2009.
- [170] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. *Human Motion—Understanding, Modeling, Capture and Animation*, pages 240–254, 2007.
- [171] T. Westerveld. Image retrieval: Content versus context. *Content-Based Multimedia Information Access*, pages 276–284, 2000.

- [172] T. Westerveld and A P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *In Proceedings of the Multimedia Information Retrieval Workshop*, Toronto, Canada, August 2003.
- [173] T. Westerveld and A.P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on.easy.data. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, Multimedia Information Retrieval Workshop*. Citeseer, 2003.
- [174] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2):251–261, 2006.
- [175] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings 17th ACM International Conference on Multimedia*, pages 175–184. ACM, 2009.
- [176] Y. Yang, Y.T. Zhuang, F. Wu, and Y.H. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008.
- [177] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. *Proc. CVPR*, 2:2126–2136, 2006.
- [178] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *Proceedings 15th ACM International Conference on Multimedia*, page 276. ACM, 2007.
- [179] X. Zhou, N. Cui, Z. Li, F. Liang, and T.S. Huang. Hierarchical gaussianization for image classification. In *IEEE 12th International Conference on Computer Vision*, pages 1971–1977. IEEE, 2009.
- [180] J. Zhu, A. Ahmed, and E.P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1257–1264. ACM, 2009.
- [181] Y. Zhuang, Y. Yang, F. Wu, and Y. Pan. Manifold learning based cross-media retrieval: a solution to media object complementary nature. *Journal of VLSI Signal Processing*, 46(2):153–164, 2007.
- [182] Y.T. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.