



**HAL**  
open science

# Handling imperfections for multimodal image annotation

Amel Znaidia

► **To cite this version:**

Amel Znaidia. Handling imperfections for multimodal image annotation. Other. Ecole Centrale Paris, 2014. English. NNT : 2014ECAP0017 . tel-01012009

**HAL Id: tel-01012009**

**<https://theses.hal.science/tel-01012009>**

Submitted on 25 Jun 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ECOLE CENTRALE DES  
ARTS  
ET MANUFACTURES  
"ECOLE CENTRALE PARIS"



# PHD THESIS

in candidacy for the degree of

**Doctor of Ecole Centrale Paris**  
**Specialty: Computer Science**

Defended by  
**Amel ZNAIDIA**

## **Handling Imperfections for Multimodal Image Annotation**

prepared at Ecole Centrale Paris, MAS Laboratory  
defended on February 11<sup>th</sup>, 2014

### **Jury:**

<i>Chairman:</i>	Pr. Henri Maître	Telecom Paritech, France
<i>Reviewers:</i>	Pr. Bernard Merialdo	Eurecom Institute, France
	Pr. Stéphane Marchand-Maillet	University of Geneva, Switzerland
<i>Examiners:</i>	Pr. Patrick Lambert	Polytech Annecy-Chambéry, France
	Pr. Matthieu Cord	UPMC-Sorbonne Universities, France
	Dr. Stéphane Ayache	Aix-Marseille University, France
<i>Advisors:</i>	Pr. Nikos Paragios	Ecole Centrale Paris, France
	Dr. Céline Hudelot	Ecole Centrale Paris, France
	Dr. Hervé Le Borgne	CEA Saclay, France

order number : 2014ECAP0017



# *Acknowledgements*

First, I would like to express my sincere gratitude to my Director Prof. Nikos Paragios and my advisors Dr. Céline Hudelot and Dr. Hervé Le Borgne for offering me the opportunity to achieve my PhD within the MAS laboratory of Ecole Centrale Paris in a collaboration with the Alternative Energies and Atomic Energy Commission (CEA). I would like to thank them for the continuous support of my Ph.D study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the committee members for accepting to review my thesis work. Specifically, I'm grateful to the chairman Pr. Henri Maître, my reviewers Pr. Bernard Merialdo and Pr. Stéphane Marchand-Maillet and the examiners Pr. Patrick Lambert, Pr. Matthieu Cord and Dr. Stéphane Ayache for their encouragement, insightful comments and questions.

Then, I would like to acknowledge all the members of the MAS and the CEA-List laboratories. I am also thankful to Annie Glomeron and Sylvie Dervin who were always concerned to provide me the necessary conditions for the completion of my thesis. I am particularly grateful for my colleagues Aymen Shabou and Adrian Popescu for our collaboration. I appreciate their feedback, constructive comments, suggestions and support. I would like to offer my special thanks to my colleagues and friends for their support and encouragement. Namely, Dhouha Bou Amor, Hichem Bannour, Aymen Shabou, Sofiene Kachroudi, and the others.

Last, but not the least, I would like to thank my family: my husband Walid HACHICHA, my parents, my sisters and my brothers for providing me all the needed support to succeed in this PhD work.

# *Abstract*

This thesis deals with multimodal image annotation in the context of social media. We seek to take advantage of textual (tags) and visual information in order to enhance the image annotation performances. However, these tags are often noisy, overly personalized and only a few of them are related to the semantic visual content of the image. In addition, when combining prediction scores from different classifiers learned on different modalities, multimodal image annotation faces their imperfections (uncertainty, imprecision and incompleteness). Consequently, we consider that multimodal image annotation is subject to imperfections at two levels: the representation and the decision. Inspired from the information fusion theory, we focus in this thesis on defining, identifying and handling imperfection aspects in order to improve image annotation.

To deal with imperfections at the representation level, we start by identifying and defining these aspects in the context of multimodal image annotation. Once these imperfections are well identified and defined, we propose two novel tag-based signatures to handle such imperfections. To tackle the problem of untagged images, we propose a novel method based on visually similar neighbors and Belief theory. Afterwards, we propose a compact semantic signature that results from the combination of textual and visual modalities.

To cope with imperfections at the decision level, we propose two multimodal frameworks which consist in the combination of classifiers from different modalities. The first framework is based on the Stack Generalization scheme where imperfection aspects are handled implicitly in the learning stage. The second one is based on classifier fusion in the Belief theory based on the Dempster-Shafer rule to handle explicitly imperfections that may occur while combining different classifiers. Extensive experimental evaluations show that our approaches achieve state-of-the-art results on several standard and challenging datasets.

**Keywords:** Multimodal Image Annotation, Supervised Image Classification, Tag Imperfections, Uncertainty, Imprecision, Incompleteness, Belief Theory, Dempster Shafer.

# Résumé

La présente thèse s'intéresse à l'annotation multimodale d'images dans le contexte des médias sociaux. Notre objectif est de combiner les modalités visuelles et textuelles (tags) afin d'améliorer les performances d'annotation d'images. Cependant, ces tags sont généralement issus d'une indexation personnelle, fournissant une information imparfaite et partiellement pertinente pour un objectif de description du contenu sémantique de l'image. En outre, en combinant les scores de prédiction de différents classifieurs appris sur les différentes modalités, l'annotation multimodale d'image fait face à leurs imperfections: l'incertitude, l'imprécision et l'incomplétude. Dans cette thèse, nous considérons que l'annotation multimodale d'image est soumise à ces imperfections à deux niveaux : niveau représentation et niveau décision. Inspiré de la théorie de fusion de l'information, nous concentrons nos efforts dans cette thèse sur la définition, l'identification et la prise en compte de ces aspects d'imperfections afin d'améliorer l'annotation d'images.

Pour traiter les imperfections au niveau de la représentation, nous commençons par identifier et définir ses aspects dans le contexte de l'annotation multimodale d'images. Une fois que ces imperfections sont bien identifiées et définies, nous proposons deux nouvelles signatures basées sur l'information textuelle pour prendre en compte de telles imperfections. Pour aborder le problème d'incomplétude des tags, nous proposons une nouvelle méthode basée d'une part sur les images visuellement voisines et sur une combinaison de leur contribution utilisant la théorie des fonctions de croyance. Ensuite, nous proposons une signature sémantique compacte qui résulte de la combinaison de modalités textuelle et visuelle.

Pour prendre en compte les imperfections au niveau de la décision, nous proposons deux modèles qui consistent à combiner des classifieurs appris sur les différentes modalités. Le premier modèle est basé sur l'algorithme de "Stack Generalization" où les aspects d'imperfections sont traités implicitement dans l'étape d'apprentissage. Le deuxième modèle est basé sur la fusion de classifieurs dans la théorie des fonctions de croyance en utilisant la règle de Dempster-Shafer pour traiter explicitement les imperfections qui peuvent exister en combinant les différents classifieurs. Les résultats expérimentaux montrent que les méthodes proposées dépassent l'état de l'art tout en restant moins coûteuses en calculs que les travaux récents dans le domaine.

**Mots Clés:** Annotation multimodale d'images, Classification supervisée d'images, Imperfections, Tags, Incertitude, Imprécision, Incomplétude, Théorie de Croyances, La règle de Dempster-Shafer.

# Contents

<b>Acknowledgements</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Résumé</b>	<b>iv</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xvi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivations . . . . .	4
1.1.1 Tagging Motivation . . . . .	6
1.1.2 Tag characteristics . . . . .	7
1.1.3 Imperfection Aspects . . . . .	9
1.2 Goals . . . . .	12
1.3 Contributions . . . . .	13
1.3.1 Identifying tag imperfections at the representation level . . . . .	13
1.3.2 Handling imperfections at the representation level . . . . .	13
1.3.3 Handling imperfections at the decision level . . . . .	14
1.4 Organization of the Dissertation . . . . .	15
<b>2 State-of-the-art</b>	<b>17</b>
2.1 Introduction . . . . .	18
2.2 The Problem of Multimodal Image Annotation . . . . .	18
2.3 Multimodal Image Annotation . . . . .	20
2.3.1 Visual Description: Bag-Of-Visual-Words . . . . .	22
2.3.2 Textual Description . . . . .	26
2.3.3 Fusion Strategy . . . . .	30
2.4 Handling Tag Imperfections . . . . .	44
2.4.1 Tag Ranking & Relevance . . . . .	44
2.4.2 Tag Refinement & Suggestion . . . . .	47
2.5 Handling Imperfections at the decision level . . . . .	50
2.5.1 Probability theory . . . . .	51

2.5.2	Fuzzy Set theory . . . . .	52
2.5.3	Possibility theory . . . . .	53
2.5.4	Dempster-Shafer theory . . . . .	53
2.6	Image Databases & Evaluation Campaigns . . . . .	57
2.6.1	Evaluation Campaigns . . . . .	57
2.6.2	Image Databases . . . . .	59
2.7	Conclusions . . . . .	63
<b>I</b>	<b>Representation Level</b>	<b>65</b>
<b>3</b>	<b>Handling Textual Imperfections</b>	<b>66</b>
3.1	Introduction . . . . .	67
3.2	Textual Imperfections in the context of Multimedia Annotation . . . . .	68
3.3	Semantic Similarity between Words . . . . .	70
3.3.1	Knowledge-Based Measures . . . . .	71
3.3.2	Corpus-Based Measures . . . . .	72
3.3.3	Discussion . . . . .	76
3.4	Problem Formalization . . . . .	77
3.5	Soft Bag-of-Concepts Signature . . . . .	80
3.5.1	Tag Modeling . . . . .	81
3.5.2	Coding/Pooling . . . . .	81
3.6	Local Soft Tag Coding Signature . . . . .	83
3.6.1	Tag Modeling . . . . .	83
3.6.2	Coding/pooling . . . . .	85
3.7	Adopted Semantic Similarities . . . . .	85
3.8	Experimental Evaluation . . . . .	87
3.8.1	Experimental Setup . . . . .	88
3.8.2	Experimental Results . . . . .	89
3.9	Conclusion and Discussion . . . . .	98
<b>4</b>	<b>Handling Full Textual Incompleteness</b>	<b>101</b>
4.1	Introduction . . . . .	102
4.2	Problem Formalization . . . . .	102
4.3	Proposed Tag Completion Method . . . . .	103
4.3.1	Finding candidate tags . . . . .	105
4.3.2	Predicting final tags . . . . .	107
4.4	Tag Suggestion Dataset Construction . . . . .	110
4.5	Experimental Evaluation . . . . .	110
4.5.1	Datasets . . . . .	111
4.5.2	Experimental Setup . . . . .	111
4.5.3	Experimental Results . . . . .	113
4.6	Conclusion and Discussion . . . . .	116
<b>5</b>	<b>Bag-of-Multimedia Words Representation</b>	<b>119</b>

5.1	Introduction . . . . .	120
5.2	Tag vs. Visual words . . . . .	121
5.3	Bag-of-Multimedia Words Model . . . . .	123
5.3.1	Multimedia codebook learning . . . . .	123
5.3.2	Multimedia signature . . . . .	125
5.4	Experimental Evaluation . . . . .	128
5.4.1	Experimental Setup . . . . .	128
5.4.2	Experimental Results . . . . .	129
5.5	Conclusion and Discussion . . . . .	134
<b>II Decision Level</b>		<b>136</b>
<b>6</b>	<b>Multimodal Late Fusion for Image Annotation</b>	<b>137</b>
6.1	Introduction . . . . .	138
6.2	Proposed Multimodal Framework for Image Annotation using Stack Generalization . . . . .	138
6.2.1	Visual Features . . . . .	139
6.2.2	Textual Features . . . . .	140
6.2.3	Stack Generalization . . . . .	140
6.3	Experimental Evaluation . . . . .	143
6.3.1	Experimental Setup . . . . .	143
6.3.2	Experimental Results . . . . .	144
6.4	Conclusion and discussions . . . . .	147
<b>7</b>	<b>Combining Classifiers Based on Belief theory</b>	<b>149</b>
7.1	Introduction . . . . .	150
7.2	Belief Theory for Large-Scale Multi-Label Image Classification . . .	150
7.2.1	Building mass functions . . . . .	152
7.2.2	Dempster's Combination Rule . . . . .	153
7.2.3	Decision Making . . . . .	154
7.3	Experimental Evaluation . . . . .	154
7.3.1	Experimental Setup . . . . .	154
7.3.2	Experimental Results . . . . .	155
7.4	Conclusion and Discussions . . . . .	158
<b>8</b>	<b>Conclusions and Future Research Directions</b>	<b>159</b>
8.1	Contributions . . . . .	159
8.2	Perspectives for future research . . . . .	161
<b>Appendices</b>		<b>163</b>
<b>A</b>	<b>Evaluation measures</b>	<b>164</b>
<b>B</b>	<b>Publications of the Author</b>	<b>166</b>

**Bibliography**

**168**

# List of Figures

1.1	An illustration of the semantic gap problem. Images (a) and (b) have similar perceptual content (in this case represented by color histograms) but different meanings. Images (a) and (c) have the same meaning and different color histograms. . . . .	2
1.2	A snapshot of the social media Website Flickr. A typical image is associated with comments, tags and ratings (mark as a favorite). . .	5
1.3	An example of images from the Flickr website with their associated user tags. Most of tags are noisy for image annotation and only few tags are related to image visual content (marked in bold). . . . .	8
2.1	Illustration of the problem of multimodal image annotation formulated as a set of binary classification tasks. . . . .	21
2.2	Categorization of image annotation approaches in the context of social media. . . . .	22
2.3	The flowchart of the BOVW generation scheme. . . . .	23
2.4	Illustration of the coding and pooling steps. . . . .	24
2.5	The general scheme of the early fusion strategy. . . . .	31
2.6	Taxonomy of work based on early fusion strategy for image annotation in social media. . . . .	32
2.7	Graphical illustrations of the various generative pLSA models. . . .	34
2.8	The general scheme of the late fusion strategy. . . . .	35
2.9	Taxonomy of the work based on the late fusion strategy for image annotation in social media. . . . .	36
2.10	The general scheme of the transmedia fusion strategy for image annotation in social media. . . . .	39
2.11	An example of image from Flickr and its associated tag list. . . . .	44
2.12	An example of tag refinement and suggestion . . . . .	47
2.13	Relation of belief and plausibility and their negation. . . . .	55
2.14	PASCAL VOC'07 dataset example images with their associated user tags (below) and labels (on top). . . . .	60
2.15	ImageCLEF'11 dataset example images with their associated user tags and labels. . . . .	61
2.16	ImageCLEF'12 dataset example images with their associated user tags and labels. . . . .	62
2.17	NUS-WIDE dataset example images with their associated user tags and labels. . . . .	63

3.1	A taxonomy of tagging motivations in Flickr [Ames and Naaman, 2007]. . . . .	68
3.2	An example of images from Flickr website with their associated user tags. Most of tags are noisy and only few tags are related to image visual content. . . . .	69
3.3	Categorization of the state-of-the-art methods on semantic similarity measures. . . . .	71
3.4	Overview of the proposed signatures: Soft-BoC and LSTC. . . . .	78
3.5	Illustration of the Tag Modeling process. . . . .	79
3.6	Illustration of the Soft-BoC signature. . . . .	80
3.7	An example of image with its associated tags. . . . .	81
3.8	An example of images with sparse Flickr user tags. . . . .	83
3.9	Illustration of the LSTC signature. . . . .	84
3.10	Illustration of the tag coding process. . . . .	84
3.11	The Wu&Palmer similarity. . . . .	86
3.12	The “maximum sense pair” illustration. . . . .	86
3.13	Classification accuracy in terms of mean Average Precision (mAP) on the ImageClef’10 dataset. . . . .	91
3.14	Performances of the LSTC signature when changing the neighborhood window size in the tag-feature-space on the ImageClef’11. . . . .	92
3.15	Several images with their associated user tags on the left and the Soft-BoC signature on the right . . . . .	94
4.1	An overview of image tagging problem that we consider in this chapter. For an untagged image, given the set of its nearest neighbors with their associated tags, our goal is to predict a set of tags that describe its visual content. . . . .	103
4.2	An example of images from Flickr on same subject from two different users with their associated user tags. . . . .	104
4.3	Overview of our tag completion approach based on local soft coding and Belief theory. . . . .	105
4.4	An example to illustrate the local soft tag coding to enrich image description. Tags such as “ <i>Sport, Challenge</i> ” are added. . . . .	106
4.5	An example to illustrate the aggregation of nearest neighbor tag descriptions to obtain the list of “candidate tags”. . . . .	107
4.6	The $k$ nearest neighbors BBA for each tag are combined using Dempster’s rule of combination to form a final BBA for each tag. . . . .	109
4.7	Example of images from the dataset of [Sigurbjörnsson and van Zwol, 2008]. . . . .	111
4.8	Performance on the PASCAL VOC’07 dataset in terms of mean Average Precision with respect to the number of nearest neighbors. . . . .	114
4.9	Examples of tag suggestion by different methods. The <b>bold</b> font indicates <b>irrelevant</b> suggested tags. Original tags are not used. . . . .	117
5.1	An example of multimedia documents from Flickr composed with images and their associated user tags. . . . .	120

5.2	An illustration of the tag imperfections. All images are associated with the tag “zebra” while having different visual content and different semantic meanings. . . . .	122
5.3	An illustration of polysemous visual words: single visual word occurring on different (but locally similar) parts on different object categories . . . . .	123
5.4	An overview of Tag-coding procedure. For example to code the tag “cat”, visual word occurrences of images tagged with “cat” are aggregated with a sum-pooling. . . . .	125
5.5	An overview of the clustering step which consists of clustering column vectors of the tag-codes matrix using K-means in order to generate the multimedia codebook ( <i>M-codebook</i> ), which is formed of relevant multimedia words. . . . .	125
5.6	BOMW signature generation consists in two steps: coding and pooling. . . . .	126
5.7	Illustration of the BOMW coding step. In this example only the three nearest multimedia words are activated and the rest are set to zero. . . . .	127
5.8	Classification performances in terms of mAP on the PASCAL VOC’07 dataset while varying the sizes of visual and multimedia codebooks. . . . .	131
5.9	Classification performances in terms of mAP on the PASCAL VOC’07 dataset. . . . .	131
5.10	Comparison of the mAP on the ImageClef’12 dataset for the BOTW and the BOVW vs the proposed BOMW. . . . .	132
5.11	Comparison of the mAP on the ImageClef’12 dataset for the BOTW and the BOVW vs the proposed BOMW. . . . .	133
5.12	Comparison of the mAP on the ImageClef’12 dataset for the BOTW and the BOVW vs the proposed BOMW. . . . .	133
6.1	The flowchart of the proposed multimodal framework for image annotation using Stack Generalization. . . . .	139
6.2	A schematic illustration of the spatial pyramid representation. . . . .	141
6.3	Flowchart of the Stack Generalization approach. Left part is the illustration of the K-fold cross-validation process for creating the meta-level training dataset, and the right part is the stacking framework in the testing stage. . . . .	142
6.4	Flowchart of the proposed multimodal framework based on the Stack Generalization algorithm. Left part is the illustration of the K-fold cross-validation process for creating the meta-level training dataset, and the right part is the stacking framework at runtime. . . . .	143
7.1	Flowchart of the proposed system based on Belief theory. A combination is performed to obtain the final mass function, used to compute the plausibility for decision making. . . . .	151

---

7.2 A qualitative comparison between individual classifiers, the proposed method and the average rule. . . . .	157
--	-----

# List of Tables

1.1	Comparison of our results, based on the proposed multimodal framework based on Stack Generalization, to the best state-of-the-art classification performances in terms of mAP. . . . .	15
2.1	A Summary of the state-of-the-art approaches on Tag-based image annotation. . . . .	29
2.2	A summary of the most related and representative work based on the early fusion strategy for multimodal image annotation in the context of social media. . . . .	41
2.3	A summary of the most related and representative work based on the late fusion strategy for multimodal image annotation in the context of social media. . . . .	43
2.4	A summary of the most related and representative work based on the transmedia fusion strategy for multimodal image annotation in the context of social media. . . . .	43
2.5	A summary of terms used to describe tag imperfections in representative work in the context of social media. . . . .	49
2.6	Dataset statistics: number of images (train/test), tags, labels and untagged images for both PASCAL VOC'07 and NUS-WIDE datasets. . . . .	63
3.1	Classification of measures of semantic similarity and relatedness based on WordNet and their relative advantages/disadvantages. . . . .	74
3.2	Summary of dictionary size for the considered datasets. Tag frequency represents the minimum tag frequency threshold used to build the dictionary. . . . .	89
3.3	Comparison for WordNet similarities on the ImageClef'12 dataset using the LSTC signature in terms of mean Average Precision (mAP). . . . .	90
3.4	Evaluation of the Soft-BoC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on PASCAL VOC'07 and NUS-WIDE datasets. . . . .	92
3.5	Evaluation of the Soft-BoC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on the ImageCLEF datasets (ImageClef'10, ImageClef'11 and ImageClef'12). . . . .	93
3.6	Evaluation of the LSTC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on PASCAL VOC'07 and NUS-WIDE datasets. . . . .	95

---

3.7	Evaluation of the LSTC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on the ImageCLEF datasets (ImageClef'10, ImageClef'11 and ImageClef'12). . . . .	96
3.8	Comparison of the two proposed signatures to the state-of-the-art for image annotation, in terms of mean Average Precision (mAP) on the five considered datasets. . . . .	99
4.1	Number and proportion of untagged images in training and test sets, for the PASCAL VOC'07 and ImageClef'11 datasets. . . . .	112
4.2	Classification performances on PASCAL VOC'07 in terms of mAP with and without tag completion. . . . .	115
4.3	Classification performances on the ImageClef'11 dataset in terms of mAP with and without tag completion. . . . .	115
4.4	The Average Precision (AP) scores on the PASCAL VOC'07 dataset per concept, without tag completion (Baseline) and with tag completion (Our model). The best classification results for each class are marked in bold. . . . .	115
4.5	Classification performances on PASCAL VOC'07 in terms of mAP, for different methods. . . . .	116
4.6	Classification performances on the ImageClef'11 dataset in terms of mAP, for different methods. . . . .	116
4.7	Comparison of our system to the state-of-the-art methods on the tag suggestion task. . . . .	116
5.1	Classification performances on the PASCAL VOC'07 dataset in terms of mean Average Precision (mAP). . . . .	129
5.2	Classification performances on the ImageClef'12 dataset in terms of mAP score. . . . .	132
6.1	Comparison of our system to previous work for PASCAL VOC'07 classification challenge in terms of mAP. . . . .	145
6.2	Comparison of our system to previous work for ImageClef'11 classification challenge in terms of mAP. . . . .	146
6.3	Our system compared to the ImageClef12 Photo Annotation best performing system [Liu et al., 2012]. . . . .	147
6.4	Our system compared to multimodal image annotation state-of-the-art approaches on the NUS-WIDE dataset. . . . .	147
7.1	Comparative Performance of individual classifiers in terms of mAP for the ImageClef'11 dataset. . . . .	155
7.2	Comparative Performance of different combination strategies in terms of mean Average Precision (mAP) for the ImageClef'11 dataset. The best results are marked in bold. . . . .	155
7.3	Comparative Performance of individual classifiers, Dempster, Average and the ImageClef 2011 Winner [Binder et al., 2011] for some challenging classes in terms of mean Average Precision (mAP). . . . .	156

---

7.4	Comparative Performance of individual classifiers in terms of mean Average Precision (mAP) for the ImageClef'10 dataset. . . . .	157
7.5	Comparative Performance of different combination strategies in terms of mean Average Precision (mAP) for the ImageClef'10 dataset. The best results are printed in bold. . . . .	158

# Abbreviations

<b>BOW</b>	<b>Bag Of Words</b>
<b>BOMW</b>	<b>Bag Of Multimedia Words</b>
<b>BOTW</b>	<b>Bag Of Tag Words</b>
<b>BOVW</b>	<b>Bag Of Visual Words</b>
<b>BBA</b>	<b>Basic Blief Assignment</b>
<b>DS</b>	<b>Dempster Shafer</b>
<b>ECC</b>	<b>Ensemble of Classifier Chains</b>
<b>ESA</b>	<b>Explicit Semantic Analysis</b>
<b>EM</b>	<b>Expectation Maximisation</b>
<b>IDF</b>	<b>Inverse Document Frequency</b>
<b>HTC</b>	<b>Histogram of Textual Concepts</b>
<b>HAL</b>	<b>Hyperspace Analogue to Language</b>
<b>GMM</b>	<b>Gaussian Mixture Models</b>
<b>LDA</b>	<b>Latent Dirichlet Allocation</b>
<b>LSTC</b>	<b>Local Soft Tag Coding</b>
<b>MKL</b>	<b>Multiple Kernel Learning</b>
<b>pLSA</b>	<b>probabilistic Latent Semantic Analysis</b>
<b>SDA</b>	<b>Syntactic Distributional Analysis</b>
<b>SIFT</b>	<b>Scale Invariant Feature Transform</b>
<b>Soft-BoC</b>	<b>Soft Bag of Concepts</b>
<b>SVM</b>	<b>Support Vector Machines</b>
<b>SPM</b>	<b>Spatial Pyramid Matching</b>
<b>TF</b>	<b>Term Frequency</b>
<b>VSM</b>	<b>Vector Space Model</b>

# Chapter 1

## Introduction

Recent years have witnessed the transition from a Web where the content was generated mainly by website owners to a more open and social Web where users are not only information consumers but also producers [Tapscott and Williams, 2006]. With the rapid development of this new age of the Web, also known as *Web 2.0*, a large number of community contributed multimedia contents have been produced and shared on the Web. Social media repositories (such as Flickr<sup>1</sup>, YouTube<sup>2</sup> and Picasa<sup>3</sup>) allow users to upload and share their personal photos or videos. For example, Flickr reaches more than eight billion photos in 2013, uploaded from more than 87 million users. In particular, more than 3.5 million new images are uploaded daily<sup>4</sup>. An important feature of online social media services is that users can annotate their photos with their own keywords called *tags*, without relying on a controlled vocabulary. This voluntary activity of users who are annotating resources with tags is called *Tagging*.

This plethora of multimedia contents raised the imperative need to address the challenge of their storage, organization and indexing for future search and access. **Image annotation**, which represents a way to address this problem, has become a core research for content-based image indexing and retrieval [Duygulu et al., 2002; He et al., 2004; Carneiro and Vasconcelos, 2005; Nowak and Huiskes, 2010; Semenovich and Sowmya, 2010; Moser and Serpico, 2013]. It consists in automatically assigning a set of keywords (also called labels or concepts) from a predefined vocabulary to describe the semantic visual content of the image. A recent review about image annotation is proposed in [Zhang et al., 2012a]. The problem of image annotation can be viewed as a learning problem and in particular a supervised classification problem [Carneiro et al., 2007], where semantic concepts are

---

<sup>1</sup><http://www.flickr.com/>

<sup>2</sup><http://www.youtube.com/>

<sup>3</sup><http://picasa.google.com/>

<sup>4</sup><http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/>

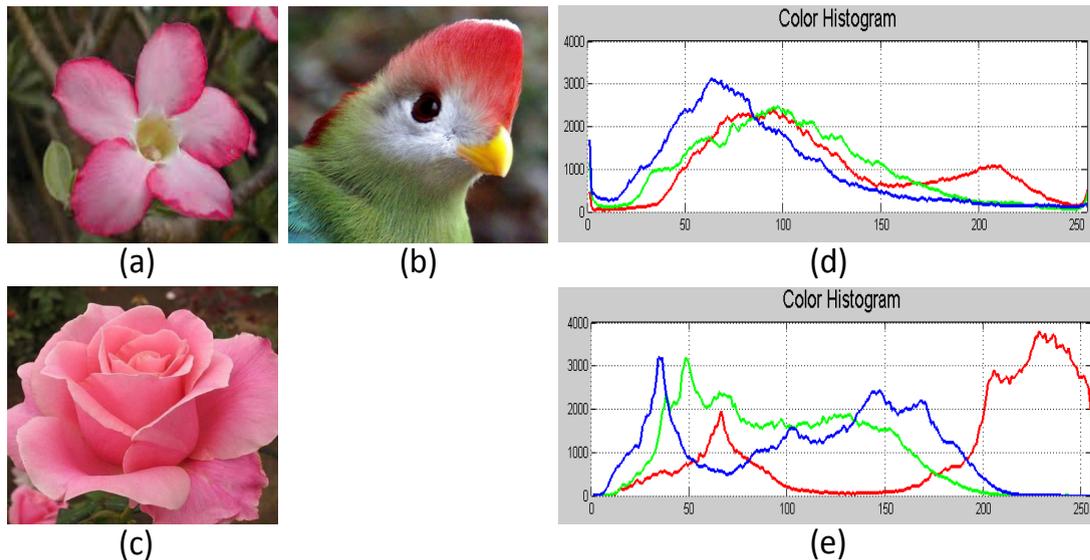


FIGURE 1.1: An illustration of the semantic gap problem. Images (a) and (b) have similar color histograms but different meanings. Images (a) and (c) have the same meaning but different perceptual contents.

learned from low-level features. In this thesis, we consider the supervised image classification scheme as the appropriate framework to study the image annotation problem in the context of social media. The problem of supervised image classification is posed as a set of classification tasks, where keywords are processed as classes and a set of classifiers are learned from low-level features to annotate an input image based on classification prediction scores. The supervised classification scheme consists of two steps: image description and concept learning. Images are commonly described using low-level features extracted from the image such as the Bag-of-Visual-Words (BOVW) representation [Sivic and Zisserman, 2003; Csurka et al., 2004]. However, these descriptions do not directly convey human understandable meaning and an important gap remains between visual descriptors and the semantic content of images. This problem is known as the *semantic gap* defined in [Smeulders et al., 2000] as “the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for an user in a given situation”. In Figure 1.1, we illustrate the semantic gap problem with a simple example in which the perceptual content is represented by color histograms. Images (a) and (b) have similar color histograms but different meanings (flower vs bird). Images (a) and (c) have the same meaning (flower) but different color histograms. Although the current state-of-the-art in content-based image annotation is progressing, it has not yet succeeded in bridging the semantic gap between semantic concepts and low-level visual features that are extracted from images. In particular, one of the main reasons is that these work have considered visual descriptors, i.e. the perceptual manifestation of the semantics, as sufficient to tackle the image annotation through supervised image classification. It seems that this is not the case.

As mentioned before, in community contributed collections, images do not appear alone but associated with various forms of textual descriptions such as tags. Tags are, as a consequence, a rich additional information to organize and access these shared multimedia contents. As opposed to pixels which do not convey semantic interpretations, tags which are directly issued from human language are notably useful when considering the semantic gap. However, these tags are generally noisy, overly personalized and only a few of them are really related to the semantic visual content of the image [Ames and Naaman, 2007]. Thus, the information extracted from tags is also not sufficient alone to narrow the semantic gap.

In the context of social media, images have, intrinsically, a multimodal nature (visual and textual). The two modalities are complementary and combining them to improve image annotation seems an appealing idea, in particular, in order to bridge the semantic gap. By the following, this will be referred by **multimodal image annotation**. Nevertheless, the two identified modalities are heterogeneous and of different nature. The main difference is related to the information vehiculated by both information sources regarding the image content. We assume that pixels can only bring information related to the visual content which is not the case of tag information that can be related to the user attention and tagging motivation. The main challenge is to take advantages of these two modalities to enhance multimodal image annotation performances.

An originality of this thesis is to consider the multimodal image annotation problem as an **information fusion** process. The latter was first defined in late 80's [F. E. White, 1987] as “*a process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats, and their significance*”. Later, [Bostrom and al., 2007] reviewed definitions proposed in the literature between 1987 and 2007 and proposed a new definition : “*Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making*”. According to the latter definition, we define multimodal annotation as the process of combining visual and tag modalities in order to improve image annotation performances. Moreover, it is common to consider that three fusion levels initially occurred in information processing namely, **information/data fusion** (low-level), **feature fusion** (intermediate-level), and **decision fusion** (high-level). Data fusion combines several sources of raw data to produce a new one that is expected to be more informative and synthetic than inputs. Feature fusion combines features extracted from different sources into an unique feature vector. Decision fusion uses a set of classifiers to provide a better and unbiased result. As highlighted by [Bloch, 2001], one important characteristic of information in fusion is its *imperfection*. This imperfection is always present and is the main reason of the fusion process. Thus, a main issue in this domain is

to define and handle the different types of imperfections (imprecision, uncertainty and incompleteness) that occur at different levels and to cope with them.

In the case of the multimodal image annotation viewed as a classification problem, according to the information fusion theory, two levels need to be considered: the **representation level** (description of multimodal data) and the **decision level** (fusion of classifiers). At the representation level, our sources of information are data descriptions of both modalities while at the decision level, prediction scores issued from classifiers are considered as sources of information.

In a perspective of data fusion, considering the fundamental difference in the nature of visual and textual information, it seems that there is no sense to perform fusion at data level (i.e. raw data). It is worth considering information imperfections at both representation and decision levels and handling them:

- At the representation level, we argue and demonstrate that only tags related to the visual content of the image are relevant for image annotation. Others are considered as imperfect in the perspective of information fusion process at the representation level. Handling such imperfections seems to be interesting to reduce the semantic gap, thus enhancing multimodal image annotation performances. However, there is no exact definition and identification of noisy tags in the state-of-the-art approaches dealing with multimedia annotation. Consequently, to handle such imperfections some definitions need to be stated clearly. Once these imperfections are well defined and identified, we focus on how to take them into account while designing tag-based signatures at the representation level.
- At the decision level, multimodal image annotation faces the problem of imperfections introduced by machine learning algorithms, when combining score predictions from different classifiers learned on different features and modalities.

## 1.1 Motivations

Recently, multimodal fusion has received an increasing attention in the multimedia analysis community [Kludas and Marchand-Maillet, 2011; Souvannavong et al., 2005; Marchand-Maillet et al., 2010; Niaz and Mérialdo, 2013]. A recent review can be found in [Atrey et al., 2010]. In social media collections, some sources of information, especially tags, offer the possibility to involve semantic evidence during the analysis of visual content in image collections. Therefore, the combination of these data sources with visual characteristics of images has received an increasing attention from the research community in multimedia fusion. Many

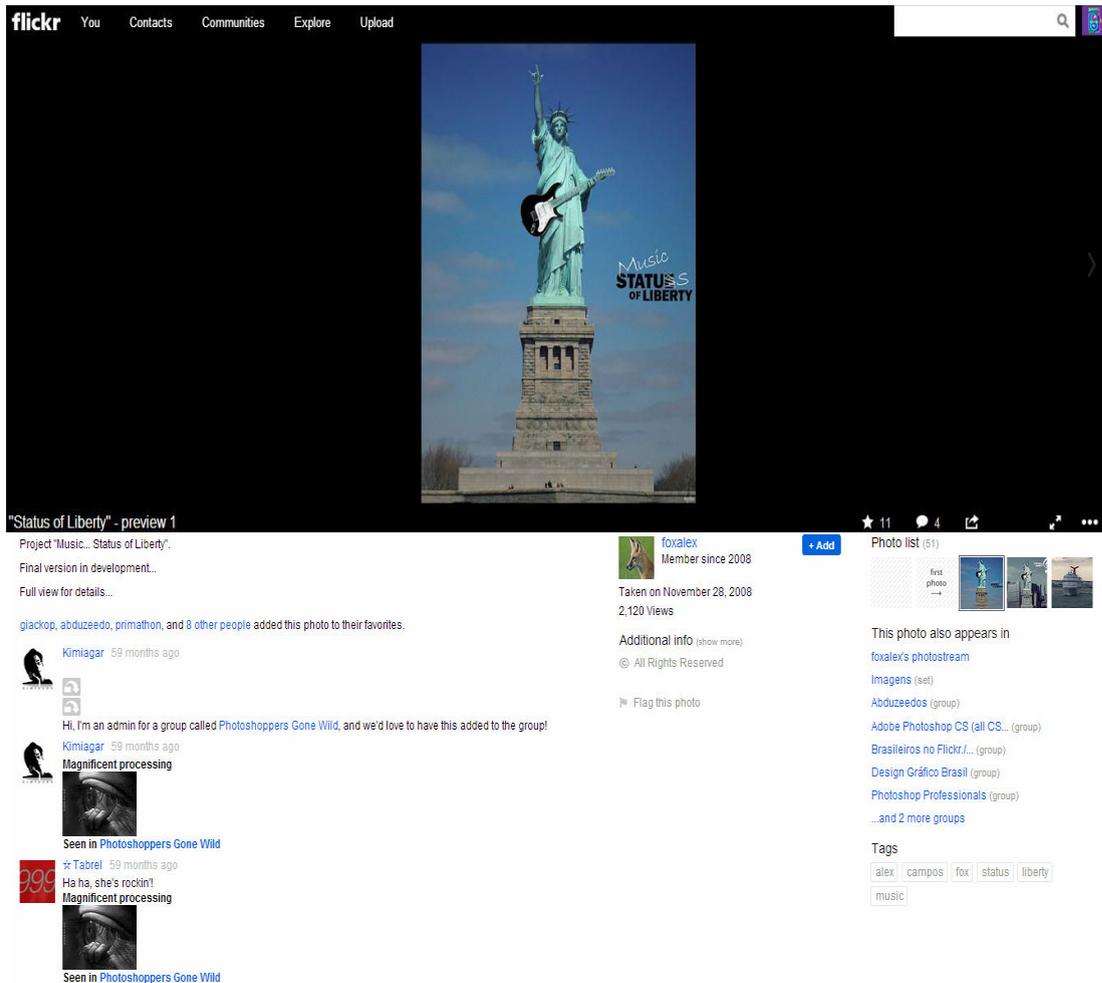


FIGURE 1.2: A snapshot of the social media Website Flickr. A typical image is associated with comments, tags and ratings (mark as a favorite).

efforts have been devoted in the literature to enhance multimodal image annotation by leveraging user tags [Guillaumin et al., 2010; Kawanabe et al., 2011]. The joint use of tag-based features with visual descriptions has consistently improved image annotation performances on challenging datasets compared to visual-only approaches. However, these approaches seem to be insufficient to narrow the semantic gap between tags and the actual visual content of the images due to the problem of tag imperfections which was partially considered. These imperfections are essentially related to the user tagging motivations which are numerous and diverse. They are not necessarily intended to precisely describe the image content. Moreover, tags are essentially personal keywords which impose a soft organization on data. As opposed to taxonomies and thesauri that are restricted by rigid definitions and relationships, tags are continuously influenced by popular trends and colloquial vocabulary. Moreover, tags are contributed from personal and often unknown motivations. They are not directed towards specific tasks such as image annotation and retrieval. Considered as one of the earliest and most popular social media sharing web sites, Flickr has been intensively studied in recent years,

especially on tagging characteristics and motivations [Marlow et al., 2006; Golder and Huberman, 2006; Ames and Naaman, 2007; Sigurbjörnsson and van Zwol, 2008]. We consider Flickr as an example to study social tagging characteristics for multimodal image annotation in the context of social media. The most salient characteristic of image collections in Flickr is that they come with a wide variety of associated data, such as text descriptions, tags, ratings and user comments as depicted in Figure 1.2. In 2004, Vander Wal<sup>5</sup> coined the term *Folksonomy* to describe the new structure of users, tags and objects (anything with a URL such as images, videos, books...). Folksonomy is defined as “*the user-created bottom-up categorical structure development with an emergent thesaurus*”.

We assume that user motivations and tag characteristics would have broad implications on the design of multimodal image annotation algorithms. Thus, it is useful to understand and identify what are the motivations of tagging in social media networks and which are the characteristics of tags in such collections.

### 1.1.1 Tagging Motivation

Tagging Motivation which aims at understanding why users tag their photos, has remained largely elusive until the first studies on this subject were conducted by [Golder and Huberman, 2006]. The authors investigate the structure of collaborative tagging systems of two snapshots taken from the *del.icio.us*<sup>6</sup> system. They found regularities in user activities, tag frequencies, tag uses and other aspects. In addition, it has been shown that user motivations strongly affect the level and the usefulness of tags in tag-based applications [Nov and Ye, 2010] such as retrieving images in Flickr using tags. [Ames and Naaman, 2007] developed a taxonomy for the revealed set of user motivations. There are two dimensions along which they place the different incentives for tagging images. The first dimension, “sociality,” relates to whether the tag intended usage is by the user who took and uploaded the photo or by others, including friends/family and strangers. The second dimension, “function,” refers to the tag’s intended uses. They found that users tagged their pictures either to facilitate later organization and retrieval or to communicate some additional context to viewers of the image (whether themselves or others).

A recent study by [Strohmaier et al., 2012] suggests that a distinction between at least two fundamental types of user motivations for tagging are important, including Categorization and Description.

---

<sup>5</sup><http://vanderwal.net/folksonomy.html>

<sup>6</sup><http://delicious.com>

- ***Categorization***: tagging is used as a tool to categorize resources according to some shared high-level characteristics. The main motivation is to build and maintain a navigational aid to the resources for later browsing <sup>7</sup>.
- ***Description***: tagging is used as a tool to accurately and precisely describe resources. The main motivation is to produce annotations that are useful for later searching.

This distinction has been found to be important because, for example, tags assigned by describers might be more useful for information retrieval (because these tags focus on the visual content of resources) as opposed to tags assigned by categorizers, which might be more useful to capture a rich variety of possible interpretations of a resource (because they focus on user-specific views on resources).

### 1.1.2 Tag characteristics

Many studies show that tags provided by Flickr users are highly “noisy” in the sense that only around 50% of them are actually related to the image visual content [Kennedy et al., 2006; Chua et al., 2009; Sigurbjörnsson and van Zwol, 2008]. An example of images from Flickr website with their associated user tags is presented in Figure 1.3. Tags which are related to the image visual content are marked in bold. Obviously, only a few tags are relevant to describe accurately the image semantic content. Let’s take the example of Figure 1.3(a). If tags such as “*flowers, red, butterfly*” are relevant to describe the image content, other tags like “*zebra*” is ambiguous since it is polysemous. The tag “*zebra*” for example is surely present in Figure 1.3(a) as a pattern but is not related to the common sense of the word “zebra” as a horse with vivid dark brown-and-white stripes. The rest of tags are user-specific such as “*d80, shieldofexcellence, flickrestrellas, anawesomeshot, Theunforgettablepictures*”. Since there is no restriction or boundary on selecting words for tagging images, user provided tags are free-style and thus are subject to many problems such as semantic ambiguity which means that the same tag has different meanings (for example the tag “*zebra*” in Figure 1.3(a), “*tiger*” in Figure 1.3(d) and “*wolf*” in Figure 1.3(b)), tag synonymy which means that different tags actually have the same meaning such as the tag “*flower*” in Figure 1.3(b) and the tag “*blossom*” in Figure 1.3(d).

Meanwhile, several other tags that can be visually significant, such as “*grass, garden*”, are missing in Figure 1.3(a). Many social image search engines are based on keyword/tag matching. Ideally, all images would have a reasonable number

---

<sup>7</sup>Photo browsing consists in the visualization of photos by their owners who upload it which is different from photo searching which consists in visualizing photos of others as a response for a query.



FIGURE 1.3: An example of images from the Flickr website with their associated user tags. Most of tags are noisy for image annotation and only few tags are related to image visual content (marked in bold).

of user generated tags, which would then enable other users to find and retrieve them. Unfortunately, in practice, only a fraction of the uploaded pictures are tagged with useful tags. This problem is referred as tag incompleteness.

Another serious problem of tags is that nearly 60% of tags are personal tags that are only used by one user [Sen et al., 2009]. According to [Cantador et al., 2011], only a set of social tags are related to the semantic visual content of images. Tags can be categorized into four categories:

- **Content-based:** Social tags that describe the visual content of the image, such as the objects and living things that appear in an image. Some examples of tags belonging to this category are “*flower, butterfly*” in Figure 1.3(a) and “*clown, rail*” in Figure 1.3(c).
- **Context-based:** Social tags that provide contextual information about the image, such as the place or the date where and when a photo was taken and camera characteristics. Examples of this kind of tags are “*mexico*” in Figure 1.3(a) and “*nikon, d200*” in Figure 1.3(b).
- **Subjective:** Social tags that express opinions and qualities of images. Some examples of these tags are “*shieldofexcellence, Theunforgettablepictures*” in Figure 1.3(a).
- **Organizational:** Social tags that define personal usages and tasks, or indicate self-references.

In this thesis, only tags that are included in the content-based category are considered as useful. Others are considered as noisy and a source of imperfection which need to be handled in order to enhance multimodal image annotation. The problem of noisy tags is assumed by [Suchanek et al., 2008]. The authors show that user-generated tags present more semantic noise than the terms extracted from Web page content usually used for image indexing and retrieval. Indeed, when tagging, users introduce not only misspellings (“*new york, new yok*”), use different synonyms (“*car, automobile*”), acronyms (“*nyc, new york city*”) and morphological derivations (“*play, players, playing*”) for a given concept, but also include tags that express personal assessments (“*funny, sad*”) [Suchanek et al., 2008].

To summarize, the main observations on using provided user tags to improve image annotation in social media are:

- Tags are often noisy;
- Tags are often not related to the semantic visual content and they are overly personalized;
- The choice of words with large variability among different users is spontaneous which causes the problem of polysemy and synonymy;
- Meaningful tags are missing;
- Concerning tag frequency, community contributed collections contain a set of most frequent tags which consists of a set of too generic tags with a high frequency. It contains also the infrequent tags with incidentally occurring terms such as misspellings and complex phrases with a low frequency.

### 1.1.3 Imperfection Aspects

The main challenge in this thesis is to improve multimodal image annotation performances by taking advantages of textual information issued from the tag modality. Consequently, we focus on how to identify imperfections both at representation and decision levels and to cope with them.

In the literature, the problem of tag imperfections has been partially considered in the context of multimodal image annotation, in the sense intended in this thesis. However, a wealth of research has been proposed to enhance the quality of tags in the context of social media for other applications. The existing work mainly focus on the following two tasks: (a) *tag ranking and relevance*, which aims at ranking and differentiating tags associated with images with various levels of relevance [Liu et al., 2009a; Li et al., 2009a; Sun and Bhowmick, 2009]; (b) *tag refinement and suggestion* with the purpose to refine the unreliable human-provided

tags by dropping the inappropriate tags and adding the missing ones [Jin et al., 2005; Weinberger et al., 2008; Xu et al., 2009].

Most of the state-of-the-art approaches address the problem of imperfections only through the notion of relevancy or tag ranking which is not sufficient. In particular, these approaches have not been interested in the different nature of imperfections. Based on such noisy and incomplete tags, existing approaches exhibit lower results than expected. Thus handling tag imperfections seems to be interesting and crucial to improve tag ranking and suggestion performances. The tag incompleteness issue in social media tagging is almost well identified in the literature [Liu et al., 2009b; Tang et al., 2009; Wang et al., 2010a]. Incomplete tags are defined as tags that describe the semantic visual content of the image but are missing in initial user tag list. However, there is no precise identification and definition of noisy tags. To handle such imperfections, an original aspect of our work is to identify and define clearly the different aspects of imperfections.

Once these imperfections are well identified and defined, we focus on how to handle them in order to build robust tag-based and multimodal signature. The next challenge is how to combine both tag and visual modalities in order to enhance multimodal image annotation performances. Since we formulate the problem of image annotation as a supervised classification problem, the fusion process, at the decision level, can be viewed as a classifier fusion problem: the predictions of the class label of images from the different classifiers are considered as information sources to be combined to make a final decision. Combining various sources of information (image and tags) for multimodal image annotation based on classifier combination allows on the fly integration of classification modules, specific to a single modality, in a classification process. These modular and extensible approaches do not require that a single method copes with every eventuality, but combine existing specialized methods to overcome their weaknesses. In the literature, the use of multiple classifiers trained on different modalities (tags and image) usually leads to better image annotation performances, due to the complementarity of the classification models [Escalante et al., 2008; Wang et al., 2009a; Xioufis et al., 2011]. However, most of the state-of-the-art approaches do not take into account imperfection aspects despite that they represent an important characteristic of information in fusion process [Bloch, 2001].

Imperfection aspects have been studied by [Bloch, 2003] who defined some of their types in the field of data and sensor fusion as follows:

- **Uncertainty** is related to the truth of some information, characterizing its adequacy to reality [Dubois and Prade, 1988]. It refers to the nature of the considered object, to its quality, or to its occurrence.

- **Imprecision** concerns the content of information and describes a quantitative defect of knowledge or measure [Dubois and Prade, 1988]. It concerns the lack of precision in quantity.
- **Incompleteness** of information issued from each source is one of the reason that motivated the fusion. Information provided by a source is generally partial, and gives only one point of view or one aspect of the observed phenomenon.

This thesis is motivated by the assumption that taking into account, explicitly, imperfection aspects both at the representation and the decision levels in the multimodal image annotation process should improve the performance of the annotation. In fact, machine learning is inseparably connected with **uncertainty**. To begin with, data presented to learning algorithms is imprecise, incomplete or noisy most of the time. It is especially the case of tags in social media as detailed before. Moreover, the generalization beyond that data, the process of induction, is still afflicted with uncertainty. Another form of imperfections is the **incompleteness** of data. This is currently an issue faced in social media where a large number of images is untagged. One primary concern of classifier learning is prediction accuracy. Handling incomplete data (images without tags) is an important issue for classifier learning since incomplete data in either the training data or test data may not only impact interpretations of the data or the models created from the data but may also affect the prediction accuracy of learned classifiers. Regarding **imprecision**, learning a classifier on uncertain and incomplete data leads to a decision function which is imprecise to decide on the real statement of an object (it belongs to a certain class or not). Unlike most of the state-of-the-art approaches, the originality of the contributions in this thesis is to take into account these imperfections at both the representation and the decision levels.

To sum up, multimodal image annotation, which consists in assigning automatically keywords to describe the image semantic content by combining both tag and visual modalities, is a very promising solution for social media collection indexing. Nevertheless, most of existing state-of-the-art approaches are still insufficient due to the following problems:

- **Tag imperfections:** Tags are often noisy, overly personalized and only a few of them are related to the semantic visual content of the image. In fact, since tags are contributed from personal, often unknown motivations they are not directed towards specific tasks such as image annotation. As it is impractical for general users to annotate comprehensively their images, many potentially useful tags may be missed. Therefore, the user-provided tags are imprecise, uncertain, and incomplete for describing the semantic visual content of the image.

- **Semantic gap problem:** In community contributed collections, a multimedia document is described with two modalities (tags and image) which are heterogeneous and complementary. Consequently, multimodal image annotation faces two views of the semantic gap problem. On one hand, the well-known semantic gap between low-level features and high concepts. On the other hand, the gap between the information extracted from user-provided tags and annotation concepts.
- **Imperfections** are introduced by machine learning algorithms at the decision level. When combining score predictions from different classifiers learned on different features and modalities, multimodal fusion faces the problem of their imperfections.

## 1.2 Goals

In this dissertation, our objective is to make advances in the field of multimodal image annotation by taking advantages of textual and visual information at the same time. Specifically, we address the following issues:

- Fusing mono-media in order to reduce imperfection aspects (uncertainty, imprecision and incompleteness): we are interested in judging the importance of document description and the potential complementarity of descriptors that are extracted from different modalities.
- Characterizing the uncertainty and imprecision on data (tags) as well as its incompleteness in order to reduce the semantic gap between extracted information and annotation concepts.
- Proposing models for multimedia fusion that exploit highly heterogeneous data in order to enrich the multimedia document description and thus enhance the performance of image annotation.
- Taking into account the available information to relate the documents at a semantic level: propose a fusion model that infers the semantic properties of a document using different modalities.
- Designing scalable methods: it is important to describe a multimedia document as compactly as possible and to develop efficient indexing strategies.

## 1.3 Contributions

Inspired from the information fusion theory, the originality of this thesis is to define, identify and handle imperfection aspects in order to improve image annotation performances. In the context of social media, we consider that image annotation is subject to imperfections at two levels: the *Representation* and the *Decision* levels. Thus, we define multimodal image annotation as “*the process of **combining** information from several modalities having different confidence levels while **handling uncertainty, imprecision and incompleteness** aspects at **representation** and **decision** levels; to obtain a consistent description and improve the accuracy of image annotation.*” Based on this definition, our contributions are the following:

### 1.3.1 Identifying tag imperfections at the representation level

Our first contribution deals with the definition of imperfection aspects at the representation level. First of all in Chapter 2, we present a survey of the state-of-the-art approaches that identify these imperfections in the context of multimodal image annotation and other related applications such as tag ranking and suggestion. In the context of image annotation, we identify and define clearly three kinds of imperfections: **Uncertainty**, **Imprecision** and **Incompleteness** in Chapter 3-Section 3.2.

### 1.3.2 Handling imperfections at the representation level

Once these aspects are well identified and defined, we propose in Chapter 3, two novel models to handle such imperfections for tag-based image annotation. Both models are extensions of the Bag-of-Words (BOW) model [Salton and McGill, 1983]. In order to build robust BOW based tag-signatures, we rely on the locality-constrained coding method [Liu et al., 2011b] that has proved to be effective for visual features when paired with max-pooling aggregation. Extensive experimental evaluation on five challenging datasets [Chua et al., 2009; Everingham et al., 2010; Nowak and Huiskes, 2010; Nowak et al., 2011; Thomee and Popescu, 2012], shows that the first model outperforms the state-of-the-art methods on three out of five datasets and the second proposed model outperforms the state-of-the-art methods on the five considered datasets on a tag-based image annotation task. Both models handle a part of imperfection aspects of tags. This contribution has been published in [Znaidia et al., 2012d, 2013b].

In a second contribution presented in Chapter 4, we address the problem of tag incompleteness. We distinguish two types of incompleteness: partial and full. Partial incompleteness is the case where the image has some tags and others are missing while full incompleteness represents the case where the image has no tag. The former has been considered in the above models, thus in this second contribution, we focus on the latter type of incompleteness. We propose a novel method named Tag Completion based on similarly visual neighbors and Belief theory [Shafer, 1976] to handle full tag incompleteness. Hence, this model supports a scheme to tackle with imprecision and uncertainty that are inherent to tag information in a social media context. Image annotation is evaluated on two well known datasets [Everingham et al., 2010; Nowak et al., 2011], on which we obtain similar or better results than the state-of-the-art. For tag suggestion, we manually annotated 241 queries to propose a new benchmark to the multimedia community. As well, we obtain competitive results on this task. This contribution has been published in [Znaidia et al., 2013a].

In a third contribution presented in Chapter 5, we propose a more integrated and compact semantic signature of multimedia documents, called Bag-of-Multimedia-Words (BOMW), than the BOVW and the Bag-of-Tag-Words (BOTW), that results from a combination of textual and visual information. It is based on *multimedia codewords* that allow the cross-coding of textual tag-words over visual-words extracted from a document. This cross-coding is used to design BOMW signatures. We exploit the recent advances in BOVW design methods [Yu et al., 2009; Boureau et al., 2010; Wang et al., 2010b; Liu et al., 2011b] in order to provide discriminative BOMW vectors suited to multimodal document classification with efficient linear classifiers. Experiments have been conducted on two well-known challenging benchmarks [Everingham et al., 2010; Thomee and Popescu, 2012]. Obtained results show the competitive performances of the BOMW, ensuring a trade-off between classification accuracy and computation cost. This work has been published in [Znaidia et al., 2012c].

### 1.3.3 Handling imperfections at the decision level

To deal with imperfections at the decision level, we propose two approaches. In a first contribution presented in Chapter 6, we propose a multimodal framework for semantic image classification which consists in the combination of the BOVW representation and the Local Soft Tag Coding (LSTC) model based on the Stack Generalization algorithm [Wolpert, 1992]. This scheme mainly includes two stages: a training stage and a testing stage. The training dataset is split into training and validation sets. The training stage consists in training classifiers through a learning algorithm on the training set and in evaluating it on the validation set. The process is repeated using cross-validation procedure. The output prediction scores from

different classifiers on validation sets are concatenated and used as input features to learn a new classifier. Extensive experimental evaluation on four challenging datasets [Chua et al., 2009; Everingham et al., 2010; Nowak and Huiskes, 2010; Nowak et al., 2011; Thomee and Popescu, 2012] shows that our framework achieves comparable and better results compared to more sophisticated state-of-the-art approaches as summarized in Table 1.1. This contribution has been published in [Znaidia et al., 2012d].

In a second contribution presented in Chapter 7, we propose a multimodal framework for image classification based on classifier combination using the Dempster-Shafer theory [Shafer, 1976]. It enables to handle the uncertainty and the conflict that can exist between different classifiers and to assess the discrepancy between them. First, we convert the classifier output probabilities into consonant mass functions using the inverse pignistic transform [Dubois et al., 2001]. Secondly, these mass functions are combined using the Dempster’s rule [Shafer, 1976]. Experimental results on two challenging datasets [Nowak and Huiskes, 2010; Nowak et al., 2011] show the effectiveness of the proposed framework. This work has been published in [Znaidia et al., 2012a].

## 1.4 Organization of the Dissertation

This dissertation is organized as follows.

In **Chapter 2**, we present a thorough survey on relevant research topics on image annotation in the context of social media. The covered topics include multimodal image annotation and the handling of tag imperfections in social media tasks. Finally, we present an overview of the datasets of the state-of-the-art, created within evaluation campaigns, and used to evaluate models proposed in this thesis.

TABLE 1.1: Comparison of our results, based on the proposed multimodal framework based on Stack Generalization, to the best state-of-the-art classification performances in terms of mAP.

Dataset	Approach	Textual	Visual	Multimodal
PASCAL VOC’07	[Guillaumin et al., 2010]	<b>53.1</b>	43.3	66.7
	Our method	52.1	<b>51.8</b>	<b>68.3</b>
ImageClef’11	[Zhang et al., 2012b]	<b>37.4</b>	34.7	<b>45.3</b>
	Our method	<u>31.2</u>	<b>38.0</b>	<u>44.8</u>
ImageClef’12	[Liu et al., 2012]	<b>34.8</b>	33.3	<b>43.6</b>
	Our method	<u>29.4</u>	<b>34.1</b>	<u>43.1</u>
NUS-WIDE	[Gao et al., 2010]	26.12	18.89	29.88
	Our method	<b>42.0</b>	<u>18.81</u>	<b>49.5</b>

In **Chapter 3**, we identify and define tag imperfections in the context of social media. Thereafter, we propose two tag-based signatures for image annotation. Both proposed approaches deal with tag imperfections: uncertainty, imprecision and incompleteness at the representation level. In these approaches only the textual information issued from the tag modality is used.

In **Chapter 4**, we address the problem of tag incompleteness. We are interested in suggesting tags for untagged images. We propose a novel method named Tag Completion based on similar visual neighbors and Belief theory [Shafer, 1976] to solve this problem introduced in Section 3.2. Based on Belief theory, this model supports a scheme to tackle with imprecision and uncertainty that are inherent to tag information in a social media context. In this approach, only the tag modality is used for tag completion. However, the visual information is used only for searching for similarly visual neighbors.

In **Chapter 5**, we focus on the combination of both visual and tag modalities into a unique and compact representation that describes well a multimedia document. We propose a more integrated and compact semantic signature for multimedia documents, that results from a combination of textual and visual information in a fusion scheme where image and tag modalities are combined at feature level (representation). It is based on *multimedia codewords* that allow the cross-coding of textual tag-words over visual-words extracted from a document. This cross-coding is used to design BOMW signatures.

To deal with imperfections at the decision level, we propose, in **Chapter 6**, a multimodal framework for semantic image classification which consists in combining visual information and tag-based signature (presented in Chapter 3) based on the Stack Generalization algorithm [Wolpert, 1992]. The Stack Generalization scheme deals implicitly with imperfections that exist at the decision level.

In **Chapter 7**, we propose a multimodal framework for image classification based on classifier combination in the Dempster-Shafer theory [Shafer, 1976]. It enables to handle the uncertainty and the conflict that can exist between different classifiers and to assess the discrepancy between them. This combination operates at decision level and permits to handle explicitly imperfection aspects based on the Belief theory formalism.

Finally, this dissertation is concluded in **Chapter 8** with a recall on our contributions and a discussion on the directions that can be inspired by the presented research topics.

# Chapter 2

## State-of-the-art

### Contents

---

<b>2.1</b>	<b>Introduction</b>	<b>18</b>
<b>2.2</b>	<b>The Problem of Multimodal Image Annotation</b>	<b>18</b>
<b>2.3</b>	<b>Multimodal Image Annotation</b>	<b>20</b>
2.3.1	Visual Description: Bag-Of-Visual-Words	22
2.3.2	Textual Description	26
2.3.3	Fusion Strategy	30
<b>2.4</b>	<b>Handling Tag Imperfections</b>	<b>44</b>
2.4.1	Tag Ranking & Relevance	44
2.4.2	Tag Refinement & Suggestion	47
<b>2.5</b>	<b>Handling Imperfections at the decision level</b>	<b>50</b>
2.5.1	Probability theory	51
2.5.2	Fuzzy Set theory	52
2.5.3	Possibility theory	53
2.5.4	Dempster-Shafer theory	53
<b>2.6</b>	<b>Image Databases &amp; Evaluation Campaigns</b>	<b>57</b>
2.6.1	Evaluation Campaigns	57
2.6.2	Image Databases	59
<b>2.7</b>	<b>Conclusions</b>	<b>63</b>

---

## 2.1 Introduction

In this chapter, we review some of the most indicative work in the literature of multimodal image annotation in the context of social media. (i.e. we consider two modalities: a visual modality and a textual modality which consists of a set of tags). First, we propose a comprehensive study of the state-of-the-art approaches on multimodal image annotation focusing at first on the representation of different modalities and secondly on the strategy of fusion used to combine these two modalities. Afterward, we are interested in the handling imperfection aspects both at the representation level for the textual information and at the decision level. Since, we are interested in classifier combination to achieve the multimodal image annotation task, we review theories that are used in the literature to deal with such imperfections at decision level aiming at highlighting the weaknesses of the existing theories.

The rest of this chapter is organized as follows. We start by a formalization of the problem of image annotation in the context of social media, in Section 2.2. In Section 2.3, we are interested in the state-of-the-art of multimodal-based approaches for image annotation. We present a review of the tag-based features used in the state-of-the-art approaches. We focus on fusion strategies used to combine both visual and tag modalities. We show that most of the state-of-the-art approaches in multimodal image annotation do not take into account imperfections neither at the representation nor at the decision levels. Thus, Section 2.4 reviews the most representative work that handle tag imperfections in other social media applications such as tag ranking and tag suggestion at representation level. In Section 2.5, we present theories that exist for reasoning with these imperfections at decision level. Finally, datasets used to evaluate the effectiveness and the robustness of the proposed approaches are presented in Section 2.6.

## 2.2 The Problem of Multimodal Image Annotation

In this dissertation, we are interested in the problem of multimodal image annotation in the context of social media where both visual and textual modalities are combined. Image annotation consists in assigning a set of keywords (called also labels or concepts) to an unknown image from a predefined vocabulary in order to describe its visual content from a high level perspective [Duygulu et al., 2002; He et al., 2004; Carneiro and Vasconcelos, 2005; Nowak and Huiskes, 2010; Semenovich and Sowmya, 2010; Zhang et al., 2012a; Moser and Serpico, 2013]. We consider the supervised image classification scheme as the appropriate framework

to tackle the image annotation problem in the context of social media. The supervised classification scheme is based on two steps: modality representation and concept learning.

**Terminology.** For a sake of clarity, we define in this section the terms used in the following and their meaning. Thereby, we refer to:

- *Label or Concept*: is a word used to annotate images. In the rest of this dissertation, label and concepts are interchangeably used. These labels are added by experts (annotators) from a predefined and controlled vocabulary to describe the semantics of image content. Labels or concepts reflect the actual visual content of the image.
- *Tags*: is a word freely associated to the image by users in social media context without relying on a controlled vocabulary. Some labels can be present among the tags but this is not mandatory;
- *Visual features*: is the set of low-level descriptors used to describe the visual content of images;
- *Textual features*: is the set of high-level descriptors extracted from the textual information (tags);
- *Signature*: is a vector of features extracted from the visual or textual information to represent their content;
- *Codebook*: is a set of predefined words extracted from the dataset and used to extract the image signature (visual or textual). In this thesis, codebook and dictionary are interchangeably used.

### Problem Formalization

**Given** a multimodal dataset composed of two subsets: a training set  $\mathcal{L}$  and a testing set  $\mathcal{T}$ , where:

- $\mathcal{L} = \left\{ ((\mathbf{I}_1, \mathbf{T}_1), \mathbf{y}_1), \dots, ((\mathbf{I}_N, \mathbf{T}_N), \mathbf{y}_N) \right\}$ ,
- $\mathcal{T} = \left\{ (\mathbf{I}_{N+1}, \mathbf{T}_{N+1}), \dots, (\mathbf{I}_{N'}, \mathbf{T}_{N'}) \right\}$ ,
- $\mathcal{I} = \left\{ \mathbf{I}_1, \dots, \mathbf{I}_N \right\}$  is the set of training images,
- $\mathcal{I}' = \left\{ \mathbf{I}_{N+1}, \dots, \mathbf{I}_{N'} \right\}$  is the set of testing images,
- $\mathbf{T}_i = \{ \mathbf{t}_{1i}, \mathbf{t}_{2i}, \dots, \mathbf{t}_{li} \}$  is the set of tags associated with the image  $\mathbf{I}_i$  which can be empty for untagged images,

- $\mathcal{W}^t = (\mathbf{w}_1^t, \dots, \mathbf{w}_M^t)$  is a textual codebook,
- $\mathcal{W}^v = (\mathbf{w}_1^v, \dots, \mathbf{w}_{M'}^v)$  is a visual codebook learned from  $\mathcal{I}$ ,
- $\mathbf{X}_i^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_M^t)$  is a textual signature of the image  $\mathbf{I}_i$  of size  $M$ ,
- $\mathbf{X}_i^v = (\mathbf{x}_1^v, \dots, \mathbf{x}_{M'}^v)$  is a visual signature of the image  $\mathbf{I}_i$  of size  $M'$ ,
- $\mathcal{C} = \langle C_1, C_2, \dots, C_k \rangle$  is the set of labels used to annotate images, where  $k$  is the number of labels,
- $\mathbf{y}_i = [y_i^1, \dots, y_i^k]$ ,  $y_i^j \in \{-1, +1\}$  is the set of labels of the image  $\mathbf{I}_i$ ,  $y_i^j = 1$  if the image  $\mathbf{I}_i$  is annotated with the label  $C_j$ , otherwise  $y_i^j = -1$ .
- $\mathcal{Y} = \{-1, +1\}^k$  is the set of all possible label sets.

**Our goal** is to build a multimodal classifier with a decision function defined as follows:

$$\begin{aligned} f : \mathcal{T} &\rightarrow \mathcal{Y} \\ f(\mathbf{X}_i^t, \mathbf{X}_i^v) &= \hat{\mathbf{y}}_i \end{aligned} \quad (2.1)$$

$f$  associates a label set  $\hat{\mathbf{y}}_i$  to each unseen image  $I_i$  from the test dataset  $\mathcal{T}$  by combining both textual and visual information as illustrated in Figure 2.1.

## 2.3 Multimodal Image Annotation

In this section, we are interested in the state-of-the-art of Multimodal-based approaches for image annotation in the context of social media. Many approaches have been proposed to solve the problem of image annotation. As depicted in Figure 2.2, these approaches can be categorized into three groups: **Content**-based where only visual information is used, **Tag**-based which exploit the tag modality alone and **Multimodal**-based approaches that leverage both modalities to improve image annotation performances. We focus in this chapter on Multimodal-based approaches.

First, we briefly review the monomedia descriptions: textual and visual. For the visual description, the BOVW approach [Sivic and Zisserman, 2003; Csurka et al., 2004] has established itself, in the last ten years, as the state-of-the-art representation for visual content description in image and video classification. Thus, in this thesis, in Section 2.3.1 we focus on the BOVW representation and their recent advances.

For the textual description, we identify two types of relations that are exploited to generate tag-based signatures. The first one, named *Tag-to-Concept* relation, relies on semantic relations between image tags and annotation concepts. The

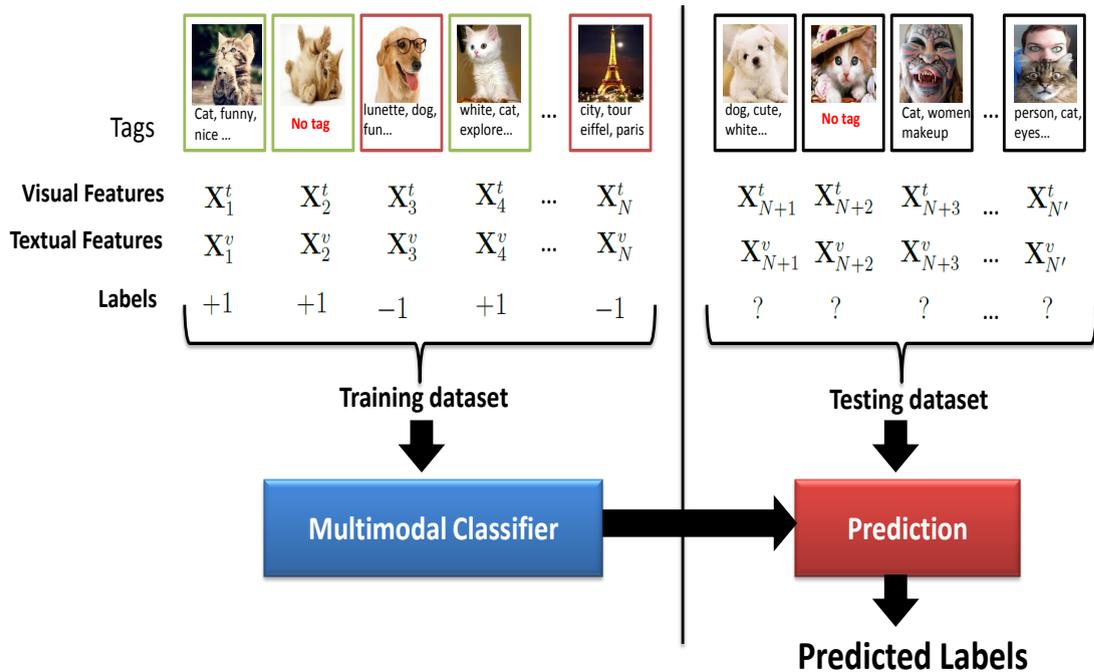
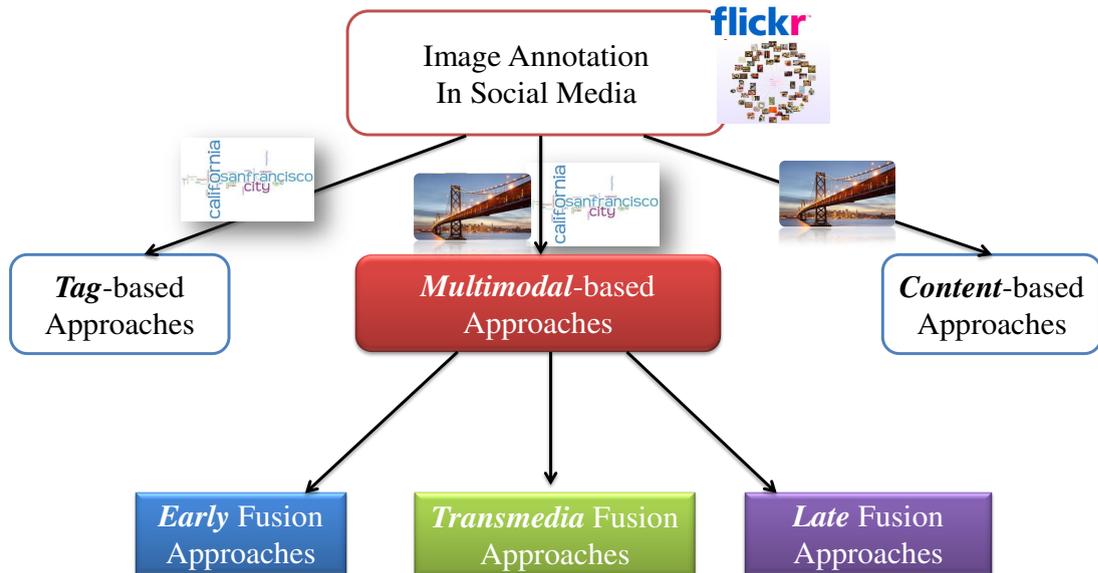


FIGURE 2.1: Illustration of the problem of multimodal image annotation formulated as a set of binary classification tasks. For example, to learn the concept “cat”, we use a training dataset comprising of both positive (in green box) and negative (in red box) examples to learn a classifier that predicts the presence or the absence of this concept in each image of the testing dataset (images in black box).

second one, named *Tag-to-Tag* relation, is based on semantic relations between image tags and a predefined dictionary of tags. These textual descriptions are reviewed in Section 2.3.2.

Finally, as far as multimodal image annotation approaches are concerned, they require a strategy to combine information from multiple modalities and features. Accordingly, in the literature, there have been many approaches covering text/image information fusion. Most of techniques developed in that context fall in three different categories: *early fusion*, *late fusion* and *transmedia fusion* as suggested by [Clinchant et al., 2011].

- **Early fusion** consists in combining both visual and textual features in a joint representation at the feature level.
- **Late fusion** consists in combining decisions (predictions) from different classifiers at the decision level.
- **Transmedia fusion** consists in using diffusion processes that act as a transmedia pseudo-relevance mechanisms. The key idea is to use one modality/feature to gather relevant documents and then to switch to the other modality and aggregate their features.



These approaches are presented and discussed in Section 2.3.3.

### 2.3.1 Visual Description: Bag-Of-Visual-Words

The BOVW approach [Sivic and Zisserman, 2003; Csurka et al., 2004] has now established itself as one of the state-of-the-art representation for visual content description. It is probably the most popular and effective image representation in supervised classification framework in the recent literature [Huang et al., 2013]. It has been inspired by the success of the BOW model for text categorization [Salton and McGill, 1983], that represents a textual document by a vector of the occurrences of each word in the document. Extended to image description, the usual BOVW design pipeline consists in learning a codebook from a large collection of local features extracted from a training dataset, then creating the global features of visual signature through coding, pooling and spatial layout. Recent work addressing this problem [Lazebnik et al., 2006; Yang et al., 2009; Boureau et al., 2010; Wang et al., 2010b; Liu et al., 2011b] proved the importance of tuning each of these steps to improve scene classification and object recognition accuracy on different benchmarks. Several extensions to the basic Bag-of-Visual-Words (BOVW) representation have been proposed including the Fisher Vector [Peronin and Dance, 2007], the Super Vector [Zhou et al., 2010] and the Vector of Locally Aggregated Descriptors [Jégou et al., 2012].

The first step in the BOVW scheme is the visual codebook learning. A visual codebook  $\mathcal{W}^v$ , as introduced in Section 2.2, is learned on a training subset of

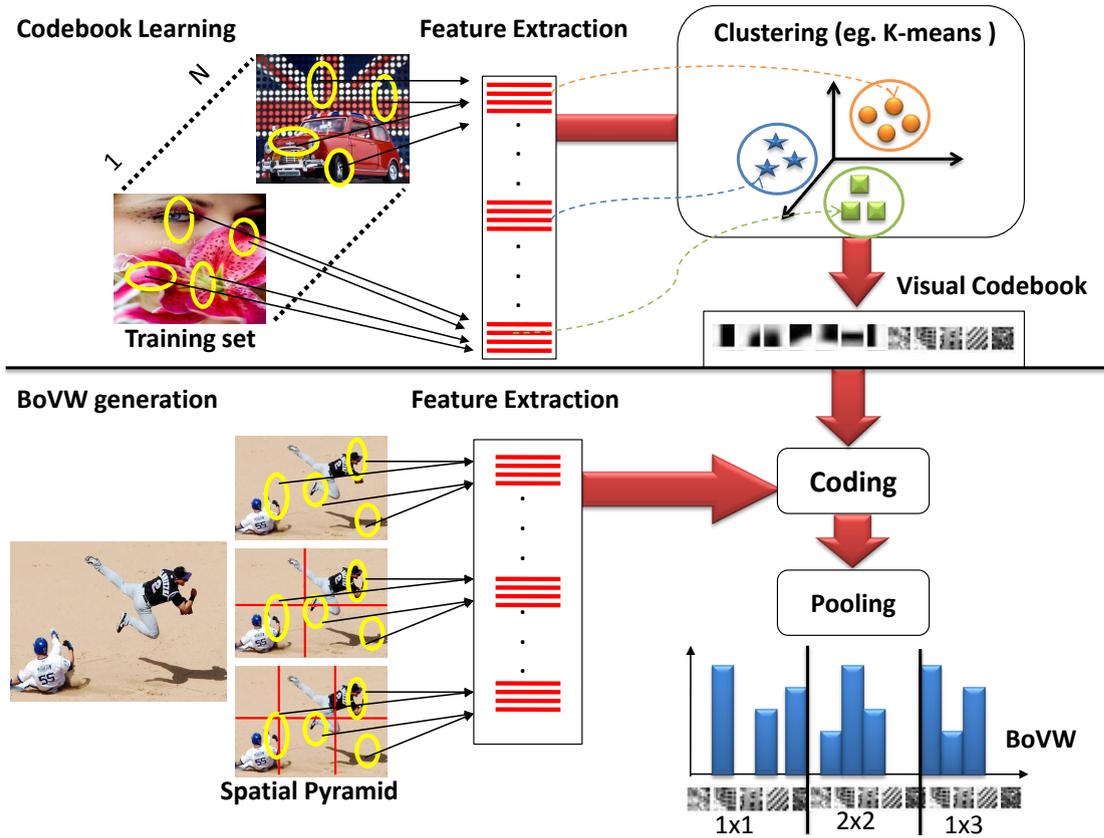


FIGURE 2.3: The flowchart of the BOVW generation scheme. (1) Codebook learning, and then given an image, its visual features are built in two steps (2) local features coding and (3) pooling.

local features extracted from the learning dataset. An example of these local features is dense SIFT [Lowe, 2004] descriptors. We denote by  $\mathbf{X}$  the set of SIFT descriptors extracted from a given image in a  $d$ -dimensional feature space, i.e.  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_P\} \in \mathfrak{R}^{d \times P}$  where  $P$  represents the number of SIFT descriptors extracted from the image.

Once the visual codebook is learned, given an image, its visual features are built in two steps (i) local feature coding and (ii) pooling, as illustrated in Figure 2.3 and Figure 2.4. In the following, we review relevant work for each one of these steps.

### 2.3.1.1 Codebook Learning

The codebook, which entries are named codewords, is a collection of basic patterns used to reconstruct the input local features. A simple way to generate the codebook is to use clustering based methods such as K-means [Sivic and Zisserman, 2003; Csurka et al., 2004] or GMM (Gaussian Mixture Models) [Dork and Schmid, 2005]. [Jurie and Triggs, 2005] propose a scalable acceptance-radius method for

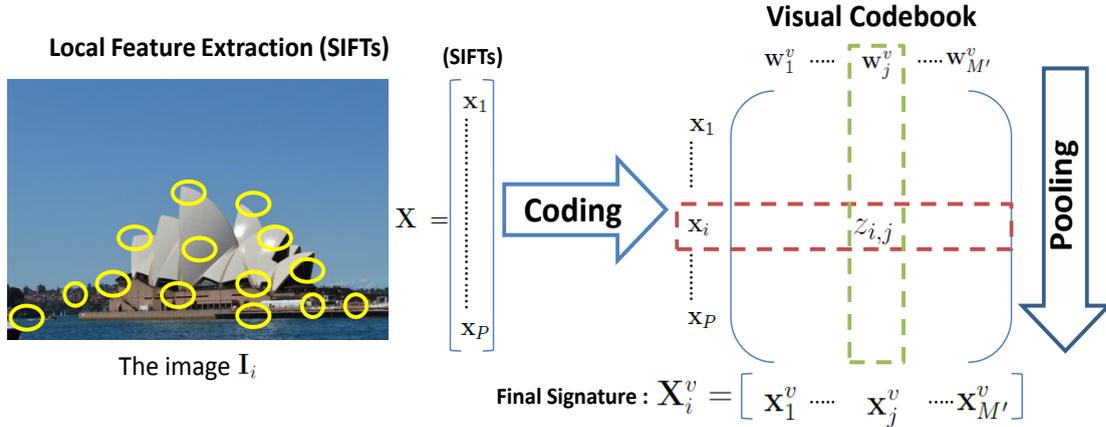


FIGURE 2.4: Illustration of the coding and pooling steps.

clustering where a local feature is assigned to the center that lies within a fixed radius. An alternative is sparse coding which has consistently yielded better results on some object recognition benchmarks [Yang et al., 2009; Boureau et al., 2010]. Even if improving the codebook generation might seem central in BOVW approaches, recent studies have shown that it turns out to be less critical than the next stages: coding, pooling and spatial layout. For instance, [Coates and Ng, 2011] empirically observed that randomly sampled local features yield to a perfectly usable codebook for object recognition challenges.

### 2.3.1.2 Coding

Different methods have been investigated in the literature in order to map local features to codes over the codebook. In the original BOVW model, hard assignment [Csurka et al., 2004] is adopted to describe an image with a frequency of codewords. It is the simplest coding scheme which assigns a local feature  $\mathbf{x}_i$  to the closest codeword, i.e.,

$$z_{i,j} = \begin{cases} 1 & \text{if } j = \underset{j \in \{1, \dots, M'\}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{w}_j^v\|_2^2, \\ 0 & \text{otherwise,} \end{cases} \quad (2.2)$$

with  $\mathbf{z}_i = \{z_{i,1}, \dots, z_{i,M'}\}$  the code of size  $M'$  related to the local descriptor  $\mathbf{x}_i$  and  $\|\cdot\|_2$  represents the  $l_2$ -norm.

However, this coding often introduces large quantization errors. To alleviate this drawback, soft coding has been proposed in [van Gemert et al., 2009], assigning a local feature to codewords depending on the distance of a descriptor to the  $j^{\text{th}}$  codeword, i.e.,

$$z_{i,j} = \frac{\exp(-\beta \|\mathbf{x}_i - \mathbf{w}_j^v\|_2^2)}{\sum_{k=1}^{M'} \exp(-\beta \|\mathbf{x}_i - \mathbf{w}_k^v\|_2^2)}, \quad (2.3)$$

with  $\beta$  the assignment softness parameter. Even if soft coding has reduced quantization errors, there is no proof that the use of the entire codebook is optimal.

Yet another alternative to soft assignment is sparse coding [Yang et al., 2009]. It is generally performed by solving the  $l_1$ -norm regularized approximation problem:

$$\mathbf{z}_i = \underset{\mathbf{z} \in \mathbb{R}^{M'}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{z}\mathcal{W}^v\|_2^2 + \lambda \|\mathbf{z}\|_1, \quad (2.4)$$

with  $\lambda$  the trade-off parameter between the reconstruction error and the sparsity of the coding and  $\|\cdot\|_1$  represents the  $l_1$ -norm. Even if sparse coding has improved state-of-the-art classification rates, it remains computationally demanding because of the optimization procedure. Additionally, recent studies [Wang et al., 2010b; Gao et al., 2011] show its non-consistency to encode similar descriptors since it might select different basis for similar descriptors to favor sparsity.

Unlike sparsity, locality, which is a property introduced in [Yu et al., 2009] and investigated in several recent work [Wang et al., 2010b; Liu et al., 2011b; Huang et al., 2011], leads to reliable sparse codes while being computationally fast. Under the assumption that descriptors approximately reside on a lower dimensional manifold in an ambient descriptor space, the use of Euclidean distances for the assignment of descriptors to codewords is only meaningful within a local region of the feature space. Hence, local bases are selected to perform the coding. For efficient implementations, authors of [Wang et al., 2010b] propose to approximate the original formulation by solving a linear system derived from their proposed criteria:

$$\begin{aligned} \mathbf{z}_i &= \underset{\mathbf{z} \in \mathbb{R}^{M'}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{z}\mathcal{W}^v\|_2^2 + \lambda \|\mathbf{d}_i \odot \mathbf{z}\|_2^2, \\ \text{s.t. } \quad & \mathbf{1}^T \mathbf{z}_i = 1, \end{aligned} \quad (2.5)$$

with  $\mathbf{d}_i = \exp(\frac{\operatorname{dist}(\mathbf{x}_i, \mathcal{W}^v)}{\sigma})$  a vector of Euclidean distances computed between  $\mathbf{x}_i$  and the basis vectors, i.e.,  $\operatorname{dist}(\mathbf{x}_i, \mathcal{W}^v) = [\operatorname{dist}(\mathbf{x}_i, \mathbf{w}_1^v), \dots, \operatorname{dist}(\mathbf{x}_i, \mathbf{w}_{M'}^v)]^T$  and  $\sigma$  a parameter controlling the weight decay speed of the locality constraint. In practice, to solve the problem rapidly, the basis vectors are the  $L$ -nearest-codewords of the local feature.

In [Liu et al., 2011b], authors propose another efficient implementation of the locality-constrained coding by restricting the probabilistic soft coding approach (2.3) to only the  $L$ -nearest-codewords to a local feature, i.e.,

$$z_{i,j} = \begin{cases} \frac{\exp(-\beta \|\mathbf{x}_i - \mathbf{w}_j^v\|_2^2)}{\sum_{k=1}^L \exp(-\beta \|\mathbf{x}_i - \mathbf{w}_k^v\|_2^2)} & \text{if } \mathbf{w}_j^v \in \mathcal{N}_L(\mathbf{x}_i), \\ 0 & \text{otherwise,} \end{cases} \quad (2.6)$$

where  $\mathcal{N}_L(\mathbf{x}_i)$  denotes the  $L$ -nearest neighborhood of  $\mathbf{x}_i$ , under the Euclidean distance for instance. Recently, [Shabou and Le Borgne, 2012] propose a coding

scheme that takes into account the local spatial context of an image into the usual coding strategies proposed in the state-of-the-art. For this purpose, given an image, dense local features are extracted and structured in a lattice. The latter is endowed with a neighborhood system and pairwise interactions. An objective function is proposed to encode local features, which preserves locality constraints both in the feature space and the spatial domain of the image. An appropriate efficient optimization algorithm is provided, inspired from the graph-cut framework [Zabih and Kolmogorov, 2004].

### 2.3.1.3 Pooling

Given the coding coefficients of all local features within one image, a pooling operation has to be performed to obtain a compact visual signature  $\mathbf{X}_k^v = (\mathbf{x}_1^v, \dots, \mathbf{x}_{M'}^v)$  for an image  $\mathbf{I}_k$ , while preserving important information and discarding irrelevant details. This operation can be formulated as the following:

$$\mathbf{x}_j^v = g\left(\{z_{i,j}; i \in \{1, \dots, P\}\}\right); \forall j \in \{1, \dots, M'\}, \quad (2.7)$$

with  $g$  a pooling function such as the average, the sum or the maximum functions.  $P$  represents the number of SIFT descriptors extracted from the image and  $M'$  denotes the size of the visual codebook. The sum-pooling is the sum of the coding coefficients obtained on local features while the average-pooling is its normalized form. Both pooling functions have been usually considered in the original BOW model. Recent work [Boureau et al., 2010; Liu et al., 2011b] show, both theoretically and empirically, that max-pooling is best suited to the recognition task. Max-pooling is obtained by selecting the maximum coding coefficient (or codeword response) over local features for each codeword. Recently, [Avila et al., 2013] propose a new pooling scheme called BossaNova which enhances image representation by keeping an histogram of distances between the local descriptors of the image and those in the codebook, preserving thus important information about the distribution of the local descriptors around each codeword. Instead of compacting all information pertaining to a codeword into a single scalar, the proposed pooling scheme produces a distance distribution.

## 2.3.2 Textual Description

In this section, we review the state-of-the-art approaches that have been proposed for image annotation based on the tag modality. Two types of relations are exploited to generate tag-based features. The first category, called *Tag-to-Concept* approaches, relies on a semantic relation between image tags and annotation concepts. The second category, called *Tag-to-Tag* approaches, is based on a semantic

relation between image tags and a predefined dictionary of tags. A comparison of these approaches is presented in Table 2.1. Let's note that other tag representations which deal with tag imperfections in other social applications such as tag ranking and tag suggestion, are presented in Section 2.4.

### 2.3.2.1 Tag-to-Concept Approaches

These approaches propose to exploit a semantic relation between tags and annotation concepts. Most of the time, the semantic relation is defined using the co-occurrence between tags and concepts. Starting from the hypothesis that “tags and concepts co-occurrences are strong for certain tags, yet weak for other noisy or related tags”, [Gao et al., 2010] propose to use this co-occurrence information to predict the probability of an image to belong to a particular concept. The concept-tag co-occurrence matrix is computed on the training dataset. To avoid the use of tags unrelated/misspelled to concepts, they consider those words which appear in WordNet as the representative tags. However, by considering only tags that appear in WordNet, many tags which have useful information are discarded. [Wang et al., 2010a; Li et al., 2010b] approaches are based on an expansion procedure of both tags and annotation concepts and a comparison on the expanded representations. [Wang et al., 2010a] propose an approach to build Semantic Fields for annotating the web images. The main idea is that the images are more likely to be relevant to a given concept, if several tags of the image belong to the same Semantic Field as the target concept. Semantic Fields are determined by a set of highly semantically associated terms with high tag co-occurrences in the image database and in different corpora and lexica such as WordNet and Wikipedia. The obtained Semantic Field is used as annotation for the image. [Li et al., 2010b] propose a method based on a document expansion procedure which assigns additional content to concepts and image tags by consulting external information resources, such as DBpedia<sup>1</sup>. After that, expanded textual metadata is compared to expanded concepts in order to make concept assumptions. Finally, additional concepts are considered by inferring affiliations and opposite relations among them.

Different from the above models, other approaches [Nagel et al., 2011; Liu et al., 2013] are based on the the BOW representation [Salton and McGill, 1983]. [Nagel et al., 2011] employ a supervised approach which learns tag frequencies on the concepts of the training set. Concept-based TF-IDF weights are assigned to each tag. A tag term frequency is detected by counting the number of times the tag occurs in images annotated with a certain concept. The document frequency term is equivalent to the number of concepts that co-occur with a tag. Finally, for each concept, the TF-IDF values of image tags are accumulated. The size of the obtained BOW feature vector is equal to the number of annotation concepts.

---

<sup>1</sup><http://dbpedia.org>

Recently, [Liu et al., 2013] propose the Histogram of Textual Concepts (HTC) model to capture the semantic relatedness of semantic concepts. The HTC model is based on the BOW representation. It is defined as an histogram of textual concepts towards a dictionary, and each bin of this histogram represents a concept of a dictionary of concepts. The bin value is the accumulation of the contribution of each tag toward the underlying concepts according to a WordNet semantic similarity [Wu and Palmer, 1994]. Let's note that this method is not based on a co-occurrence relation.

### 2.3.2.2 Tag-to-Tag Approaches

These approaches propose to exploit a semantic relation between image tags and a predefined dictionary of tags to generate the tag-based signature. In this category, all reported approaches are based on the BOW representation.

[Guillaumin et al., 2010] proposed a model based on the classic BOW representation where the textual signature is defined as a binary vector representing the presence or the absence of image tags towards a predefined dictionary. This dictionary is built by keeping the most frequent tags. However, as properties of tag data are completely different from those of text documents, the chosen BOW representation is not convenient for tag representation due to the number of tags which is very small compared to that of words in a document, which introduces the problem of sparse feature representation.

In order to overcome this shortcoming, [Kawanabe et al., 2011] were interested in improving the tag-based signature generation step through a new smoothing technique of the final tag-signature. This approach relies on Markov random walks on a graph of tags. [Xioufis et al., 2011] propose a textual signature based on a binary BOW representation including word stemming, stop words removal, and feature selection using the chi-squared-max method in order to remove irrelevant or redundant features. The learning step is achieved with an Ensemble of Classifier Chains (ECC) [Read et al., 2009] using Random Forests as base classifier. [Zhang et al., 2012b] propose the semantic BOW model in order to capture the semantic information between tags which is hardly described with a classic BOW model. WordNet-based distance between tags is used for dictionary construction and histogram assignment.

### 2.3.2.3 Discussion

In most of these approaches, BOW model represents the dominant approach for tag representation using different variants of word frequency (TF, TF-IDF...), using

TABLE 2.1: A Summary of the state-of-the-art approaches on Tag-based image annotation.

Method	Principle	Relation	BOW	Dictionary	Learning	Knowledge Resource	Handling Imperfections
[Wang et al., 2010a]	Semantic Fields based on the tag-concept co-occurrence.	Tag-Concept	No	No	No	WordNet, Wikipedia and the training set.	No
[Gao et al., 2010]	Probability based on the tag-concept co-occurrence.	Tag-Concept	No	No	No	The training set.	No
[Li et al., 2010b]	Compare Tag and annotation concepts expansion vectors.	Tag-Concept	No	No	No	DBpedia.	No
[Nagel et al., 2011]	BOW based on the tf-idf values of tags.	Tag-Concept	Yes	Annotation concepts	SVM	The training set.	No
[Liu et al., 2013]	Histogram of Textual Concepts based on the Tag-to-Concept similarity.	Tag-Concept	Yes	Annotation concepts	SVM	WordNet.	The incomplete data problem.
[Guillaumin et al., 2010]	Binary BOW representation representing the presence/absence of tags.	Tag-Tag	Yes	Frequent tags.	SVM	No	No
[Kawanabe et al., 2011]	Binary BOW representing the presence/absence of tags with random walks over tags.	Tag-Tag	Yes	Frequent tags	SVM	No	The incomplete data problem.
[Xioufis et al., 2011]	Binary BOW representation with feature selection.	Tag-Tag	Yes	Frequent tags.	ECC	No	No
[Zhang et al., 2012b]	Semantic BOW based on the Tag-to-Tag similarity.	Tag-Tag	Yes	Frequent tags	SVM	WordNet.	The incomplete data problem.
[Romberg et al., 2012]	Probabilistic latent semantic analysis on tags co-occurrence matrix.	Tag-Tag	Yes	Frequent tags	PLSA	No	No

some pre-processing techniques (stemming, stop words removal ..) or smoothing techniques such as the random walks over a graph of tags. We identify two main differences between the BOW-based approaches. The first difference concerns the considered dictionary. In the first family of approaches, concept annotations are used as entries of the dictionary, however, the second type of approaches opts for frequent tags to build the dictionary. Certainly, the size of the BOW representation with annotations concepts as a dictionary is more compact, whereas, the obtained signature with a BOW on frequent tags is richer and seems to be more suitable for tags. The second difference concerns the histogram assignment step in the BOW representation. As introduced in the previous chapter, tags and annotation concepts are different and a simple Tag-to-Concept matching, called also hard coding, is not the best coding scheme in the histogram assignment step. Thus, using external knowledge resource such as WordNet can be useful as shown in [Liu et al., 2013]. In [Xioufis et al., 2011; Guillaumin et al., 2010; Kawanabe et al., 2011] approaches, the mapping of image tags and dictionary tags is a simple Tag-to-Tag matching, however, [Zhang et al., 2012b] exploit semantic similarity using WordNet in order to match more tags, called also soft coding. However, considering the contribution of all tags, and specifically those with small similarities, in the BOW representation introduces much noise. In [Kawanabe et al., 2011] approach, the smoothing step handles a part of the incomplete data problem. In fact, the random walk on a graph of tags enables to add semantically tags in the tag representation. A similar effect is obtained in both [Zhang et al., 2012b; Liu et al., 2013] by using the soft coding scheme in the histogram assignment. As most of these tag representations are based on classic BOW model, they do not take into account tag imperfections and fail to capture semantic tag relatedness.

### 2.3.3 Fusion Strategy

In this section, we are interested in the fusion strategy used to combine both textual and visual modalities. We focus on the combination of both tag and image modalities in the context of multimodal image annotation. Let's note that the combination of different mono-media features are out of the scope of this dissertation. The work of [Snoek, 2005] was the first to introduce and identify two general fusion strategies within the machine learning trend to semantic video analysis: early fusion and late fusion. We refer the reader to a review on multimodal fusion strategies for multimedia analysis in [Atrey et al., 2010]. As introduced in Section 2.1, multimodal approaches can be categorized into three categories depending on the level of fusion.

### 2.3.3.1 Early fusion

This category of fusion is performed at the feature level. After processing of both visual and textual modalities, the extracted features are combined into a single representation. The general scheme of early fusion strategy is illustrated in Figure 2.5. We categorize most of approaches that fall in early fusion strategy into three groups: Concatenation-based, Dictionary-based and Topic-based approaches. The taxonomy of the work based on early fusion strategy for image annotation is presented in Figure 2.6.

- ***Concatenation-based approaches***

The simple and widely used method in the early fusion is the concatenation of both visual and tag features. This method has been used by [Li et al., 2009b] for landmark annotation, where visual and tag features are simply concatenated. A Support vector machine (SVM) is then learned on the combined feature vector. Nevertheless, this simple concatenation do not take into account the correlation that may occur between visual and textual modalities.

- ***Dictionary-based approaches***

These approaches are based on the learning of the correlation between visual features and textual words. The translation model [Duygulu et al., 2002] is a representative work in which images are segmented into regions. Then the words and blobs (segmented regions) are considered as two equivalent languages. After training, the translation model can attach words to a

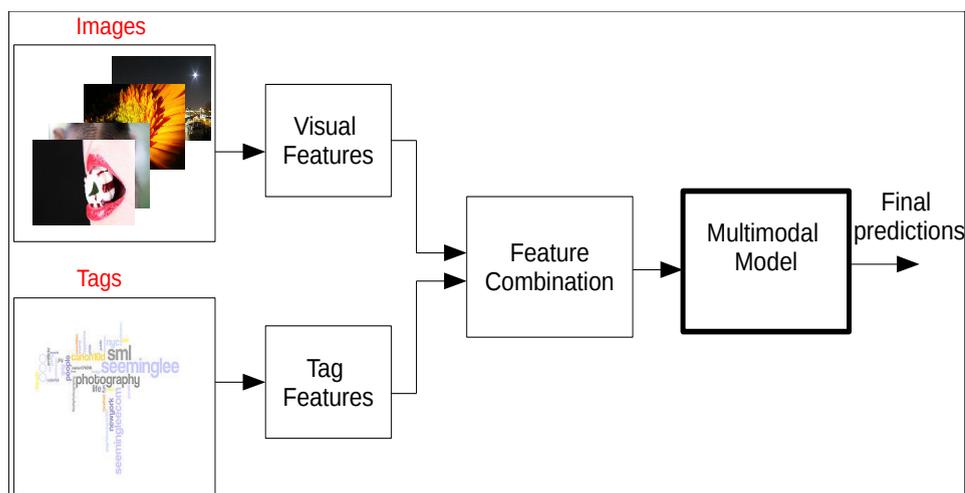


FIGURE 2.5: The general scheme of the early fusion strategy. After analysis of both visual and textual modalities, the extracted features are combined into a single representation.

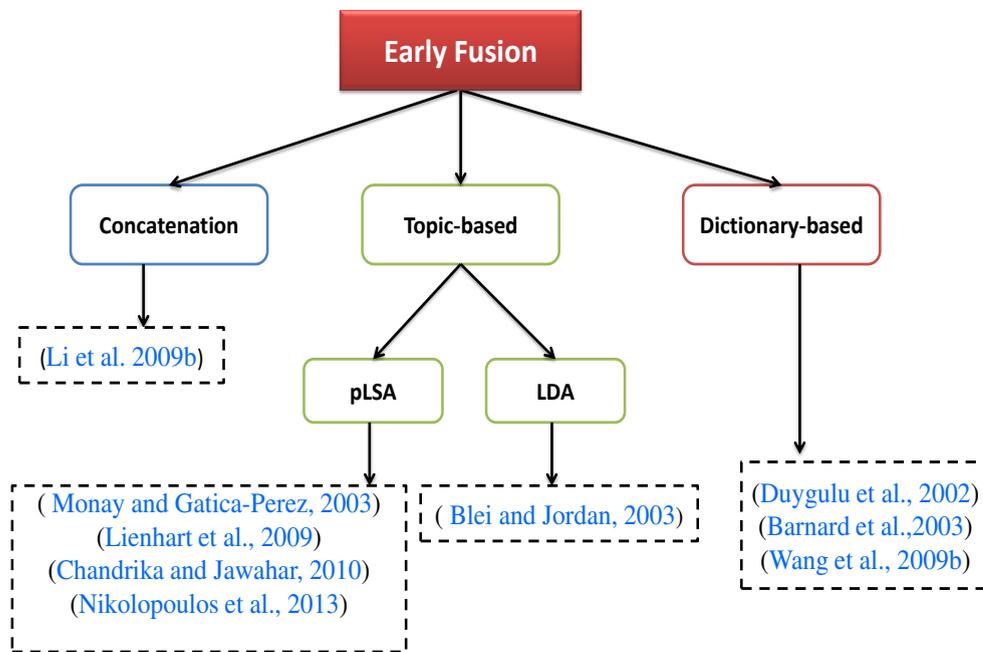


FIGURE 2.6: Taxonomy of work based on early fusion strategy for image annotation in social media.

new image region through learning the correlations using Estimation Maximization (EM) algorithm. Similarly, [Barnard et al., 2003] discuss several models to represent the joint distribution of words and blobs. Once the joint distribution has been learned, the annotation problem is converted into a likelihood problem relating blobs to words. However, the performance of these models is strongly affected by the quality of image segmentation.

[Wang et al., 2009b] propose to construct a visual tag dictionary by mining community-contributed media corpus. For each specific tag, a Gaussian Mixture Model (GMM) is built based on the images annotated with it and their visual-word representations. A set of GMM parameters needs to be learnt.

- ***Topic-based approaches***

These approaches rely on the use of aspect or topic models and the definition of a latent semantic space. The key idea in latent semantic analysis is to map high-dimensional count vectors, such as term frequency vectors arising in the vector space representation of text documents [Salton and McGill, 1983], to a lower dimensional representation, a so-called *latent semantic space*. This latter is composed of a set of variables called hidden aspects or topics that bridges the semantic gap between high-level concepts that human perceives and low-level features that usually describe images. Probabilistic Latent

Semantic Analysis (pLSA) [Hofmann, 1999a] and Latent Dirichlet Allocation (LDA) [Blei and Jordan, 2003] are the popular techniques in this direction. For instance, [Blei and Jordan, 2003] employ correspondence LDA model to build a language-based correspondence between words and the whole image. This model assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. This latter is then used to generate words and regions. EM algorithm is used to estimate this model.

The key concept of the pLSA model is to map high dimensional word distribution vector of a document to a lower dimensional topic vector or aspect vector. Thus, it introduces an unobservable latent topic between documents and words. Each document consists of mixture of multiple topics and thus the occurrences of words is a result of the topic mixture. One of the aspects of this model is that word occurrences are conditionally independent from the document given the unobservable aspect.

Both pLSA and LDA are topic-based approaches but have some differences. In LDA, each document may be viewed as a mixture of various topics. This is similar to pLSA, except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the pLSA model is equivalent to the LDA model under an uniform Dirichlet prior distribution [Girolami and Kabán, 2003]. A graphical illustration of the various generative pLSA models is presented in 2.7. [Monay and Gatica-Perez, 2003] apply a pLSA on a concatenated representation of the textual and the visual modalities of a set of annotated images. Using a concatenated representation, this approach attempts to simultaneously model visual and textual modalities. Assuming that no particular importance is given to any modality, the amount of visual and textual information need to be balanced in the concatenated representation of an annotated image which can limit the size of the visual representation. [Lienhart et al., 2009] use a pLSA-based model to support multi-modal image retrieval in Flickr, using both visual content and tags. They propose to extend the standard single-layer pLSA model to multiple layers by introducing not just a single layer of topics, but a hierarchy of topics. First pLSA is applied to each modality (image and tags) separately, and then the derived topic vectors of each modality are concatenated. pLSA is applied on the top of the derived vectors to learn the final document concept relation. This is equivalent to forming an alternative dictionary of concepts, one for each modality, and merging them into a single dictionary on which pLSA is performed. This approach has the intrinsic problem of having to merge dictionaries of the different modalities. This method does not place importance to interactions between the different modalities and suffers from a high computational complexity.

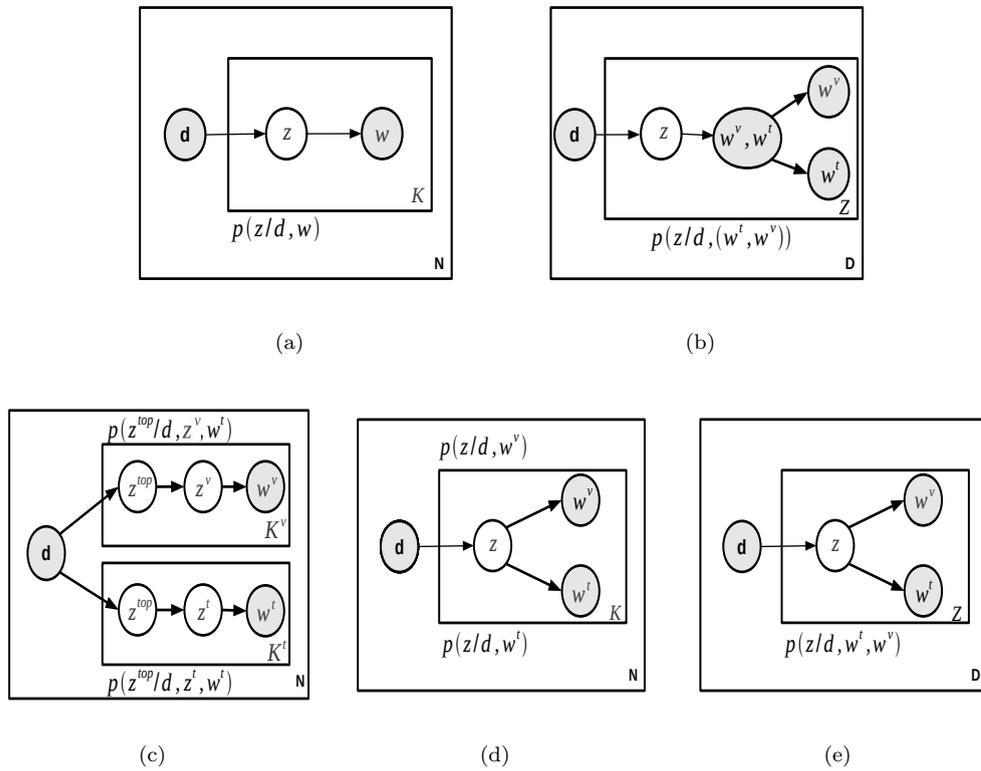


FIGURE 2.7: Graphical illustrations of the various generative pLSA models. (a) Standard monomodal pLSA [Hofmann, 1999b], (b) pLSA on a concatenated representation [Monay and Gatica-Perez, 2003], (c) multilayer monomodal pLSA of [Lienhart et al., 2009], (d) multimodal pLSA of [Chandrika and Jawahar, 2010]. (e) High order pLSA of [Nikolopoulos et al., 2013].  $N$  (resp.  $K$ ) is the number of observed documents  $d$  (resp. words  $w$ ),  $z$  corresponds to topic and  $v$  (resp.  $t$ ) denotes the visual (resp. textual) mode. Gray color indicates an observed (non-latent) variable.

To overcome this shortcoming, [Chandrika and Jawahar, 2010] propose a multimodal pLSA that captures the patterns between images (i.e. text words and visual words) using the EM algorithm to determine the hidden layers connecting them. Although the authors goal is to exploit the interactions between the different modes when defining the latent space, they eventually implement a simplified model where they assume that a pair of different words are conditionally independent given the respective image.

Recently, [Nikolopoulos et al., 2013] propose an extension of pLSA to become applicable for more than two observable variables. Then, by processing images, visual features and tags as the three observable variables of an aspect model, a space of latent topics is learnt such that it incorporates the semantics of both visual and tag information. This approach is based on using the cross-modal dependencies learned from a corpus of images to approximate

the joint distribution of the observable variables. In [Chandrika and Jawahar, 2010] approach, hidden topics generated from visual and tag words are considered separately, whereas, in [Nikolopoulos et al., 2013] images, visual and tag words are considered as the three observable variables of pLSA.

### 2.3.3.2 Late fusion

The late fusion is performed at the decision level. Approaches that use late fusion strategies also start with the extraction of unimodal features. In contrast to early fusion, where features are combined into a multimodal representation, late fusion approaches learn models directly from unimodal features. The general scheme of the late fusion strategy is illustrated in Figure 2.8.

Most of these approaches fall into two categories: *Rule-based* and *Classification-based* approaches. The taxonomy of the work based on late fusion strategy for image annotation is presented in Figure 2.9.

- ***Rule-based approaches***

This category of approaches is based on the application of a rule to combine information from different modalities. In Rule-based approaches, we distinguish two categories: Statistical-based and Dempster-Shafer-based approaches.

- **Statistical-based approaches**

In this category of approaches, rules used for combining multimodal information are statistical rules such as linear weighted fusion (sum, average), MAX, MIN, AND, OR, majority voting. Linear weighted

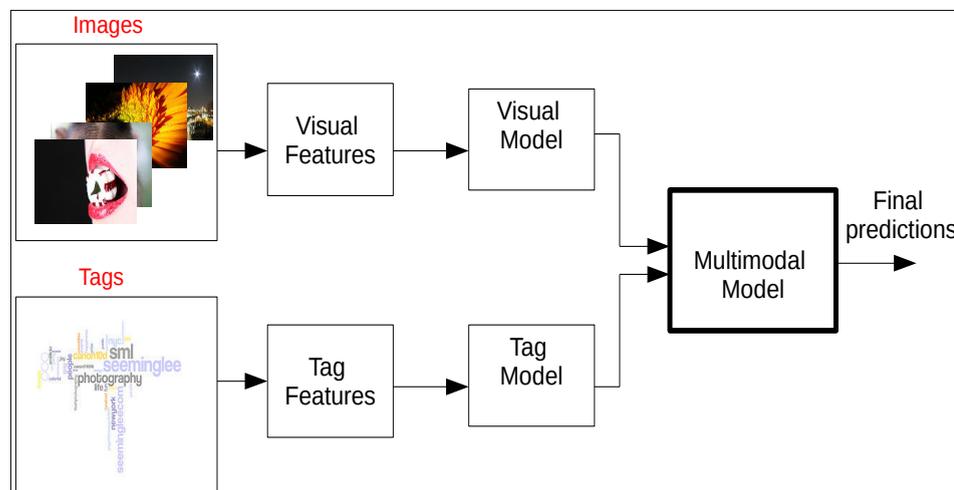


FIGURE 2.8: The general scheme of the late fusion strategy. Each modality is processed separately and then combined at decision level.

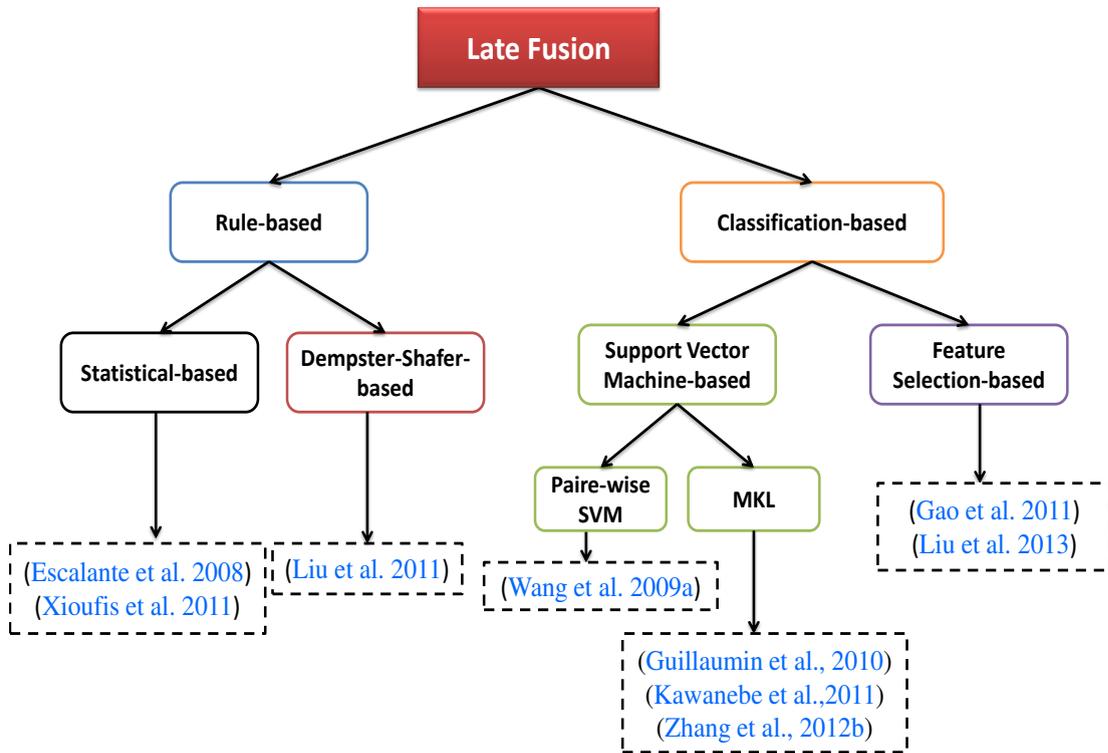


FIGURE 2.9: Taxonomy of the work based on the late fusion strategy for image annotation in social media.

fusion rule represents one of the simplest and most widely used methods for image annotation. In [Escalante et al., 2008], predictions obtained from different classifiers learned on both textual and visual modalities are combined in a linear way. Similarly, [Xioufis et al., 2011] propose a multimodal scheme based on a hierarchical late fusion. In a first stage, predictions obtained from different classifiers learned on several visual features are averaged. At a second stage, obtained scores from the visual modality are averaged with textual predictions obtained from tag-based signatures.

#### – Dempster-Shafer-based approaches

Although statistical rule-based approaches are simple and give good performances in most of cases on multimodal image annotation, some researchers have chosen to use the Dempster-Shafer (DS) evidence theory [Shafer, 1976], that is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers. This theory will be introduced in details in Section 2.5.4.

In this vein, [Liu et al., 2011] propose to combine textual and visual classifier predictions based on the Dempster’s rule of combination to improve the classification accuracy. Each feature was used to train a classifier, which produces a measurement vector as a degree of belief that the input image belongs to different classes. Classifiers were trained

based on adjusting the evidence of different classifiers, by minimizing the Mean Square Error (MSE) of training data according to [Al-Ani and Deriche, 2002]. In our knowledge, this is the first attempt to apply Dempster-Shafer theory to combine both visual and tag information for image annotation in the context of social media. However, this theory has been applied in many other research fields.

Although the Dempster-Shafer fusion method has been found more suitable for combining both visual and textual classifiers compared to simple rule-based approaches, this method suffers from the combinatorial explosion when the number of frames of discernment is large. This latter corresponds to the number of annotation labels for the image annotation task. This point will be discussed in details in Section 2.5.4.

- ***Classification-based approaches***

In this category of approaches, the optimal combination of different modalities and features is learned using a range of classification techniques. There have been many classification-based approaches in the literature. We categorize these approaches in two groups: Support Vector Machines-based (SVM) and Feature selection-based approaches as shown in the taxonomy presented in Figure 2.9. SVM-based approaches include both linear SVM classifiers (called Paire-wise SVM) and MKL classifiers.

### **Support Vector Machines-based approaches**

SVM [Cortes and Vapnik, 1995] have become increasingly popular for data classification and related tasks. More specifically, SVMs are being used to combine different modalities for multimedia annotation.

- **Paire-wise SVM**

From the perspective of multimodal fusion, SVM is used to solve image classification problem, using different modalities, where the input of this classifier are the scores given by the individual classifiers. In this way, [Wang et al., 2009a] propose to build two separate classifiers, one for the text features and the other one for the visual features. A third classifier is then trained to combine the confidence values of the two initial classifiers into a final prediction. This final classifier uses logistic regression and is trained on a validation set.

- **Multiple Kernel Learning**

The basic SVM method is extended to create a non-linear classifier by using the kernel concept, where every dot product in the basic SVM formalism is replaced using a non-linear kernel function. In this vein, one representative method is to consider the features as multiple kernel matrices and then combine them in the kernel space. One of the most successful feature fusion methods is MKL [Lanckriet et al., 2004], which

learns the optimal kernel mixture and the model parameters of the SVM simultaneously. Many approaches have been proposed in the literature to combine different modalities using the MKL framework [Guillaumin et al., 2010; Kawanabe et al., 2011; Zhang et al., 2012b].

[Guillaumin et al., 2010] propose a semi-supervised learning approach to leverage the information contained in tags associated with unlabeled images in a two-step process. First, labeled images are used to learn a strong classifier that uses both the image content and tags as features. The MKL framework (more precisely, the simple MKL method of [Rakotomamonjy et al., 2008]) is adopted to combine a kernel based on the image content with a second kernel that encodes the tags associated with each image. This MKL classifier is used to predict labels of unlabeled training images with associated tags to obtain additional examples to train a classifier. In the second step, both the labeled data and the output of the classifier on unlabeled data are used to learn a second classifier, that uses only visual features as input.

A similar approach is used in [Kawanabe et al., 2011]. However, for simplicity, authors deployed uniform kernel weights and trained SVMs with the averaged kernels, which achieved comparable results to MKL. Based on a tag refinement step using markov random walk on the tag graph, this method outperforms the one of [Guillaumin et al., 2010]. The same framework has been applied by [Zhang et al., 2012b] to combine two kernels learned on both visual and textual features.

### **Feature selection-based approaches**

Feature selection is a technique commonly used in machine learning to find the best subset of features or experts in classifier combination that enhance classification performances. It was applied to combine tag and visual features for image annotation.

By studying the characteristics of tag and visual features, [Gao et al., 2010] propose the Grouping-Based-Precision & Recall-Aided (GBPRA) feature selection strategy for concept annotation. More specifically, authors define the Grouping Semantic Depth as the mode of semantic depth of all concepts in WordNet belonging to this grouping. Based on the level of Grouping Semantic Depth, a selection is performed to decide either to use tag or visual features for concept annotation. By studying the tag distribution, they adopt precision and recall as a complementary indicator for feature selection.

[Liu et al., 2013] proposed a fusion scheme, called Selective Weighted Late Fusion (SWLF), that selectively chooses and weights the best discriminative features for each visual concept to be predicted in optimizing the overall mean average precision. The proposed approach mainly includes two stages: a training stage and a testing stage. The training stage consists of training

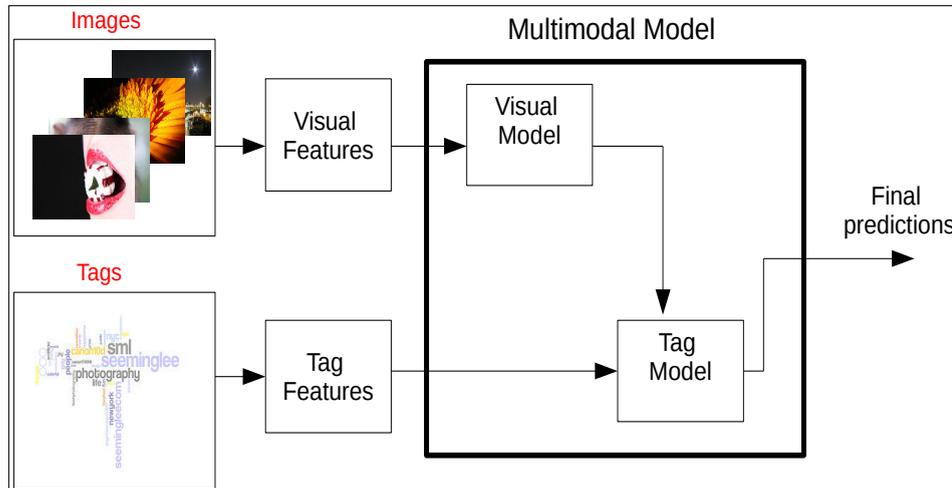


FIGURE 2.10: The general scheme of the transmedia fusion strategy for image annotation in social media.

experts through SVM for each pair of concept and type of features using a training set. These experts are then evaluated using a validation set to learn SWLF. The testing stage proceeds to extract various types of features (textual and visual) from an input image, and then to apply the corresponding fusion scheme learned by SWLF for each concept to obtain a recognition decision.

### 2.3.3.3 Transmedia fusion

This level of fusion is also called cross-media or intermediate level fusion. This kind of approaches views image annotation as a process of transferring tags from nearest neighbors. The basic idea is to use visual features to gather relevant images (visually nearest neighbors) and then to switch to the textual modality to aggregate tag features of these neighbors. We call these methods nearest neighbor-based approaches. The general scheme of the transmedia fusion strategy is illustrated in Figure 2.10.

Recently, nearest neighbor approaches have been investigated in the annotation community with promising results. Notably, [Torralba et al., 2008] collected about 80 million tiny images, each of which is labeled with one of the 75,062 abstract nouns from WordNet. By fully leveraging on the redundancy of information on the Web, they claimed that with sufficient number of samples, the simple nearest neighbor classifier can achieve reasonable performance for several object/scene detection tasks, when compared with the more sophisticated state-of-the-art techniques.

In the same vein, [Makadia et al., 2008] have developed the joint equal contribution (JEC) technique, where they used a combination of multiple features and

distance metrics to find the nearest neighbors of the input image and used a greedy algorithm for transferring tags from visually similar images. Although these methods [Makadia et al., 2008; Torralba et al., 2008] show good performances, they do not allow the integration of metric learning. This latter defines the nearest neighbors in order to maximize the predictive performance of the model.

[Guillaumin et al., 2009] have proposed the tag propagation (TagProp), a discriminative metric learning approach, to annotate an input image by propagating tags of weighted nearest neighbors of that input image. Neighbor weights are determined based on the neighbor rank or its distance, and set automatically by maximizing the likelihood of annotations in a set of training images. The weighted nearest neighbors were identified by optimally integrating several image similarity metrics.

In the same direction, [Li et al., 2009a] proposed an algorithm that learns tag relevancy by accumulating votes from visually similar neighbors. This approach relies on the intuition that if different persons label visually similar images using the same tags, these tags are likely to reflect the objective aspects of the visual content. In fact, given an user-tagged image, they first perform a  $k$ NN search to find its visual neighbors. The tag relevance is determined as the probability that this tag is used to annotate the neighborhood images minus the probability of that tag used in the entire collection.

#### 2.3.3.4 Discussion

This section provides a critical look on the fusion strategies presented in Section 2.3.3. A summary of the most related work using the early fusion strategy described above is provided in Table 2.2. While some approaches such as the one proposed by [Li et al., 2009b] are straightforward and simply consist in concatenating the features extracted from both visual and tag modalities into a single representation, their disadvantage is also well known: the fusion of features usually results into a large feature vector, which becomes a bottleneck for the learning task. This is known as the *curse of dimensionality* [Bellman, 1961]. It suffers also from the difficulty in combining features of different natures into a common homogeneous representation. To overcome this shortcoming, many approaches propose to process the image annotation problem as a translation task from image instances to tags, and it is usually accomplished based on some models that exploit the co-occurrence of images and tags [Duygulu et al., 2002; Barnard et al., 2003]. However, the performance of these models is strongly affected by the quality of image segmentation. Topic-based approaches [Blei and Jordan, 2003; Monay and Gatica-Perez, 2003; Lienhart et al., 2009; Chandrika and Jawahar, 2010] represent an alternative to LDA-based approach. However, most of these approaches do not consider the correlation between both visual and tag modalities. Thus,

TABLE 2.2: A summary of the most related and representative work based on the early fusion strategy for multimodal image annotation in the context of social media.

The work	Fusion method	Fusion level	Handling Imperfections
[Li et al., 2009b]	Simple concatenation of visual and tag representations.	Early Fusion	No
[Duygulu et al., 2002]	Translation model to link tags and blobs.	Early Fusion	No
[Barnard et al., 2003]	Learn the joint distribution of tags and blobs.	Early Fusion	No
[Blei and Jordan, 2003]	LDA on tags and images.	Early Fusion	No
[Monay and Gatica-Perez, 2003]	pLSA on a concatenation of tag and image features.	Early Fusion	No
[Lienhart et al., 2009]	Multilayer and multimodal pLSA.	Early Fusion	No
[Chandrika and Jawahar, 2010]	Multimodal pLSA	Early Fusion	No
[Nikolopoulos et al., 2013]	High Order pLSA.	Early Fusion	No
[Wang et al., 2009b]	Visual tag dictionary using GMM.	Early Fusion	No

some topic-based approaches [Nikolopoulos et al., 2013], that exploit cross-modal dependencies learned from a corpus of images to approximate the joint distribution of the observable variables, give better results and seem to be more adapted to multimodal data. However, most of the early fusion approaches do not take into account tag imperfections at the feature level (uncertainty, imprecision and incompleteness) that have been introduced in Chapter 1.

Contrary to early fusion approaches where the combination process is performed at the feature level, late fusion approaches process each modality separately and combine them at the decision level. A summary of work using the late fusion strategy described above is provided in Table 2.3. It is clear that many fusion methods such as Linear weighted fusion and SVM have been used more often in comparison to the other methods. Linear weighted fusion method has been used due its simplicity as well as it is computationally less expensive than other approaches and it can be easily used to prioritize one modality or the other. This method performs well if the weights of different modalities are appropriately determined, which has been a major issue. Feature selection-based methods [Gao et al., 2010; Liu et al., 2013] have been proposed as an alternative to learn which modality is more discriminative for each concept. Among others, MKL has been used recently

to combine image and tag features. Although, state-of-the-art MKL-based fusion approaches [Guillaumin et al., 2010; Kawanabe et al., 2011; Zhang et al., 2012b] show good results, they still suffers from high computational complexity, compared to the pair-wise SVM approach [Wang et al., 2009a]. To reduce the computational cost of MKL-based approaches, [Kawanabe et al., 2011] show that uniform kernel weights and a SVM trained with the averaged kernels, achieve comparable results to MKL. Although these approaches give good overall performances, imperfection aspects at decision level (uncertainty, imprecision and incompleteness), that have been introduced in Chapter 1, are occasionally considered. In fact, the decision cannot be estimated with absolute certainty using the classification models. Most of the above methods consider the fusion as a score aggregation task, except the work of [Liu et al., 2011]. In this approach the Dempster-Shafer (DS) evidence theory [Shafer, 1976] seems to be particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers. However, in [Liu et al., 2011] approach, Dempster-Shafer theory was applied for a small dataset ( $\approx 1,200$  images) and only for six classes of emotions. Although, the Dempster-Shafer theory has been found to be effective in combining different classifiers, this method suffers from the combinatorial explosion in particular when the number of annotation concepts (i.e the frames of discernment) is large. Unfortunately, this is precisely the case one must handle for the considered multimedia collections. A review of some representative work that have used Dempster-Shafer theory for various multimedia analysis tasks such as segmentation of satellite images, video classification and finger print classification can be found in [Atrey et al., 2010].

A summary of work using the transmedia fusion strategy described above is provided in Table 2.4. These approaches are based on a classic neighbor voting algorithm that uses information from the nearest neighbors to predict tags. Unfortunately, in the context of social tagging, tags are freely assigned by users, with various motivations and different judgments on the relevance between a tag and an image. Consequently, tags in social tagging setting are much more uncertain compared to labels in traditional classification problems. In the original voting  $k$ NN algorithm, the image is assigned to the majority class according to its  $k$ -nearest neighbors, independently of the relevance of each neighbor. Moreover, the classical  $k$ NN methods does not deal with ambiguous and imprecise information because of the limitation of the probabilistic framework. Moreover, there is no explicit use of a formalism which is able to handle neighbors conflict and to deal with tag imperfections.

To the best of our knowledge, handling imperfections both in representation and decision levels has never been explicitly considered in image annotation methods in the context of social media. Nevertheless, these imperfections have been identified and studied in other related fields. In Section 2.4, we review the most representative work that consider the relatively low quality of tags in tag-based applications

TABLE 2.3: A summary of the most related and representative work based on the late fusion strategy for multimodal image annotation in the context of social media.

The work	Fusion method	Fusion level	Handling Imperfections
[Escalante et al., 2008]	Linear combination of classifier predictions.	Late Fusion	No
[Xioufis et al., 2011]	Hierarchical late fusion using average rule.	Late Fusion	No
[Wang et al., 2009a]	SVM classifier using the concatenation of predictions as input feature.	Late Fusion	No
[Guillaumin et al., 2010]	Multiple Kernel Learning.	Late Fusion	No
[Kawanabe et al., 2011]	Multiple Kernel Learning.	Late Fusion	No
[Zhang et al., 2012b]	Multiple Kernel Learning.	Late Fusion	No
[Gao et al., 2010]	Grouping-Based-Precision & Recall-Aided (GBPRA) feature selection.	Late Fusion	No
[Liu et al., 2013]	Selective weighted late fusion.	Late Fusion	No
[Liu et al., 2011]	Dempster’s rule to combine classifier predictions	Late Fusion	<b>Yes</b>

TABLE 2.4: A summary of the most related and representative work based on the transmedia fusion strategy for multimodal image annotation in the context of social media.

The work	Fusion method	Fusion level	Handling Imperfections
[Makadia et al., 2008]	Joint Equal Contribution of nearest neighbors and tag transfer.	Transmedia fusion	No
[Torralba et al., 2008]	Leveraging tags from neighbors.	Transmedia fusion	No
[Guillaumin et al., 2009]	Tag propagation using metric learning.	Transmedia fusion	No
[Li et al., 2009a]	Tag relevance by accumulating votes from neighbors.	Transmedia fusion	No

such as tag ranking and suggestion. In Section 2.5, we review the state-of-the-art theories related to classifier combination that handle imperfections at decision level.

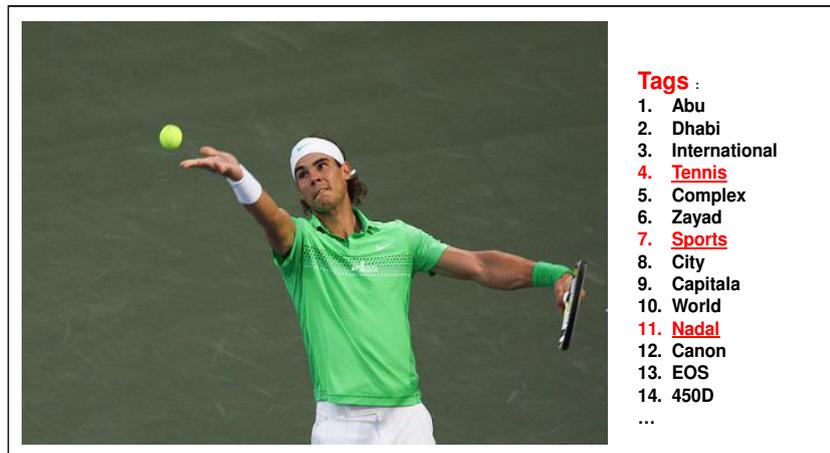


FIGURE 2.11: An example of image from Flickr and its associated tag list. The most relevant tags such as “Tennis, Sport, Nadal” are not at the top positions.

## 2.4 Handling Tag Imperfections

As introduced in Chapter 1, many online media repositories, such as Flickr, support tag-based multimedia search. However, since these tags are freely assigned by users, they are often noisy and incomplete and there is still a gap between these tags and the actual visual content of images [Kennedy et al., 2006; Liu et al., 2011a].

Recently, many research efforts have been proposed to enhance the quality of tags in the context of social media. The existing work mainly focus on the following two social media applications: (a) *tag ranking and relevance*, which aims to re-order tags associated with images with various levels of relevance; (b) *tag refinement and suggestion* which aims at refining the unreliable human-provided tags by dropping inappropriate tags and adding new missing tags.

### 2.4.1 Tag Ranking & Relevance

The relevance levels of tags associated with a social image cannot be distinguished from the tag list. An example is illustrated in Figure 2.11, from which we can see that the most relevant tags to describe the visual content are “Tennis, Sport, Nadal”. Their relevance can not be discovered from the tag list directly, by considering the order of the tags for instance. Indeed, the order of the different tags in the tag list is just based on the manual input and carries little information about their importance or relevance. Further, this limits the effectiveness of tags in search-based applications.

Tag ranking is defined as the process of assigning the right order or weight to each tag associated to an image. Many approaches have been devoted to solve this

problem. As suggested by [Liu et al., 2011a; Ballan et al., 2013], these approaches can be distinguished into two broad categories: *Statistical modeling* and *Data-driven* approaches.

#### 2.4.1.1 Statistical modeling approaches

These approaches consist in learning a statistical model used to determine tag relevance. As a pioneering work, [Liu et al., 2009a] propose a *tag ranking* scheme, aiming at automatically ranking tags associated with a given image according to their relevance to the image content. First, initial relevance scores for tags are estimated based on probability density estimation, and then a random walk over a tag similarity graph is performed to refine the relevance scores. Although, this method achieves a better result than the initial rank order given by users, it is limited in several aspects. The proposed method has to be trained on a very large database to construct a convenient tag similarity graph. Moreover, each tag graph is used for only one corresponding image. A new tag graph has to be learned for ranking tags of another image. In addition, the method of [Liu et al., 2009a] works in a transductive manner and only the already tagged images can be tag ranked, thus it is unable to deal with untagged images. To overcome this shortcoming, [Wang et al., 2010c] propose a semi-supervised learning model, called *Learning To Rank Tags*, which learns a ranking projection from visual word distribution to the relevant tag distribution using a simple linear regression model, and then use it for ranking new image tags. Similarly, [Feng et al., 2010] propose a novel *tag saliency ranking* scheme, which aims at automatically ranking tags associated with a given image according to their saliency to the image content. The proposed method combines both visual attention model and multi-instance learning algorithm to investigate the saliency ranking order information of tags with respect to the given image. Specifically, tags annotated on the image-level are propagated to the region-level via an efficient multi-instance learning algorithm. Then, visual attention model is used to measure the importance of regions in the given image. And finally, tags are ranked according to the saliency values of the corresponding regions.

#### 2.4.1.2 Data-driven approaches

These approaches consist in using an external data collection to determine tag relevance. As a pioneering work, [Li et al., 2009a] proposed an algorithm that learns *tag relevancy* by accumulating votes from visually similar neighbors. In fact, given a user-tagged image, they first perform a  $k$ NN search to find its visual neighbors. The tag relevance is determined as the probability that this tag being

used to annotate the neighborhood images minus the probability of the tag being used in the entire collection. [Sun and Bhowmick, 2009] propose a method to calculate the *Normalized Image Tag Clarity* score that evaluates the effectiveness of a tag in describing the visual content of its annotated images. It is measured by computing the zero-mean normalized distance between the tag language model estimated from images annotated by this tag and the collection language model. [Zhuang and Hoi, 2011] propose a two-view learning approach to discover the relationship between tags and images by exploiting both textual and visual contents of social images. The tag ranking task is formulated as a problem of learning a tag weighting matrix that encodes the relevance relationship between images and tags. Similarly, [Li et al., 2012] propose a two-view learning approach for tag ranking scheme through a two-stage graph-based relevance propagation approach. The first stage builds a tag graph on each image and implements a random walk process on it in order to get the initial relevance of each tag for one image and the second stage builds a kNN-sparse image graph and propagates the relevance of tags among the web images. [Sun et al., 2013] propose a tag ranking scheme to automatically rank tags with respect to given images by taking into account both the irrelevance to the image visual content and their relationships. First, given a tag query, a set of web images is collected from multiple searching engines to cover the semantic space. Second, initial relevance scores of tags with respect to a given image visual content are estimated in a Bayesian framework, in which a fused visual similarity is adopted. Third, tag graph is build by mining the relationship among tags.

### 2.4.1.3 Discussion

Most of tag ranking approaches assume that tags are noisy and the quality of tags associated with images is still far from satisfactory. The common goal of current work is to re-order the tag position (tag rank) according to the relatedness between each tag and an image. As highlighted in Chapter 1, tagging is not controlled, thus some initially assigned tags on the image may not be relevant to the image visual content. Therefore, the performance of tag ranking approaches using initially assigned tags without handling tag imperfections may not be satisfactory. Statistical-based approaches achieve good results but suffer from the drawback that the learning and the building of tag graph must be applied periodically as new images and tags are added, which is somehow impractical in large-scale evolving collections such as the case of Flickr collections. Moreover, most of these approaches usually ignore the tag imperfection issue. Data-driven approaches have shown to be easier to apply than statistical-based approaches. Similarly to statistical-based approaches, most of the Data-driven approaches do not take into account tag imperfections while collecting tags from visually nearest neighbors. In fact, as introduced in Chapter 1, the original tags associated

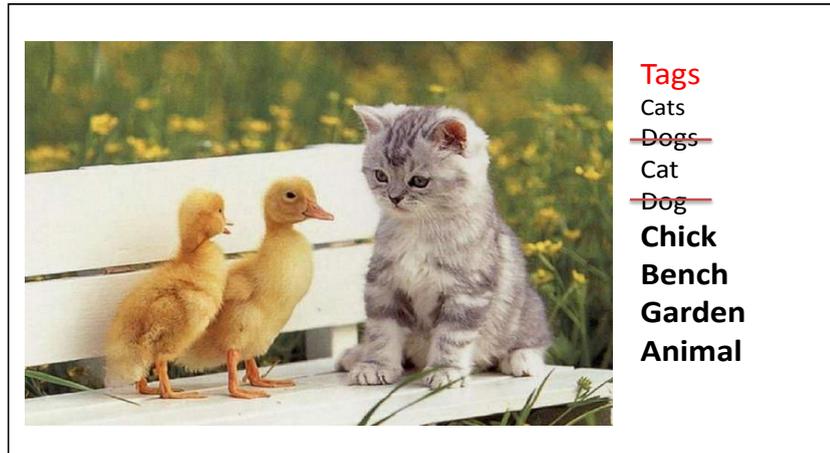


FIGURE 2.12: An example of tag refinement and suggestion: some tags are not related to the visual content of the image such as “dogs, dog”, some other are missing such as “bench, chick, animal, garden” and should be added (bold).

with images in social media websites are expected to be incomplete. Based on such noisy and incomplete tags, existing approaches can hardly acquire satisfying results. Thus, handling tag imperfections both for tag correlation and neighbor voting steps seems to be interesting and crucial to improve tag ranking performances.

## 2.4.2 Tag Refinement & Suggestion

Tag refinement and enrichment methods are proposed to refine unreliable human-provided tags by dropping inappropriate tags and adding new missing ones. These approaches deal with tag imperfection problem, especially tag incompleteness. An example in Figure 2.12 illustrates the tag refinement and suggestion task. We can see that some tags such as “*dogs, dog*” are not related to the image content, however, other are missing such as “*chick, bench, garden, animal*”. Numerous research efforts have been devoted to solve such a problem. As suggested by [Liu et al., 2011a], these approaches can be distinguished into two broad categories: *Statistical modeling* and *Data-driven* approaches.

### 2.4.2.1 Statistical modeling approaches

These approaches are based on the statistics of tags (tag co-occurrence) from knowledge resources or data collections. As a pioneering work, [Jin et al., 2005] used WordNet [Fellbaum, 1998] to estimate the semantic correlation among the annotated keywords and then those weakly-correlated ones are removed. To identify irrelevant keywords, they investigate various semantic similarity measures between keywords and fuse outcomes of all these measures together to make a final decision

using Dempster-Shafer evidence combination. [Weinberger et al., 2008] proposed a method that measures *tag ambiguity*, and suggests new tags that best reduce this ambiguity. This ambiguity is based on the co-occurrence of a tag set in two different contexts. [Xu et al., 2009] propose to solve the problem of tag refinement from the angle of topic modeling and present a novel graphical model, regularized Latent Dirichlet Allocation (rLDA). Both tag similarity and tag relevance are jointly estimated in an iterative manner, so that they can benefit from each other, and the multi-wise relationships among tags are explored. Moreover, both the statistics of tags and visual affinities of images in the corpus are explored to help topic modeling. However, similar to other topic-based analysis approaches, iterative estimation of probabilities for topic modeling requires high computational costs. [Wang et al., 2010a] propose an approach to build Semantic Fields for annotating the web images. The main idea is that images are more likely to be relevant to a given concept, if several tags of the image belong to the same Semantic Field as the target concept. Semantic Fields are determined by a set of highly semantically associated terms with high tag co-occurrences in the image corpus and in different corpora and lexica such as WordNet and Wikipedia.

#### 2.4.2.2 Data-driven approaches

These approaches rely on visually similar images from external data collections in order to refine initial tag list or to suggest new ones. [Liu et al., 2009b] propose a scheme to improve poorly annotated tags associated with social images. Two properties are exploited and integrated in a unified optimization framework: (1) consistency between visual and semantic similarities, where the semantic similarity is estimated using tags; (2) compatibility of tags before and after improvement, since the initial user provided tags carry valuable information. This is posed as an optimization problem and an iterative bound method is derived to solve it. [Kennedy et al., 2009] propose a framework for gathering reliable image tags. In this work, authors consider as reliable tags that are related to the image visual content. They leverage a large database of tagged photographs and discover pairs of visually similar images. [Tang et al., 2009] propose a tag refinement strategy within a graph based learning framework to handle the noise in tags, by bringing in a dual regularization for both the quantity and sparsity of the noise. In addition, a compact concept space with small semantic gap to infer the semantic concepts is constructed. The sparse graph is build by datum-wise. A one-vs-all sparse reconstructions of all samples can remove most of the concept-unrelated links among data. [Liu et al., 2010] propose a social image “retagging” scheme that aims at assigning images with better content descriptors. The refining process, including denoising and enriching, is formulated as an optimization framework based on the consistency between visual similarity and semantic similarity in social images, that is, the visually similar images tend to have similar semantic descriptors, and vice

versa. In the same way, [Zhu et al., 2010] propose a tag refinement approach which is referred as low-rank and error sparsity approximation. This method is based on several assumptions: visually similar images are similarly tagged; tags are often correlated and interact at the semantic level; the semantic space spanned by all tags can be approximated by a smaller subset of them; user tags are sufficiently accurate so that the image tag matrix has error sparsity condition. Following these assumptions tag refinement was cast into the problem of decomposing the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix. [Yang et al., 2011] propose an automatic scheme called *tag tagging* to supplement semantic image descriptions by associating a group of property tags with each existing tag. The tagging scheme mainly consists of two steps: tag to region and property tag generation. Tag to region consists in finding each tag corresponding image region through lazy diverse density. Property tag generation consists in deriving property tags based on the image regions found in the first step. Recently, [Wu et al., 2012] proposed a framework for tag refinement and suggestion. They represent the image-tag relation by a tag matrix, and search for the optimal tag matrix consistent with both the observed tags and the pairwise visual similarity between images. This optimization problem is solved using a sub-gradient descent based approach. Although this approach shows to give good results, it still suffer from the computational complexity due to the matrix optimization step and thus the scalability to large datasets can be a problem.

TABLE 2.5: A summary of terms used to describe tag imperfections in representative work in the context of social media.

The work	Terms used for tag imperfections
[Jin et al., 2005]	Noisy
[Weinberger et al., 2008]	Noisy, Ambiguous
[Xu et al., 2009]	Noisy, Ambiguous
[Wang et al., 2010a]	Noisy, Ambiguous, Incomplete
[Liu et al., 2009b]	Noisy, Imprecise, Incomplete
[Kennedy et al., 2009]	Noisy, Unreliable
[Tang et al., 2009]	Noisy, Incomplete, Incorrect
[Liu et al., 2010]	Imprecise, Biased, Incomplete
[Zhu et al., 2010]	Noisy
[Yang et al., 2011]	Noisy, Ambiguous
[Wu et al., 2012]	Noisy, Unreliable, Inconsistent, Incomplete

### 2.4.2.3 Discussion

There has been a rich state-of-the-art on Tag refinement and suggestion. These approaches have been proposed to solve the problem of noisy and incomplete tags. The tag incompleteness issue in social media tagging is almost well identified in the literature [Liu et al., 2009b; Tang et al., 2009; Wang et al., 2010a]. However, there is no precise identification and definition of noisy tags. Table 2.5 presents a summary of the most used terms in the literature that try to handle tag imperfections. In fact, this notion of noise covers many aspects. Some authors consider as noisy tags those who are *ambiguous* [Weinberger et al., 2008; Xu et al., 2009; Wang et al., 2010a; Yang et al., 2011], due to the well-known polysemy and synonymy nature of words (tags). To the best of our knowledge, [Weinberger et al., 2008] was the only work where authors define ambiguity and underly the intuition that “a tag set is ambiguous if it can appear in at least two different tag contexts. These could be defined by geographic locations, word senses, languages or temporal events, etc.”. Other work consider as noisy, tags that are *unreliable* [Kennedy et al., 2009; Yang et al., 2011; Wu et al., 2012]. [Kennedy et al., 2009] define unreliable tags as “tags that are not related to the image visual content”. To describe noisy tags other terms have been used in the literature such as *Imprecise*, *Biased* and *Inconsistent*, however there are no explicit definitions of these notions. To handle such imperfections some definitions need to be stated clearly.

## 2.5 Handling Imperfections at the decision level

It has been theoretically and empirically demonstrated that combining multiple classifiers can substantially improve the classification performances. In the context of social media, the use of multiple classifiers trained on different modalities and several types of features usually leads to better performances in image annotation task, due to the complementarity of the classification models [Guillaumin et al., 2010; Kawanabe et al., 2011; Duin, 2002]. In this thesis, we consider the problem of classifier combination as an information fusion process where predictions from different classifiers can be viewed as information sources to be combined to make a final decision. Thus, a particular attention must be paid to the fusion process, in order to take advantage of the complementarity while minimizing potential issues due to the conflict between the sources of information (different classifier predictions).

In the literature, different frameworks exist for reasoning with imperfections of information at the decision level. In this section, as highlighted in [Bellenger, 2013], we list the most well-known ones: the *Probability* theory, the *Fuzzy-Set* theory, the *Possibility* theory and the *Dempster-Shafer* theory. Other theories dealing with

such imperfections can be mentioned such as Imprecise Probability theory [Walley, 1991] or Rough set theory [Pawlak et al., 1995]. In the following, we present only the principle of each theory and its advantages as well as its limitations. In this thesis, we choose to use the Dempster-Shafer theory as a formalism in order to deal with imperfections in multimodal image annotation. Thus, Section 2.5.4.1 is devoted to give more in depth details about the fundamentals of the Dempster-Shafer theory.

### 2.5.1 Probability theory

The Probability theory is surely the most well known mathematical theory dealing with imperfections such as uncertainty. Input data are modeled using probabilities or likelihood numbers which allow to model measurement of the uncertainties. More precisely, as highlighted by [Dubois, 2007], the roles of probabilities are two folds. On one hand, through repeated observations, probabilities are capturing variability and randomness. Thus, probabilities can be considered as objective quantities that can be interpreted as frequencies. On the other hand, probabilities are considered as subjective quantities that have to be interpreted as degrees of belief. Let us recall that often degrees of belief in classifier combination are provided by classifiers learned on imperfect data and thus may be erroneous. A major drawback of the probability theory resides in the requirement of a perfect knowledge of the probabilities and especially the apriori probabilities as stated in [Bellenger, 2013]. Unfortunately, when knowledge on the problem is imperfect which is the case of different classifier predictions, probabilities can not be estimated correctly.

At the core of the probability theory lies the “Bayesian fusion” which uses the Bayes rule to combine decisions from different classifiers. The Bayesian inference fusion method is briefly described as follows. Let  $(s_1, s_2, \dots, s_n)$  a set of decisions scores obtained from  $n$  different classifiers to be combined. Assuming that these classifiers are statistically independent, the joint probability of an hypothesis  $A$  based on the fused decisions can be computed as suggested in [Papandreou et al., 2009]:

$$p(A|s_1, s_2, \dots, s_n) = \frac{1}{N} \prod_{k=1}^n p(s_k|A)^{w_k} \quad (2.8)$$

where  $N$  is used to normalize the posterior probability estimate  $p(A|s_1, s_2, \dots, s_n)$ . The term  $w_k$  is the weight of the  $k^{th}$  classifier, and  $\sum_{j=1}^n w_j = 1$ . This posterior probability is computed for all the possible hypotheses,  $\Omega$ . The hypothesis that has the maximum probability is determined using the maximum aposteriori probability rule defined as follows:

$$\hat{A} = \operatorname{argmax}_{A \in \Omega} p(A|s_1, s_2, \dots, s_n) \quad (2.9)$$

Bayesian fusion method has been successfully used to fuse multimodal information at the decision level for performing various multimedia tasks. For instance, [Meyer et al., 2004] and [Xu and Chua, 2006] have used the Bayesian inference method, respectively, for spoken digit recognition and sports video analysis.

## 2.5.2 Fuzzy Set theory

The Fuzzy set theory has been introduced by [Zadeh, 1965] as an extension of the classic notion of sets. It allows the representation and the gradual assessment of truth about vague information. In classic set theory, the membership of elements in a set is assigned in a binary way (1 or 0), i.e. it belongs or not to the considered set. By extension, Fuzzy set theory allows the fuzzy assessment of the membership of an element in a set through a membership function valued in the real unit interval  $[0, 1]$ . As in the Probability theory, the degree of truth is a value between 0 and 1. However, the degree of truth in Fuzzy sets represents a membership of elements in vaguely defined set, whereas a probability represents the likelihood of the membership itself. Formally, a fuzzy set  $F \subseteq \Omega$  is defined by the gradual membership function  $\mu_F(A)$  in the interval  $[0, 1]$  as follows:

$$\mu_F(A) \in [0, 1] \quad \forall \quad A \in \Omega \quad (2.10)$$

where the higher the membership degree is, the more  $A$  belongs to  $F$ . Fuzzy information can be combined using fuzzy rules to produce fuzzy fusion outputs. Examples of fusion rules for two fuzzy sets  $F_1$  and  $F_2$ , are the following:

$$\mu_{F_1, F_2}(A) = \min[\mu_{F_1}(A), \mu_{F_2}(A)] \quad \forall \quad A \in \Omega \quad (2.11)$$

$$\mu_{F_1, F_2}(A) = \max[\mu_{F_1}(A), \mu_{F_2}(A)] \quad \forall \quad A \in \Omega \quad (2.12)$$

A general framework for combining information from several individual classifiers for the classification of urban remote sensing images have been proposed by [Fauvel et al., 2006]. It is based on the definition of two measures of accuracy. The first one is a point-wise measure which estimates for each pixel the reliability of the information provided by each classifier. By modeling the output of a classifier as a fuzzy set, this point-wise reliability is defined as the degree of the fuzzy set uncertainty. The second measure estimates the global accuracy of each classifier. Finally, the results are aggregated with an adaptive fuzzy operator rule by these two accuracy measures. [Fakhar et al., 2012] propose a multi-biometric identification system based on fusion at the decision level using fuzzy set theory. The fusion system is based on face and iris modalities where output classifiers from each modality are modeled as a fuzzy set.

### 2.5.3 Possibility theory

The Possibility theory has been introduced by [Zadeh, 1978] as a generalization of the Fuzzy Set theory. The aim was to enable the management of imperfections by defining the concept of a possibility distribution as a fuzzy restriction which acts as an elastic constraint on the values that may be assigned to a variable [Belenger, 2013]. It does not model a degree of belief or truth, but rather the reference we have for a hypothesis. Contrary to the probability theory, which associates a unique probability to each statement, in Possibility theory we have a possibility distribution  $\pi_B(A) \in [0, 1] \forall A \in \Omega$ . This possibility distribution characterizes the uncertain membership of an element  $A$  in a well-defined class  $B$ . Given the possibility distribution  $\pi(U)$ , the possibility measure  $\Pi(U)$  and the necessity measure  $N(U)$  of an event  $U$  are defined as follows:

$$\Pi(U) = \max_{A \in U} \{\pi_B(A)\} \quad \forall U \subseteq \Omega \quad (2.13)$$

$$N(U) = \min_{A \in U} \{1 - \pi_B(A)\} \quad \forall U \subseteq \Omega \quad (2.14)$$

As stated in [Khaleghi et al., 2013], a possibility degree  $\Pi(U)$  quantifies to what extent the event  $U$  is plausible, while the necessity measure  $N(U)$  quantifies the uncertainty of  $U$ , in the face of incomplete information expressed by a possibility distribution  $\pi(A)$  [Destercke et al., 2009]. The possibility and the necessity measures can be also interpreted as a special case of upper and lower probabilities, in connection with the probability theory [Dubois and Prade, 1992]. Although possibility theory has not been commonly used in data fusion applications, its performance has been compared to probabilistic and evidential fusion approaches for active object recognition [Borotschnig et al., 1999]. Possibilistic fusion is argued to be most appropriate in poorly informed environments as well as in fusion of heterogeneous data sources [Dubois and Prade, 1994]. [Oussalah et al., 2001] propose a fusion framework based on the possibility theory applied to a robotic application. This latter deals with a mobile robot equipped with ultrasonic sensors and odometry whose measurements are combined using the possibility theory.

### 2.5.4 Dempster-Shafer theory

The Dempster-Shafer (DS) theory has been introduced by [Shafer, 1976] taking support on the work made by [Dempster, 1967]. The Dempster-Shafer theory, also known as Evidential theory or Belief theory, offers a theoretical framework for modeling uncertainty, and provides a method for combining distinct items of evidence collected from different sources. It is based on two ideas: obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on

independent items of evidence. This evidence combination rule provides an interesting operator to integrate multiple pieces of evidence from different sources. Thus, it is very useful when combining multiple pieces of information that come from different classifiers, as in classifier combination task.

The advantage of Dempster-Shafer theory is that it allows coping with absence of preference, due to limitations of the available information. The theory is often viewed as a generalization of the Probability theory, by providing a coherent representation for ignorance (lack of evidence) and also by discarding the insufficient reasoning principle. However, these two approaches differ significantly and the extent of their applicability to data fusion is still being debated [Braun, 2000]. The Probability theory is based on the classical ideas of probability, while Dempster-Shafer theory allows more interpretation of what uncertainty is all about. One of the major advantages of the Dempster-Shafer theory over probability is thus to allow one to specify a degree of ignorance in a situation instead of being forced to supply prior probabilities. Moreover, probabilistic approaches reason only on singletons while Dempster-shafer theory enables not only to affect belief on elementary hypothesis but also on composite ones. The Dempster-Shafer theory contains two new ideas that are foreign to the Probability theory. These are the notions of *belief* and *plausibility*.

As pointed by [Dubois, 2007], the Dempster-Shafer theory can be considered as an extension of both Probability theory, Possibility and the Sets theory. As a matter of fact, DS theory includes extensions of probabilistic notions (conditioning), of the possibility theory and the set-theoretic notions (intersection, union ...).

Considering the advantages presented above, we have chosen to rely on the Dempster-Shafer theory of evidence as a framework to handle imperfections for image annotation. The following section is devoted to give more in depth details about the fundamentals of the Dempster-Shafer theory. In Dempster-Shafer theory, evidence is represented in terms of evidential functions and ignorance. These functions include mass functions (or basic belief assignment function), belief and plausibility functions [Shafer, 1976].

#### 2.5.4.1 Belief functions on finite domains

In Dempster-Shafer (DS) theory [Shafer, 1976], a *frame of discernment*  $\Omega$  is defined as the set of all hypothesis in a certain domain. A basic belief assignment (BBA), called also mass function, is a function  $m$  that defines the mapping from the power set of  $\Omega$  to the interval  $[0, 1]$  and verifies:

$$m : 2^\Omega \rightarrow [0, 1] \quad (2.15)$$

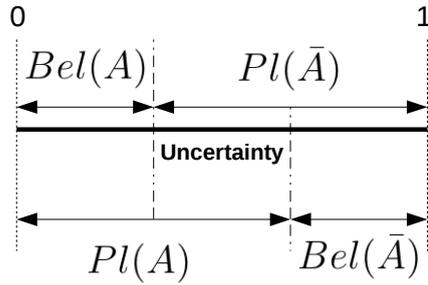


FIGURE 2.13: Relation of belief and plausibility and their negation.

$$\sum_{A \in \Omega} m(A) = 1, \quad m(\emptyset) = 0 \quad (2.16)$$

The quantity  $m(A)$  can be interpreted as a measure of the belief that is committed *exactly* to  $A$ , given the available evidence. A subset  $A \in \Omega$  with  $m(A) > 0$  is called a *focal element* of  $m$ . The major difficulty relies on assigning belief masses to each hypothesis. The objective is to model expert opinions (in our case classifier predictions) using belief functions. Most of existing modeling depends on the considered application.

In DS theory, two functions of evidence can be deduced from  $m$  and its associated focal elements, *belief* function  $Bel$  and *plausibility* function  $Pl$ .

The belief function, called also credibility, is defined as a mapping  $Bel : 2^\Omega \rightarrow [0, 1]$  that satisfies  $Bel(\emptyset) = 0$ ,  $Bel(\Omega) = 1$  and for each focal element  $A$ , we have:

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \quad (2.17)$$

$Bel(A)$  represents the measure of the *total* belief committed to a set  $A$ . The *plausibility* of  $A$ ,  $Pl(A)$ , represents the amounts of belief that could *potentially* be placed in  $A$  and defined as: The *plausibility* of  $A$ ,  $Pl(A)$ , represents the amounts of belief that could *potentially* be placed in  $A$  and defined as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \quad (2.18)$$

The duality between belief and plausibility is depicted through the following equation and graphically illustrated in Figure 2.13:

$$Pl(A) = 1 - Bel(\bar{A}) \quad \forall A \subseteq \Omega \quad (2.19)$$

The difference  $Pl(A) - Bel(A)$  quantifies the uncertainty about a specific hypothesis  $A$ . The belief can be considered as a kind of loose lower limit to the uncertainty. On the other hand, the plausibility is viewed as a loose upper limit to the uncertainty.

### 2.5.4.2 Combination process

The Dempster-Shafer theory offers a framework to combine different items of evidence from different sources. We propose in the following the major rules developed and used in the fusion community working with belief functions. We refer the reader to the survey of [Martin et al., 2008] for other combination rules.

#### Unnormalized Dempster's combination rule

Let  $m_1$  and  $m_2$  be two mass functions on  $\Omega$  induced by two independent items of evidence. We denote  $m_{1-2} = m_1 \oplus m_2$ , the combined mass distribution issued from the combination of the two distributions  $m_1$  and  $m_2$  and defined as follows:

$$m_{1-2}(A) = m_1 \oplus m_2 = \sum_{B \cap C = A} m_1(B)m_2(C) \quad (2.20)$$

#### Normalized Dempster's combination rule

The mass functions  $m_1$  and  $m_2$  are combined under the normalized Dempster's combination rule [Shafer, 1976] as follows:

$$m_{1-2}(A) = m_1 \oplus m_2 = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{1-k}, & \forall A \subseteq \Omega, A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \quad (2.21)$$

where  $k = \sum_{B \cap C = \emptyset} m_1(B)m_2(C)$  represents the degree of conflict between the two sources. If  $k$  is close to 0, the two sets of masses are not in conflict, while if  $k$  is close to 1, they are almost in conflict and they can not be combined.

### 2.5.4.3 Decision making

Main decision processes are through the maximum of credibility or the maximum of plausibility. In the first case, we choose the hypothesis whose credibility (belief) is the higher. In other words, we choose subsets whose implications for this subset are maximal. In the second case, we choose the hypothesis with the highest plausibility. That's to say, we choose the subset that contradicts the less the whole available information.

Another possibility is to choose the hypothesis with the maximum of pignistic probability proposed by [Smets, 1989]. Let  $m$  be a mass function, its pignistic probability distribution is defined as follows:

$$P_m(A) = \sum_{\emptyset \neq B \subseteq \Omega} m(B) \frac{|A \cup B|}{B} \quad \forall A \subseteq \Omega \quad (2.22)$$

#### 2.5.4.4 Discussion

The Dempster-Shafer (DS) theory provides an interesting and useful computational scheme for representing and integrating (or fusing) uncertain information. DS theory has been widely used in many applications, e.g., information fusion and pattern recognition [Atrey et al., 2010]. However, high computational cost of evidence combination is a drawback which is often raised against the Dempster-Shafer theory. It is well known that the computational cost of evidence combination increases exponentially with respect to cardinality of the frame of discernment. Unfortunately, this is precisely the case one must handle for the considered multimedia collections. To encounter this limitation, authors in [Younes et al., 2009] proposed a method to reduce the complexity of manipulating and combining mass functions, when belief functions are defined over a suitable subset of the frame of discernment equipped with a lattice structure. This approach was applied for multi-label classification based on the Evidential KNN classifier [Denoeux, 1995]. For a problem with  $\mathbf{k}$  classes, this method reduces the complexity from  $2^{2^{\mathbf{k}}}$  to  $3^{\mathbf{k}} + 1$ . Although such a reduction is impressive, the problem remains intractable when  $k$  is above 10, that is quite common for a multimedia classification problem, for which  $\mathbf{k}$  can reach 100 or 1000. To the best of our knowledge, there is no attempt to apply Dempster theory for a multimodal image annotation in the context of social media for a large dataset ( $\approx 20k$  images) and a large variety of categories simultaneously (scene, event, objects, image quality and emotions  $\approx 99$  classes).

## 2.6 Image Databases & Evaluation Campaigns

In this section, we describe the datasets used to evaluate the effectiveness and the robustness of the proposed approaches for multimodal image annotation in the context of social media. We employ real-world social images with human annotated tags. Images and their associated user tags are downloaded from the photo sharing website Flickr<sup>2</sup>. All the datasets used for evaluation, except NUS-WIDE dataset [Chua et al., 2009], are created within photo annotation challenges in evaluation campaigns.

### 2.6.1 Evaluation Campaigns

Often, multimedia annotation systems are evaluated on different test collections with different performance measures, which makes the comparison to state-of-the-art approaches limited. Benchmarking campaigns counteract these tendencies and

---

<sup>2</sup><http://www.flickr.com>

establish an objective comparison among the performance of different approaches by posing challenging tasks and by distributing test collections and measures. In this thesis, we focus on multimodal image annotation. In the following, we present in details two evaluation campaigns for image annotation: **ImageCLEF** [Clough et al., 2010] and **PASCAL VOC** [Everingham et al., 2010].

- Image Retrieval in Cross-Language Evaluation Forum (**ImageCLEF**) [Clough et al., 2010] is an initiative for evaluating cross-language image retrieval systems in a standardized manner. It was launched for the first time in 2003 as part of the Cross-Language Evaluation Forum (CLEF). The main goal of ImageCLEF is to support the advances of the field of visual media analysis, indexing, classification, and retrieval, by developing the necessary infrastructure for the evaluation of visual information retrieval systems operating in both monolingual, cross-language and language-independent contexts. A major outcome of ImageCLEF has been the creation of a number of publicly accessible evaluation resources. These benchmarks have helped researchers to develop new approaches to visual information retrieval and automatic annotation by enabling the performance of various approaches to be assessed. Tasks and datasets used in ImageCLEF changed over the years while the objectives broadly remained the same:
  - To investigate the effectiveness of combining textual and visual features for crosslingual image retrieval.
  - To collect and provide resources for benchmarking image retrieval systems.
  - To promote the exchange of ideas to help improve the performance of future image retrieval systems.

To meet these objectives a number of tasks have been organized by ImageCLEF within two main domains: (1) medical image retrieval and (2) non medical image retrieval, including historical archives, news photographic collections and Wikipedia pages. Broadly speaking the tasks fell within the following categories: ad-hoc retrieval, object and concept recognition, and interactive image retrieval.

- **Ad-hoc retrieval.** This simulates a classic document retrieval task: given a query describing an user information need, find as many relevant documents as possible and rank the results by relevance. In the case of cross-lingual retrieval the language of the query is different from the language of the metadata used to describe the image.
- **Object and concept recognition.** Although ad-hoc retrieval is a core image retrieval task, a common precursor is to identify whether

an object is contained in an image (object recognition), assign labels to an image (automatic image annotation) or classify images into one or many classes (image classification). In this thesis, we are specifically interested in the ***Visual Concept Detection Task*** (VCDT) which is a multi-label classification challenge. It aims at the automatic annotation of a large number of images with multiple annotations.

- **Interactive image retrieval.** Image retrieval systems are commonly used by people interacting with them. Interaction in image retrieval can be studied with respect to how effectively the system supports users with query formulation, query translation (in the case of crosslingual IR), document selection and document examination.

A major contribution of ImageCLEF has been to collect a variety of datasets for use in the different tasks. Since 2003, ImageCLEF has created (and/or acquired and adapted) almost a dozen document collections to support its various evaluation tasks. Some collections are freely available for download from the ImageCLEF website<sup>3</sup>, while others are subject to signing an end-user agreement with the task organizers and/or original copyright holders. We describe, in Section 2.6.2, the collections which are used to evaluate our methods.

- The Pattern Analysis Statistical Modeling and Computational Learning Visual Object Classes (**PASCAL VOC**) challenge [Everingham et al., 2010] is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotations, and standard evaluation procedures. Organized annually from 2005 to present, the challenge and its associated dataset has become accepted as the benchmark for object detection. The PASCAL<sup>4</sup> Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available dataset of images and annotation, together with standardized evaluation software; and (ii) an annual competition and workshop.

## 2.6.2 Image Databases

Specifically, five publicly available Flickr image datasets are used to evaluate the proposed approaches. Although all these datasets are collected from Flickr website, they differ significantly based on many criterion summarized in Table 2.6. In the following, we present a brief description of each dataset considered in this dissertation.

---

<sup>3</sup><http://www.imageclef.org/>

<sup>4</sup>PASCAL stands for pattern analysis, statistical modeling and computational learning. It is an EU Network of Excellence funded under the IST Program of the European Union.

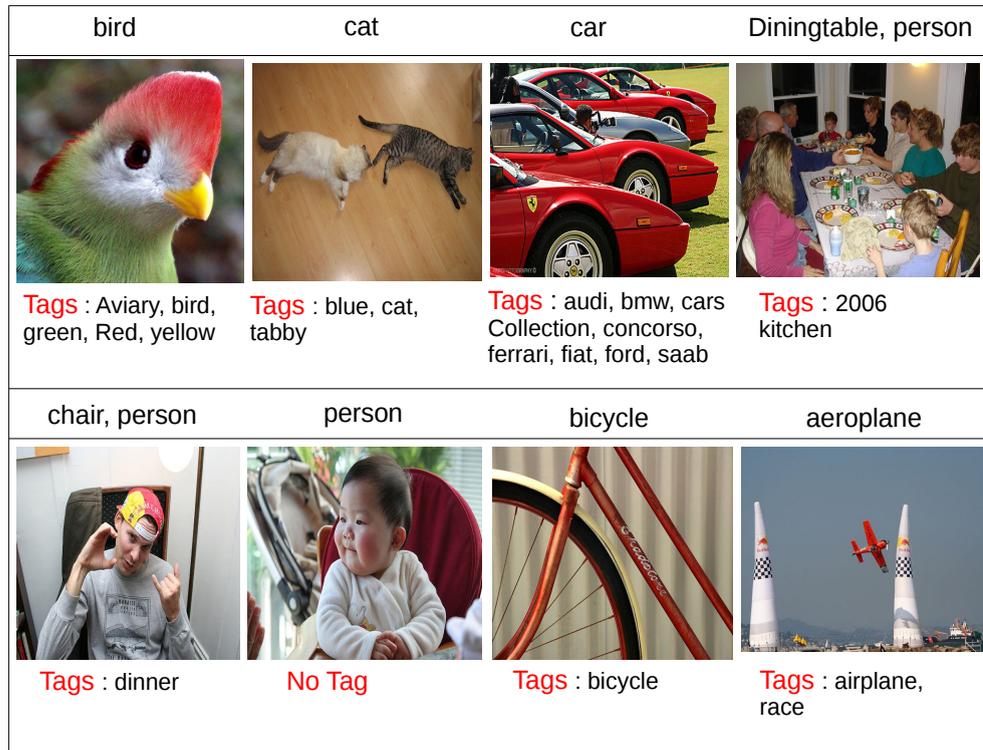


FIGURE 2.14: PASCAL VOC'07 dataset example images with their associated user tags (below) and labels (on top).

- The **PASCAL VOC'07** dataset contains around 10,000 images annotated according to 20 concepts. These concepts describe vehicles (car, bus, bicycle...), animals (cat, dog, horse...), household (sofa, tv/monitor, chair ...) and persons. In the PASCAL VOC challenge, this dataset is not multimodal and only images are available. [Guillaumin et al., 2010] adapt the PASCAL VOC'07 dataset to be used in multimodal image annotation. Using the image identifiers, they downloaded the user tags for the 9,587 images that were still available on Flickr at time of download, and assumed complete absence of tags for the remaining ones. By keeping the tags that appear at least 8 times (a minimum of 4 times in the training and test sets), the dataset results with a list of 804 unique tags. Example images with their associated user tags and class labels are given in Figure 2.14.
- **ImageClef'10**<sup>5</sup> is used to refer to the subset of the MIR-Flickr that was used within the ImageCLEF 2010 photo annotation challenge [Nowak and Huiskes, 2010]. The dataset consists of 8,000 images for training and 10,000 for testing belonging to 93 concepts. The concepts describe the scene (indoor, outdoor, landscape, mountains ...), objects (dog, car, animal, person, building..), event (holidays, sport, work ...) and image quality (overexposed, underexposed, blurry).

<sup>5</sup>This dataset is available at: [http://www.idmt.fraunhofer.de/de/projects/expired\\_publicly\\_financed\\_research\\_projects/photo\\_annotation.html#tabpanel-4](http://www.idmt.fraunhofer.de/de/projects/expired_publicly_financed_research_projects/photo_annotation.html#tabpanel-4)

		
<b>Tags</b> : Sports, dodgers, dodger Stadium, Coloradorockies, chavezravine, Baseball, russell, martin, troy, Tulowitzki,2ndbase	<b>Tags</b> : Dal, cinnamoroll	<b>No Tag</b>
Sports Outdoor Day Neutral_Illumination No_Blur Small_Group Cute Male Adult	Summer Outdoor Day Still_Life Neutral_Illumination No_Blur No_Persons Fancy cute	No_Visual_Season Neutral_Illumination Partly_Blurred No_Blur Single_Person No_Persons Aesthetic_Impression Overall_Quality Natural, cute
<b>Ground Truth Labels</b>		

FIGURE 2.15: ImageCLEF'11 dataset example images with their associated user tags and labels.

- **ImageClef'11** is the same dataset used within the ImageCLEF 2010 photo annotation challenge [Nowak et al., 2011] with a small difference is that images are annotated with 99 concepts. In 2011, nine novel sentiment concepts were added to the test collection (happy, funny, euphoric, nice, cute ..). In this collection there are 1,386 tags which occur at least in 20 images, with an average total number of 8.94 tags per image.
- **ImageClef'12** is used to refer to the subset of the MIR-Flickr that was used within the ImageCLEF 2012 photo annotation challenge [Thomee and Popescu, 2012]. It consists of 15,000 images for training and 10,000 for testing belonging to 94 concepts. The concepts are very diverse and range across categories such as people (e.g. teenager, female), scenery (e.g. lake, desert), weather (e.g. rainbow, fog) and even impressions (e.g. unpleasant, euphoric). The dataset contains 45,408 unique tags.

Concepts in ImageCLEF datasets are variable. They contain both well defined objects such as "lake, river, plants, trees, flowers", as well as many rather ambiguously defined concepts such as "winter, boring, architecture, macro, artificial, motion blur", however, those concepts might not always be connected to objects present in an image. This makes it highly challenging for any recognition system.

			
<b>Tags</b> : 365days, day316 Self-portrait, candy Close-up, mouth, a piece of me, christmas Holiday, christmas candy, red lipstick, green, white	<b>No Tag</b>	<b>Tags</b> : light, dark, sun, flower, gerbera, macro, yellow, orange, FlickrBest	<b>Tags</b> : London, uk, night, skyline, river, thames, lights, Bridge, Southwark Millenium, city, skyscraper, highrise England, longexposure, 15 secs, Flow, pier, Tourist, 5PhotosaDay, The Perfect Photographer COTC:mostinteresting
quantity_none quality_partialblur view_closeupmacro sentiment_happy sentiment_funny	quantity_none quality_partialblur	flora_flower quantity_none quality_noblur view_closeupmacro sentiment_calm	celestial_moon scape_city quantity_none quality_noblur setting_citylife
<b>Labels</b>			

FIGURE 2.16: ImageCLEF'12 dataset example images with their associated user tags and labels.

- The **NUS-WIDE** dataset includes 269,648 images crawled from Flickr, and about 425,000 unique original tags. Authors [Chua et al., 2009] of this dataset set several rules to filter the original tag set. They delete tags with too low frequency (number of occurrence in the dataset is less than a threshold). The low frequency threshold is set to 100. They also remove tags that does not exist in WordNet. At the end, they provide a list of 5,018 unique tags. Let's note that original tags are also available. Images in NUS-WIDE corpus are manually labeled to provide ground-truth for 81 concepts. This forms 161,789 images for training, and 107,859 images for testing. The 81 concepts are divided into six categories: people, objects, scene or location, event or activities, program and graphics.

A summary of dataset statistics is presented in Table 2.6. Considered datasets are very challenging because of the large variation on view size, illumination, scale, deformation and clutter, as well as complex backgrounds. As we can see, all images were collected from Flickr but they differ significantly. For instance, the number of classes and unique tags varies from a dataset to another. Moreover, PASCAL VOC'07 dataset contains only object classes (e.g. dog, car, sofa, train ...) while ImageClef and NUS-WIDE concepts are very diverse and range across different categories. In fact, ImageClef contains classes such as people (e.g. male, female), quality issues (e.g. overexposed, underexposed, blurry...), nature (e.g.

			
<b>Tags :</b> monochrome car vintage mercedes benz cilest kurt vehicle	<b>Tags :</b> dog husky wolf perro lobo roja caperucita siberiano vorfas	<b>Tags :</b> blue summer sky umbrella hoildays ysplix flickelite colorartaward platinumheartaward	<b>Tags :</b> ocean sunset sea water animals whale whales orca whalewatching whaletail orcawhale whalephotos
<b>Labels :</b> vehicle	<b>Labels :</b> dog	<b>Labels :</b> Sky	<b>Labels :</b> Ocean Sunset Water Whales

FIGURE 2.17: NUS-WIDE dataset example images with their associated user tags and labels.

lake, beach), weather (e.g. rainbow, fog) and even sentiments (e.g. unpleasant, euphoric).

## 2.7 Conclusions

In this chapter, we presented a survey of the state-of-the-art approaches on multimodal image annotation. Our aim was not to provide an exhaustive survey of the state-of-the-art approaches, nor to state that an approach is better than the others, but to introduce the different categories of approaches for image annotation in the context of social media. We underlined the benefits and limits of each of them. Indeed, handling imperfections in multimodal image annotation seems to be crucial to enhance annotation performances. Although, the problem of tag

TABLE 2.6: Dataset statistics: number of images (train/test), tags, labels and untagged images for both PASCAL VOC'07 and NUS-WIDE datasets.

	# images	# unique tags	# labels	# Untagged images
PASCAL VOC'07	5k/5k	804	20	3,764
NUS-WIDE	160k/100k	425k	81	0
ImageClef'10	8k/10k	21k	93	1,740
ImageClef'11	8k/10k	21k	99	1,740
ImageClef'12	15k/10k	45k	94	2,128

imperfection is known in community contributed collections, it still not well exploited to enhance tag-based applications. In fact, most of the state-of-the-art tag representations based on classic BOW representation do not take into account tag imperfections and fail to capture semantic tag relatedness. Moreover, even if a wealth of research has been proposed to enhance the quality of tags in other tasks such as tag ranking and refinement, there is no precise identification and definition of noisy tags. We believe that to handle such imperfections, some definitions need to be stated clearly. Most of approaches do not take into account explicitly imperfections at decision level while combining different classifier predictions. From the state-of-the-art theories that deal with imperfections in data fusion process, the Dempster-Shafer (DS) theory seems to be an interesting framework. It provides an interesting and rigorous computational and theoretical scheme for representing and integrating (or fusing) imperfect information. DS theory has been widely used in many applications, e.g., information fusion and pattern recognition [Atrey et al., 2010]. However, high computational cost of evidence combination is a drawback which is often raised against the Dempster-Shafer theory. It is well known that the computational cost of evidence combination increases exponentially with respect to cardinality of the frame of discernment. Unfortunately, this is precisely the case one must handle for the considered multimedia collections.

# Part I

## Representation Level

# Chapter 3

## Handling Textual Imperfections

### Contents

---

<b>3.1</b>	<b>Introduction</b>	<b>67</b>
<b>3.2</b>	<b>Textual Imperfections in the context of Multimedia Annotation</b>	<b>68</b>
<b>3.3</b>	<b>Semantic Similarity between Words</b>	<b>70</b>
3.3.1	Knowledge-Based Measures	71
3.3.2	Corpus-Based Measures	72
3.3.3	Discussion	76
<b>3.4</b>	<b>Problem Formalization</b>	<b>77</b>
<b>3.5</b>	<b>Soft Bag-of-Concepts Signature</b>	<b>80</b>
3.5.1	Tag Modeling	81
3.5.2	Coding/Pooling	81
<b>3.6</b>	<b>Local Soft Tag Coding Signature</b>	<b>83</b>
3.6.1	Tag Modeling	83
3.6.2	Coding/pooling	85
<b>3.7</b>	<b>Adopted Semantic Similarities</b>	<b>85</b>
<b>3.8</b>	<b>Experimental Evaluation</b>	<b>87</b>
3.8.1	Experimental Setup	88
3.8.2	Experimental Results	89
<b>3.9</b>	<b>Conclusion and Discussion</b>	<b>98</b>

---

## 3.1 Introduction

As introduced in Chapter 1, the textual information issued from the tag modality in the context of social media, represents an interesting source of information for semantic image annotation. However, many studies have shown that tags represent an interesting source of information for semantic image annotation but subject to many imperfections [Kennedy et al., 2006; Chua et al., 2009; Sigurbjörnsson and van Zwol, 2008; Cantador et al., 2011].

This thesis aims at annotating multimedia content and images in particular. Tags that will be considered as relevant are those that are directly related to the visual content for multimedia image annotation and others are considered as imperfect and noisy. Our first objective is to clearly identify and define these imperfections in the context of image annotation. Second, our goal is to handle these aspects at the representation level in order to enhance image annotation performances. In this context, we propose two novel signatures to handle such imperfections for tag-based image annotation. Both signatures are based on the BOW representation [Salton and McGill, 1983] with the same coding scheme. This latter consists in three steps: Tag modeling, feature coding and pooling. In order to build robust BOW based tag-signatures, we rely on the locality-constrained coding method [Liu et al., 2011b] that has proved to be effective for visual features when paired with max-pooling aggregation. These tag-based signatures have been published in [Znaidia et al., 2012b,d, 2013b]. We rely on semantic similarities to achieve the coding step for tag-based signature generation. Computing semantic similarities requires the use of external knowledge resources. In our work, we consider two knowledge resources: WordNet and Flickr to compute semantic similarities between words.

The rest of this chapter is organized as follows. In Section 3.2, we identify and define different tag imperfections in the context of content-based annotations of social media. A review of the state-of-the-art on semantic similarities between words is presented in Section 3.3. We introduce respectively, in Section 3.5 and Section 3.6, the Soft Bag-of-Concepts signature and the Local Soft Tag Coding signature, to handle tag imperfections and improve tag-based image annotation. Adopted semantic similarity measures for our models are detailed in Section 3.7. Section 3.8 reports our experimental results on several publicly datasets. The chapter is concluded in Section 3.9.

## 3.2 Textual Imperfections in the context of Multimedia Annotation

As introduced in Chapter 1, tags in online social media services, such as Flickr<sup>1</sup>, represent an important resource for facilitating multimedia information processing and management for future search and sharing. The main purpose of the users being to make their picture popular to the public, it conflicts with an objective description of image semantics [Ames and Naaman, 2007]. Consequently, only a few set of user generated tags are related to the image semantic visual content [Kennedy et al., 2006]. Indeed, as explained in Chapter 1, the motivations of tagging are multiple and can be classified according to the ‘**sociality**,’ and the ‘**function**,’ of the incentives for tagging images as shown in Figure 3.1. This is consistent with the categorization of tags presented in [Cantador et al., 2011]. Tags can be categorized into four categories: Content-based, Context-based, Subjective and Organizational. In our case, only tags that are Content-based are considered as relevant for semantic image annotation.

In order to identify and define tag imperfections, we present in Figure 3.2 an example of images from Flickr and their associated user tags. Let’s take the example in Figure 3.2(a). If tags such as “*bear, panda, baby, endangeredspecies*” are relevant to describe the image content, other tags like “*giant, precious*” are ambiguous. Meanwhile, several other tags that can be useful, such as “*tree, animal*”, are missing. In Figure 3.2(b), only the tag “*lotus*” actually describes the image visual content while others such as “*confucianism, buddha, breathtaking, taoism*” are noisy. Moreover, if we consider the concepts lexical variability and the hierarchy of semantic expressions, tags such as “*flower, leaf*” also need to be added. In Figure 3.2(c), tags such as “*religious, god, pray, jesus*” are noisy. In Figure 3.2(d), tags such as “*longexposure, efs1022, abigfave*” are actually related

<sup>1</sup><http://www.flickr.com/>

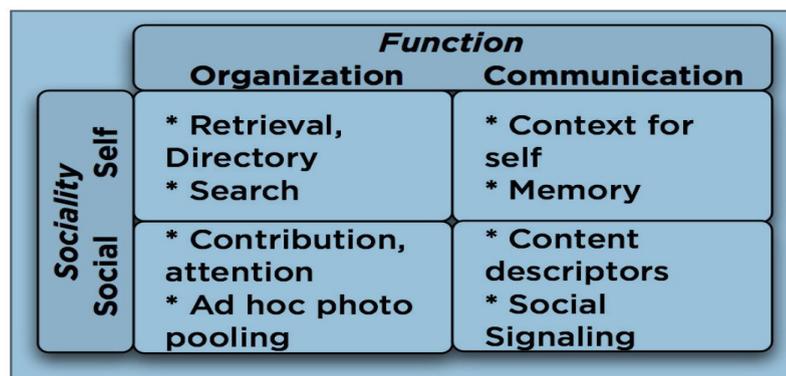


FIGURE 3.1: A taxonomy of tagging motivations in Flickr [Ames and Naaman, 2007].

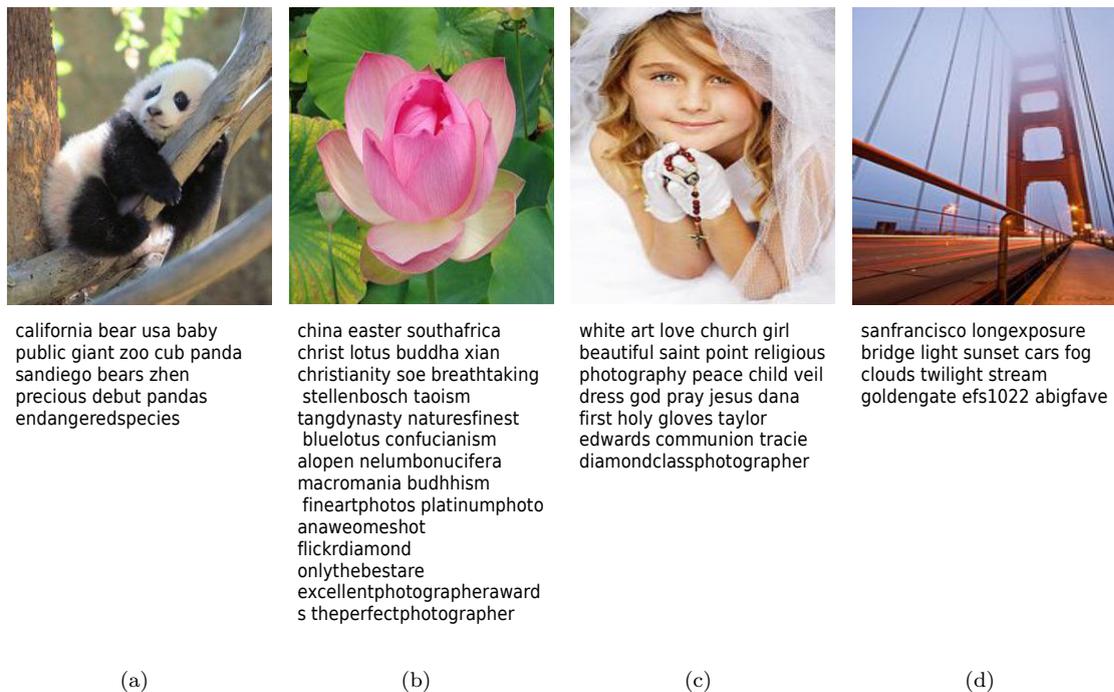


FIGURE 3.2: An example of images from Flickr website with their associated user tags. Most of tags are noisy and only few tags are related to image visual content.

to the user subjectivity. Based on consensual definitions of the different forms of information imperfections [Bloch, 2008], we identify and define three types of textual imperfections in the context of content-based image annotation:

- **Imprecision:** Imprecision is *related to the lack of knowledge and accuracy* on the available textual information. In our case, it corresponds to tags for which there is no precise definition and thus can be interpreted differently depending on the context where they appear. In the case of textual information, the imprecision is highly related to linguistic issues such as homonymy, synonymy, different lexical forms of alternate spellings, and misspellings of tags. For example, tags such as “*giant, precious*” in Figure 3.2(a) are imprecise.
- **Uncertainty:** Uncertainty is *related to the degree of truth of a piece of information*. In our case, it is related to the relevancy of a given tag to describe the image content. Indeed, many tags are irrelevant to describe the image content due to the motivation and the subjectivity of the users during the tagging process which leads to the problem of uncertainty about the relevance of the tag in describing the image visual content. For example tags such as “*sandiego*” in Figure 3.2(a) are uncertain.
- **Incompleteness:** Incompleteness is *related to the absence of a piece of information*. In our case, it corresponds to missing tags, i.e. tags that are

relevant to the visual content but does not appear in the user tag list. For example tags such as “*animal, tree*” are missing in Figure 3.2(a). We distinguish two types of incompleteness: partial and full. Partial incompleteness corresponds to the case where the image has some tags and others are missing while full incompleteness represents the case where the image has no tag. In this chapter, only the partial incompleteness is considered while the full incompleteness will be handled in Chapter 4.

To improve the accuracy of social media retrieval and management systems, these textual imperfections need to be taken into account. Imprecise and uncertain tags will introduce false positives into user’s search results and thus degrading precision and recall rates in tag-based applications, while incomplete tags will make the actually related images inaccessible.

In this chapter, we propose two novel tag-based signatures bearing such imperfections. Both proposed methods need a coding step of a given tag over a codebook which requires a word semantic similarity measure. Given two input words (tags or concepts), our objective is to automatically derive a score that indicates their similarity at a semantic level, thus going beyond the simple word matching method traditionally used in the classic BOW model.

### 3.3 Semantic Similarity between Words

Measures of semantic similarity between words are widely used in Natural Language Processing. Measuring the semantic similarity (or distance) between words is a process of quantifying the relatedness between the words and which often implies the use of background knowledge (external resources of information). These information sources can be: (i) lexical resources such as dictionaries, thesauri and semantic networks (e.g. WordNet); (ii) collections of documents such as corpus (e.g. Wikipedia), and (iii) the web.

In the literature on semantic similarity, some authors [Budanitsky and Hirst, 2006] emphasize a difference between measures of *semantic similarity* and measures of *semantic relatedness*. Semantic relatedness is a more general notion of the relatedness of concepts, while similarity is a special case of relatedness that is tied to the likeness of the concepts. Semantic relatedness refers to human judgments of the degree to which a given pair of concepts is related. For example, “cars” and “gasoline” would seem to be more closely related than, say, “cars” and “bicycles”, but the latter pair are certainly more similar because both are “wheeled vehicle”.

There exists a significant body of literature on semantic similarity measures [Meng et al., 2013; Panchenko, 2013]. Most approaches use an external source of information to derive a similarity score between words. Prior research suggests that

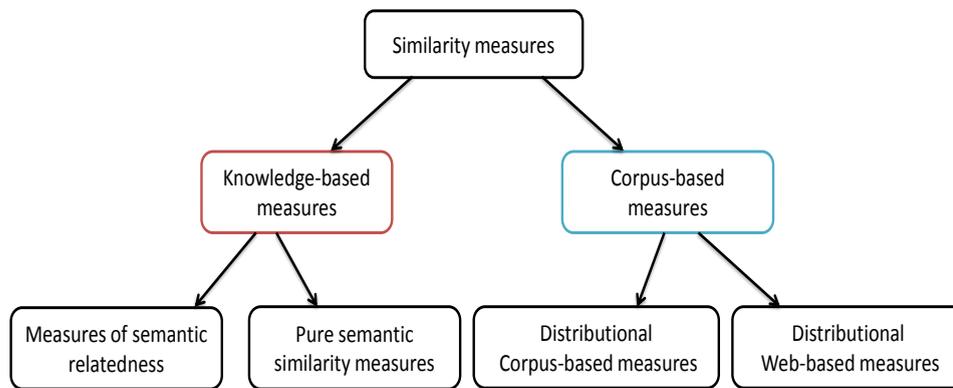


FIGURE 3.3: Categorization of the state-of-the-art methods on semantic similarity measures.

measures based on these sources of information are complementary [Heylen et al., 2008; Panchenko, 2013]. According to the type of the source of information, we categorize them into two groups: Knowledge-based and Corpus-based measures (presented in Figure 3.3).

### 3.3.1 Knowledge-Based Measures

Knowledge-based measures, also called thesaurus-based, use the structure of semantic networks in order to compute semantic similarity measures between words. An example of semantic network is WordNet, which is organized in such a way that synsets<sup>2</sup> and word-senses are the nodes of the network, and relations among the synsets and word-senses are the edges of the network. WordNet has been commonly used to measure semantic similarity among words since it has the inherent advantages of being structured in the way of simulating human recognition behaviors [Fellbaum, 1998]. On the whole, similarity measures based on WordNet can be grouped into four categories: Path length based, Information Content based, Feature based and Hybrid based measures as presented in Table 3.1. Based on the difference between *similarity* and *relatedness*, knowledge based measures are categorized into two groups: measures of semantic relatedness and Pure semantic similarity measures. The first group includes Feature based measures while the second includes the rest of semantic similarity measures detailed in Table 3.1. Path based approaches are based on the path length linking both concepts in the “IS-A” taxonomy of WordNet. Although these approaches are simple they are not so accurate and ignore most of the structure of WordNet. Information content approaches attempt to avoid problems of path-based approaches by incorporating

<sup>2</sup>The basic object in WordNet is a set of strict synonyms called a synset. By definition, each synset in which a word appears has a different sense of that word.

an additional, and qualitatively different, knowledge source, namely information from a corpus. These approaches depend on the amount of information that both concepts have in common. These approaches are based on the intuition of [Resnik, 1995] that states that “*the more information two concepts share in common, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in the taxonomy*”. The information content of a concept  $c$  can be quantified as the negative of the log likelihood,  $-\log(p(c))$ , where  $p(c)$  is the probability of encountering an instance of concept  $c$ . Feature-based approaches are based on the assumption that “concepts with more common features and less non common features are more similar”. For example, [Banerjee and Pedersen, 2003] define common features by counting the number of shared words (overlaps) in the word senses of the concepts, as well as in the glosses of words that are related to those concepts according to the dictionary. These related concepts are explicitly encoded in WordNet as relations, but can be found in any dictionary via synonyms, antonyms, or also references provided for a word sense.

### 3.3.2 Corpus-Based Measures

Corpus-based measures try to identify the degree of similarity between words using statistical information derived from a large corpus. Since the creation of corpus databases is expensive, labor-intensive and time-consuming, the Web which is an information resource with virtually unlimited potential sometimes used as a corpus. Thus, in Figure 3.3, we distinguish two types of Corpus-based similarities: corpus-based and web-based similarities. They are computed differently but both are based on the distributional hypothesis [Harris, 1954] which states that “words that appear in the same contexts tend to be semantically similar”. As a pioneering work, [Schutze, 1993] presented a word as a vector in a multidimensional space of its context in a corpus. For example, the word “Tennis” can be represented as a vector composed of its context defined with words such as “ball, player, racket, sport ...”. The meaning of words in this vector is modeled using spatial word proximity. In the simplest case, the distributional analysis relies on the context of window approach. It is based on the hypothesis that words are semantically similar if they appear within similar context windows. In fact, for each word  $w$  in the dataset, each window  $W$  centered around this word is collected and added to the vector together with its frequency (the total number of times we saw a window  $W$  around the word  $w$  on the whole corpus).

### 3.3.2.1 Distributional Corpus-Based Measures

Corpus-based measures compute the similarity between words based on statistics derived from a specific corpus. The most successful approaches in this category are based on the Vector Space Model (VSM) [Salton and McGill, 1983] and include the Syntactic Distributional Analysis (SDA) proposed by [Grefenstette, 1994] based on the hypothesis that “Words are semantically similar if they appear in similar syntactic contexts”. Syntactic analysis allows to know which words modify other words and to develop contexts from this information. Starting from the hypothesis that “Words with similar meaning repeatedly occur closely”, [Lund and Burgess, 1996] propose the Hyperspace Analogue to Language model (HAL). The basic idea is to develop a matrix of word co-occurrence values for a given vocabulary. A “window” of a certain size ( e.g. ten words) is defined which is slid over the corpus. The co-occurrence values are inversely proportional to the number of words separating a specific pair of words. The Latent Semantic Analysis (LSA) [Landauer and Dutnais, 1997] is based on the idea that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determine the similarity of meaning of words and set of words to each other. Recently, several prominent approaches based on Wikipedia were proposed. In [Strube and Ponzetto, 2006], a measure is proposed by exploiting the article abstracts and the network from Wikipedia categories. In fact, given a word pair, the Wikipedia pages which they refer to are retrieved. By extracting the categories the pages belong to, the category tree is determined. Finally, the semantic similarity between words is computed based on the extracted page and the found paths along the category taxonomy. In [Gabrilovich and Markovitch, 2007], authors proposed the Explicit Semantic Analysis (ESA) where a concept is represented in a vector space of all Wikipedia articles. Specifically, in ESA, a word is represented as a column vector in the TD-IDF matrix of Wikipedia article text. The semantic similarity between two words can be obtained using the cosine similarity between their corresponding vectors. A comprehensive comparison of some corpus-based measures can be found in [Ferret, 2010].

### 3.3.2.2 Distributional Web-Based Measures

Web-based measures use the Web as a corpus in order to compute semantic similarity measures. They use Web text search engines in order to compute the similarities. They rely on the number of times words co-occur in the documents indexed by an information retrieval system. Many web-based similarities have been proposed in the literature including the Point-wise Mutual Information Information Retrieval (PMI<sub>IR</sub>) [Turney, 2001], Normalized Google Distance (NGD) [Cilibrasi and Vitanyi, 2007], WebJaccard, WebDice and WebOverlap [Bollegala et al., 2007].

TABLE 3.1: Classification of measures of semantic similarity and relatedness based on WordNet and their relative advantages/disadvantages.

Type	Method	Principle	Advantages/disadvantages
Path based	[Rada et al., 1989], [Wu and Palmer, 1994], [Leacock and Chodorow, 1998]	Function of path length linking concepts in the “IS-A” taxonomy of WordNet	(+) simplicity, (-) “IS-A” relations only, (-) No so accurate and ignore most of the structure of WordNet.
Information Content (IC) based	[Resnik, 1995], [Lin, 1998], [Jiang and Conrath, 1997]	Depends on the amount of information that both concepts have in common. The more common the concepts share, the more they are similar.	(+) Uses empirical information from corpora, (-) “IS-A” relations only, (-) Need an additional corpus.
Hybrid based	[Zhou et al., 2008]	Takes the path length between two concepts and IC value of each concept as its metric.	(-) Weights of both metrics need to be settled.
Feature based (Relatedness)	[Tversky, 1977], [Hirst and St Onge, 1998], [Banerjee and Pedersen, 2003], [Patwardhan, 2003]	Concepts with more common features and less non common features are more similar.	(+) Measures relatedness of all parts of speech more than IS-A relations, (+) Uses empirical knowledge implicit in a corpus of data, (-) Needs complete attribute features.

In particular, web-based measures rely on the number of documents (hits)  $h_i$  returned by the system for the query “ $w_i$ ”, the number of hits  $h_{ij}$  returned by the query “ $w_i$  AND  $w_j$ ” and  $M$  the number of documents indexed by the system. The Point-wise Mutual Information using data collected by information retrieval (PMI-IR) was suggested by [Turney, 2001] as an unsupervised measure for the evaluation of the semantic similarity of words. It is based on word co-occurrence using counts collected over very large corpora (e.g. the Web). Given two words  $w_i$  and  $w_j$ , the PMI-IR is measured as:

$$PMI - IR(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i) * p(w_j)} = \log \frac{\frac{h_{ij}}{\sum_{i,j} h_{ij}}}{\frac{h_i}{\sum_{i,j} h_{ij}} * \frac{h_j}{\sum_{i,j} h_{ij}}} \approx \frac{h_{ij}}{h_i * h_j} \quad (3.1)$$

where  $p(w_i, w_j)$  is the joint probability (the probability that  $w_i$  and  $w_j$  co-occur), and  $p(w_i)$  and  $p(w_j)$  represents respectively the probability that  $w_i$  and  $w_j$  occur in the documents returned as a result to the query.

[Cilibrasi and Vitanyi, 2007] propose the Normalized Google Distance (NGD) based on Google page counts and it is defined as:

$$NGD(w_i, w_j) = \frac{\max(\log(h_i); \log(h_j)) - \log(h_{ij})}{\log(|M|) - \min(\log(h_i); \log(h_j))} \quad (3.2)$$

where  $h_i$  denotes the number of pages containing  $w_i$ , and  $h_{ij}$  denotes the number of pages containing both  $w_i$  and  $w_j$ , as reported by Google results. Words with the same or similar meanings in a natural language sense tend to be “close” in the sense of Normalized Google Distance, while words with dissimilar meanings tend to be farther apart. [Bollegala et al., 2007] propose three similarity measures based on search engine, defined as follows:

$$WebJaccard(w_i, w_j) = \begin{cases} 0 & \text{if } h_{i,j} \leq c, \\ \frac{h_{i,j}}{h_i + h_j - h_{i,j}} & \text{otherwise,} \end{cases} \quad (3.3)$$

Given the scale and noise in Web data, it is possible that two words may appear on some pages purely accidentally. In order to reduce the adverse effects attributable to random co-occurrences, [Bollegala et al., 2007] set the WebJaccard coefficient to zero if the page count for the query ( $w_i$  and  $w_j$ ) is less than a threshold  $c$ .

$$WebDice(w_i, w_j) = \begin{cases} 0 & \text{if } h_{i,j} \leq c \\ \frac{2 * h_{i,j}}{h_i + h_j} & \text{otherwise,} \end{cases} \quad (3.4)$$

$$WebOverlap(w_i, w_j) = \begin{cases} 0 & \text{if } h_{i,j} \leq c \\ \frac{h_{i,j}}{\min(h_i, h_j)} & \text{otherwise,} \end{cases} \quad (3.5)$$

A comprehensive study of web-based similarity measures is presented in [Lindsey et al., 2007]. Recently, [Popescu and Grefenstette, 2011] proposed to apply the ESA approach (see Section 3.3.2.1) for Flickr website. In fact, this approach uses Flickr as a corpus and each Flickr tag is considered as a concept and thus can be represented as a vector of co-occurring Flickr tags.

### 3.3.3 Discussion

As we have seen, there is a significant literature on semantic similarity measures. Both models proposed in this chapter require a semantic similarity for the tag modeling step. For that, we rely on the use of two different knowledge resources: WordNet and Flickr to define two semantic similarities. WordNet is used to derive a Knowledge-based similarity measure while Flickr is exploited to derive a distributional Web-based measure. Both knowledge sources are different in the nature of the language used and the type of conceptual relations that we can extract.

In fact, WordNet offers a variety of different semantic relations to weave its word-senses together. Although the vocabulary of WordNet is very extensive, most of tags used to annotate social media are not included in WordNet. Consequently, it has important limitations due to the resource limited coverage. Moreover, in some WordNet-based semantic similarities, important semantic relations are discarded.

In order to overcome this shortcoming, Flickr is exploited. Since tags are added by users in Flickr Website, this second data source covers a larger panel of concepts than WordNet and is inherently multilingual.

As presented in Section 3.3, many similarity measures have been proposed in the literature to determine the semantic relation between two words. The question is how can we reason about and evaluate the superiority of a semantic measure over other ones?

Generally, there is no standard to evaluate the effectiveness of a semantic similarity measure. On the whole, three kinds of methods are identified in the literature to evaluate semantic similarities [Meng et al., 2013].

- The first one is a theoretical examination of a semantic measure for those mathematical properties thought desirable, such as whether it is a metric, whether its parameter projections are smooth functions, and so on.
- The second one is to compare measures by calculating their coefficients of correlation with human judgments [Zhou et al., 2008; Seco et al., 2004a]. Insofar as human judgments of similarity and relatedness are deemed to be correct by definition.

- The third approach is to evaluate the measures with respect to their performance in the framework of a particular application [Budanitsky and Hirst, 2006]. In other words, if we have a framework which requires a measure of semantic similarity, we compare the performance of different measures to find the most effective one, while holding all other aspects of the system unchanged.

While comparison with human judgments is the ideal way to evaluate a measure of similarity or semantic relatedness, it is difficult in practice to obtain a large set of reliable and objective judgments and it is extremely time-consuming and labor-intensive. It is especially true in the context of social media where there exists a huge number of tags and obtaining human judgments on such data would be a very large task. The third approach seems to be an alternative in our tag-based annotation system which requires a semantic similarity to compute Tag Models. An experimental evaluation of WordNet based similarities is conducted to evaluate the most effective similarity for multimodal image annotation and presented in Section 3.8.2.1.

### 3.4 Problem Formalization

Once tag imperfections are well identified and defined clearly in the context of social media, we focus on how to handle these imperfections to improve image annotation. The problem posed in this chapter is an instance of the problem formulated in Chapter 2-Section 2.2 where only the tag modality is exploited. Our goal is thus to build a tag-based classifier with a decision function defined as follows:

$$\begin{aligned} f : \mathcal{T} &\rightarrow \mathcal{Y} \\ f(\mathbf{X}_i^t) &= \hat{y}_i \end{aligned} \tag{3.6}$$

$f$  associates a label set  $\hat{y}_i$  to each unseen image  $I_i$  from the test dataset  $\mathcal{T}$  using only textual information issued from the tag modality.

In this chapter, only textual features are used for training and testing. Therefore, we propose two approaches to compute tag-based signature,  $\mathbf{X}_i^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_M^t)$ , while handling tag imperfections. The first model, called Soft Bag-of-Concepts (Soft-BoC), represents a new variant of the TF-IDF of the Vector Space Model (VSM) [Salton and McGill, 1983], where term weights are computed using a thresholded semantic similarity between tags and annotation concepts. The second model is based on the BOW model [Salton and McGill, 1983] with a locality-constrained coding method [Liu et al., 2011b] for tags.

Given an image with its associated tags, tag-based signatures for both approaches are built in three steps as shown in Figure 3.4.

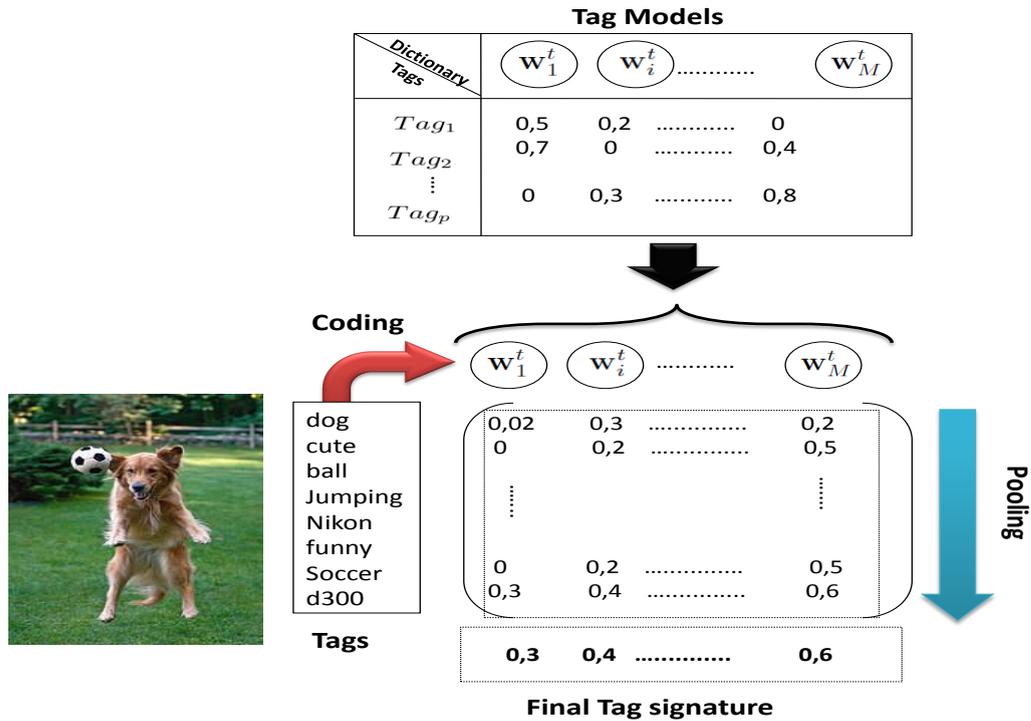


FIGURE 3.4: Overview of the proposed approaches: Soft-BoC and LSTC. Given an image with its associated tags, tag-based signatures for both models are built in three steps: (1) tag modeling, (2) coding and (3) pooling.

### 1. Tag Modeling

Let's note that tag modeling is different from tag signature and it consists in representing each tag from the dataset as a weighted vector of dictionary words. This vector is called a *Tag Model* where weights represent semantic similarity scores between the tag and dictionary words using an external knowledge resource as illustrated in Figure 3.5. Given a textual codebook  $\mathcal{W}^t = (\mathbf{w}_1^t, \dots, \mathbf{w}_M^t)$ , each tag  $t_k$  from the dataset is represented with a vector  $S_k = \langle s(t_k, \mathbf{w}_1^t), \dots, s(t_k, \mathbf{w}_M^t) \rangle$ , where  $s(t_k, \mathbf{w}_i^t)$  is the semantic similarity score between the tag  $t_k$  and the  $i^{\text{th}}$  word  $\mathbf{w}_i^t$  in the textual codebook. In our case, the semantic similarity designates a pair  $\langle \text{Information resource, similarity measure} \rangle$ . An example of semantic similarity used in our models is  $\langle \text{WordNet, Wu\&Palmer} \rangle$  by using WordNet as knowledge resource and Wu\&Palmer [Wu and Palmer, 1994] as a similarity measure. The two signatures are generic in the sense they can rely on any semantic similarity.

### 2. Coding

Once similarity measures are computed, we perform a coding for each tag in order to achieve an assignment step which consists in activating only dictionary words which are semantically similar to the considered tag and others are set to zero. Given an image  $\mathbf{I}_i$  with a set of associated tags  $\mathbf{T}_i = \{t_{1i}, t_{2i}, \dots, t_{li}\}$ , each tag from the set  $\mathbf{T}_i$  is transformed into a Tag

Model vector as presented above. The process is repeated for all the tags associated with the image. The number of words in the dictionary which are semantically similar to a given tag is fixed differently in each model. For the Soft-BoC approach, the number of similar tags is determined with a threshold (diameter of the neighborhood) on the semantic similarity scores while for the LSTC signature, this number is set to a neighborhood window size. Both are optimally determined with a cross-validation on the training dataset. The obtained vectors after the coding step are called *Tag-related codes*.

### 3. Pooling

In order to obtain the final tag-signature vector, all the tag-related codes within one image are aggregated with a pooling function such as the average, the sum or the maximum functions.

In our case, separate signatures are generated considering each similarity measure.

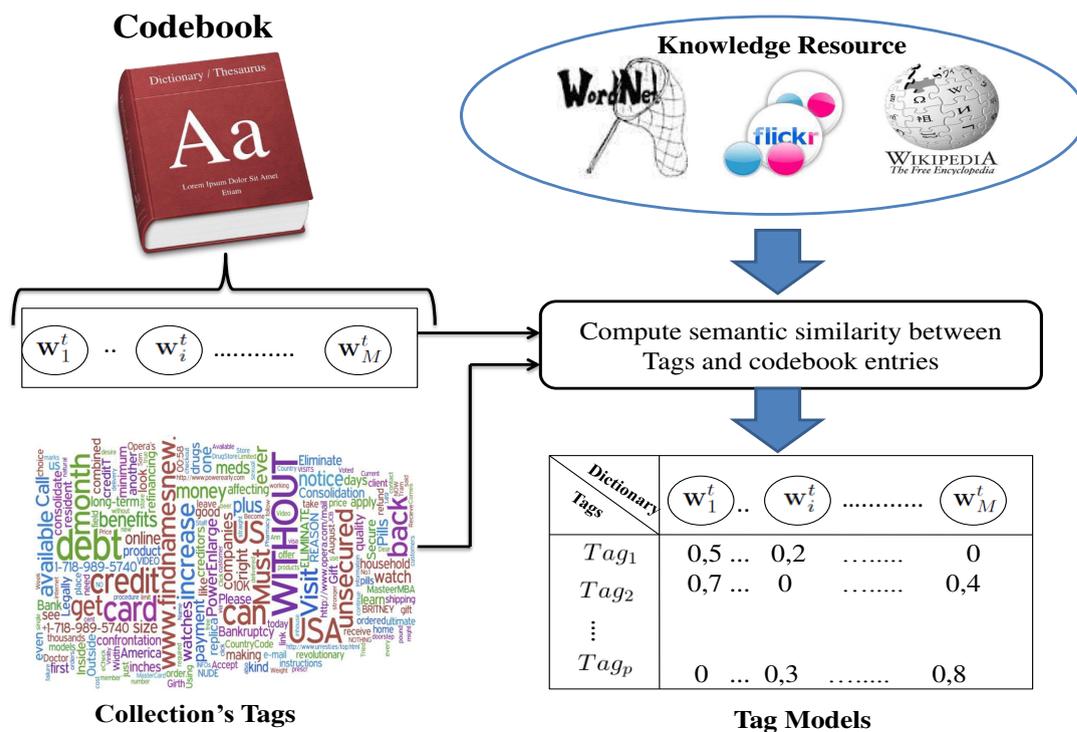


FIGURE 3.5: Illustration of the Tag Modeling process. Each tag from the collection is represented with a vector of scores defined as the semantic similarity measures between the tag and the codebook words.

### 3.5 Soft Bag-of-Concepts Signature

In this section, we propose a novel textual signature, namely the Soft Bag-of-Concepts (Soft-BoC) illustrated in Figure 3.6. The model represents a new variant of the TF-IDF of the Vector Space Model (VSM) [Salton and McGill, 1983], where term weights are computed using a thresholded semantic similarity measure between tags and annotation concepts. This model is inspired from a model that we can call componential space model, such as conceptual vector [Schwab et al., 2002], which describes the meaning of a word by its atoms, components, attributes, behavior, related ideas. Specifically, the proposed Soft-BoC signature of an image is defined as an histogram of annotation concepts. Each bin of this histogram represents the accumulation of the contribution of each tag from the image tags toward the underlying concept according to a predefined semantic similarity. In the rest of this section, we give a detailed description of this tag-based signature by describing the three steps described before: tag modeling, coding and pooling.

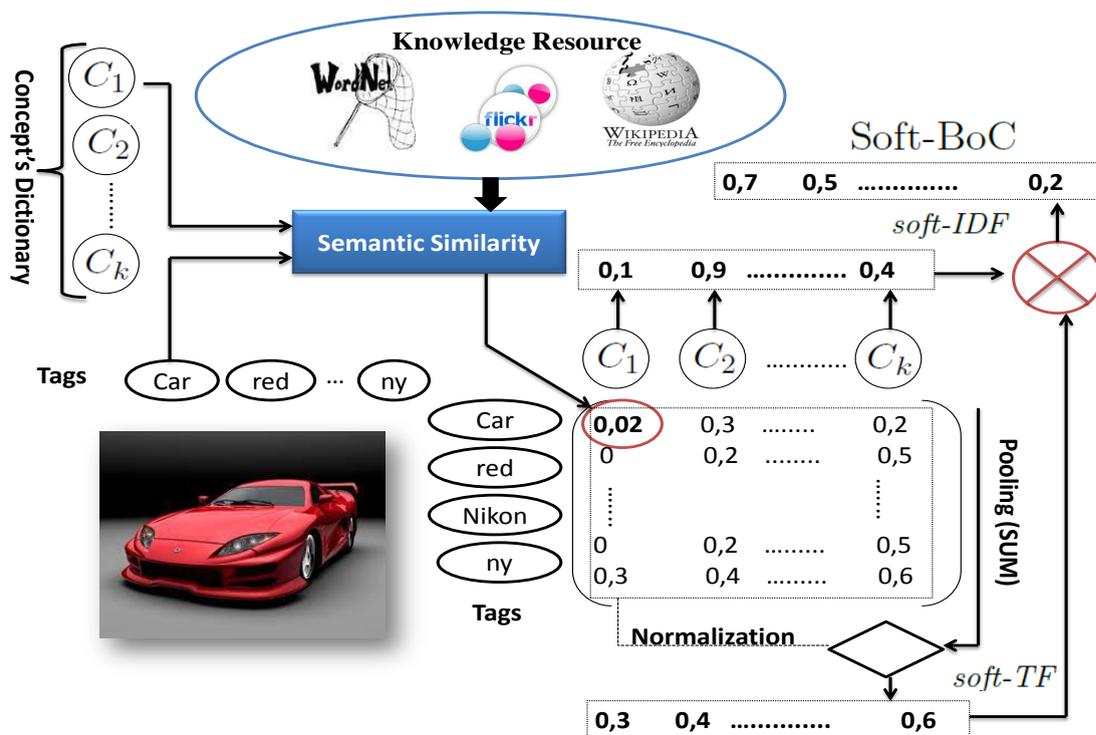


FIGURE 3.6: Illustration of the Soft-BoC signature which represents a new variant of the the TF-IDF, where term weights are computed using a thresholded semantic similarity between tags and annotation concepts.

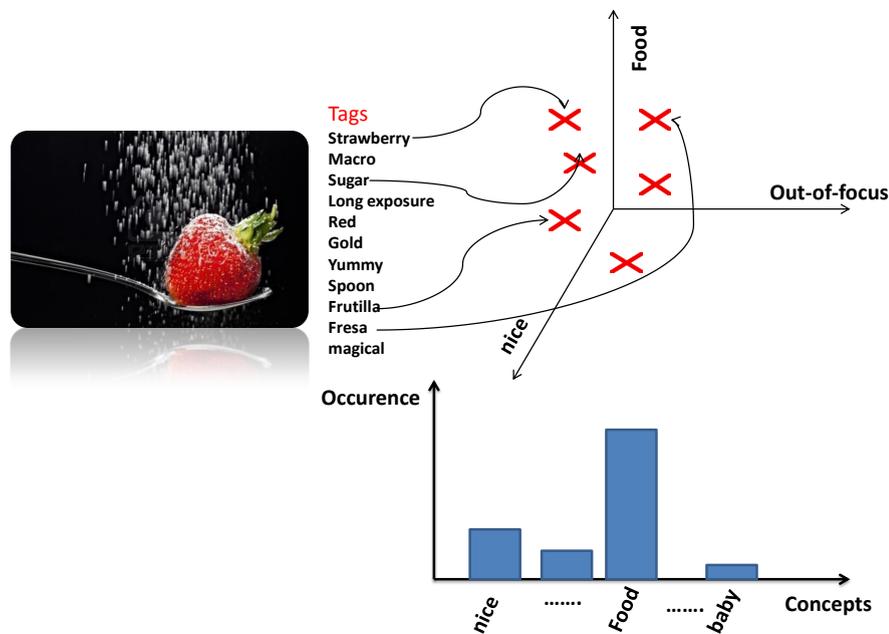


FIGURE 3.7: An example of image with its associated tags. Tags such as “*strawberry, sugar, spoon, frutella, fresa*” will be associated with the concept “*food*” which will be relevant to describe the image content. The Term Frequency is illustrated with the histogram in the bottom. The bin value for the concept “*food*” is higher than others. The image will be described with this concept.

### 3.5.1 Tag Modeling

In this approach, the words of the codebook  $\mathcal{W}^t = (\mathbf{w}_1^t, \dots, \mathbf{w}_M^t)$ , introduced in Chapter 2-Section 2.2, are the annotation concepts (labels). Consequently, each tag is represented with a vector whose size is the number of annotation concepts of the considered dataset. The key idea is to project the tags of a given image in the annotation concept space using a semantic similarity measure. Each tag will be associated with one or more concepts according to its semantic similarity with the concepts. As illustrated in Figure 3.7, tags such as “*strawberry, sugar, spoon, frutella, fresa*” will be associated with the concept “*food*” which will be relevant to this image. In this manner, the concept voted by several tags is then considered appropriate to describe the image content. This is in clear contrast to the classic BOW approaches where the relatedness of textual concepts is simply ignored as word terms are statistically counted using simple word matching.

### 3.5.2 Coding/Pooling

As detailed in Chapter 2, in the classic TF-IDF, an hard assignment is performed to determine the presence or the absence of a concept (1 or 0) in order to compute TF and IDF values. This hard assignment is generally achieved using simple

word-to-word matching. Nevertheless, tags are different from traditional classification labels. In fact, tags are in different languages, with different spellings and meanings. This idea has been investigated in the literature on the differences between folksonomies and traditional taxonomies or ontologies, e.g, [Quintarelli, 2005]. In general, tags are freely chosen keywords leading to a big variability in the set of user tags. Therefore, it is more appropriate to proceed on a soft assignment in which a tag is matched to a concept with some confidence value. This confidence value represents the semantic similarity between a tag  $t_k$  and a concept  $C_i$ . Ideally, if the user tagged his photo with a concept included in the set of classification concepts, this value is equal to 1. Else, it is a value between 1 and 0 depending on how similar they are. In this way, we take into account the similarity score between tags and annotation concepts. We introduce the *soft Term Frequency (soft-TF)* and the *soft Inverse Document Frequency (soft-IDF)* which are computed as follows:

$$soft - TF_{i,j} = \frac{\sum_{t_k \in T_j} F_\alpha(s(t_k, C_i))}{\sum_{C_i \in \mathcal{C}} \sum_{t_k \in T_j} F_\alpha(s(t_k, C_i))} \quad (3.7)$$

$$soft - IDF_i = \log\left(\frac{|\mathcal{I}|}{\sum_{I_j \in \mathcal{I}} \frac{\sum_{t_k \in T_j} F_\alpha(s(t_k, C_i))}{n_{i,j}}}\right) \quad (3.8)$$

where  $T_j$  represents the set of tags associated with an image  $I_j$ .  $\mathcal{C}$  and  $\mathcal{I}$  represent respectively the set of annotation concepts and images in the dataset,  $n_{i,j}$  is the number of occurrences of the considered concept  $C_i$  in image  $I_j$ . And  $F_\alpha$  represents a strictly increasing function defined by:

$$F_\alpha : \quad [0, 1] \longrightarrow [0, 1]$$

$$s(t_k, C_i) \longmapsto \begin{cases} 0 & \text{if } s(t_k, C_i) < \alpha \\ s(t_k, C_i) & \text{if } s(t_k, C_i) \geq \alpha \end{cases} \quad (3.9)$$

where  $s(t_k, C_i)$  represents the semantic similarity between a tag  $t_k$  and a concept  $C_i$ . For an image  $I_j$  with a set of user provided tags  $T_j$ , the tag-based signature vector  $\mathbf{X}_j^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_M^t)$ , called *soft - TF - IDF*, is obtained by the product of the *soft Term Frequency (soft-TF)* and the *soft Inverse Document Frequency (soft-IDF)* values, as follows:

$$\mathbf{x}_i^t = (soft - TF_{i,j}) * (soft - IDF_i) \quad (3.10)$$

In case  $F_\alpha(s(t_k, C_i))$  is equal to 1, we found the same formula as the classic *TF - IDF*. To compute the Soft-BoC signature, we consider only concepts that are similar to the considered tag in a neighborhood defined with a threshold.  $\alpha$  represents the diameter of this neighborhood and is determined by cross-validation.



The idea is to enrich image content description by expanding initial tag list using tags from the collection. Thus, the Tag Modeling step consists in mapping each tag in the collection to the tag dictionary using semantic similarity measures. Once Tag Models for all tags in the collection are computed, we perform a coding step to map a tag to only its semantically related tags.

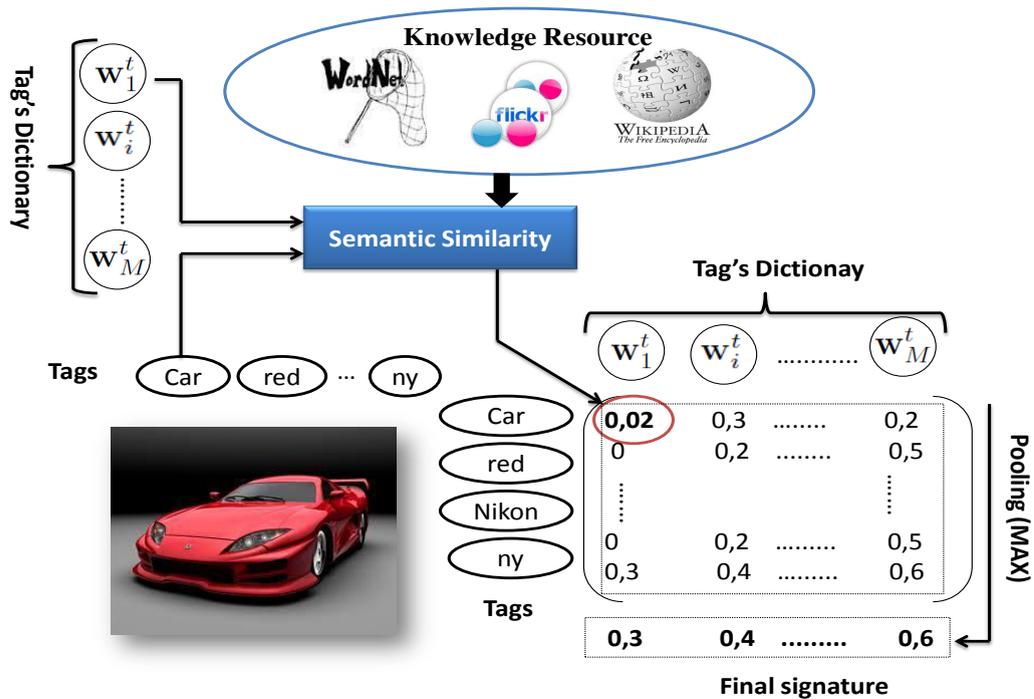


FIGURE 3.9: Illustration of the LSTC signature.

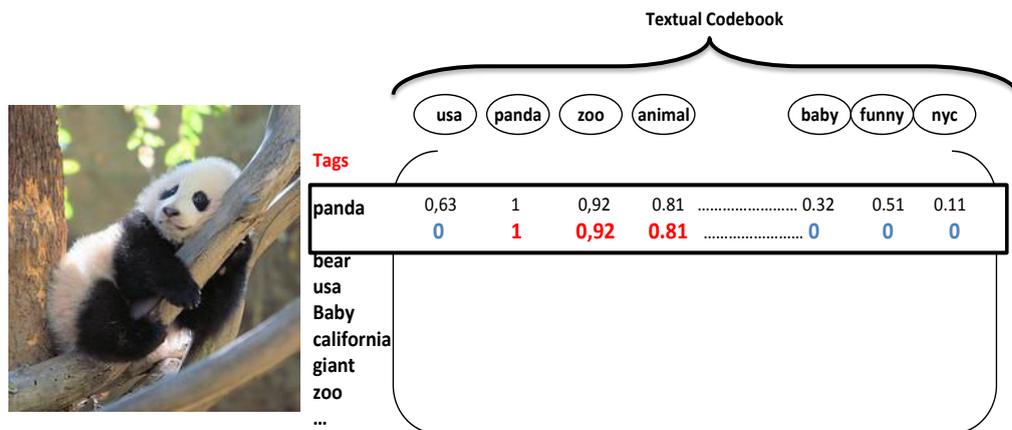


FIGURE 3.10: Illustration of the tag coding process. Only semantically similar tags to the tag “panda” , such as “panda, zoo, animal” , are kept while others are set to zero.

### 3.6.2 Coding/pooling

Contrary to the classic BOW representation where an hard assignment is used, we rely on the locality-constrained coding method [Liu et al., 2011b] in order to activate only semantically related tags and set others to zero. Given an image  $I_j$  with its associated tags, each tag  $\mathbf{t}_k$  is mapped to only its  $L$ -nearest tags under a similarity measure.

$$z_{k,i} = \begin{cases} s(\mathbf{t}_k, \mathbf{w}_i^t) & \text{if } \mathbf{w}_i^t \in \mathcal{N}_L^t(\mathbf{t}_k), \\ 0 & \text{otherwise,} \end{cases} \quad (3.11)$$

where  $\mathbf{w}_i^t$  represents the dictionary entries and  $\mathcal{N}_L^t(\mathbf{t}_k)$  denotes the  $L$ -nearest neighbor tags of  $\mathbf{t}_k$ , under the considered semantic similarity denoted by  $s(\mathbf{t}_k, \mathbf{w}_i^t)$ .  $L$  represents the window size of this neighborhood and is determined by cross-validation on the training dataset. Figure 3.10 illustrates the coding step. In this example, only semantically similar tags of the tag “panda”, such as “panda, zoo, animal”, are activated while others are set to zero.

Given the tag-related codes within one image, a max-pooling is performed in order to obtain the final tag-signature vector. The superiority of max-pooling over other pooling methods, combined with such coding scheme, can be explained probabilistically as being the lower bound of the probability of occurrence of a tag with the image [Liu et al., 2011b].

For the pooling step, for an image  $I_j$  with a set of user provided tags  $T_j$ , the element of the tag-based feature vector  $\mathbf{X}_j^t = (\mathbf{x}_1^t, \dots, \mathbf{x}_M^t)$  are defined as follows:

$$\mathbf{x}_i^t = \max_{\mathbf{t}_k \in T_j} z_{k,i} \quad \forall i = 1, \dots, M \quad (3.12)$$

In our case, separate signatures are generated considering each similarity measure.

## 3.7 Adopted Semantic Similarities

Both similarity measures used to compute Tag Models are detailed below.

### WordNet-based Similarity

WordNet concepts are structured as synsets (sets of synonyms) that are arranged as a hierarchy whose main structural axis is defined by conceptual inheritance. Wu-Palmer measure [Wu and Palmer, 1994] gives a similarity between two concepts as their distance in the WordNet hierarchy as shown in Figure 3.11. Given two

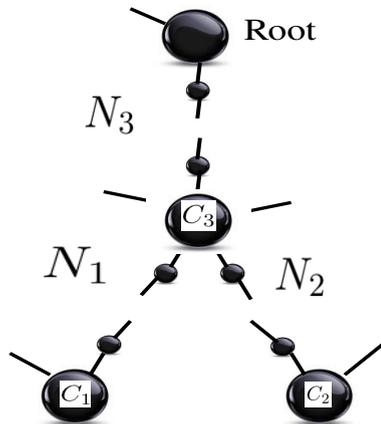


FIGURE 3.11: The Wu&amp;Palmer similarity.

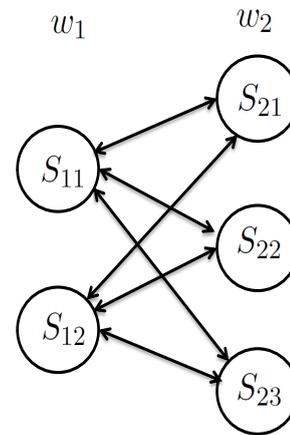


FIGURE 3.12: The “maximum sense pair” illustration.

concepts  $C_1$  and  $C_2$ , this similarity measure considers the position of these two concepts in the taxonomy relative to the position of the most specific common concept  $C_3$ . As there may be multiple parents for each concept, two concepts can share parents by multiple paths. The most specific common concept  $C_3$  is the common parent related with the minimum number of ‘IS-A’ links with concepts  $C_1$  and  $C_2$ .

$$sim_{wup}(C_1, C_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (3.13)$$

where  $N_1$  and  $N_2$  is the number of ‘IS-A’ links from  $C_1$  and  $C_2$  respectively to the most specific common concept  $C_3$ , and  $N_3$  is the number of ‘IS-A’ links from  $C_3$  to the root of the taxonomy. It scores between 1 (perfect similarity) and 0 (no similarity).

A word may have several senses, so it may appear in multiple synsets. While computing similarity between two words  $w_1$  and  $w_2$ , all the senses for a word are considered against all the senses of the second. An example is given in Figure 3.12, where  $w_1$  has two senses  $S_{11}$  and  $S_{12}$ ;  $w_2$  has three senses  $S_{21}$ ,  $S_{22}$  and  $S_{23}$ . Given all the senses of  $w_1$  are considered against all the senses of  $w_2$ , we will get 6 sense pairs similarity values. The sense pair with the maximum similarity value is called the “maximum sense pair” and the similarity is computed as follows:

$$sim_{WordNet}(w_1, w_2) = \max \{ sim_{wup}(s_1, s_2); (s_1, s_2) \in \text{syns}(w_1) \times \text{syns}(w_2) \}, \quad (3.14)$$

where  $sim_{wup}$  is the Wu-Palmer similarity.

## Flickr-based Similarity

In [Popescu and Grefenstette, 2011], an adaptation of the TF-IDF model to the social space is proposed in order to compute the social relatedness of two tags.

Let  $\mathbf{S}$  be the matrix of size  $N \times M$  defined by:

$$\mathbf{S}(i, j) = \text{users}(\mathbf{t}_i, \mathbf{t}_j) \times \log\left(\frac{\text{users}_{\text{collection}}}{\text{users}_{\text{collection}}(\mathbf{t}_j)}\right), \quad (3.15)$$

where  $\mathbf{t}_i$  is the target tag,  $\mathbf{t}_j$  is an element of the codebook,  $\text{users}(\mathbf{t}_i, \mathbf{t}_j)$  is the number of distinct users who associate the tag  $\mathbf{t}_i$  to the tag  $\mathbf{t}_j$  among the top results returned by the Flickr API for  $\mathbf{t}_i$ ;  $\text{users}_{\text{collection}}(\mathbf{t}_j)$  is the number of distinct users from a pre-fetched subset of Flickr users that have tagged photos with tag  $\mathbf{t}_j$ , and  $N$  is the number of unique tags associated to photos of the dataset and  $M$  is the size of the codebook. Note that some of the tags can have entries on both dimensions of matrix  $\mathbf{S}$ . In the current work, we consider a fixed set of tags, that is a tag-codebook.

Relying on this matrix, a Flickr model for a given tag  $\mathbf{t}_i$  is proposed in [Popescu and Grefenstette, 2011] as the following vector of weights:

$$\mathbf{w}_i = [w_{i,1}, w_{i,2}, \dots, w_{i,M}]^T, \quad (3.16)$$

with  $w_{i,j}$  the normalized social weight defined by:

$$w_{i,j} = \frac{\mathbf{S}(i, j)}{\max\{\mathbf{S}(i, k), k = 1, \dots, M\}}. \quad (3.17)$$

Thereby, given two tag-Flickr models  $\mathbf{w}_i$  and  $\mathbf{w}_j$ , we compute the contextual similarities between their related tags  $\mathbf{t}_i$  and  $\mathbf{t}_j$  using the cosine similarity:

$$\text{sim}_{\text{Flickr}}(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{w}_i^T \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}. \quad (3.18)$$

## 3.8 Experimental Evaluation

As part of this chapter, we evaluate both proposed models for tag-based image annotation. Only the tag modality is used here to generate tag-based features. Multimodal-based image annotation experiments within a multimodal framework combining visual and textual features, are presented in Chapter 6 and Chapter 7.

To evaluate the effectiveness and the robustness of the proposed methods on user-provided noisy/missing tags, we employ the real-world social images with human

annotated tags. Specifically, five publicly available Flickr<sup>3</sup> image datasets are used for the experiments. We refer the reader back to Section 2.6.2 for dataset statistic details (number of images, number of labels, number of tags...).

Our tag models depend on two parameters: the WordNet similarity and the neighborhood window/diameter which need to be studied. First, we evaluate different WordNet similarity, in Section 3.8.2.1, to choose empirically the best one for our tag-based image annotation task. In Section 3.8.2.2, we study the influence of the neighborhood window/diameter on tag-based image annotation performance. Second, we present results obtained using the Soft-BoC signature in Section 3.8.2.3 and those of the LSTC signature in Section 3.8.2.4. Finally, a comparison to the state-of-the-art is presented in Section 3.8.2.5.

### 3.8.1 Experimental Setup

For the Soft-BoC approach, the tag signature size is equal to the number of class labels. For the LSTC approach, the textual codebook is obtained by keeping only the  $N$  most frequent tags in each dataset. For all datasets except for the NUS-WIDE, we do not perform any pre-processing to clean tags.

For PASCAL VOC'07 dataset, we use the experimental setting of [Guillaumin et al., 2010]. A dictionary of size 804 is obtained by keeping only tags that appear at least eight times. In the case of ImageClef'10 and ImageClef'11 datasets, we keep tags that were used at least three times in the collection, resulting in a textual codebook of 2,500 tags. For the ImageClef'12 dataset, we keep only tags that appear at least four times leading to a tag dictionary of size 5,134.

For the NUS-WIDE dataset, among the 425,059 unique tags, there are 9,325 tags that appear more than 100 times. Authors of [Chua et al., 2009] propose to check these tags against WordNet and keep only tags that exist resulting in a dictionary of size 5,018.

Dataset dictionary sizes are summarized in Table 3.2. Tag frequency represents the minimum tag frequency (number of occurrences of a tag in the whole dataset) used as a threshold to build the dictionary. Tag models are computed for each dataset using the two chosen similarities detailed in Section 3.7. The tag-based signature vector of an untagged image is set to zero. Let  $\mathbf{x}_i$  be a signature vector of an image  $I_i$ , we train for each concept  $C_j$  a classifier that can associate this concept with its feature vector. For this, we use  $N$  binary linear kernel based Support Vector Machines (SVM) [Cortes and Vapnik, 1995] (One-Versus-All). Given a training dataset  $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid y_i \in \{-1, +1\}\}_{i=1}^N$  where the  $y_i$  is either +1 or -1,

---

<sup>3</sup><http://www.flickr.com/>

indicating the class to which the point  $\mathbf{x}_i$  belongs. We want to find the maximum-margin hyperplane that divides the points having  $y_i = +1$  from those having  $y_i = -1$ . Any hyperplane can be written as the set of points  $\mathbf{x}$  satisfying

$$\mathbf{w} \cdot \mathbf{x} - b = 0, \quad (3.19)$$

where  $(\cdot)$  denotes the dot product and  $\mathbf{w}$  the normal vector to the hyperplane. The parameter  $\frac{b}{\|\mathbf{w}\|}$  determines the offset of the hyperplane from the origin along the normal vector  $\mathbf{w}$ . The distance between these two hyperplanes is equal to  $\frac{2}{\|\mathbf{w}\|}$ . So to find the optimal hyperplane we need to minimize  $\|\mathbf{w}\|$ , i.e, to solve the following optimization problem which deals with the parameters  $\mathbf{w}$  and  $b$ :

$$\arg \min_{(\mathbf{w}, b)} \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \quad (i = 1, \dots, n) \quad (3.20)$$

This optimization problem is known as the primal problem formulation.

## 3.8.2 Experimental Results

In our experiments, we measure performance using the Average Precision (AP) criterion for each class, and also using the mean AP (mAP) which computes the mean over the average precision of the image ranking per label.

### 3.8.2.1 Evaluation of WordNet semantic similarities

In this section, we aim at comparing the performance of the tag-based image annotation using different WordNet similarities in order to choose the one which gives better performance score. We use WordNet::Similarity [Pedersen et al.,

TABLE 3.2: Summary of dictionary size for the considered datasets. Tag frequency represents the minimum tag frequency threshold used to build the dictionary.

	PASCAL VOC'07	NUS-WIDE
Dictionary size	804	5,018
Tag frequency	8	100

	ImageClef'10	ImageClef'11	ImageClef'12
Dictionary size	2,500	2,500	5,134
Tag frequency	3	3	4

TABLE 3.3: Comparison for WordNet similarities on the ImageClef'12 dataset using the LSTC signature in terms of mean Average Precision (mAP).

	path	lch	wup	res	lin	jcn	hso	lesk	vector
mAP	27.7	28.3	<b>31.0</b>	26.3	27.1	26.2	27.5	27.9	28.2

2004] and WordNet-3.0<sup>4</sup> to construct a semantic similarity measure between words. Specifically, we consider three measures based on path lengths: **lch** [Leacock and Chodorow, 1998], **wup** [Wu and Palmer, 1994] and **path** [Rada et al., 1989], three based on additional information content: **res** [Resnik, 1995], **lin** [Lin, 1998] and **jcn** [Jiang and Conrath, 1997] and three other feature-based measures: **hso** [Hirst and St Onge, 1998], **lesk** [Banerjee and Pedersen, 2003] and **vector** [Patwardhan, 2003]. We choose, the ImageClef'12 dataset for comparison using the LSTC signature.

From the results presented in Table 3.3 for the nine considered similarities, we can see that the Wu&Palmer similarity (**wup**), surprisingly, performs better than other similarity measures. This is not the case in some evaluations [Sebti and Barfroush, 2008; Seco et al., 2004b] where semantic similarities are compared to human judgment. Performances of similarity measures are highly related to the concerned task such as word sense disambiguation [Altintas et al., 2005] or the domain of application such as biomedical [Garla and Brandt, 2012]. Given these results, we select the Wu&Palmer similarity for the WordNet-based similarity for the rest of this dissertation.

### 3.8.2.2 Influence of the neighborhood diameter/window

In this section, our goal is to study the influence of the neighborhood diameter (respectively window) in the classification accuracy for the Soft-BoC (respectively for the LSTC) signature.

We choose the ImageClef'10 dataset to study the influence of the neighborhood diameter corresponding to the parameter  $\alpha$  defined in Equation 3.9. We split training data into 20-fold with cross-validation. Figure 3.13 shows the classification accuracy in terms of mean Average Precision (mAP) on the ImageClef'10 while varying the neighborhood diameter size from 0.6 to 0.95 with a step of 0.05. The bottom curve is obtained with a cross validation on the training dataset to determine the optimal diameter value. The top curve represents the obtained scores on the testing dataset. As we can see the best classification score on the training dataset (the bottom curve) reaches 28.5% of mAP score and is obtained with a value of diameter equal to 0.8. This value achieves the best classification score (32% mAP) on the test dataset.

<sup>4</sup><http://wordnet.princeton.edu/wordnet/download/>

The number of semantically similar tags in the LSTC signature is an important parameter. Our goal in this experiment is to analyze the impact of neighborhood window size. We tried various values of  $L \in \{1, 5, 10, 50, 100, 1000\}$  on the ImageClef'11 dataset. Figure 3.14 shows the influence of the neighborhood size for locality coding in classification accuracy on ImageClef'11 dataset. The top curve represents results obtained with Flickr similarity while the bottom one corresponds to WordNet similarity. Clearly, restricting the neighborhood leads to better tag coding results than considering the nearest neighbor for hard assignment or all the codewords for the soft assignment. The optimal size of the neighborhood has been estimated by cross-validation on the training dataset leading to a number of 5 (respectively 50) neighboring codewords for the WordNet (respectively Flickr) tag-based similarity measures and is also empirically validated in Figure 3.14.

### 3.8.2.3 Evaluation of the Soft Bag-of-Concepts Approach

In this section, we evaluate the performance of the Soft-BoC signature for tag-based image annotation. As a baseline, we consider the TF-IDF model [Salton and McGill, 1983] to evaluate the effectiveness of handling tag imperfections. A separate tag-based signature is obtained using each of the two considered semantic similarities detailed in Section 3.7. Our first set of experimental results, presented

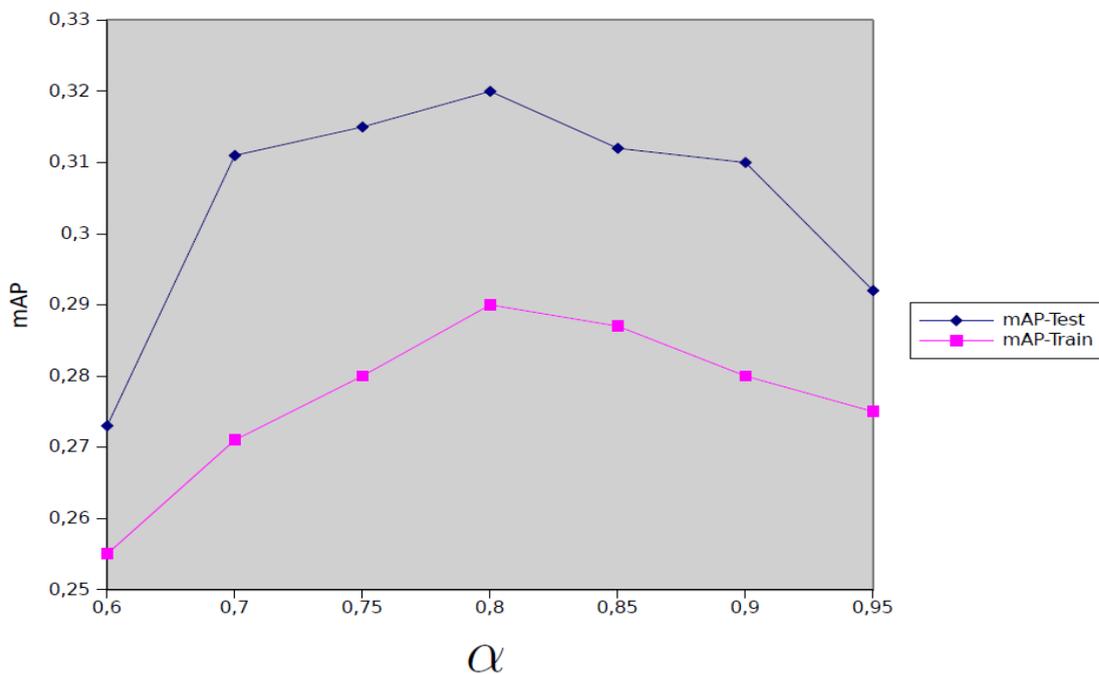


FIGURE 3.13: Classification accuracy in terms of mean Average Precision (mAP) on the ImageClef'10 while varying the neighborhood diameter size based on the Soft-BoC signature. The bottom curve is obtained with a cross validation on the training dataset to determine the optimal diameter value. The top curve represents the obtained scores on the testing dataset.

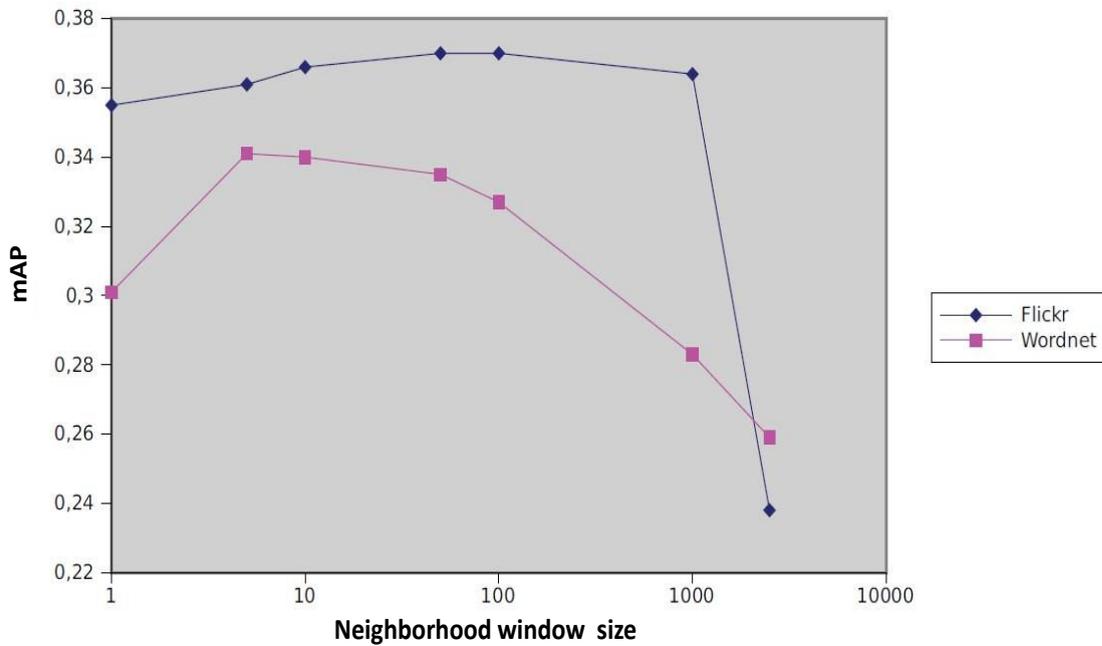


FIGURE 3.14: Performances of the LSTC signature when changing the neighborhood window size in the tag-feature-space on the ImageClef'11.

TABLE 3.4: Evaluation of the Soft-BoC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on PASCAL VOC'07 and NUS-WIDE datasets.

Method	PASCAL VOC'07	NUS-WIDE
TF-IDF(WordNet)	26.6	11.5
TF-IDF(Flickr)	46.8	37.4
<b>Soft-BoC(WordNet)</b>	<b>32.9</b>	<b>21.9</b>
<b>Soft-BoC(Flickr)</b>	<b>49.3</b>	<b>38.7</b>
<b>Soft-BoC(WordNet&amp;Flickr)</b>	<u>49.6</u>	<u>39.4</u>

in Table 3.4 compares the classification performances based on tag-signatures in terms of mean Average Precision (mAP) on PASCAL VOC'07 and NUS-WIDE datasets. Results on the ImageCLEF datasets (ImageClef'10, ImageClef'11 and ImageClef'12) are presented in Table 3.5. For the five considered datasets, we can observe that the Soft-BoC signature outperforms the classic TF-IDF. For the PASCAL VOC'07, the mAP score increases by more than 6% (respectively 3%) with WordNet (respectively with Flickr) similarity. The effectiveness of this model is confirmed with the NUS-WIDE dataset with which we obtain a gain of 10% in terms of mAP scores using WordNet and an improvement of 2% by exploiting Flickr similarity.

As one can observe, on the ImageCLEF datasets (ImageClef'10, ImageClef'11 and ImageClef'12), the Soft-BoC signature obtains better results than the classic TF-IDF. For ImageClef'10 dataset, an improvement of 10% in terms of mAP

TABLE 3.5: Evaluation of the Soft-BoC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on the ImageCLEF datasets (ImageClef’10, ImageClef’11 and ImageClef’12).

Method	ImageClef’10	ImageClef’11	ImageClef’12
TF-IDF (WordNet)	14.1	14.7	16.3
TF-IDF (Flickr)	15.3	13.6	23.9
<b>Soft-BoC (WordNet)</b>	<b>24.2</b>	<b>29.2</b>	<b>20.8</b>
<b>Soft-BoC (Flickr)</b>	<b>26.8</b>	<b>32.8</b>	<b>27.2</b>
<b>Soft-BoC (WordNet&amp;Flickr)</b>	<u>29.6</u>	<u>34.6</u>	<u>27.9</u>

scores is achieved with WordNet similarity and 11% with Flickr similarity. For ImageClef’11 dataset, we obtain a gain of 15% (respectively 19%) with WordNet (respectively Flickr). An improvement of about 4% is obtained both with WordNet and Flickr for the ImageClef’12 dataset.

In this experiment, our goal is to illustrate the effectiveness of the Soft-BoC signature. We show in Figure 3.15 several images with their associated user tags on the left and the Soft-BoC signature on the right. In the first example, concepts such as “*reflection, water, city, night*”, which are relevant to describe the image visual content, are successfully captured. The obtained signature shows a high value for the bin associated with these concepts. In the second image, most of tags are subjective, however our approach is able to predict some relevant concepts such as “*portrait*”. In the third example, again the proposed model successfully discovers most of the relevant concepts such as “*sun, flower, reflection*” and even “*plant*” concept is inferred by exploiting Flickr similarity.

To summarize, the Soft-BoC signature outperforms the classic TF-IDF with a superiority of the Flickr similarity compared to the WordNet similarity. This can be explain by the limitations of WordNet due to the resource’s limited coverage, to its availability in English only. In fact, although the vocabulary of WordNet is very extensive, lots of tags are not included in WordNet. To show the complementarity of WordNet and Flickr resources we present, in the last row of Tables 3.4 and 3.5, the obtained results with their combination. The latter is computed by averaging the classification predictions obtained using WordNet and Flickr. Obviously, we can observe that the combination of both resources gives better results than using only WordNet or Flickr for all the considered datasets. Obtained results confirm our hypothesis that both knowledge sources enable the capture of complementary facets of tags and their combination improves the quality of predicted tags.

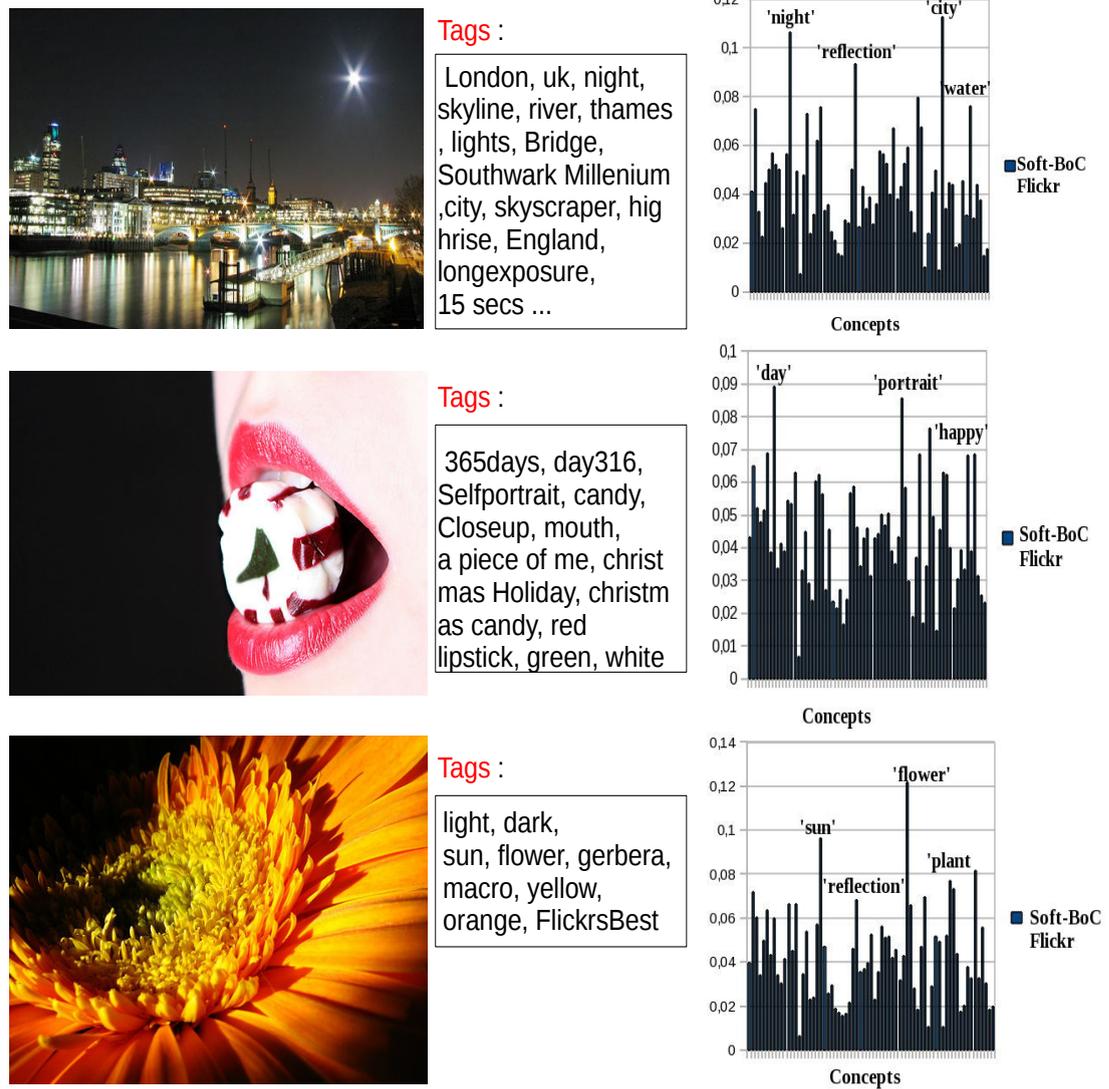


FIGURE 3.15: Several images with their associated user tags on the left and the Soft-BoC signature on the right

### 3.8.2.4 Evaluation of the Local Soft Tag Coding Approach

In this section, we evaluate the performance of the LSTC signature for tag-based image annotation. We compare our model to various coding scheme: hard coding, Soft coding (WordNet) and Soft coding (Flickr). In the hard coding, only tags which are present in the codebook are set to one, while others are set to zero. It corresponds to a tag-to-tag matching process. In the case of the soft coding, all codebook entries have values which are the similarity between tags and codewords. There is no selection of semantically similar tags and the size of the neighborhood window is set to the size of the considered tag dictionary. Table 3.6 presents a comparison between various coding scheme for the PASCAL VOC 07 and NUS-WIDE datasets. Obtained results on the ImageCLEF datasets (ImageClef'10,

ImageClef’11 and ImageClef’12) are presented in Table 3.7. As shown, our method based on local soft tag coding strategy outperforms both the hard coding and the soft assignment coding, indicating the effectiveness of the “*early cut-off*” strategy, which removes the adverse impact of unreliable tags that are semantically far and thus decreasing tags uncertainty.

For the PASCAL VOC 07 dataset, we observe that our two tag-versions (WordNet and Flickr) relying on the local soft assignment coding outperform the hard assignment based coding scheme (51.1% vs 43.3% of mAP scores). For the NUS-WIDE dataset, our model gives again better results than the other coding schemes. We obtain a gain of 10% of mAP scores compared to the hard coding and improvement of 12 to 15% of mAP compared to the soft assignment. The effectiveness of our model is confirmed on the ImageCLEF datasets. For the ImageClef’10 dataset, our model gives better results than the other coding schemes. For the ImageClef’11 dataset, the LSTC approach outperforms both the hard and soft coding. As shown in Table 3.7, a gain of  $\approx 15\%$  of mAP is achieved compared to the hard coding. The proposed LSTC signature using WordNet improves the obtained mAP score of the soft coding by  $\approx 9\%$ . An improvement of 2% of mAP is obtained for the ImageClef’12.

To summarize, our LSTC signature outperforms both the hard coding and soft coding schemes for the five considered datasets. Unlike the classic BOW coding where a tag is mapped to only itself (tag-to-tag matching), the LSTC approach allows the soft contribution of its  $L$ -semantic nearest tags in a dictionary under a similarity measure. Consequently, the initial user’s tag list can be enriched with semantically related tags which tackle the incomplete tag problem. Moreover, considering only the  $L$ -nearest tags produces a selection effect, which removes the adverse impact of unreliable tags that are semantically far and thus decreasing tag uncertainty. Our tag representation is based on a predefined dictionary built with the most frequent tags. Thereby, we eliminate rare, misspelled and subjective tags to handle a part of the imprecision and uncertainty aspects of tags.

TABLE 3.6: Evaluation of the LSTC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on PASCAL VOC’07 and NUS-WIDE datasets.

Coding Scheme	PASCAL VOC’07	NUS-WIDE
Hard assignment	43.3	30.1
Soft assignment (WordNet)	43.0	25.4
Soft assignment (Flickr)	51.0	28.3
<b>LSTC (WordNet)</b>	<b>49.4</b>	<b>40.4</b>
<b>LSTC (Flickr)</b>	<b>51.6</b>	<b>39.2</b>
<b>LSTC (WordNet&amp;Flickr)</b>	<u>51.8</u>	<u>42.0</u>

TABLE 3.7: Evaluation of the LSTC signature for tag-based image annotation in terms of mean Average Precision (mAP%) on the ImageCLEF datasets (ImageClef’10, ImageClef’11 and ImageClef’12).

Coding Scheme	ImageClef’10	ImageClef’11	ImageClef’12
Hard assignment	22.5	21.3	30
Soft assignment (WordNet)	27.9	25.9	25
Soft assignment (Flickr)	35.8	23.8	28.0
<b>LSTC (WordNet)</b>	<b>35.4</b>	<b>34.7</b>	<b>31</b>
<b>LSTC (Flickr)</b>	<b>37.6</b>	<b>37.0</b>	<b>32.9</b>
<b>LSTC (WordNet&amp;Flickr)</b>	<u>38.8</u>	<u>38.1</u>	<u>34.1</u>

### 3.8.2.5 Comparison to the state-of-the-art

In this section, we compare our two models to the state-of-the-art approaches on the five considered datasets with an uniform experimental setup. Detailed results are shown in Table 3.8, the best mAP score for each dataset is marked in bold. Obviously, we can observe that for the five considered datasets, our models give in most cases better results than the state-of-the-art methods with a superiority for the LSTC approach.

On the PASCAL VOC’07, both proposed tag models outperform the state-of-the-art methods [Guillaumin et al., 2010; Wang et al., 2009a]. These methods are based on classic BOW representation where the textual features are defined as an histogram of occurrences of image tags towards a predefined dictionary, as detailed in Chapter 2. The BOW representation has shown to be effective in text categorization which is not the case in the context of social media. In fact, images generally have only few tags [Sigurbjörnsson and van Zwol, 2008] which need to be enriched. Thus, the use of external knowledge resources is crucial and effective to expand the initial tag list of the image and to enrich its semantic description. Moreover, we introduce the locality coding constraint in the BOW representation which shows to be more suitable than the hard coding for tags in the context of social media.

On the NUS-WIDE dataset, the proposed method in [Gao et al., 2010], gives a low mAP score compared to both proposed models and the one of [Wang et al., 2010a]. Our results on the NUS-WIDE dataset are in line with those of [Wang et al., 2010a] (41.9 vs 42 of mAP %) which confirms the advantage of using knowledge resources such as WordNet, Flickr and Wikipedia to handle tag imperfections. While authors of [Wang et al., 2010a] combine information from WordNet, Flickr, Wikipedia and co-occurrence, we use only two resources which make our signatures more simple and efficient.

On the ImageClef'10 dataset, our signatures outperform the method proposed in [Li et al., 2010b]. We obtain a gain of 7% of mAP using the Soft-BoC signature. An improvement of 16% of mAP score is obtained with the LSTC approach. Certainly, [Li et al., 2010b] uses an external knowledge resource to expand initial tag list as we do, however these expanded tags are compared directly to concepts to predict final concepts. This is not the case of the proposed methods where a tag-signature is extracted and used as an input feature for an SVM classifier. Again, handling tag imperfections shows the effectiveness of the proposed signatures.

On the ImageClef'11 dataset, our signatures outperform the state-of-the-art methods [Liu et al., 2011; Xioufis et al., 2011; Nagel et al., 2011]. The obtained results using the proposed Soft-BoC model are in line with [Zhang et al., 2012b]. An improvement of 4% of mAP score is achieved using the proposed LSTC signature. The Histogram of Textual Concepts (HTC) [Liu et al., 2011] is closely related to our Soft-BoC approach that uses WordNet as an external knowledge to capture semantic tag relatedness. The difference is that we apply a selection procedure to keep only semantically similar concepts to a given tag and others are discarded. Again, we confirm the importance of the local soft coding of tags in the BOW representation. Both [Xioufis et al., 2011; Nagel et al., 2011] are based on classic BOW representation. There is no exploitation of an external knowledge resource to enrich initial tag list.

On the ImageClef'12, authors in [Liu et al., 2012] use the Histogram of Textual Concepts model. This approach outperforms our Soft-BoC signature. It consists of a combination of 6 textual features for each concept based on the selective weighted late fusion (SWFL) scheme, while we combine only two textual features (WordNet and Flickr). However, we obtain the best score on this dataset using our LSTC signature.

To summarize, the BOW model represents the dominant approach for tag representation in the state-of-the-art methods using different variants of word frequency (TF, TF-IDF, BOW..) and some pre-processing techniques (stemming, stop words removal ..). As most of these tag representations are based on classic BOW model, they do not take into account tag imperfections and fail to capture semantic tag relatedness. In addition, the shortcoming in the classic BOW representations lies in the valuable semantic information can not be captured. The classic BOW approach has two main drawbacks: (1) The BOW is sensitive to the changes in vocabulary, that occur when training data can not be reasonably expected to be representative of all the potential testing data; (2) The BOW considers only the word frequency information, thus disregards tag semantic information.

### 3.9 Conclusion and Discussion

In this chapter, we considered the problem of tag imperfections in tag-based image annotation. We have introduced two novel textual signatures for tag-based image annotation in the context of social media. We reported extensive experimental results on five challenging datasets. From these results, we conclude that the Soft-BoC signature outperforms the state-of-the-art methods on three out of five datasets. The size of the obtained tag-signature with the Soft-BoC signature is very compact (the number of annotation labels). Thus, this model seems to be suitable for large scale datasets. The second proposed signature, LSTC, outperforms the state-of-the-art methods on the five considered datasets on the tag-based image annotation task.

It seems interesting to focus on the resemblance and differences between the two models of tags presented in this chapter in order to interpret how they handle tag imperfections. Both signatures are based on the BOW representation with a coding step for each tag in order to achieve the assignment step. This latter consists in activating only dictionary entries which are semantically similar to the considered tag. These neighbor tags are with a fixed window represented by the  $L$ -nearest neighbor for the LSTC signature and represented by a diameter value for the Soft-BoC signature.

In both approaches, considering only the  $L$ -nearest words from the textual codebook produces an “*early cut-off*” effect, which removes the impact of unreliable tags/concepts that are semantically far and thus handling a part of the tag uncertainty problem. Moreover, the soft contribution of nearest neighbors allows to the initial user tag list to be enriched by adding semantically related tags. This tackles a part of the problem of partial incompleteness.

In the LSTC approach, the tag modeling step is based on a predefined dictionary built with the high frequent tags. In this manner, we eliminate rare, misspelled and subjective tags to handle a part of the imprecision and uncertainty aspects of tags. In the case of polysemy, the Soft-BoC model helps disambiguate textual concepts according to the context. For instance, the concept of “*jaguar*” can refer not only to car manufacturer but also to the “*Panthera*” species. However, when the tag “*jaguar*” comes with a photo showing a car, correlated tags such as “*car*”, “*automobile*” and “*industry*”, are very likely to be used, thereby, clearly distinguishing the concept “*jaguar*” in automobile industry from that of an animal where correlated tags can be “*animal*”, “*panthera*”, “*tiger*”, etc. Similarly, in the case of synonyms, the Soft-BoC signature will reinforce the concept related to the synonym as far as the semantic similarity measurement takes into account the phenomenon of synonyms. Note that the introduction of the threshold in the TF-IDF model allows, in comparison with the classic TF-IDF model, to “share” the weights of a tag on the word dictionary, but not all of them.

TABLE 3.8: Comparison of the two proposed signatures to the state-of-the-art for image annotation, in terms of mean Average Precision (mAP) on the five considered datasets.

Method	PASCAL VOC'07	NUS-WIDE	ImageClef'10	ImageClef'11	ImageClef'12
[Guillaumin et al., 2010]	43.5	–	–	–	–
[Wang et al., 2009a]	43.5	–	–	–	–
[Gao et al., 2010]	–	18.8	–	–	–
[Wang et al., 2010a]	–	41.9	–	–	–
[Li et al., 2010b]	–	–	22.8	–	–
[Liu et al., 2011]	–	–	–	32.1	–
[Xioufis et al., 2011]	–	–	–	32.5	–
[Nagel et al., 2011]	–	–	–	32.6	–
[Zhang et al., 2012b]	–	–	–	34.7	–
[Liu et al., 2012]	–	–	–	–	33.3
<b>Soft-BoC(WordNet&amp;Flickr)</b>	49.6	39.4	29.6	34.6	27.9
<b>LSTC (WordNet&amp;Flickr)</b>	<b>51.8</b>	<b>42.0</b>	<b>38.8</b>	<b>38.1</b>	<b>34.1</b>

In this chapter only the problem of partial incompleteness was taken into account. The problem of full incompleteness is not yet handled and the tag-based feature vector of an untagged image is set to zero. Thus, untagged images are considered as belonging to the same cluster in the SVM classifier which can decrease classification performances. To handle missing features, we can ignore untagged images from the training set or using a process generally referred to as *imputation*. A simple imputation method is just to use the average value of feature vectors of the dataset, but there are more robust techniques based on nearest neighbors that can be used. A novel method is presented in the next chapter in order to predict tags for untagged images based on nearest neighbors and Belief theory.

# Chapter 4

## Handling Full Textual Incompleteness

### Contents

---

<b>4.1</b>	<b>Introduction</b>	<b>102</b>
<b>4.2</b>	<b>Problem Formalization</b>	<b>102</b>
<b>4.3</b>	<b>Proposed Tag Completion Method</b>	<b>103</b>
4.3.1	Finding candidate tags	105
4.3.2	Predicting final tags	107
<b>4.4</b>	<b>Tag Suggestion Dataset Construction</b>	<b>110</b>
<b>4.5</b>	<b>Experimental Evaluation</b>	<b>110</b>
4.5.1	Datasets	111
4.5.2	Experimental Setup	111
4.5.3	Experimental Results	113
<b>4.6</b>	<b>Conclusion and Discussion</b>	<b>116</b>

---

## 4.1 Introduction

Ideally, images in social media websites would have a reasonable number of user generated tags, which would then enable other users to find and retrieve them. Unfortunately, in practice, only a part of the uploaded pictures are tagged with useful tags, others are not tagged at all, making a huge number of images unavailable in tag-based search applications. This problem is presented in both Chapter 1 and Chapter 3 as tag incompleteness. We distinguish two types of incompleteness: partial and full. Partial incompleteness is the case where the image has some tags and others are missing while full incompleteness represents the case where the image has no tag. The former has been studied in Chapter 3, thus in this chapter, we focus on the latter type of incompleteness. We are interested in the problem of tag completion defined as the process that automatically assigns a set of tags to an untagged image without any contribution from humans. We propose a novel method named Tag Completion based on similar visual neighbors and Belief theory to handle full tag incompleteness introduced in Chapter 3-Section 3.2. In this chapter, we demonstrate that local tag coding is an effective coding scheme for handling tag imperfections. We use the same coding scheme to enrich visual neighbors tag description. Our approach differs from existing techniques on two main points. The first difference and novelty is that we use tag corpus knowledge (Flickr) to enrich nearest neighbors description from existing tags. The second and most important difference is that we explicitly use the Belief theory, which is able to handle neighbors conflict and deal with tag imperfections. This work has been published in [Znaidia et al., 2013a].

The rest of this chapter is structured as follows. First, in Section 4.2, we present a formalization of the problem of image tagging. In Section 4.3, we introduce the proposed method to handle full tag incompleteness. We present the tag suggestion dataset building in Section 4.4. Experimental results for both image classification and tag suggestion are presented in Section 4.5. The chapter is concluded in Section 4.6.

## 4.2 Problem Formalization

Automatic image tagging is defined as the process that automatically assigns a set of tags to an untagged image. The problem of automatic tagging is usually posed as a tag propagation procedure from visually similar images. An overview of image tagging problem that we consider in this chapter is presented in Figure 4.1. In the following, we introduce some notations used in the rest of this chapter.

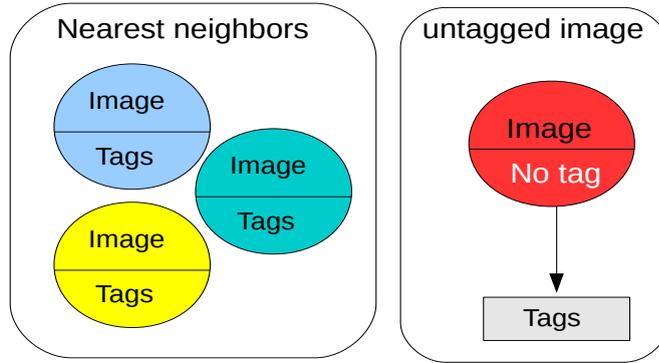


FIGURE 4.1: An overview of image tagging problem that we consider in this chapter. For an untagged image, given the set of its nearest neighbors with their associated tags, our goal is to predict a set of tags that describe its visual content.

- $\mathcal{N}^v = \{I^1, \dots, I^k\}$  is the set of images that are the nearest visual neighbors of an image  $I_s$ ,
- $\mathbf{T}^i = \{\mathbf{t}_1^i, \mathbf{t}_2^i, \dots, \mathbf{t}_l^i\}$  is the set of tags associated with the neighbor image  $I^i$ ,
- $d(I_s, I^i)$  is a metric to determine the similar visual neighbors,
- $\mathcal{W}^t = (\mathbf{w}_1^t, \dots, \mathbf{w}_M^t)$  is a textual codebook.

Our goal is to predict a set of tags to describe the image content of an untagged image  $I_s$  using tag information from its nearest neighbors. The visual feature is used for the determination of the set of visual neighbors. Moreover, the set of visual neighbors ( $\mathcal{N}^v$ ) is searched in an image dataset different from the training dataset  $\mathcal{L}$  introduced in Chapter 2-Section 2.2.

### 4.3 Proposed Tag Completion Method

Tags provided by users reflect the personal perspective and context that are important to the photo owner as introduced in Chapter 3-Section 3.2. This implies that if another user tags the same photo, it is possible that he will use different tags. In Flickr, one can find many photos on the same subject from many different users, which are described by a wide variety of tags. An example of images from Flickr on the same subject from two different users with their associated user tags is shown in Figure 4.2. The first photo is described by its owner using the tags “*la dame de fer, paris, Tour Eiffel*” while the second is tagged with “*b&W, architecture, tour, towe, paris, autumn, city, europe, france...*”. Using the collective knowledge that resides in Flickr community, we can extend the description of these photos with tags such as “*Tourism, France, Landmark*”. This extension

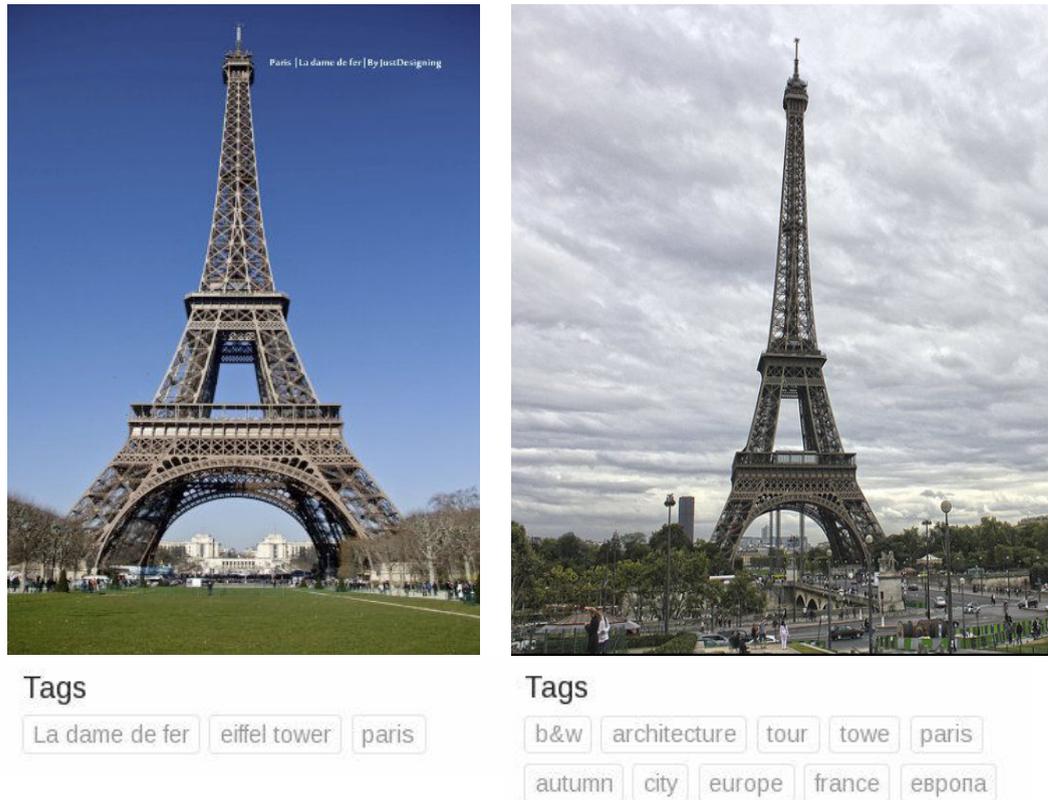


FIGURE 4.2: An example of images from Flickr on same subject from two different users with their associated user tags.

provides a rich semantic description of photos. Last trends to generate tags for images without any annotation rely on the idea that “if many distinct users use the same tags to label visually similar images, then these tags are likely to reflect the visual content of the annotated images”. Starting from this intuition, classic neighbor voting algorithms use information from the nearest neighbors to predict tags. In the original voting  $k$ NN algorithm, an image is assigned to the majority class according to its  $k$ -nearest neighbors, independently of the relevance of each neighbor. Moreover, the classical  $k$ NN method does not deal with ambiguity and imprecise information due to the limitation of the probabilistic framework. We propose a method named Tag Completion to tackle these problems of robustness and effectiveness.

The flowchart of the proposed method using visually similar images is presented in Figure 4.3. It consists in two main steps: (1) creating a list of “candidate tags” from the visual neighbors of the untagged image and then (2) using them as pieces of evidence to be combined to provide the final list of predicted tags. More precisely, given an untagged image  $I_s$ , we start by searching the  $k$ -nearest neighbors using visual information (color, texture). Then, we compute a BOW signature for each neighbor based on the LSTC approach presented in Chapter 3-Section 3.6. Second, a sum-pooling operation across the BOW of the  $k$ -nearest neighbors is

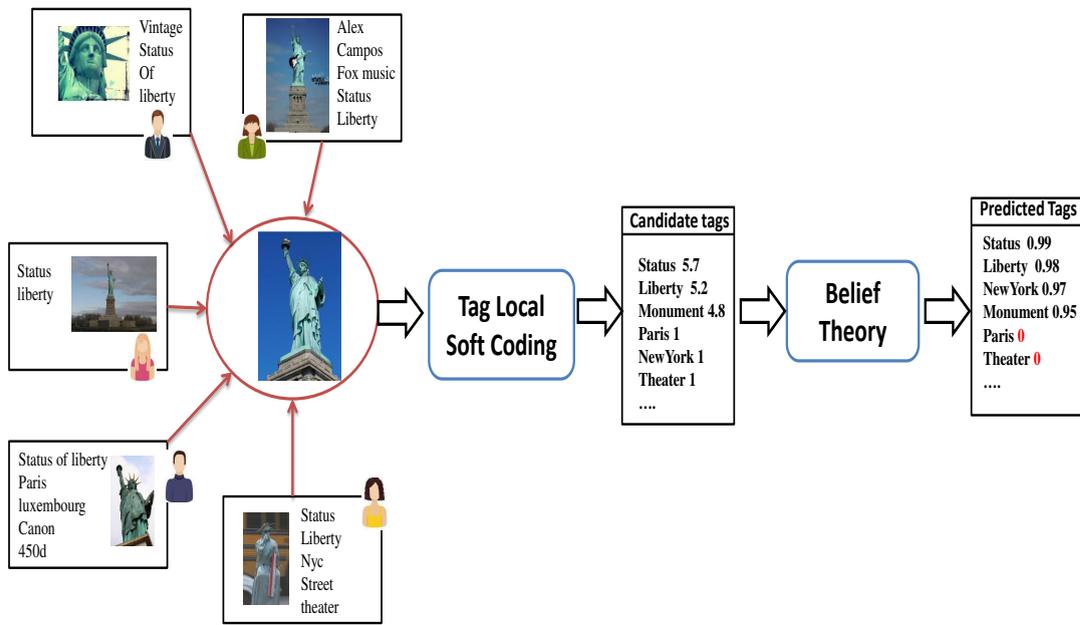


FIGURE 4.3: Overview of our Tag Completion approach based on local soft coding and Belief theory. First, we compute a BOW signature for each neighbor based on LSTC presented in Chapter 3-Section 3.6. Second, a sum-pooling operation across the BOW of the  $k$ -nearest neighbors is performed to obtain the list of the candidate tags. Finally, pieces of evidence from neighbors are combined using Dempster’s rule to obtain the set of predicted tags.

performed to obtain the list of “candidate tags” (the high frequent). Finally, basic belief masses are obtained for each nearest neighbor using the distances between this pattern and its neighbors. Their fusion leads to the list of final predicted tags.

### 4.3.1 Finding candidate tags

Let  $I_s$  be an untagged image and  $\mathcal{N}^v = \{I^1, \dots, I^k\}$  the set of its  $k$ -nearest neighbors according to a given measure  $d$ , within an image database. These resources (image database, visual features and similarity function) are not specified at this point but their importance will be discussed later (Section 4.5.3). Each neighbor image  $I^r$  has a set of tags  $T^r = \{t_1^r \dots t_l^r\}$ . Let consider as well a textual codebook  $\mathcal{W}^t = (\mathbf{w}_1^t, \dots, \mathbf{w}_M^t)$  that has been built previously (detailed in Chapter 3-Section 3.8). Each tag  $\mathbf{t}_p^r \in T^r$  is then coded according to its  $L$ -nearest neighbors in the codebook:

$$z_{p,q} = \begin{cases} \text{sim}_{\text{Flickr}}(\mathbf{t}_p^r, \mathbf{w}_q^t) & \text{if } \mathbf{w}_q^t \in \mathcal{N}_L^t(\mathbf{t}_p^r), \\ 0 & \text{otherwise,} \end{cases} \quad (4.1)$$

where  $\mathcal{N}_L^t(\mathbf{t}_p^r)$  denotes the  $L$ -nearest neighbors of the tag  $\mathbf{t}_p^r$ , using the Flickr-based similarity detailed in Chapter 3-Section 3.7. This step is motivated by the fact that images have only a set of few tags which need to be completed to enrich their content description. The final tag-signature vector  $\mathbf{c}^r = [c_1^r \dots c_M^r]$  of the neighbor image  $I^r$  results from an aggregation of the maximal values of the coded tags as follows:

$$c_q^r = \max_{p=1}^{\text{Card}(T^r)} (z_{p,q}) \quad (4.2)$$

Figure 4.4 shows an example which illustrates this step.

To obtain the list of “candidate tags”, a sum-pooling operation is performed on the tag signature of the visual neighbors. For a given image, according to its  $k$  nearest visual neighbors, we obtained the following vector  $\mathbf{C} = [C_1 \dots C_M]$  where  $C_q$  is defined as follows:

$$C_q = \sum_{r=1}^k (c_q^r) \quad \forall \quad q = 1 \dots M \quad (4.3)$$

The tags corresponding to the entries of  $\mathbf{C} = [C_1 \dots C_M]$  with the highest values constitutes the list of “candidate tags” that we note  $\Omega$  in the next section. The number of “candidate tags” retained is empirically set to 10 in our experiments. Figure 4.5 illustrates with an example how to obtain the list of “candidate tags”.

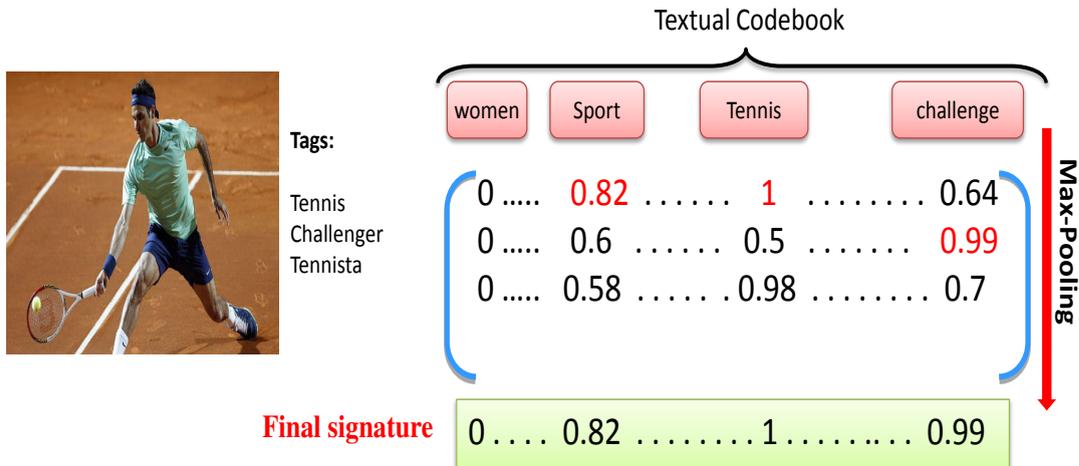


FIGURE 4.4: An example to illustrate the local soft tag coding to enrich image description. Tags such as “Sport, Challenge” are added.

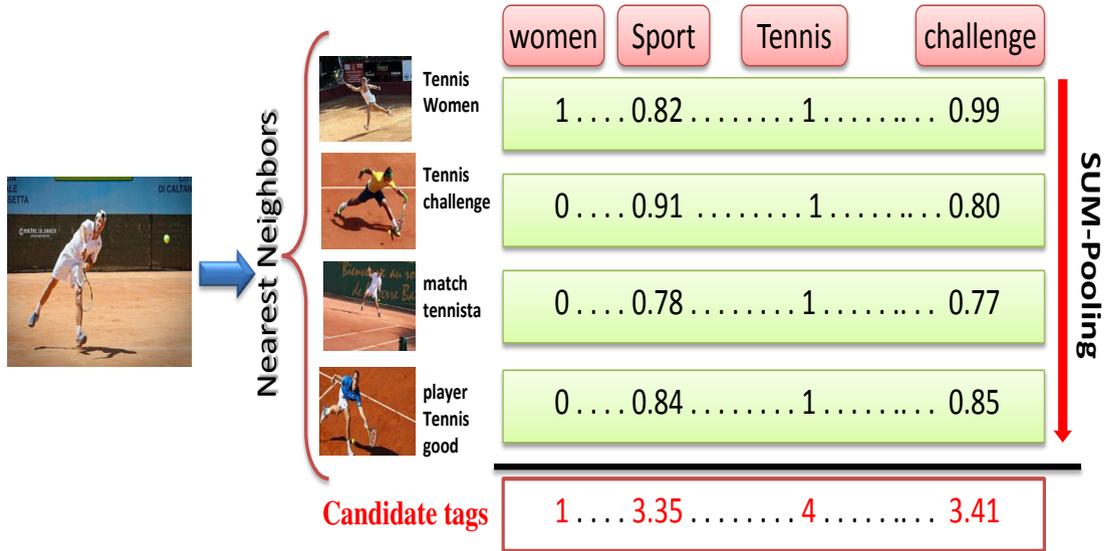


FIGURE 4.5: An example to illustrate the aggregation of nearest neighbor tag descriptions to obtain the list of “candidate tags”.

### 4.3.2 Predicting final tags

We use Belief theory to predict the final tag list. We refer the reader back to Chapter 2-Section 2.5.4 for details about fundamentals of this theory. In the following, we denote  $\Omega$  the set of “candidate tags”.  $\Omega$  represents the frame of all possible hypotheses, called *frame of discernment* in Belief theory. Each pair  $(I^i, t_j)$ , where  $I^i \in \mathcal{N}^v, t_j \in \Omega$ , constitutes a distinct item of evidence regarding the relevance of the tag  $t_j$  to describe the visual content of the untagged image  $I_s$ . If  $I_s$  is “close” to  $I^i$  according to the relevant metric  $d$ , then one will be inclined to believe that both images can be tagged with the same tag. On the contrary, if  $d(I_s, I^i)$  is very large, then the consideration of  $I^i$  will leave us in a situation of almost complete ignorance concerning the tag  $t_j$ . Consequently, this item of evidence may be postulated to induce a Basic Belief Assignment (BBA)  $m(\cdot|I^i)$  over the  $k$ -nearest neighbors.  $m_j(\{t_j\}|I^i)$  represents **the degree of belief associated to the tag  $t_j$** , induced from the image  $I^i$ , about its relevancy in describing the image content of the untagged image.  $m_j(\Omega|I^i)$  represents the degree of belief associated to the whole frame of discernment (other candidate tags). The BBA function  $m(\cdot|I^i)$  is defined as follows:

$$m_j(\{t_j\}|I^i) = \alpha\phi_j(d^i) \tag{4.4}$$

$$m_j(\Omega|I^i) = 1 - \alpha\phi_j(d^i) = 1 - m_j(\{t_j\}|I^i) \tag{4.5}$$

where  $d^i = d(I_s, I^i)$  is the distance between the untagged image  $I_s$  and a neighbor  $I^i$ ,  $\alpha$  is a parameter such that  $0 < \alpha < 1$ . The strength of this evidence,  $m_j(\{t_j\}|I^i)$ , decreases with the distance  $d^i$  and thus  $\phi_j$  is chosen as a decreasing

function verifying  $\phi_j(0) = 1$  and  $\lim_{d \rightarrow \infty} \phi_j(d) = 0$ . As presented in [Denoeux, 1995], one possible choice for the function  $\phi_j$  can be :

$$\phi_j(d) = \exp(-\gamma_j d^2) \quad (4.6)$$

In [Denoeux, 1995], it was proposed to set  $\alpha = 0.95$  and  $\gamma_j$  to the inverse of the mean distance between images tagged with the tag  $t_j$ . This heuristic yields good results on average. These parameters can be determined also by optimizing a performance criterion as shown in [Denoeux, 1995]. For simplicity, we choose the first alternative. For each nearest neighbor of  $I_s$ , a BBA depending on both the tag  $t_j$  and the distance between  $I^i$  and  $I_s$  can therefore be defined resulting in a set of  $k$  BBAs. In order to make a decision regarding the tag assignment of  $I_s$ , these BBA can be combined using Dempster's rule to form a final BBA for each tag contained in this neighborhood. We note  $\mathcal{N}_j^v$  the subset of neighbors from  $\mathcal{N}^v$  tagged with the tag  $t_j$ . Let us first consider two neighbors  $I^i$  and  $I^{i'}$  from  $\mathcal{N}_j^v$ . The BBA resulting from the combination of  $m_j(\cdot|I^i)$  and  $m_j(\cdot|I^{i'})$ , using equations (4.4) and (4.5), is given by:

$$m_j(\{t_j|I^i, I^{i'}\}) = 1 - (1 - \alpha\phi_j(d^i))(1 - \alpha\phi_j(d^{i'})) \quad (4.7)$$

$$m_j(\Omega|I^i, I^{i'}) = m_j(\Omega|I^i) * m_j(\Omega|I^{i'}) = (1 - \alpha\phi_j(d^i))(1 - \alpha\phi_j(d^{i'})) \quad (4.8)$$

Considering the set  $\mathcal{N}_j^v$  of neighbors tagged with  $t_j$ , the combination of the corresponding BBAs can be done as follows:

$$m_j(\{t_j|\mathcal{N}_j^v\}) = 1 - \prod_{i \in \mathcal{N}_j^v} (1 - \alpha\phi_j(d^i)) \quad (4.9)$$

$$m_j(\Omega|\mathcal{N}_j^v) = \prod_{i \in \mathcal{N}_j^v} (1 - \alpha\phi_j(d^i)) \quad (4.10)$$

Now that the BBAs are obtained from each subset of neighbors, a global BBA for all "candidate tags" can be obtained using the normalized Dempsters's rule of combination presented in equation (2.21), as follows:

$$m(\{t_j\}) = \frac{1}{K} (1 - \prod_{i \in \mathcal{N}_j^v} (1 - \alpha\phi_j(d^i))) \prod_{l \neq j} \prod_{i \in \mathcal{N}_l^v} (1 - \alpha\phi_l(d^i)) \quad (4.11)$$

$$m(\Omega) = \frac{1}{K} \prod_{l=1}^n \prod_{i \in \mathcal{N}_l^v} (1 - \alpha\phi_l(d^i)) \quad (4.12)$$

where  $n$  the cardinality of  $\Omega$  and  $K$  is a normalization factor.  $m(\{t_j\})$  represents the degree of belief associated to the tag  $t_j$ , induced from all nearest neighbor images. This step is illustrated in Figure 4.6.

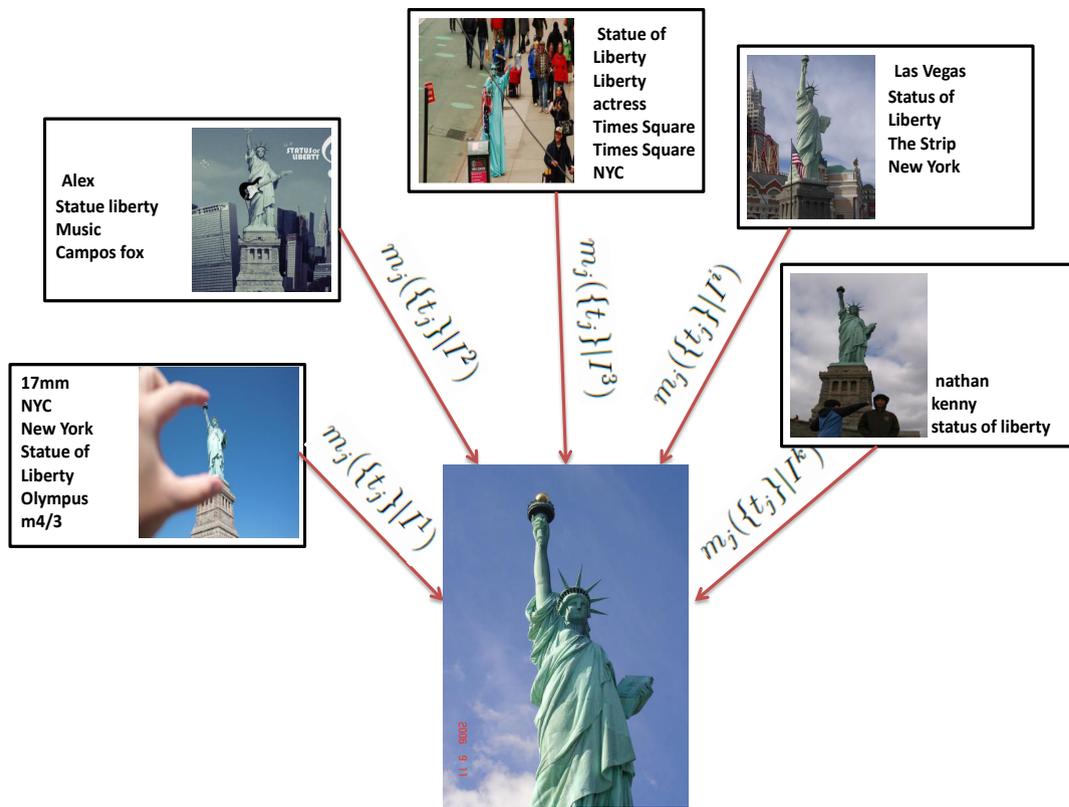


FIGURE 4.6: The  $k$  nearest neighbors BBA for each tag are combined using Dempster's rule of combination to form a final BBA for each tag.

Consequently, the credibility and the plausibility can be defined as follows:

$$Bel(\{t_j\}) = m(\{t_j\}) \quad (4.13)$$

$$Pl(\{t_j\}) = m(\{t_j\}) + m(\Omega) \quad (4.14)$$

Let's note  $A$  the hypothesis "that the tag  $t_j$  is relevant to describe the image content". The credibility  $Bel(\{t_j\})$  quantifies the total amount of *justified specific support* given to the hypothesis  $A$ . We say "justified" because we include in  $Bel(\{t_j\})$  only the basic belief masses given to  $A$ . The plausibility  $Pl(\{t_j\})$  can be viewed as the maximum amount of *potential specific support* that could be given to the hypothesis  $A$ . These two functions can be used to decide if the tag is relevant to describe the image content. In our case, for fair comparison with the state-of-the-art, we choose to sort the list of "candidate tags" by decreasing credibility values and keep only the  $p$  tags with highest values.

## 4.4 Tag Suggestion Dataset Construction

For tag suggestion task, we want to evaluate our method on the dataset used in [Sigurbjörnsson and van Zwol, 2008; Li et al., 2009a]. It consists of 331 images downloaded from Flickr. The selected photos are based on a series of high level topics, for example “basketball”, “Iceland”, and “sailing”, that were chosen by the assessors to ensure that they have the necessary expertise to judge the relevancy of the recommended tags in the photo context. This dataset is created by manually assessing the relevance of user tags with respect to images. Four relevance scale are fixed: *very good*, *good*, *not good*, and *don't know*.

An example of images, given in Figure 4.7, shows that the proposed ground truth do not reflect perfectly the image visual content and thus former evaluation of some systems lead to quite poor results (*e.g* [Li et al., 2009a] obtained below 0.15 mAP and 0.1 Precision@5). Thus, we decide to manually re-annotate the dataset to better reflect the image visual content. For this, we follow a protocol inspired from the collaborative annotation tool of TrecVid [Ayache and Quénot, 2007] showing that annotating a small part of carefully chosen samples of a collection is enough to achieve similar performance (or even better) compared to those obtained with the entire collection. We downloaded all images available on Flickr among the 331, resulting in a collection of 241 images. We run our method as well as two recent ones [Wang et al., 2009a; Li et al., 2009a] on these queries to collect potential tags. Then, we manually annotated the queries by keeping tags that reflect the image visual content.

## 4.5 Experimental Evaluation

The proposed automatic image tagging approach is evaluated in the context of two applications: image classification and tag suggestion. In the former, predicted tags for untagged images are used to compute tag signature based on the LSTC approach presented in Chapter 3-Section 3.6, to learn a SVM classifier. In the latter, tags are used directly to evaluate the method performances. We start by studying the influence of the neighborhood size in Section 4.5.3.1. For image classification, we report results based on two widely used datasets: the PASCAL VOC'07 and ImageClef'11. As part of this chapter, we evaluate only the tag-based feature for the classification task in Section 4.5.3.2. For tag suggestion, a third database is derived from the one used in [Sigurbjörnsson and van Zwol, 2008; Li et al., 2009a] for which we created a new ground truth<sup>1</sup>, by manually annotating 241 queries as explained in Section 4.4. The database used for visual neighbors

<sup>1</sup><http://perso.ecp.fr/~znaidiaa/dataset.html>



<b>Initial ground truth</b>	2005, february	2006 , costume october	2006, Asia, chinese, City travel	2006
<b>Our ground truth</b>	Music, concert Live, Show, Lights, night	People, portrait Makeup, Girl	Baby, sleeping, Bicycle, man Market, Asia	Girl, music Party, Food

FIGURE 4.7: Example of images from the dataset of [Sigurbjörnsson and van Zwol, 2008]. First row represents ground truth proposed by [Sigurbjörnsson and van Zwol, 2008] and the second row represents our annotations used as ground truth for tag suggestion evaluation.

search contains 1.2 million socially tagged images<sup>2</sup> extracted from Flickr. Tag suggestion experiments are presented in Section 4.5.3.3.

### 4.5.1 Datasets

Both the PASCAL VOC'07 [Everingham et al., 2010] and ImageClef'11 [Nowak et al., 2011] datasets were collected from Flickr but they differ significantly. Table 4.1 gives an overview of the proportion of untagged images. As we can see about 38% (respectively 10%) images are not tagged at all in the PASCAL VOC'07 (respectively ImageClef'11) dataset. We refer the reader back to Section 2.6.2 for dataset statistic details (number of images, number of labels, number of tags...).

**Flickr 1.2 million** consists of 1.2 million socially tagged images downloaded from Flickr having no overlap with the untagged images used for test. This collection is used for visual neighbors searching.

### 4.5.2 Experimental Setup

We compared our method, for both image classification and tag suggestion experiments, with two approaches: the Tag Frequency [Wang et al., 2009a] and the Tag

<sup>2</sup><http://staff.science.uva.nl/~xirong/software/tagrel/>

TABLE 4.1: Number and proportion of untagged images in training and test sets, for the PASCAL VOC'07 and ImageClef'11 datasets.

Dataset	# untagged Train	# untagged Test
PASCAL VOC'07 (prop. total)	1917 (38.3%)	1847 (37.3%)
ImageClef'11 (prop. total)	812 (10.1%)	930 (9.3%)

Relevancy [Li et al., 2009a].

- **Tag Frequency** [Wang et al., 2009a]: for a query image, its  $k$ -nearest neighbor images are retrieved from the auxiliary dataset (Flickr 1.2 million dataset) using visual features. Tags associated with these nearest neighbors are treated as an individual item in the text representation. The text signature is a normalized histogram of tag counts from the  $k$ -nearest neighbor images.
- **Tag Relevancy** [Li et al., 2009a]: this approach consists in accumulating votes from visually similar neighbors. In fact, given user-tagged image, they first perform a  $k$ NN search to find its visual neighbors. The tag relevance is determined as the probability that this tag being used to annotate the neighborhood images minus the probability of the tag being used in the entire collection. These probabilities are based on the number of occurrence that the tag appears in the neighborhood, respectively, in the whole dataset.

For the sake of fair comparison, the same processing chain is considered, following literature settings to ensure consistency.

**Searching Visual Neighbors:** The visual similarity between two images is measured by the similarity between their corresponding visual features. Though numerous work have been done for visual feature representation, it is still a challenging problem for content-based image retrieval [Torralba et al., 2008]. For fair comparison, the set of  $k$ -nearest neighbors used as a starting point of our method is determined according to the visual similarity computed between the same visual features as in [Li et al., 2009a]. These visual features are used in our method and those of [Li et al., 2009a; Wang et al., 2009a] to search for visual neighbors. It consists in a combined 64-dimensional global feature for its empirically success in searching millions of web images [Wang et al., 2008]. It consists of a 44-dimensional color correlogram in the 44-bin HSV color space [Huang et al., 1997], 14-dimensional color texture moments [Yu et al., 2003], and 6-dimensional RGB color moments. The three features are normalized to unit length and concatenated into a final 64-dimensional signature. The dissimilarity between images is measured using the Euclidean distance between signature vectors. To look for visual

neighbors, we adopt K-means clustering. First for indexing, the whole dataset is divided into a set of smaller blocks by K-means clustering. Then for a query, we find neighbors within fewer blocks closest to the query. The search space is reduced and thus we decrease the computation cost. For both visual feature extraction and neighbors searching, we use the implementation of [Li et al., 2009a]. We fixed the number of visual neighbors to 100 for both applications (this number is discussed in Section 4.5.3.1).

**BOW-signature:** For the PASCAL VOC'07 textual codebook, we kept only tags that appear at least 8 times, leading to a dictionary of size 804, following the same setting used in [Guillaumin et al., 2010]. In the case of ImageClef'11, we kept tags used at least 3 times, resulting into a textual codebook of 2500 tags. For local soft coding, the neighborhood in the tag feature space was set to 50. The number of tag neighbors has been studied in Chapter 3-Section 3.8.2.2.

**Tag suggestion experiment:** for each method, we select the top 5 tags as final suggestion for each untagged image. For tag suggestion, we evaluate directly the performance on these tags using the precision at rank 5.

**Image classification experiment:** we built a BOW based signature as explained above. A one-versus-all linear kernel based SVM classifier is learned for each method and we compare their performances in terms of mAP.

### 4.5.3 Experimental Results

Before presenting results to both targeted applications (image classification and tag suggestion), we present one experiment to study how results may vary according to the number of visual neighbors considered.

#### 4.5.3.1 Impact of visual neighborhood size

The number of nearest neighbors is an important parameter in tag suggestion methods based on nearest neighbors. To analyze the impact of the neighborhood size, we tried various values of  $k \in \{50, 100, 200, 500\}$  on the PASCAL VOC'07 dataset. We compare the performances of different methods on tag-based image classification in terms of mAP. As shown in Figure 4.8, our method outperforms the two considered methods for all neighborhood size. Both considered methods tend to suggest tags occurring frequently in the neighborhood and treat all neighbors equally while the distances between the image and its neighbors are ignored. By contrast, our method starts by predicting new tags by the LSTC approach to enrich neighbors description and uses the distances to promote the most closed neighbors. In fact, our method reaches the best score (55.2% mAP) with only

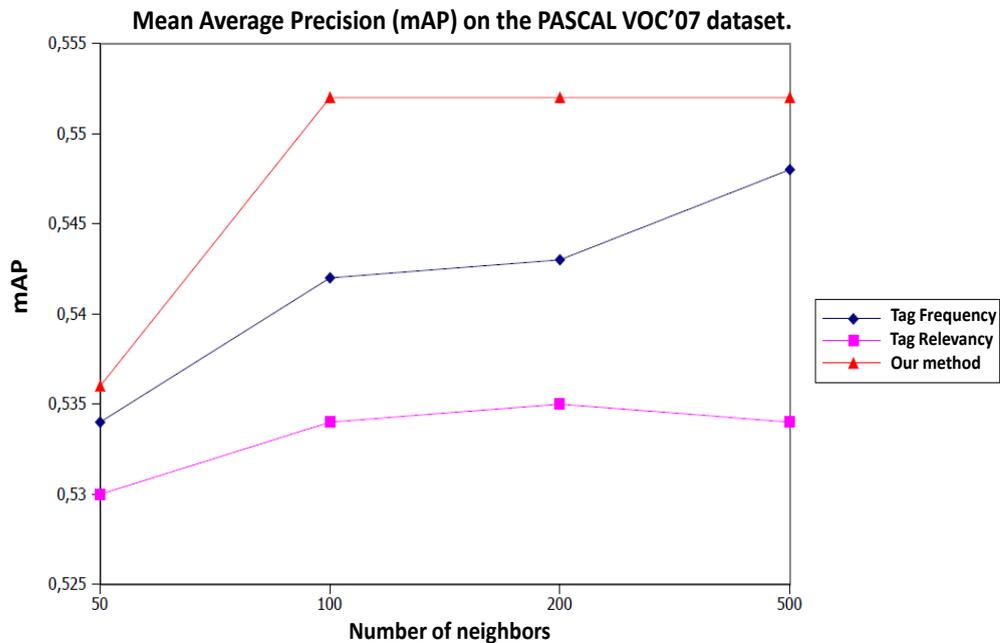


FIGURE 4.8: Performance on the PASCAL VOC'07 dataset in terms of mean Average Precision with respect to the number of nearest neighbors.

100 neighbors and remains stable while varying the neighborhood size. Hence, our method is more effective and stable.

#### 4.5.3.2 Image classification

To show the effectiveness of the tag completion method, we compare in Table 4.2 the classification performances with and without tag completion in terms of mAP for the PASCAL VOC'07 dataset. Results on the ImageClef'11 dataset are shown in Table 4.3. As we can see, the proposed method improves the classification performances for the PASCAL VOC'07 dataset with a gain of 4% in terms of mAP while for the ImageClef'11 we observe that there is no significant improvement. It can be explained by the proportion of untagged images in the PASCAL VOC'07 dataset which reaches 38% of the whole dataset against only 10% for the ImageClef'11 dataset. In fact, on one hand if the number of untagged images is high it can damage the classification performances and thus suggesting tags for untagged images helps to improve the overall classification performances. On the other hand, we admit that not all suggested tags are relevant and far to be perfect. Consequently, if the number of untagged images is small (only 10% of the dataset), we can not improve significantly the classification scores, since we introduce also some noisy tags.

Table 4.4 presents the Average Precision (AP) scores on the PASCAL VOC'07 dataset per concept, without tag completion (Baseline) and with tag completion

TABLE 4.2: Classification performances on PASCAL VOC’07 in terms of mAP with and without tag completion.

Method	mAP
Without tag completion	51.6
With tag completion	<b>55.2</b>

TABLE 4.3: Classification performances on the ImageClef’11 dataset in terms of mAP with and without tag completion.

Method	mAP
Without tag completion	37.0
With tag completion	<b>37.2</b>

TABLE 4.4: The Average Precision (AP) scores on the PASCAL VOC’07 dataset per concept, without tag completion (Baseline) and with tag completion (Our model). The best classification results for each class are marked in bold.

	aeroplane	bicycle	bird	boat	bottle	bus	car
Baseline	75.2	52.0	66.6	47.9	26.5	48.6	61.0
Our model	<b>82.0</b>	<b>52.1</b>	<b>69.2</b>	<b>54.6</b>	<b>26.9</b>	<b>54.7</b>	<b>64.5</b>
	cat	chair	cow	diningtable	dog	horse	motorbike
Baseline	68.0	23.6	50.7	13.7	65.1	73.1	62.6
Our model	<b>68.0</b>	<b>26.7</b>	<b>55.5</b>	<b>20.3</b>	<b>68.3</b>	<b>76.7</b>	<b>64.0</b>
	person	pottedplant	sheep	sofa	train	tvmonitor	<b>average</b>
Baseline	70.8	34.3	54.7	24.5	74.6	38.2	51.6
Our model	<b>74.1</b>	<b>37.7</b>	<b>60.8</b>	<b>26.9</b>	<b>78.3</b>	<b>42.0</b>	<b>55.2</b>

(Our model). The best classification results for each class are marked in bold. As we can see, our model improves the classification performances for all classes. Only for some classes such as “*bicycle, cat*”, there is no improvement compared to the baseline.

In Table 4.5, we compare the results of the textual classifier based on suggested tags for the three methods on the PASCAL VOC’07 dataset. Results on the ImageClef’11 dataset are shown in Table 4.6. By comparing results on both datasets, we can see that the proposed method based on LSTC signature and Belief theory gives better results than the considered methods based on only tag frequency. Our method leads to scores 3% (respectively 1%) better than the considered methods on ImageClef’11 (respectively PASCAL VOC’07) dataset. Over the two datasets, our method clearly dominates the considered methods.

#### 4.5.3.3 Tag Suggestion

In Table 4.7, we report the precision at rank 5 (P@5) on the manually annotated 241 queries. Precision at rank  $k$  is defined as the proportion of suggested tags

TABLE 4.5: Classification performances on PASCAL VOC’07 in terms of mAP, for different methods.

Method	mAP
Tag Relevancy [Li et al., 2009a]	53.4
Tag Frequency [Wang et al., 2009a]	54.2
Our method	<b>55.2</b>

TABLE 4.6: Classification performances on the ImageClef’11 dataset in terms of mAP, for different methods.

Method	mAP
Tag Relevancy [Li et al., 2009a]	33.7
Tag Frequency [Wang et al., 2009a]	34.3
Our method	<b>37.2</b>

TABLE 4.7: Comparison of our system to the state-of-the-art methods on the tag suggestion task.

Method	Average Precision@5
Tag Relevancy [Li et al., 2009a]	0,349
Tag Frequency [Wang et al., 2009a]	0,387
Our method	<b>0,413</b>

that is relevant, averaged over all photos. We consider a predicted tag as relevant with respect to a test image if the tag is from the ground truth tags of the image. As well, we obtain competitive results in tag suggestion task. The tag frequency [Wang et al., 2009a] results are surprisingly better than those of tag relevancy [Li et al., 2009a] on average. It can be explained by the accuracy of visual search which is query-dependent as observed in [Li et al., 2009a].

An example of images with suggested tags by the three methods is illustrated in Figure 4.9. As we can see, original tags are imperfect and most of them are subjective. Let’s note that these tags are not used in the three methods. Obviously, we can observe that tags predicted by our method are more relevant than those predicted by the two considered methods. Our approach is more likely to rank relevant tags ahead of irrelevant ones (shown in bold in Figure 4.9) which is not the case for both tag relevancy and tag frequency methods.

## 4.6 Conclusion and Discussion

In this Chapter, we introduced a novel approach for tag suggestion based on LSTC approach and Belief theory. First, a list of “candidate tags” is created from the visual neighbors of the untagged image, using both local soft coding and

			
<b>Original tags</b>	<i>Cape cod Bass river lighthouse</i>	<i>plants nature cornwall stonehenge 2005.05.03 xato Vwhiz philip anderson</i>	<i>Iceland Reykjavík 2000.09.03</i>
<b>Tag Relevancy</b> [Li et al. 2009a]	architecture house <b>tower</b> <b>flag</b> building	<b>food</b> flower <b>salad</b> <b>strawberry</b> red	boat blue <b>city</b> travel boats
<b>Tag Frequency</b> [Wang et al. 2009a]	architecture <b>water</b> house street <b>car</b>	<b>food</b> flower red nature <b>salad</b>	blue <b>street</b> <b>city</b> <b>canon</b> sky
<b>Our Method</b>	architecture house houses blue sky	flowers flower nature red <b>food</b>	blue boat sky Cloud <b>street</b>

FIGURE 4.9: Examples of tag suggestion by different methods. The **bold** font indicates **irrelevant** suggested tags. Original tags are not used.

two consecutive pooling steps. Then, these tag-signatures are used as pieces of evidence to be combined to provide the final list of predicted tags. This fusion is based on the Dempster's rule of combination, in accordance with the Evidential  $k$ NN framework. Hence, both steps support a scheme to tackle with imprecision and uncertainty, that are inherent to this type of information in a social media context. The experiments that we carried out for image classification on two publicly available datasets show that we obtain comparable or better results than the state-of-the-art methods: on PASCAL VOC'07, results are improved of 3.5% of mAP for textual-only descriptions. On ImageClef'11, our method leads to scores that are above recent state-of-the-art methods. For tag suggestion, we manually annotated 241 queries to propose a new benchmark to the community. For that application as well, we obtained competitive results, with a score 2% of mAP better than the best recent state-of-the-art method.

In fact, in the considered methods, an image is assigned to the tags with the majority votes according to its nearest neighbors, independently of the relevance

of each neighbor. When nearest neighbors have been tagged subjectively by users, noisy tags will be inevitably assigned to the untagged image due to conflicts or lack of knowledge. First, in the local soft coding step, our method gives a degree of confidence about each tag. Second, by exploiting the distance between the untagged image and its nearest neighbors based on Belief theory, we are able to reduce the risk of assigning wrongly some tags to an image when the degrees of confidence are not high. That explains the good performances of our method.

In summary, all experiments show the effectiveness and the robustness of the proposed algorithm for tag-based image classification and automatic tag suggestion for untagged images. Predicting tags for untagged images enables to tackle the problem of full incompleteness making them accessible in tag-based image applications. So far, only tag information is used to handle tag imperfections. The visual information is only used to search for similar visual neighbors. In the next chapter, pixel-based information will be integrated in order to improve image annotation performance while taking into account tag imperfections.

# Chapter 5

## Bag-of-Multimedia Words Representation

### Contents

---

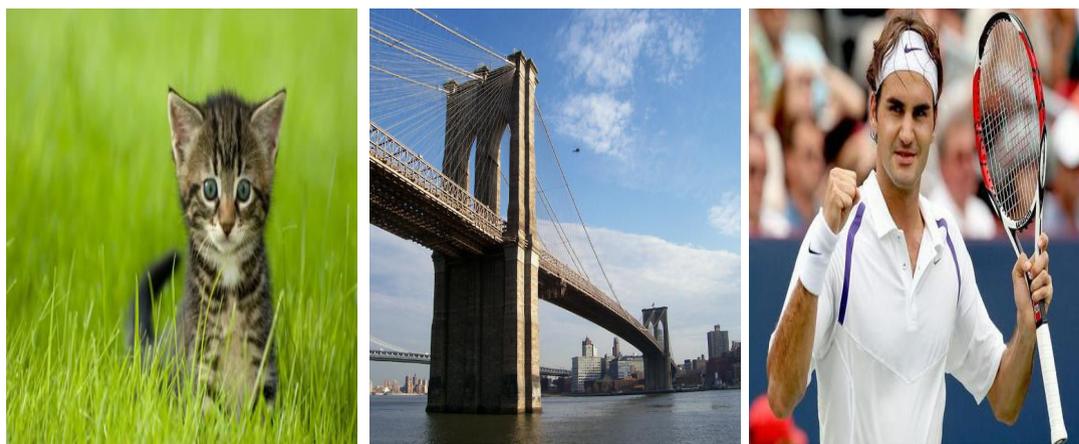
<b>5.1</b>	<b>Introduction</b>	<b>120</b>
<b>5.2</b>	<b>Tag vs. Visual words</b>	<b>121</b>
<b>5.3</b>	<b>Bag-of-Multimedia Words Model</b>	<b>123</b>
5.3.1	Multimedia codebook learning	123
5.3.2	Multimedia signature	125
<b>5.4</b>	<b>Experimental Evaluation</b>	<b>128</b>
5.4.1	Experimental Setup	128
5.4.2	Experimental Results	129
<b>5.5</b>	<b>Conclusion and Discussion</b>	<b>134</b>

---

## 5.1 Introduction

As introduced in Chapter 1, images in social media do not appear alone but associated to various forms of textual metadata such as tags, as shown in Figure 5.1. Our objective in this chapter is to propose a compact multimodal representation which combines both tag and visual information. This representation need to be more appropriate to describe a multimedia document than individual modalities, by taking into account tag characteristics and imperfection aspects.

Describing a document containing both text and pixel-based information faces the problem of their heterogeneous nature. The textual modality is mapped to a dictionary that reflects a language or a sub-part of it in a particular domain, while the visual modality is usually transformed to feature vectors that form a low-level visual description. As presented in Chapter 2, a common approach to tackle the problem of information heterogeneity is to process each modality separately and combine them at the decision level (**late fusion**). An alternative is to work on the description to make it more homogeneous (**early fusion**). A popular representation for document description is the BOW model, introduced in the text community [Salton and McGill, 1983]. In its simplest form, it consists in making an histogram of occurrences of words within a document (term counts). Many refinements have been proposed, such as taking into account the occurrences of words within the collection (inverse document frequency), the length of each document, and so on. This model has been introduced in the image community

**Tags:**

Pussy, young pussy, cute pussy

**Tags:**New York, bridge, brooklyn bridge,  
Manhattan, Bronx, Hudson river,  
East river, helicopter.**Tags:**Roger Federer, popular, tennis,  
Player, Swiss, ranking.

FIGURE 5.1: An example of multimedia documents from Flickr composed with images and their associated user tags.

ten years ago [Sivic and Zisserman, 2003], and its numerous extensions make it one of the most efficient representations used in image classification and retrieval. Visual words are then derived from local features such as SIFT [Lowe, 2004], and the model is then named BOVW. Nevertheless, these descriptions do not directly convey human understandable meaning and a gap remains between them and the semantic content of images [Smeulders et al., 2000].

In this chapter, we propose a semantic signature, called Bag-of-Multimedia-Words (BOMW), for multimedia documents. This signature results from the combination of textual and visual modalities. It is based on *multimedia codewords* that allow on the one hand the cross-coding textual tag-words over visual-words extracted from a document; and on the other hand designing BOMW signature. We exploit the recent advances in BOVW design methods [Yu et al., 2009; Boureau et al., 2010; Wang et al., 2010b; Liu et al., 2011b] in order to provide discriminative BOMW vectors appropriate to multimodal document classification with efficient linear classifiers. This work has been published in [Znaidia et al., 2012c].

The rest of this chapter is organized as follows. Section 5.2 presents a comparison of tags vs visual words. The proposed BOMW signature is presented in Section 5.3. Section 5.4 reports our experimental results on several publicly datasets. The chapter is concluded in Section 5.5.

## 5.2 Tag vs. Visual words

A multimedia document, such as images with their associates user tags in social media, specifically consists of two sources of information: pixel-based and tag-based information. Both information can be used to describe the semantics of image content.

Tags provide contextual and semantic information which can be used to improve the accuracy of image classification [Wang et al., 2009a; Guillaumin et al., 2010]. Such improvements, however, depend on the availability and the quality of tags. As introduced in Chapter 1, tags in community contributed collections are imperfect. An example of images from Flickr website is given in Figure 5.2. As we can see all the images are associated with the tag “zebra” while having different visual content and different semantic meanings. In the first image, the tag “zebra” is related to the concept “animal” while the rest of images are related to the concept zebra as “a striped or cross-hatch pattern”.

Among the recent advances made in image classification, perhaps the most significant one is the representation of images by the statistics of local features, in particular through the BOVW representation [Sivic and Zisserman, 2003]. In the BOVW model, local features extracted from images are first mapped to a set



FIGURE 5.2: An illustration of the tag imperfections. All images are associated with the tag “zebra” while having different visual content and different semantic meanings.

of visual words obtained by a clustering of local feature descriptors (e.g. with k-Means). An image is then represented as an histogram of visual word occurrences. Visual words provide a low-level information to design BOVW signatures. However, the size of a visual-word vocabulary involves a trade-off between the discriminatory power and the computation cost. Indeed, with a small vocabulary, BOVW signature would not be discriminative enough because of assignment ambiguities of local features to codewords. As the size of the learned vocabulary increases, the signature becomes more discriminative, but meanwhile less generative and forgiving to noise, since similar descriptors would be mapped to different codewords. Furthermore, the computational cost for designing BOVW signatures and classifying them grow. Currently, there is no consensus to the appropriate size of a visual vocabulary which varies from several hundreds [Lazebnik et al., 2006], to tens of thousands [Zhao et al., 2006] and even more. Moreover, the same visual word, no matter how local it is, is likely to exhibit quite different visual appearances under different lighting conditions, views, scales and partial occlusions. Although a visual dictionary of a finite collection of visual words may be forcefully obtained by clustering those primitive visual features (e.g., by vector quantization or k-Means clustering), such visual words tend to be much more ambiguous than texts. Specifically, the ambiguity lies in two aspects: *synonymy* and *polysemy* as highlighted in [Yuan et al., 2007]. A synonymous visual word shares the same semantic meaning with other visual words, because the corresponding semantics is split and represented by multiple visual words. On the other hand, a polysemous



FIGURE 5.3: An illustration of polysemous visual words: single visual word occurring on different (but locally similar) parts on different object categories

visual word may mean different things under different contexts. This is the case of the “zebra pattern” in Figure 5.2 which is related to the concept “animal” in the first image and related to the pattern of the “socks”, “road” and “the background” respectively for the other images. An illustration of polysemous visual words is presented in Figure 5.3. As we can see, single visual word occurs in different (but locally similar) parts of different object categories.

Let’s note that we do not take into account visual word imperfections and we focus on the problem of tag imperfections. We define a multimedia word as the elementary part of a multimedia document similar to visual and tag words as elementary parts of an image and its corresponding caption.

## 5.3 Bag-of-Multimedia Words Model

In this section, we propose a more integrated semantic signature for multimedia documents, called Bag-of-Multimedia-Words (BOMW), that results from a combination of textual and visual information. Given an image and its associated tags, its BOMW signature is built in two steps (i) a Multimedia codebook learning, (ii) a Multimedia signature generation.

### 5.3.1 Multimedia codebook learning

The multimedia codebook, which entries are named *multimedia words*, is a collection of basic patterns used to reconstruct the input local features. A simple way to build the multimedia codebook is to perform two steps (1) a tag-coding and (2) a clustering.

### 5.3.1.1 Multimedia word

We denote by  $\mathbf{T}_i$  the set of textual tags associated with an image  $I_i$  and  $\mathbf{T}$  is the set of all textual tags of the training dataset, with  $\mathbf{T}_i \subset \mathbf{T}$  for each image  $I_i$ .

The first step of the multimedia codebook learning consists in expressing each tag of  $\mathcal{W}^t$  over a discrete visual codebook  $\mathcal{W}^v$ . This mapping, that we call **tag-coding**, relies on the fact that textual tag-words are semantically more consistent than visual-words, as it has also been observed in [Monay and Gatica-Perez, 2004].

With a few exceptions [Blei and Jordan, 2003], most previous work assume that words and visual features should have the same importance [Barnard et al., 2003; Wang et al., 2009b]. There are limitations with this assumption. First, the semantic level of textual words is much higher than the one of visual features. Second, in practice, visual feature co-occurrences across images often do not imply a semantic relation between them. This results in a severe degree of visual ambiguity that in general cannot be well handled by existing joint models. Therefore, coding tags over the visual codebook is much more interesting and coherent than the opposite way.

Formally, let  $\mathbf{V}$  be the visual word occurrence matrix learned on the training dataset  $\mathcal{L}$  composed of  $N$  images.  $\mathbf{V}$  is of size  $M' \times N$ , with  $M'$  is the size of a visual codebook  $\mathcal{W}^v$ . The tag-coding matrix  $\mathbf{X}$  has the size  $M' \times M$ , with  $M$  is the size of the textual tag-codebook  $\mathcal{W}^t$ . To build  $\mathbf{X}$ , we sum for each textual tag the visual word occurrences across the list of images tagged with it, i.e.,

$$\mathbf{X}(i, j) = \sum_{I_k \in \mathcal{L}, t_j \in \mathbf{T}_k} \mathbf{V}(i, k), \quad (5.1)$$

with  $I_k$  is the  $k^{th}$  image in the training dataset  $\mathcal{L}$ ,  $t_j$  a tag in  $\mathbf{T}_k$  and  $\mathbf{V}(i, k)$  is the occurrence of the  $i^{th}$  visual word in the image  $I_k$ . This step is illustrated in Figure 5.4. For example, to construct the tag-code for the tag “cat”, we sum the occurrences of visual words of images tagged with “cat”.

The obtained Tag-Codes matrix is then  $l_1$  column normalized, expressing the frequency of a visual word relatively to a tag within the whole training dataset.

The second step, depicted in Figure 5.5, consists in clustering column vectors of the tag-coding matrix  $\mathbf{X}$ , using K-means for instance, in order to generate the multimedia codebook ( $M$ -codebook), which is formed of relevant multimedia words. This step results in the following M-codebook :

$$\mathcal{W}^m = \{\mathbf{m}_i; \mathbf{m}_i \in \mathfrak{R}^{M'}; i = 1, \dots, K^m\}, \quad (5.2)$$

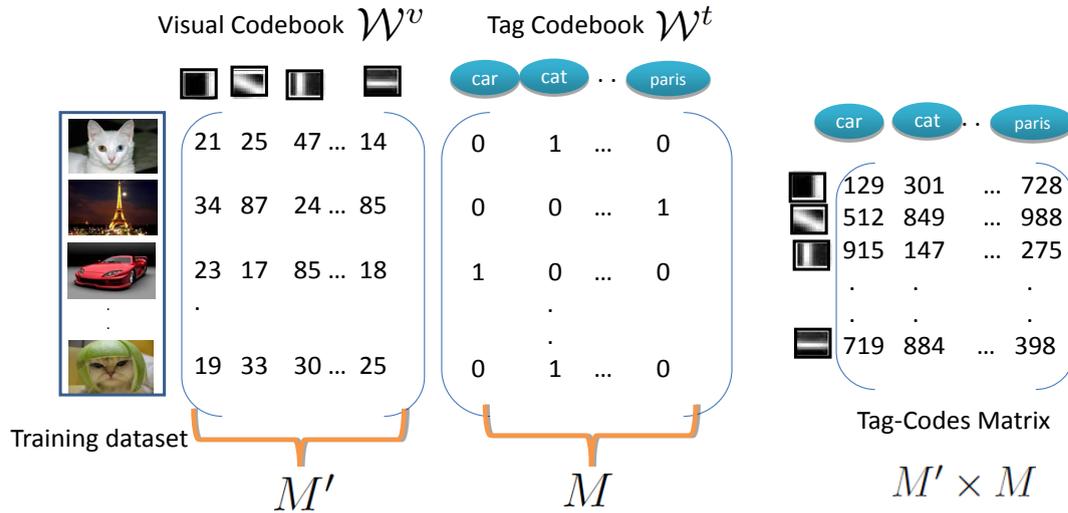


FIGURE 5.4: An overview of Tag-coding procedure. For example to code the tag “cat”, visual word occurrences of images tagged with “cat” are aggregated with a sum-pooling.

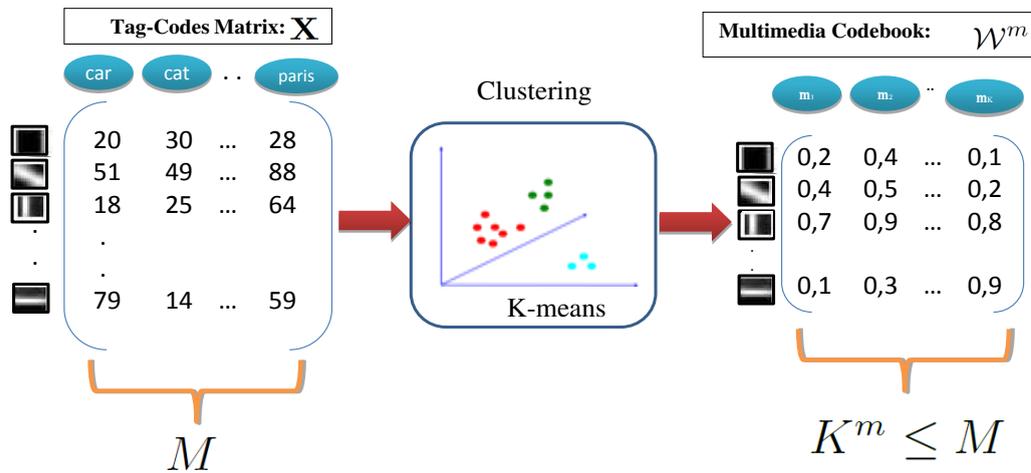


FIGURE 5.5: An overview of the clustering step which consists of clustering column vectors of the tag-codes matrix using K-means in order to generate the multimedia codebook ( $M$ -codebook), which is formed of relevant multimedia words.

with  $K^m$  is the size of the  $M$ -codebook ( $K^m \leq M$ ).

### 5.3.2 Multimedia signature

From the obtained Multimedia codebook, we generate the BOMW signature in two steps: *coding* and *pooling*, as shown in Figure 5.6.

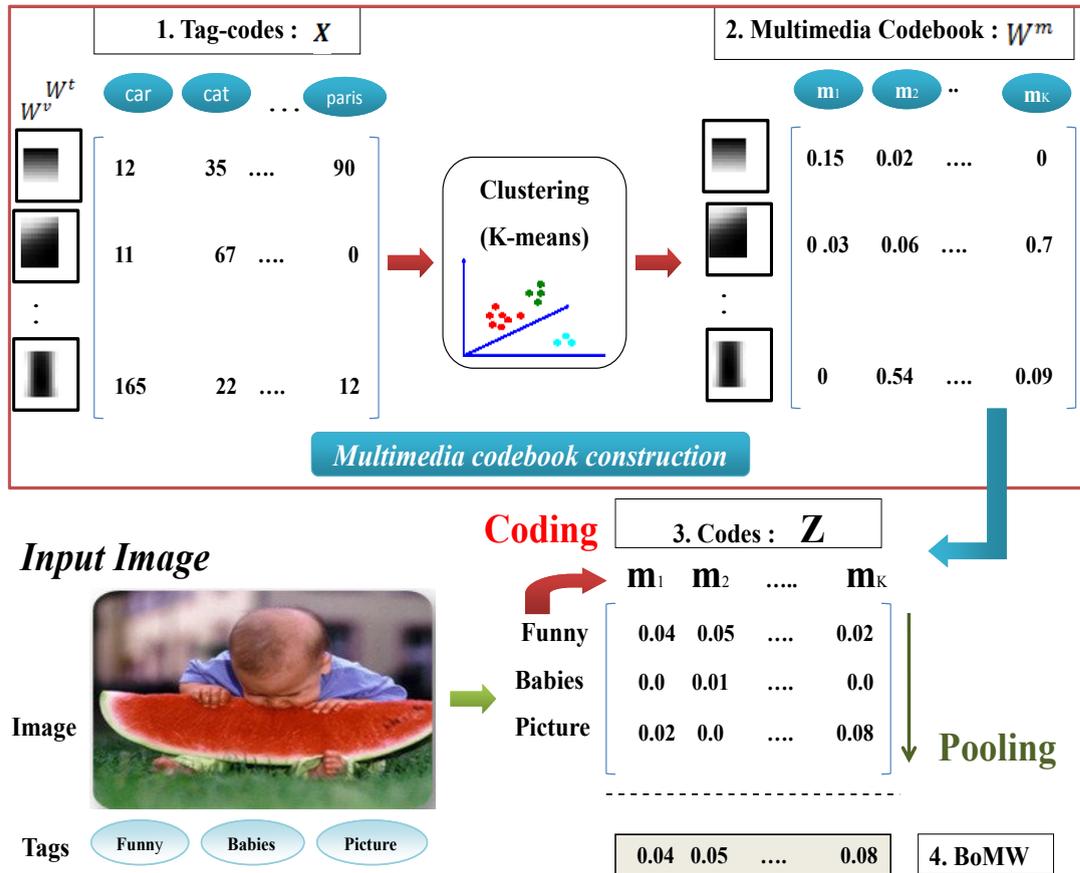


FIGURE 5.6: BOMW signature generation consists in two steps: coding and pooling. Given an image with its associated tags, each tag is represented with its tag-code using the tag coding matrix  $X$  and coded based on the locality constraint coding using the learned multimedia codebook. Finally, given the coding coefficients of all tags within one image, a pooling operation is performed to obtain the BOMW signature.

### 5.3.2.1 Coding

As introduced in Chapter 2-Section 2.3.1, different methods have been investigated in the literature in order to map local features to codes over the visual codebook, preserving some interesting properties such as sparsity [Yang et al., 2009], locality in the feature space [Yu et al., 2009], saliency [Huang et al., 2011], etc. These coding schemes alleviate the main drawbacks of classic coding ones, namely hard and soft assignments [Sivic and Zisserman, 2003; van Gemert et al., 2009]. The locality based coding is currently the most interesting technique in terms of trade-off between robustness and computational complexity. In [Liu et al., 2011b] for instance, authors propose an efficient implementation of the locality-constrained coding presented in [Yu et al., 2009] by restricting the probabilistic soft coding of the approach of [van Gemert et al., 2009] to the  $L$ -nearest-codewords to a descriptor in the feature space.

In our case, a tag-code  $\mathbf{x}_k$  (a column of  $\mathbf{X}$ ) of a given image is the descriptor to be coded over the M-codebook  $\mathcal{W}^m$  as the following:

$$z_{k,i} = \begin{cases} \frac{\exp(-\beta \|\mathbf{x}_k - \mathbf{m}_i\|_2^2)}{\sum_{r=1}^L \exp(-\beta \|\mathbf{x}_k - \mathbf{m}_r\|_2^2)} & \text{if } \mathbf{m}_i \in \mathcal{N}_L(\mathbf{x}_k), \\ 0 & \text{otherwise,} \end{cases} \quad (5.3)$$

where  $\mathbf{z}_k$  is a vector of size  $K^m$ . It is the obtained code associated to the tag-code  $\mathbf{x}_k$ ,  $\mathcal{N}_L(\mathbf{x}_k)$  denotes the set of  $L$ -nearest neighbors to the vector  $\mathbf{x}_k$  within the tag-code set of column vectors in tag-coding matrix  $\mathbf{X}$  and  $\beta$  is a parameter controlling the weight decay speed of the locality. An illustration of the BOMW coding step is presented in Figure 5.7. The “Dog” tag-code represents the column vector corresponding to the tag “Dog” and obtained from the tag-coding matrix  $\mathbf{X}$ . In this example, we consider the number of neighbors equal to three. Thus, only the three nearest multimedia words are activated and the rest are set to zero.

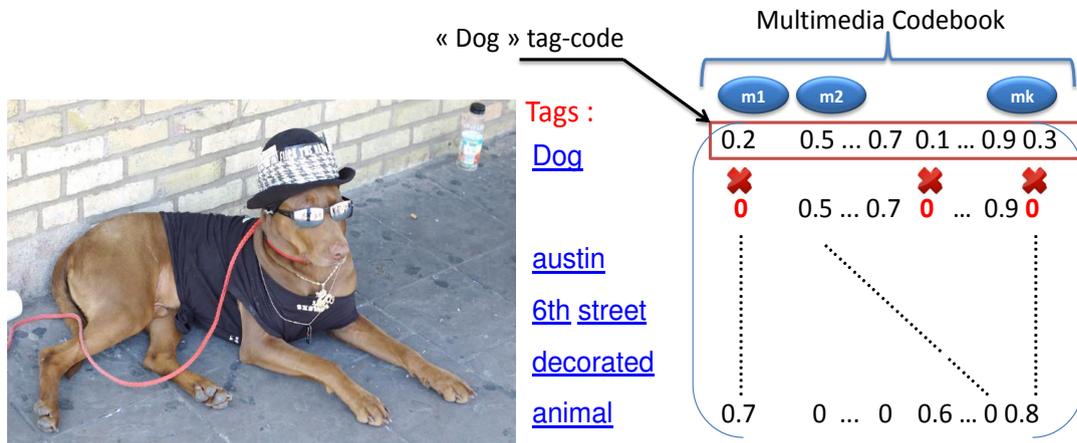


FIGURE 5.7: Illustration of the BOMW coding step. In this example only the three nearest multimedia words are activated and the rest are set to zero.

### 5.3.2.2 Pooling

Given the coding coefficients of all tags within an image  $I_j$ , the pooling step is performed to obtain the BOMW signature  $\mathbf{X}_j^m = (\mathbf{x}_1^m, \dots, \mathbf{x}_{K^m}^m)$ . The element of the BOMW signature are defined as follows:

$$\mathbf{x}_i^m = \max_{t_k \in T_j} z_{k,i} \quad \forall i = 1, \dots, K^m \quad (5.4)$$

where  $T_j$  is the set of user tags associated to the image  $I_j$ . Recent work [Boureau et al., 2010; Liu et al., 2011b] show, theoretically and empirically, that max-pooling is well suited to the recognition task. It is performed by selecting the maximum

coding coefficient (or the salient codeword response) over tag-codes for each multimedia word as shown in Figure 5.6.

## 5.4 Experimental Evaluation

The proposed method is evaluated in the context of image classification. We report results based on two widely used datasets: PASCAL VOC'07 and ImageClef'12 described in Section 2.6.2.

Classification performances of the proposed method on the considered datasets are evaluated in terms of three criterion: 1) classification performance (mAP), 2) computation cost for both signature design and classifier training and testing and 3) stability of results toward codebook size. These three issues are the most challenging ones in the classification task, since they involve robustness of the recognition system and its scalability on large scale datasets.

### 5.4.1 Experimental Setup

We consider the following setup. Dense SIFT features are extracted from images within a regular spatial grid at only one scale. The step-size is fixed to 6 pixels and the patch size to  $16 \times 16$  pixels. Visual codebooks of various sizes have been generated using the K-means clustering method on randomly selected SIFTs from the training set. For the PASCAL VOC'07 dataset, a textual codebook is also generated using the same experiment setting of [Guillaumin et al., 2010] leading to a dictionary of size 804. For the ImageClef'12 dataset, we keep only tags that appear at least 4 times leading to a textual dictionary of size 5,134. For each dataset, once the tag-coding matrix has been created, the multimedia codebook (M-codebook) is generated by clustering columns of the tag-codes matrix using K-means. In order to analyze the robustness of the BOMW signature toward codebook size, we tried different visual and multimedia codebook sizes. When designing BOMW, coding tag-codes over the M-codebook is performed using the locality-constrained soft assignment with a neighborhood of size 5 and the softness parameter  $\beta$  is set to 10, as it has been also considered in [Liu et al., 2011b]. Finally, we used linear SVM for classification.

As a baseline, we consider two models: the BOTW and the BOVW models.

- **Bag-of-Tag Words (BOTW)** represents an histogram of occurrences of each tag according to a fixed textual dictionary. Generally, in the case of tags associated with images, a tag is present once. Consequently, this histogram is reduced to a binary vector which encodes the presence or the absence of

each tag in a fixed vocabulary. This BOTW signature has the same size as the textual codebook. This vector is used as an input vector for a linear SVM classifier.

- **Bag-of-Visual Words (BOVW)** represents the image with an histogram of visual words. In our case, we use the locality-constrained soft assignment with a neighborhood of size 5 for the coding step. The max-pooling operation is performed to aggregate the obtained codes and a spatial pyramid decomposition into 3 levels ( $1 \times 1$ ;  $2 \times 2$ ;  $4 \times 4$ ;) is adopted to generate the visual signature. The size of the BOVW signature is equal to  $(4096 \times 21)$ . This vector is used as an input vector for a linear SVM classifier.

## 5.4.2 Experimental Results

In this section, we present classification performances obtained for each model (BOTW, BOVW and BOMW) in terms of mAP. We choose the PASCAL VOC'07 dataset to show the stability of our BOMW model towards the size of both visual and multimedia codebooks.

### 5.4.2.1 Experiments on the PASCAL VOC'07 dataset

In Table 5.1, we compare the classification performances of our method to the two considered baselines in terms of mAP on the PASCAL VOC'07 dataset. We observe that the proposed BOMW model outperforms both the BOTW and the BOVW models. We obtain a gain of 12% in terms of mAP compared to the textual modality. An improvement of 6% of mAP scores is achieved compared to the visual modality. The signature size in Table 5.1 shows that our BOMW signature is more compact than the two other BOWs (tags and image).

To show the stability of the proposed BOMW model towards the size of the visual and multimedia codebooks, we tried various values of the visual codebook size  $K^v \in \{256, 512, 1024, 4096\}$  and the multimedia codebook size  $K^m \in \{128, 256, 512\}$  on the PASCAL VOC'07. We compare the performances of different methods in terms of mAP. Figure 5.8 shows classification performances using

TABLE 5.1: Classification performances on the PASCAL VOC'07 dataset in terms of mean Average Precision (mAP).

Method	mAP	Signature size
Bag-of-Tag Words (BOTW)	43.3	804
Bag-of-Visual Words (BOVW)	49.3	4096*21
Bag-of-Multimedia Words (BOMW)	<b>55.5</b>	512

either BOVW or BOMW, while changing the sizes of both visual and multimedia codebooks. For the BOVW, we note that the spatial pyramid matching technique of [Lazebnik et al., 2006] is additionally performed with three pyramid levels, as often done in the literature to take into account the spatial information.

We note that for all codebook sizes, classification results using BOMW outperform by about 6% ~ 10% those obtained with BOVW (depicted in Figure 5.8) or by the BOTW obtained by [Guillaumin et al., 2010] which is 43.3%. This is expected since the proposed multimedia words are semantically higher than the low-level visual local features and the tag words separately. Multimedia words are more consistent to encode the content of a multimedia document through an effective fusion of visual and text modes.

The important point is that contrary to classic BOW signatures, classification results remain stable, with a very low fluctuation, when changing sizes of both visual or multimedia codebooks. This is an interesting property useful to reduce the complexity of the classification system in both training and testing, which is obtained at the cost of a small pre-processing step for signature design (building tag-coding matrix and clustering it). The best classification scores obtained with the BOVW and BOTW are 49.36% and 43.3% respectively, using visual signatures of size 86,016 and textual signatures of size 804. The best classification score obtained with the BOMW is 55.54% using a signature of size 512 (see Figure 5.8). The performance gain is due to the fact that the proposed multimedia signatures lie on a structured space, well appropriate to describe multimedia documents. Therefore, BOMW are probably much more class-discriminative than other types of BOW.

The proposed multimedia signature could also be integrated in late fusion classifiers as it captures complementary information to those already used in the literature. Let's note that once the tag-coding matrix and the M-codebook are generated, the proposed BOMW signature is obtained using only tags. Thus, we choose to combine it with visual features (BOVW). Obtained results with late fusion, by averaging classifier predictions, are presented in Figure 5.9. Obviously, we observe that the combination of BOMW with BOVW outperforms the best scores reported on the state-of-the-art by [Guillaumin et al., 2010; Durand et al., 2013] on the PASCAL VOC'07 dataset.

#### 5.4.2.2 Experiments on the ImageClef'12 dataset

Table 5.2 shows a comparison of the proposed BOMW model and the two considered baselines: BOTW using only tags and the BOVW using only visual information. We can see that the BOMW outperforms both the BOTW and BOVW models. An improvement of 10% in terms of mAP scores is achieved compared to

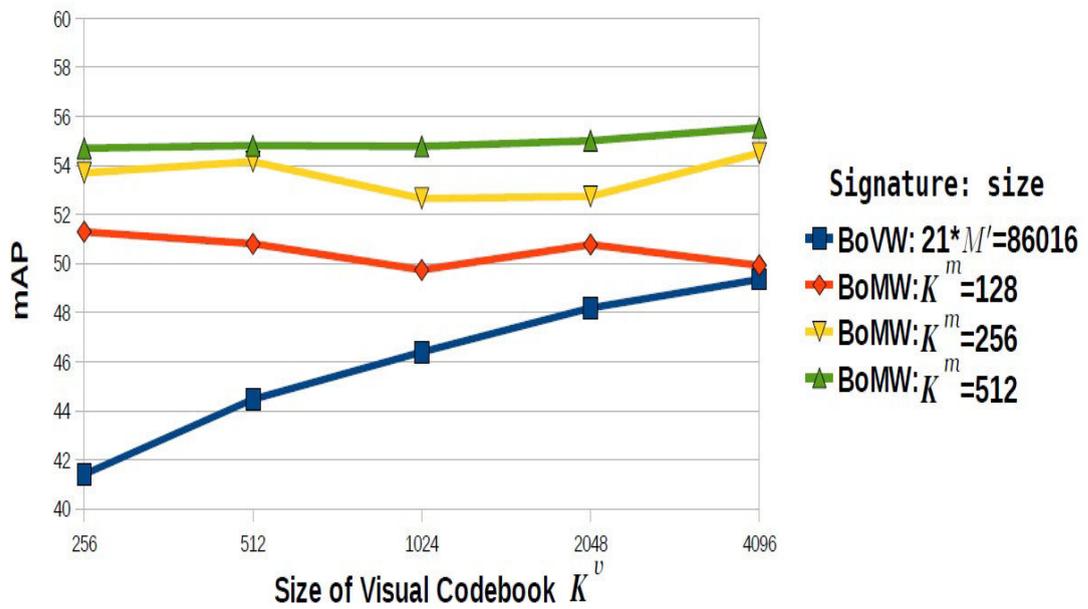


FIGURE 5.8: Classification performances in terms of mAP on the PASCAL VOC'07 dataset while varying the sizes of visual and multimedia codebooks.

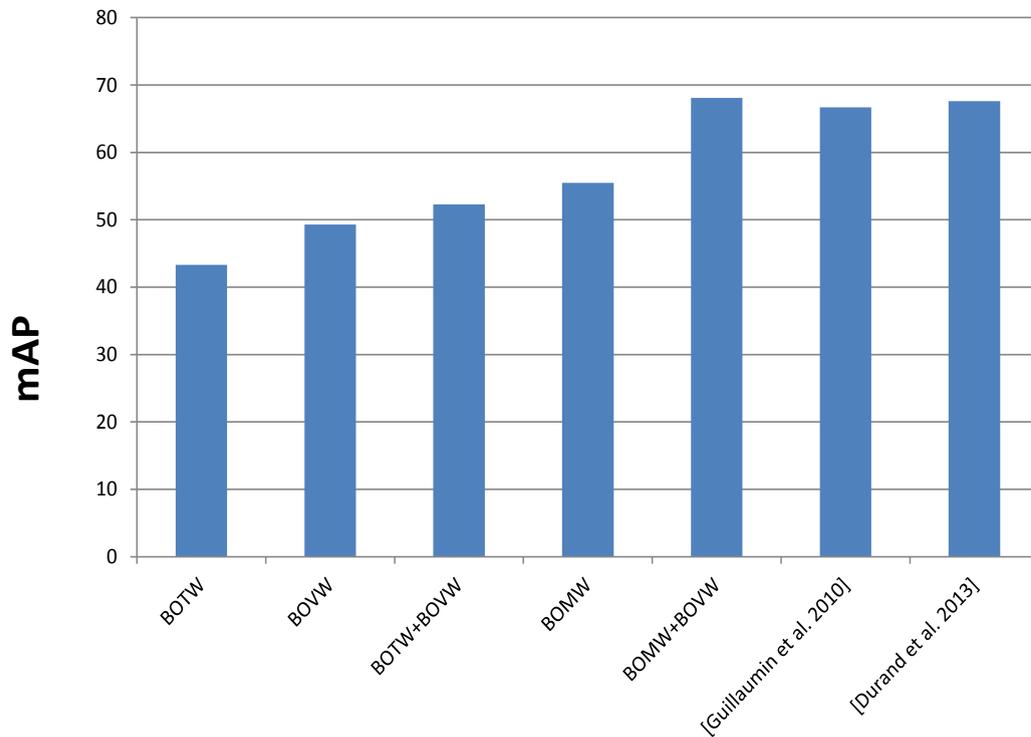


FIGURE 5.9: Classification performances in terms of mAP on the PASCAL VOC'07 dataset.

the textual modality only. We obtain a gain of 11% compared to the BOVW model which uses only visual information. In addition the proposed BOMW signature is more compact than the BOTW and BOVW models.

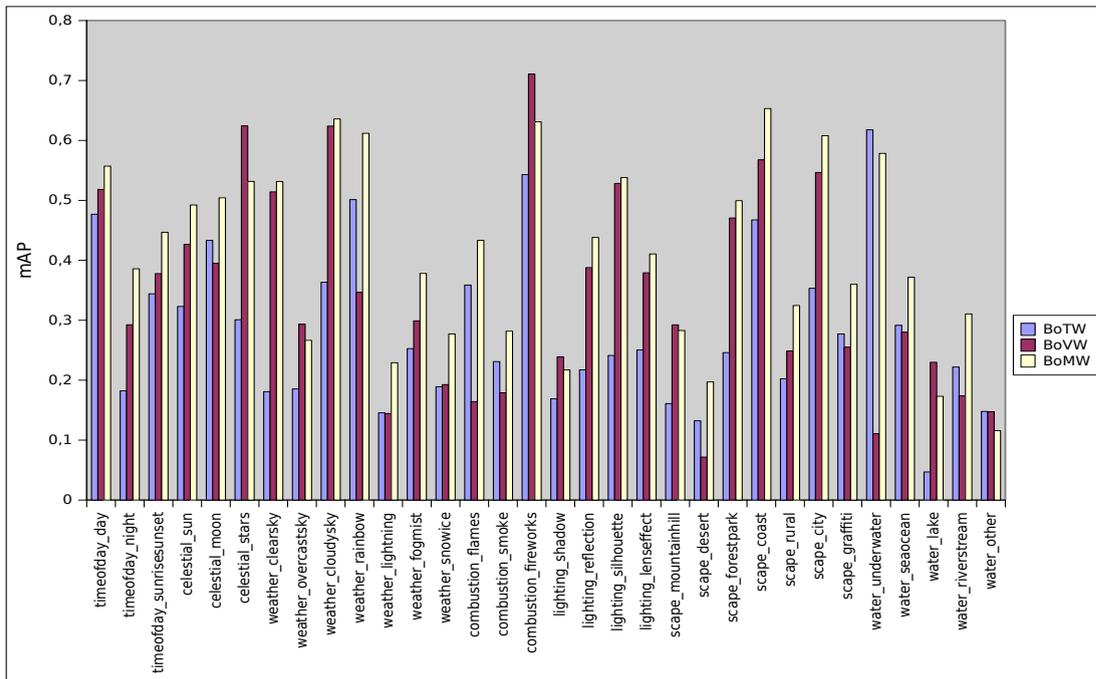


FIGURE 5.10: Comparison of the mAP on the ImageClef’12 dataset for the BOTW and the BOVW vs the proposed BOMW.

As presented in Figures 5.10, 5.11 and 5.12, the proposed BOMW outperforms both BOTW and BOVW models on the classification performances on 80 concepts out of 94. Only for 14 concepts, it fails to improve the Average Precision score. These concepts are “celestial\_stars, weather\_overcastsky, combustion\_fireworks, lighting\_shadow, water\_underwater, water\_seaocean, water\_lake, water\_riverstream, water\_other, fauna\_spider, age\_teenager, age\_elderly, quality\_completeblur, style\_pictureinpicture, style\_circularwarp, transport\_truckbus”. Six of these concepts are not visual concepts such as “water\_other, style\_pictureinpicture, age\_teenager, age\_elderly, quality\_completeblur, style\_circularwarp”. Our BOMW signature is based on the visual representation of images tagged with a certain tag, thus, it can fail for this kind of concepts.

TABLE 5.2: Classification performances on the ImageClef’12 dataset in terms of mAP score.

Method	mAP	Signature size
Bag-of-Tag Words (BOTW)	30.0	5134
Bag-of-Visual Words (BOVW)	29.4	4096*21
Bag-of-Multimedia Words (BOMW)	<b>40.8</b>	2500

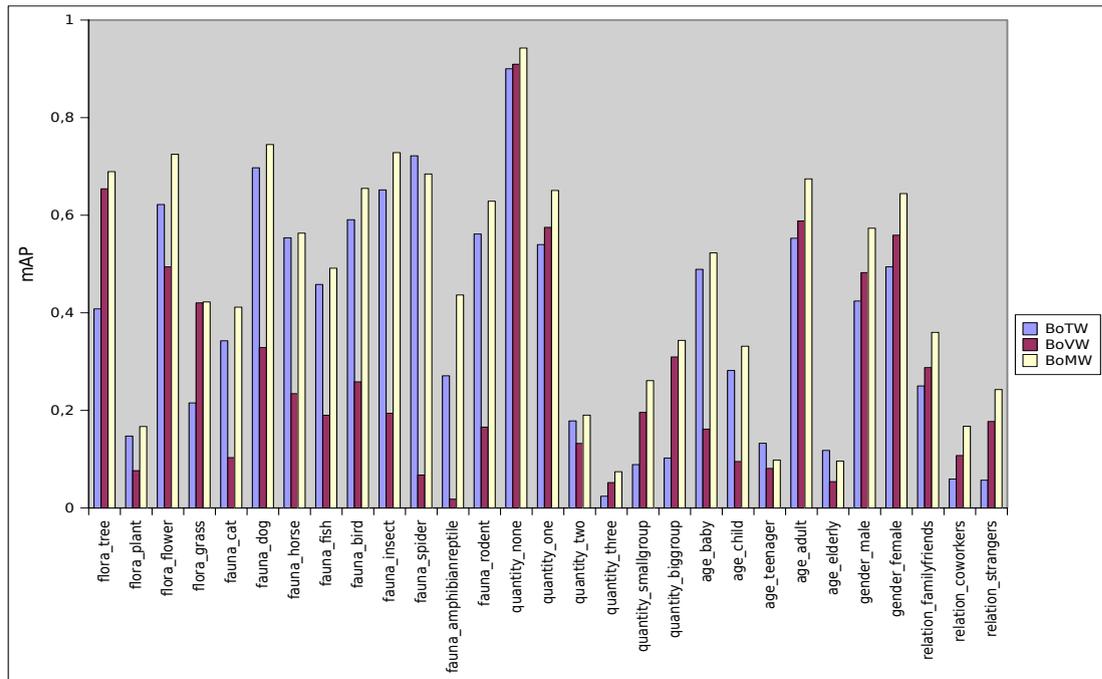


FIGURE 5.11: Comparison of the mAP on the ImageClef'12 dataset for the BOTW and the BOVW vs the proposed BOMW.

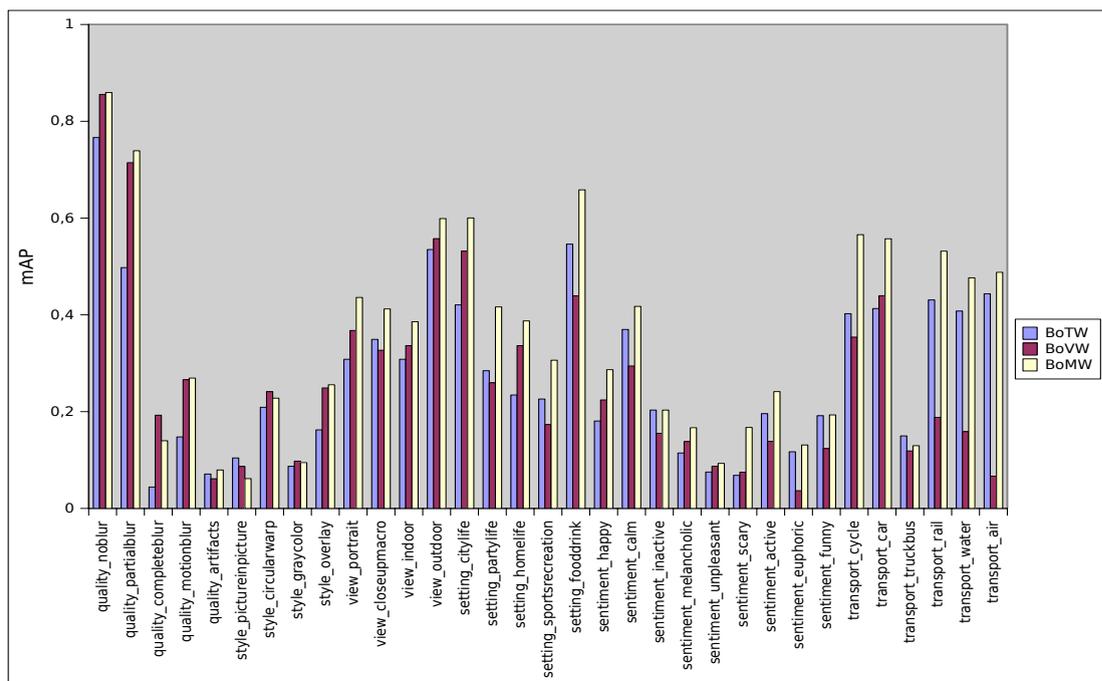


FIGURE 5.12: Comparison of the mAP on the ImageClef'12 dataset for the BOTW and the BOVW vs the proposed BOMW.

## 5.5 Conclusion and Discussion

We introduced a new BOW based signature, called BOMW, that is appropriate to describe multi-modal documents. It represents a more integrated semantic signature for multimedia documents than the classic BOW signatures, that results from a combination of textual and visual information. It is based on *multimedia code-words* that allow on the one hand cross-coding textual tag-words over visual-words extracted from a document; and on the other hand designing BOMW signature. We exploit the recent advances in BOVW design methods in order to provide discriminative BOMW vectors well suited to multimodal document classification with efficient linear classifiers.

Experiments have been conducted on two well-known challenging benchmarks: PASCAL VOC'07 and ImageClef'12. Obtained results show the competitive performances of the BOMW, ensuring a trade-off between classification accuracy and computation cost. In opposition to classic BOW signatures, classification results remain stable, with a very low fluctuation, when changing sizes of the visual or multimedia codebooks. This is an interesting property useful to reduce the complexity of the classification system in both training and testing, which is obtained at the cost of a small pre-processing step for signature design (building tag-coding matrix and clustering it). The performance gain is due to the fact that the proposed multimedia signature lies on a structured space, well appropriate to describe multimedia documents. Therefore, BOMW are probably much more class-discriminative than other types of BOW. The tag-coding matrix step is based on a predefined dictionary built with the high frequent tags. In this manner, we eliminate rare, misspelled and subjective tags to handle a part of the imprecision and uncertainty aspects of tags. The clustering step to obtain the multimedia codebook can be seen as a reduction of the space of tags into topics as in the pLSA model [Monay and Gatica-Perez, 2004]. In fact, tags in different languages such as “*cat*” in English and “*chat*” in French will have similar tag-codes. Thus, the clustering step is very useful to reduce the tag-codes space by gathering similar tag-codes together. This step tackles also a part of the uncertainty and imprecision problems.

The difference between our BOMW approach and some recent state-of-the-art approaches [Li et al., 2010a; Torresani et al., 2010], is that we apply an early fusion to combine both modalities and only tags are used in the test stage which is not the case of Object Bank approach [Li et al., 2010a] and Classemes method [Torresani et al., 2010]. In fact, in [Li et al., 2010a], authors propose a high-level image representation where an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors. In [Torresani et al., 2010], authors propose a visual descriptor which is the output of a large number of weakly trained object category classifiers on the image. Both methods are based on a learning stage of concepts.

Up to now, only tag imperfections have been handled at the representation level by using either the tag modality alone or by combining both tag and visual modalities. As introduced in the Chapter 1, image annotation is subject to other type of imperfections at the decision level. These imperfections need to be handled to improve the image annotation performances.

## Part II

### Decision Level

# Chapter 6

## Multimodal Late Fusion for Image Annotation

### Contents

---

<b>6.1</b>	<b>Introduction</b>	<b>138</b>
<b>6.2</b>	<b>Proposed Multimodal Framework for Image Annotation using Stack Generalization</b>	<b>138</b>
6.2.1	Visual Features	139
6.2.2	Textual Features	140
6.2.3	Stack Generalization	140
<b>6.3</b>	<b>Experimental Evaluation</b>	<b>143</b>
6.3.1	Experimental Setup	143
6.3.2	Experimental Results	144
<b>6.4</b>	<b>Conclusion and discussions</b>	<b>147</b>

---

## 6.1 Introduction

As introduced in Chapter 1, multimodal image annotation is subject to imperfection at two possible levels: representation and decision. Up to now, only imperfection aspects at the representation level are handled. In this chapter, we are interested in handling imperfections at the decision level, that can exist at the fusion process when combining information from different classifiers. Although, tags represent an important resource to improve multimodal image annotation, they are generally imperfect and only a few of them are really related to the visual content of the image. These “imperfections” recovers different problems including imprecision, uncertainty and incompleteness. In fact, learning from imperfect tags can decrease classification performances. But even if observations are perfect, the generalization beyond that data and the process of induction, are still afflicted with **uncertainty**. Another form of imperfection is **incompleteness** of data. Handling incomplete data (images without tags) is an important issue for classifier learning since incomplete data may affect the prediction accuracy of learned classifiers. Regarding **imprecision**, learning a classifier on uncertain and incomplete data leads to an imprecise decision function.

In this chapter, we introduce an unified multimodal framework for semantic image classification based on a two novel textual representations presented in Chapter 3 along with visual features through an effective and robust scheme of late fusion based on the Stack Generalization algorithm [Wolpert, 1992].

The rest of this Chapter is organized as follows. In Section 6.2, we present the proposed method for classifier combination based on the Stack Generalization algorithm. Section 6.3 reports our experimental results on several publicly datasets compared to the state-of-the-art approaches. The chapter is concluded in Section 6.4.

## 6.2 Proposed Multimodal Framework for Image Annotation using Stack Generalization

Figure 6.1 depicts the flowchart of the proposed multimodal framework for image annotation using Stack Generalization algorithm, which mainly includes two stages: a training stage and a testing stage. The training dataset is split into training and validation sets. The training stage consists in training classifiers through a learning algorithm on the training set and in evaluating it on the validation set. The process is repeated using a cross-validation procedure. The output prediction scores from different classifiers on validation sets are concatenated and used as input features to learn a new classifier. This latter is represented with multimodal

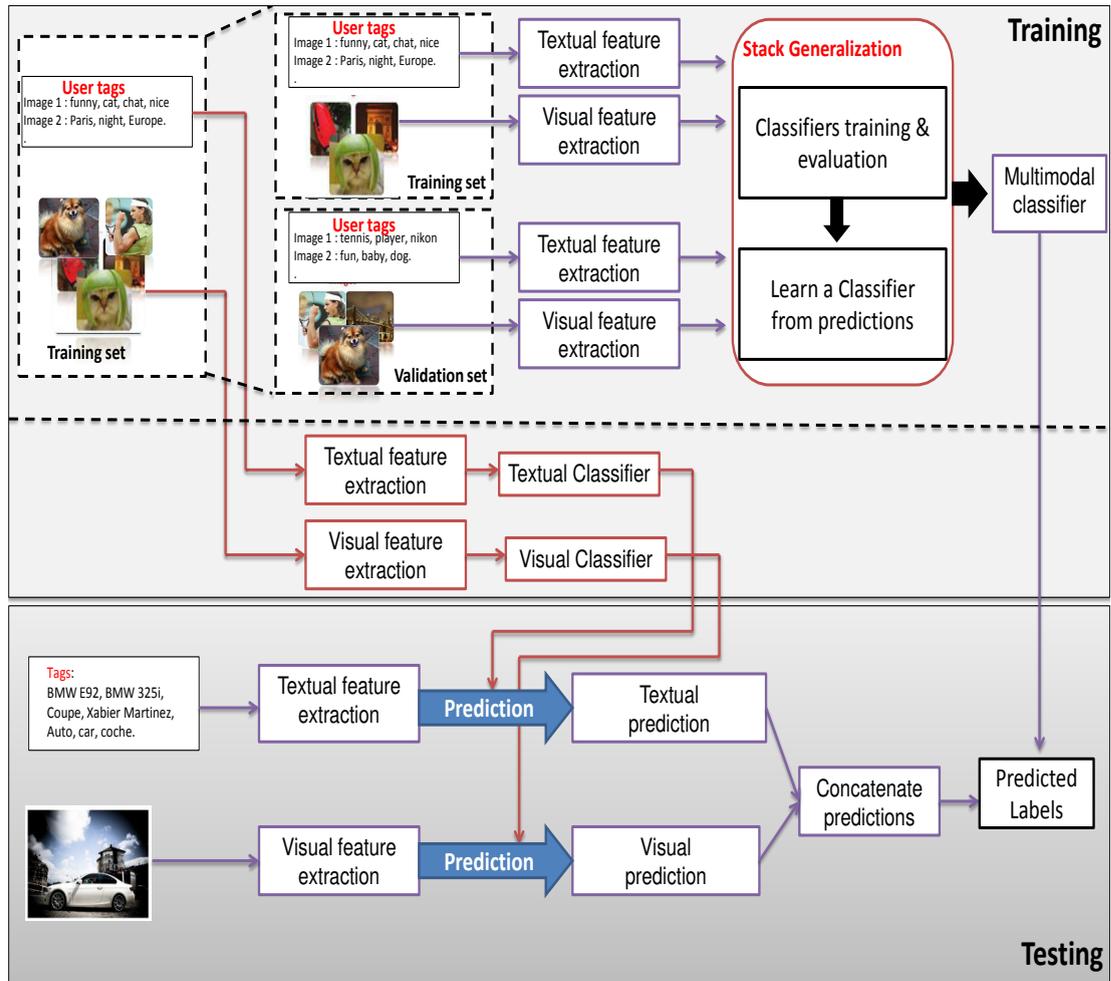


FIGURE 6.1: The flowchart of the proposed multimodal framework for image annotation using Stack Generalization.

classifier in Figure 6.1. Classifiers using the whole training set are also learned and used in the testing stage to provide predictions from different modalities. Finally, obtained predictions from different classifiers are concatenated and used as a test feature in the multimodal classifier to obtain the final predicted labels. This scheme of combining classifiers is called Stack Generalization or Stacking [Wolpert, 1992]. Stack generalization represents a general method of using a high-level model to combine lower-level models to improve accuracy of single classifiers.

### 6.2.1 Visual Features

As introduced in Chapter 2-Section 2.3.1, the BOVW approach [Sivic and Zisserman, 2003; Csurka et al., 2004] has now established itself as the state-of-the-art for generic image classification. It commonly consists of feature extraction, codebook creation, feature coding, and feature pooling.

First, a visual codebook is constructed using K-means algorithm. Then, for each image, dense local descriptors (such as SIFT [Lowe, 2004]), are extracted and mapped to codes. Recent research shows for a given visual codebook, how to code each local feature and how to pool the coding coefficient to obtain an image-level representation, have a significant impact on classification performance. Following these observations, we chose to implement the locality-constraint coding based on local soft coding [Liu et al., 2011b], because of its effectiveness and robustness against quantization errors. Final codes result from a max-pooling aggregation. The superiority of max-pooling over other pooling methods, combined with such coding scheme, can be explained probabilistically as being the lower bound of the probability of occurrence of a visual word in the image [Liu et al., 2011b].

For the coding step, we use the locality-constrained coding which restricts the probabilistic soft coding approach [Liu et al., 2011b] to only the  $L$ -nearest-codewords to a local feature.

Furthermore, since the classic BOVW is an orderless signature that disregards the location of the visual words in the image, the spatial pyramid matching (SPM) [Lazebnik et al., 2006] is an interesting way to incorporate some global spatial contextual information into the signature. An image  $\mathbf{I}_k$  is divided into  $P$  different regions and a pooling is conducted in each of them. The final signature  $\mathbf{X}_k^v$ , is then obtained by a concatenation of all the region-relative  $R_i$  signatures, as follows:

$$\mathbf{X}_k^v = [\mathbf{X}_{(k,R_1)}^v, \mathbf{X}_{(k,R_2)}^v, \dots, \mathbf{X}_{(k,R_P)}^v] \quad (6.1)$$

where  $\mathbf{X}_{(k,R_i)}^v$  is the BOVW signature of the image  $\mathbf{I}_k$  in the region  $R_i$ . A schematic illustration of the spatial pyramid representation is depicted in Figure 6.2.

## 6.2.2 Textual Features

In this chapter, we use the LSTC signature presented in Chapter 3 to produce two tag-based signatures using two external knowledge resources: WordNet and Flickr. We refer the reader back to Section 3.6 for more details.

## 6.2.3 Stack Generalization

Stacked Generalization or stacking was introduced by [Wolpert, 1992] as an approach for combining multiple classifiers. The key idea is to learn a meta-level (or level-1) classifier based on the output of base-level (or level-0) classifiers, estimated via cross-validation as follows.

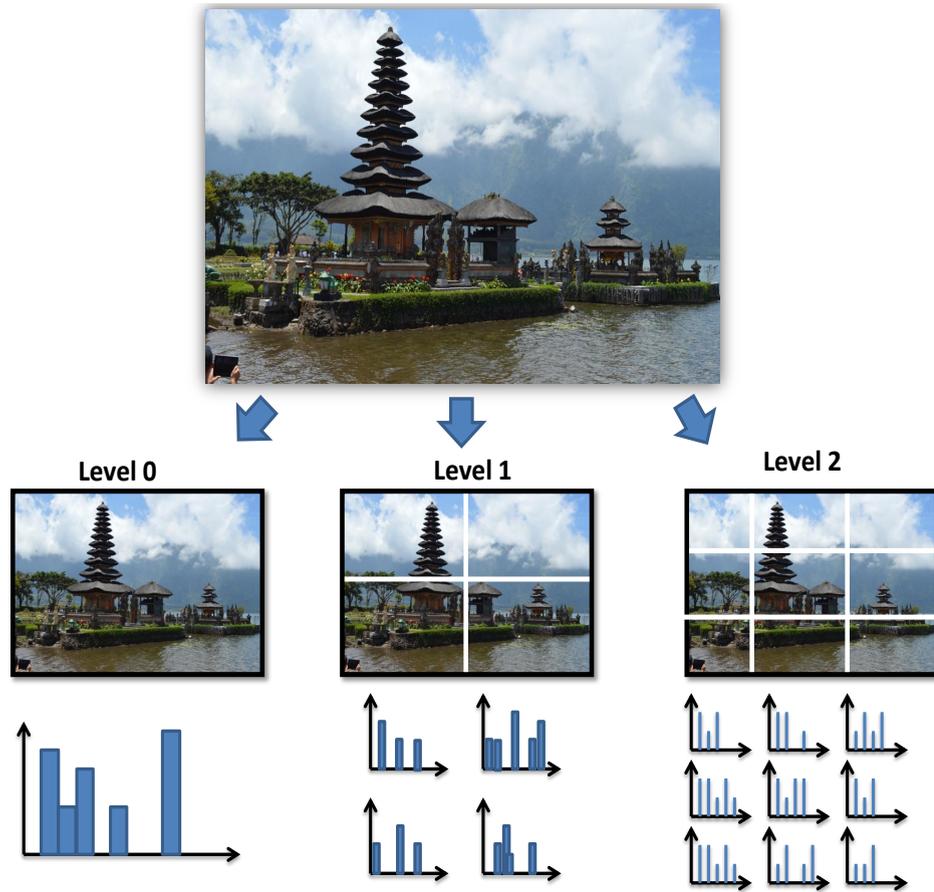


FIGURE 6.2: A schematic illustration of the spatial pyramid representation. A spatial pyramid is a collection of order-less feature histograms computed over cells defined by a multi-level recursive image decomposition. At level 0, the decomposition consists of just a single cell, and the representation is equivalent to a standard BOVW. At level 1, the image is subdivided into four quadrants, yielding four feature histograms, and so on.

Given a dataset  $D = \{(x_i, y_i), i = 1, \dots, n\}$ , also referred to as level-0 data, where  $x_i$  is a vector representing the attribute values of the  $i^{\text{th}}$  instance and  $y_i$  is the associated class label, the algorithm operates as follows:

- A  $K$ -fold cross-validation process randomly splits  $D$  into  $K$  disjoint parts of almost equal size  $D_1, \dots, D_K$ . At each  $k^{\text{th}}$  fold,  $D_k$  and  $D^{(-k)} = D - D_k$  are used as the test part and the training part, respectively.
- $N$  learning algorithms  $L_1, \dots, L_N$ , are applied to the training part  $D^{(-k)}$  to induce  $N$  level-0 classifiers  $C_1(k), \dots, C_N(k)$ . The concatenated predictions of the  $N$  level-0 classifiers on each sample in  $D_k$ , together with the original class label, form a new set  $MD_k$  of meta-level vectors.
- At the end of the entire cross-validation process, the union  $\mathbf{MD} = \cup_{k=1}^K \mathbf{MD}_k$  constitutes the full meta-level dataset, also referred to as level-1 data, which

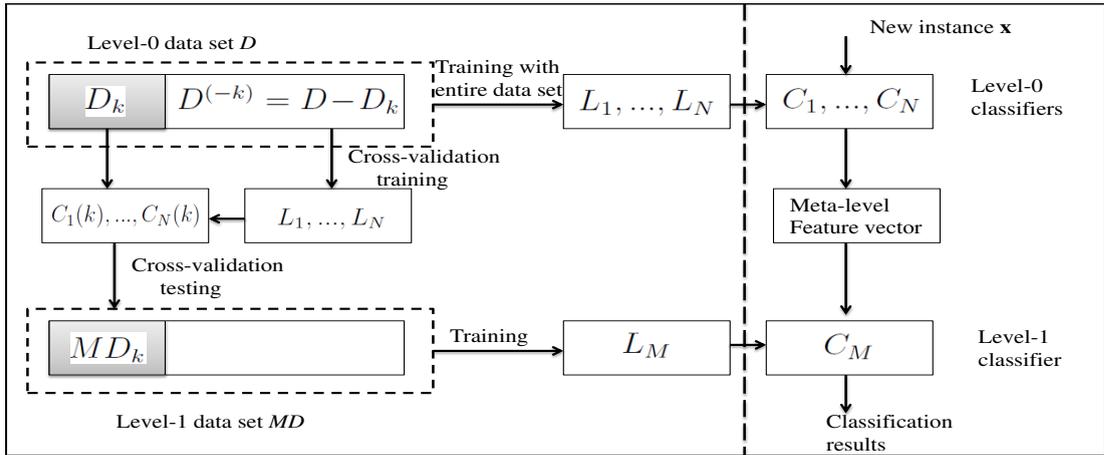


FIGURE 6.3: Flowchart of the Stack Generalization approach. Left part is the illustration of the K-fold cross-validation process for creating the meta-level training dataset, and the right part is the stacking framework in the testing stage.

is used for applying a learning algorithm  $L_M$  and inducing the meta-level classifier  $C_M$ .

- At meta-level, the learning algorithm  $L_M$  could be one of  $L_1, \dots, L_N$  or a different one. It is worth noting that, after forming the full meta-level dataset, the learning algorithms  $L_1, \dots, L_N$  are trained on the entire dataset  $D$  to induce the final base-level classifiers  $C_1, \dots, C_N$  to be used in the testing stage.

In order to classify a new instance, the concatenated predictions of all level-0 classifiers form a meta-level vector, which has  $N$  components. Then, the vector will be assigned a class label by the level-1 classifier. The class label is the final classification result of the input instance. The left part of Figure 6.3 is the illustration of the K-fold cross-validation process for creating the meta-level training dataset, while the right part is the stacking framework in the testing stage.

Figure 6.4 depicts the flowchart of the Stack Generalization in our context of image classification using both visual and tag-based features. We use one visual feature detailed in Section 6.2.1 and two tag signatures based on the LSTC signature presented in Chapter 3-Section 3.6. As learning algorithm for both level-0 and level-1 classifiers, a one-versus-all linear kernel based SVM classifier is used. At each  $k^{th}$  fold, three linear SVMs are applied to  $\mathbf{D}^{(-K)}$  giving three level-0 confidence matrix for the visual and the two tag-based features, denoted by  $\mathbf{C}_{Visual}^k$ ,  $\mathbf{C}_{WordNet}^k$  and  $\mathbf{C}_{Flickr}^k$  respectively. At the end of the cross-validation process, the union  $\mathbf{MD} = \cup_{k=1}^K \mathbf{MD}_k$  constitutes the meta-level dataset that is used to train the meta-level classifier  $\mathbf{C}_M$ . The three based linear SVMs are now trained on the entire dataset to induce the final base-classifiers  $\mathbf{C}_{Visual}$ ,  $\mathbf{C}_{WordNet}$  and  $\mathbf{C}_{Flickr}$  required by the classification task. Finally, given a new instance, the concatenated

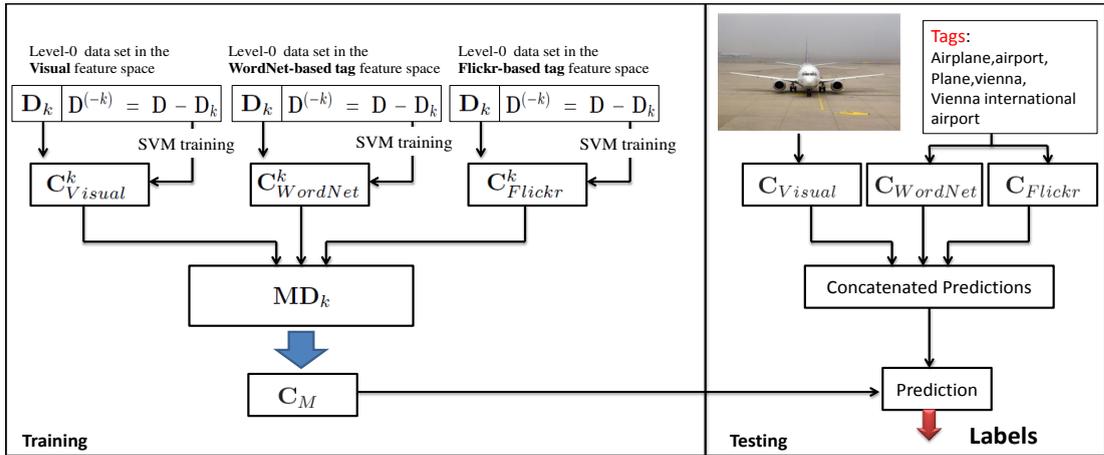


FIGURE 6.4: Flowchart of the proposed multimodal framework based on the Stack Generalization algorithm. Left part is the illustration of the K-fold cross-validation process for creating the meta-level training dataset, and the right part is the stacking framework at runtime.

predictions of all level-0 classifiers are used as input for the level-1 classifier  $C_M$  to compute the final prediction scores. As presented in [Ting and Witten, 1999], we propose to use class probabilities instead of the single predicted class as input attributes for higher level learning.

## 6.3 Experimental Evaluation

To evaluate the effectiveness and the robustness of the proposed methods on user-provided noisy/missing tags, we employ the real-world social images with human annotated tags. Specifically, four publicly available Flickr image datasets are used for the experiments. We refer the reader back to Section 2.6.2 for dataset statistic details (number of images, number of labels, number of tags...).

First, the pipeline used to compare our system to the state-of-the-art approaches is detailed. Then we present results of the proposed approach on the four considered benchmarks. Finally, we discuss the effectiveness and robustness of the proposed method.

### 6.3.1 Experimental Setup

For all datasets, the same processing chain is considered, following the literature settings to ensure consistency. The pipeline is as follows:

- **Local visual descriptors:** dense SIFTs of size 128 are extracted within a regular spatial grid and only one scale. The patch size is fixed to  $16 \times 16$  pixels and the step size for dense sampling to 6 pixels;
- **Visual Codebook:** a visual codebook of size 4,000 is created using the K-means clustering method on a randomly selected subset of SIFTs from the training dataset ( $\sim 10^5$  SIFTs).
- **Textual codebook:** for PASCAL VOC'07 dataset, we use the same experimental setting as in [Guillaumin et al., 2010]. A dictionary of size 804 is obtained by keeping only tags that appear at least 8 times. In the case of ImageClef'11 datasets, we kept only tags that were used at least 3 times in the collection, resulting in a textual codebook of 2,500 tags. For the ImageClef'12 dataset, we keep only tags that appear at least 4 times leading to a tag dictionary of size 5,134. For the NUS-WIDE dataset, among the 425,059 unique tags, there are 9,325 tags that appear more than 100 times. Authors of [Chua et al., 2009] propose to check these tags using WordNet and keep only tags that exist resulting in a dictionary of size 5,018.
- **Coding/pooling:** for coding the local visual descriptors SIFTS, we fix the patch size to  $16 \times 16$  pixels and the step size for dense sampling to 6 pixels. Then for the extracted visual (respectively tags) descriptors associated to one image, we consider a neighborhood in the visual (respectively tag) feature space of size 5 (respectively 50 in Flickr and 5 in Wordnet) for local soft coding and the softness parameter  $\beta$  is set to 10. The max-pooling operation is performed to aggregate the obtained codes. A spatial pyramid decomposition into 3 levels ( $1 \times 1, 2 \times 2, 3 \times 3$ ) is adopted for the visual-signature.
- **Base classifier:** for each modality, a one-versus-all linear kernel based SVM classifier is used, since it has shown good performances in scene categorization task when paired with the max-pooling operation on local features [Wang et al., 2010b; Liu et al., 2011b].
- **Classifier fusion:** base classifiers are trained on the considered modalities (visual, WordNet and Flickr) and combined by the stack generalization approach using five-cross-validations on the training set.

We use the mAP to evaluate the classification performances.

### 6.3.2 Experimental Results

In this section, we report the obtained results based on the mAP score for multimodal image annotation. We present separately each dataset, to compare our results to the best obtained classification score in the state-of-the-art.

### 6.3.2.1 Experiments on the PASCAL VOC 07 dataset

In Table 6.1, we compare the obtained results of our multimodal image annotation to the state-of-the-art approaches. We notice that our two tag-based feature relying on the local soft assignment coding outperform the hard assignment based coding scheme (51.8% vs 43.3% of mAP). Combined with a BOVW feature, these tag features reach better performances than that reported in [Guillaumin et al., 2010] (itself already significantly better than [Wang et al., 2009a]), while reducing drastically the computational complexity of the learning system, both in terms of the number of generated features and the learning algorithm. Our learning system is a simple combination of linear SVM based output classifiers, less computationally demanding than the Multiple Kernel Learning (MKL) one. The performances of our visual signature are slightly below those reported in the state-of-the-art system while the performances of the tag based classification are consistently better. The fusion of both modalities improves the state-of-art of about 2% in terms of mAP (68.3% vs. 66.7%).

TABLE 6.1: Comparison of our system to previous work for PASCAL VOC’07 classification challenge in terms of mAP.

Approach	Visual	Textual	Multimodal
[Wang et al., 2009a]	45.4	43.5	49.0
[Guillaumin et al., 2010]	<b>53.1</b>	43.3	66.7
Our method	52.1	<b>51.8</b>	<b>68.3</b>

### 6.3.2.2 Experiments on the ImageClef’11 dataset

In Table 6.2, we compare results of our multimodal method and that of [Binder et al., 2011] and [Liu et al., 2013]. We notice that our multimodal classification results outperforms those of [Liu et al., 2013] and are similar to those of [Binder et al., 2011]. The authors of [Binder et al., 2011] do not report separate results for textual annotation, focusing on visual representation. Their visual annotation framework clearly outperforms ours but this gain is obtained mainly through the use of significantly much complex visual signature. In particular, they use 5 different local color features (size 2048) while we use only classical SIFTs (size 128). Their BOVW has a size 160,000 while ours is 60,000. Above all, we use simple linear SVMs when they use a complex MKL. The authors of [Liu et al., 2013] use a SWLF scheme to automatically select and weight scores from the best features. Their visual annotation outperforms ours by  $\approx 4\%$  of mAP. This can be explained by the number of visual feature used which is fixed to 34 signatures including local and global descriptors. In the textual annotation, they make use of 10 textual features. Our textual model which relies on the combination of only two textual features achieved a gain of 6% of mAP compared to those of [Liu

TABLE 6.2: Comparison of our system to previous work for ImageClef’11 classification challenge in terms of mAP.

Method	Visual	Textual	Multimodal
[Xioufis et al., 2011]	31.1	32.5	40.16
[Liu et al., 2013]	35.5	32.1	43.6
[Binder et al., 2011]	<b>38.2</b>	- -	44.3
[Zhang et al., 2012b]	<b>37.4</b>	34.7	<b>45.3</b>
Our method	<u>31.2</u>	<b>38.0</b>	<u>44.8</u>

et al., 2013]. Finally, our strategy enhances the relevance of our textual modeling, that compensates the performance loss of visual modeling while maintaining a low computational complexity. The best score for this dataset is obtained by [Zhang et al., 2012b]. Their visual model outperforms our visual features by 6% of mAP score. This can be explained by the number and the size of visual features used. While we use only a BOVW, [Zhang et al., 2012b] opt for a combination of more than 5 global and local features. However, our textual models outperform textual models of all considered approaches. We obtain a gain of 4% of mAP score compared the textual model of [Zhang et al., 2012b]. We obtain approximately the same performances for the fusion (0.5% of difference on mAP score). Our learning system is a simple combination of linear SVM based output classifiers, less computationally demanding than the MKL one used in [Zhang et al., 2012b]. Again, we show that even if our visual model score is below the best visual models, our framework based on the stacked generalization algorithm gives similar results to the best results on the state-of-the-art while reducing the computational cost of learning.

### 6.3.2.3 Experiments on the ImageClef’12 dataset

In Table 6.3, we compare results of our multimodal method to that of [Liu et al., 2012]. We observe that the multimodal performance scores are similar. The visual model of [Liu et al., 2012] outperforms ours by 5% of mAP. This is due to the number of visual feature used which reaches 24. Five groups of features have been considered: color, texture, shape, local descriptor and mid-level features [Liu et al., 2011]. We obtain a gain of 1% of mAP by using only 2 textual features whereas [Liu et al., 2012] use a combination of 11 textual features. Similar to the previous experiments, the performances of our visual signature are slightly below those reported in the state-of-the-art system while the performances of the tag based classification are consistently better. The textual modality compensates the performance loss of visual modeling while maintaining a low computational complexity. This confirms the effectiveness and the robustness of the stack generalization scheme on the combination of different features and modalities.

TABLE 6.3: Our system compared to the ImageClef12 Photo Annotation best performing system [Liu et al., 2012].

Method	Visual	Textual	Multimodal
[Liu et al., 2012]	<b>34.8</b>	33.3	<b>43.6</b>
Our method	<u>29.4</u>	<b>34.1</b>	<u>43.1</u>

### 6.3.2.4 Experiments on the NUS-WIDE dataset

Table 6.4 shows the obtained results for the NUS-WIDE dataset. We compare our results to the approach of [Gao et al., 2010] which is based on feature selection aided with precision and recall scores of both tag and visual modalities. Our visual model gives similar results, however, it is obvious that our textual model gives better performances in term of mAP scores. We obtain a gain of 16% of mAP score compared to [Gao et al., 2010]. As highlighted in Chapter 2, it can be explained by the fact that the tag model of [Gao et al., 2010] is based on a simple tag-concept co-occurrence and do not take into account tag imperfections. The combination using Stack Generalization algorithm shows its effectiveness and robustness compared to other fusion strategies.

TABLE 6.4: Our system compared to multimodal image annotation state-of-the-art approaches on the NUS-WIDE dataset.

Method	Visual	Textual	Multimodal
[Gao et al., 2010]	18.89	26.12	29.88
Our method	<u>18.81</u>	<b>42.0</b>	<b>49.5</b>

## 6.4 Conclusion and discussions

We introduced a novel multimedia feature generation framework which makes use of different modalities in order to obtain efficient image classification. Feature generation is performed in an unified manner and the framework is easy to extend to other potentially useful image representations. Also, we applied computer vision techniques to text modeling, showing that inspiration between these two domains can be reciprocal. The experiments that we carried on four publicly available datasets show that we obtain comparable or better results than the state-of-the-art methods while decreasing the complexity of image representations and concept learning.

First of all, we have to stress that our visual representation is simple compared to those presented in related work, since our focus here is to introduce a scalable multimodal fusion framework. We are using only one BOVW feature based on simple SIFTs, while results reported in [Guillaumin et al., 2010] or [Binder et al.,

2011] are obtained thanks to the use of 15, respectively 72, different features. Global features such as GIST [Oliva and Torralba, 2001] and BOVW based features similar to the one we are using are combined in [Guillaumin et al., 2010; Binder et al., 2011]. We chose to focus on the generation of textual features from two complementary sources and on their combination with a relatively compact BOVW description of the images. We show that an appropriate combination of textual and visual features, based on the Stack Generalization scheme, produces competitive results when compared to much more complex state-of-the-art frameworks which focus on the visual aspect. The Stack Generalization framework shows to be interesting to deal with imperfections and the conflict that can exist in the classifier combination process.

From a computational point of view, the benefits of our approach are twofold. First, we introduce complementary information sources early in the text features extraction process. This allows us to exploit much simple visual features than those of the state-of-the-art, while obtaining comparable overall classification performances. Second, the generation of robust signatures for tag and content representation, allows us to replace classifier with high computational complexity classifier (such as MKL), used in the state-of-the-art, with a linear SVM classifier that has a lower complexity. The proposed learning system is easier to scale for large-scale datasets. In spite of the attention given to scalability, it remains a hard research problem and we show that a good trade-off between performances can be obtained through appropriate modality fusion.

Although, the proposed model in this chapter achieves good results, imperfection aspects are handled implicitly in the learning stage. As highlighted in Chapter 2, another interesting alternative to deal with these imperfections explicitly is to use Belief theory which is the subject of the model presented in the next chapter.

# Chapter 7

## Combining Classifiers Based on Belief theory

### Contents

---

<b>7.1</b>	<b>Introduction</b>	<b>150</b>
<b>7.2</b>	<b>Belief Theory for Large-Scale Multi-Label Image Classification</b>	<b>150</b>
7.2.1	Building mass functions	152
7.2.2	Dempster's Combination Rule	153
7.2.3	Decision Making	154
<b>7.3</b>	<b>Experimental Evaluation</b>	<b>154</b>
7.3.1	Experimental Setup	154
7.3.2	Experimental Results	155
<b>7.4</b>	<b>Conclusion and Discussions</b>	<b>158</b>

---

## 7.1 Introduction

As presented in the last chapter, learning based approaches such as Stacking show to be effective and robust approaches to enhance multimodal image annotation performances. However, imperfection aspects at the decision level are not handled explicitly. As highlighted in Chapter 2, Belief theory allows to deal with such imperfections. However, high computational cost of evidence combination is a drawback which is often raised against the Dempster-Shafer theory. It is well known [Smets, 1999; Smarandache and Dezert, 2009] that the computational cost of evidence combination increases exponentially with respect to the frame of discernment cardinality. Unfortunately, this is precisely the case one must handle for the considered multimedia collections.

In this Chapter, we propose a multimodal framework for image classification based on classifier fusion relying on the Dempster-Shafer theory to handle the uncertainty and the conflict that can exist between different classifiers and to assess the conflict between various sources of information. First, we convert the classifier output probabilities into consonant mass functions using the inverse pignistic transform [Dubois et al., 2001]. Secondly, these mass functions are combined in the Belief theory using Dempster's rule [Shafer, 1976]. To encounter the limitation of Belief theory due to its complexity cost, we focus on only the most probable hypothesis instead of considering all the power set of the frame of discernment. This work has been published in [Znaidia et al., 2012a].

The remainder of this Chapter is organized as follows. The proposed approach for large scale multi-label image classification is presented in Section 7.2, and experimental results are reported and discussed in Section 7.3. Section 7.4 concludes this Chapter.

## 7.2 Belief Theory for Large-Scale Multi-Label Image Classification

In this section, we propose a multimodal framework for image classification based on classifier fusion relying on the Dempster-Shafer theory to handle imperfection aspects at the decision level. In fact, a classifier is learned on each modality or signature. We aim at combining prediction scores from different classifiers in order to make a final decision. To deal with imperfection aspects at the decision level, we rely on the Dempster-Shafer theory. We refer the reader back to Chapter 2-Section 2.5.4 for details about fundamentals of this theory. To make a decision that an image can be annotated with a given label/concept, the frame of discernment is defined as  $\Omega = \{C_1, C_2, \dots, C_k\}$ , the set of annotation concepts. The frame of

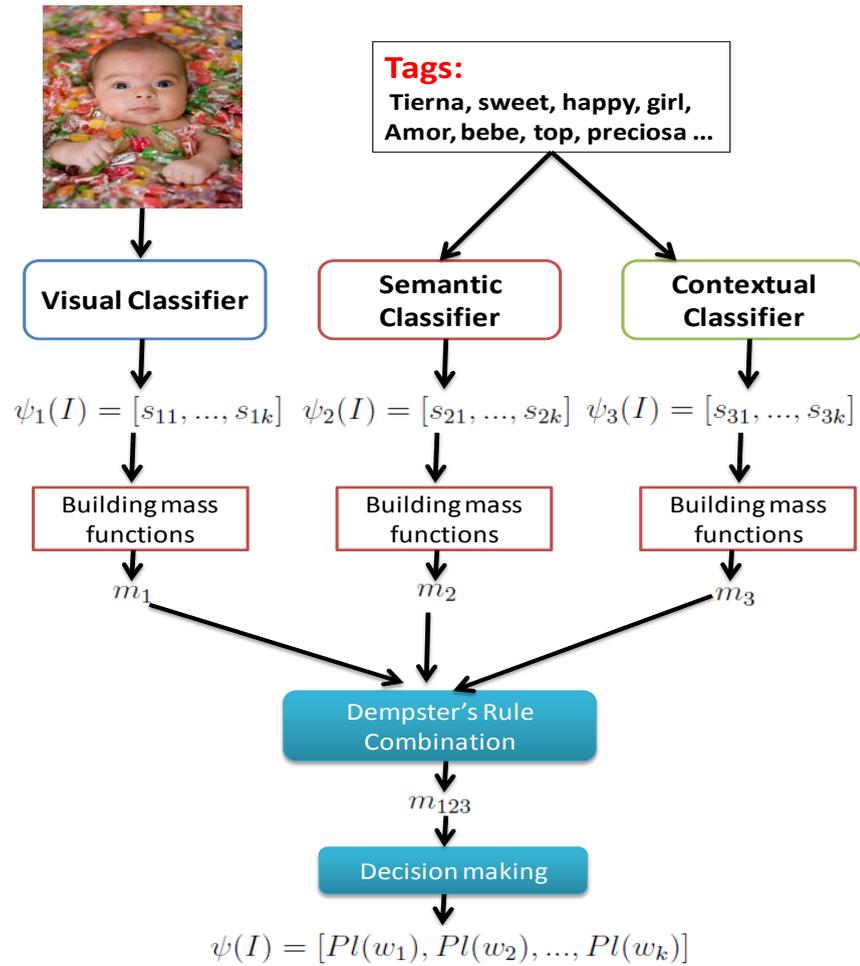


FIGURE 7.1: Flowchart of the proposed system. First, the classifier output scores  $\psi_i$  are normalized to sum to one. Secondly, the obtained probabilities are transformed into mass function using the inverse pignistic transform. A combination is performed to obtain the final mass function, used to compute the plausibility for decision making.

discernment of the multi-label extended Dempster Shafer theory is not the set of all possible single hypotheses but its power set  $\Theta = 2^\Omega$ . Thus, Dempster-Shafer theory suffers from a high computational cost, due to the frames of discernment cardinality. Unfortunately, this is precisely the case one need to handle for the considered multimedia collections. To overcome this limitation, we focus on only the most probable hypothesis instead of considering all the power set of the frame of discernment. The flowchart of the proposed system is presented in Figure 7.1. The proposed model consists in three steps:

1. **Building mass functions:** This step consists in transforming classifier output prediction scores into mass functions. First, the classifier output scores are normalized to sum to one. Secondly, the obtained probabilities are transformed into mass function using the inverse pignistic transform [Dubois et al., 2001].

2. **Combination using Dempster's rule:** The obtained mass functions need to be combined to produce a final mass function. Many rules of combination can be used. In our model, we use the Dempster's rule [Shafer, 1976] of combination.
3. **Decision making:** To decide on the image label set, obtained final mass function from the combination step is used to compute plausibility function.

### 7.2.1 Building mass functions

Assume that we have a set of  $Q$  classifiers, denoted by  $\Psi = \{\psi_1, \psi_2, \dots, \psi_Q\}$  to be combined. Given an unseen image  $I$ , each classifier  $\psi_i$  produced an output  $\psi_i(I)$  defined as :

$$\psi_i(I) = [s_{i1}, \dots, s_{ik}] \quad (7.1)$$

where  $s_{ij}$  indicates the degree of confidence in stating that “the image  $I$  belongs to the class  $C_j$  according to the classifier  $\psi_i$ ”.

First, classifier outputs are normalized to obtain a probability distribution  $p_i$  over  $\Omega$  as follows:

$$p_i(C_j) = \frac{s_{ij}}{\sum_{j=1}^k s_{ij}}, \quad \text{for } j = 1, \dots, k \quad (7.2)$$

For each classifier  $\psi_i$ , the element of  $\Omega$  are ranked by decreasing probabilities to obtain an ordered frame of discernment  $\Omega_{ord} = \{w_1, w_2, \dots, w_{|\Omega|}\}$  where  $w_1$  corresponds to the concept with the highest probability value.

The class label of an unseen image  $I$  may be represented by a variable  $\hat{y}$  taking values in  $\Theta = 2^\Omega$ . Thus, expressing partial knowledge of  $\hat{y}$  in the Dempster-Shafer framework may imply storing  $2^{2^{|\Omega|}}$  numbers. Given an image  $I$ , based on this ordering, instead of considering the whole power set of  $\Omega$ , we will focus on a smaller subset  $R(\Omega)$  defined by:

$$R(\Omega) = \{\cup_{j=0}^{|\Omega|-1} R_j(\Omega_{ord})\} \quad (7.3)$$

$$R_j(\Omega_{ord}) = \{w_1, \dots, w_{j+1}\}, \forall j = 0, \dots, |\Omega| - 1 \quad (7.4)$$

The size of this subset is equal to  $|\Omega|$ , it is thus much smaller than  $2^{2^{|\Omega|}}$  while being rich enough to express evidence because we consider only the most probable subsets.

We take an example of a multi-label classification problem with three classes where  $\Omega = \{dog, baby, flower\}$ . The frame of discernment is not the set of all possible single hypotheses but its power set  $\Theta = \{\emptyset, \{dog\}, \{baby\}, \{flower\}, \{dog, baby\}, \{dog, flower\}, \{baby, flower\}, \{dog, baby, flower\}\}$ . For a test image, the classifier outputs are normalized to obtain the following probabilities:

$p(dog) = 0.3$ ,  $p(baby) = 0.5$  and  $p(flower) = 0.2$ . Based on the ordering of these probabilities ( $p(baby) \geq p(dog) \geq p(flower)$ ), we obtain an ordered frame of discernment  $\Omega_{ord} = \{baby, dog, flower\}$ . Instead of considering the whole power set  $\Theta$ , we focus on only a small subset  $R(\Omega) = \{A_0 = \{baby\}, A_1 = \{baby, dog\}, A_2 = \{baby, dog, flower\}\}$  which represents the most possible hypothesis.

Secondly, we convert the obtained probabilities into consonant mass functions using the inverse pignistic transform [Dubois et al., 2001]. The consonant mass function derived from these probabilities verifies :

$$m : R(\Omega) \rightarrow [0, 1], \quad \sum_{R_j(\Omega_{ord}) \in R(\Omega)} m(R_j(\Omega_{ord})) = 1 \quad (7.5)$$

$$\begin{aligned} m(\{w_1, w_2, \dots, w_i\}) &= i \times [p(w_i) - p(w_{i+1})] \quad \forall i < |\Omega| \\ m(\{w_1, w_2, \dots, w_{|\Omega|}\}) &= |\Omega| \times p(w_{|\Omega|}) \\ m(X) &= 0 \quad \forall X \notin R(\Omega). \end{aligned} \quad (7.6)$$

We take the above example with  $\Omega = \{baby, dog, flower\}$ , the mass functions for the considered subset  $R(\Omega)$ , based on the probability ordering, are computed as follows:

$$\begin{aligned} m(\{baby\}) &= 1 \times [p(baby) - p(dog)] = 1 \times [0.5 - 0.3] = 0.2 \\ m(\{baby, dog\}) &= 2 \times [p(dog) - p(flower)] = 2 \times [0.3 - 0.2] = 0.2 \\ m(\{baby, dog, flower\}) &= 3 \times p(flower) = 3 \times 0.2 = 0.6 \end{aligned}$$

## 7.2.2 Dempster's Combination Rule

In the proposed method, we choose to combine the obtained consonant mass functions from different classifiers using the normalized Dempster's rule [Shafer, 1976]. Other combination rules can be used [Quost et al., 2011]. Let  $m_i$  be the mass function of the source  $i$ , the combination of  $n$  mass functions (corresponding to  $n$  classifiers) is defined according to Dempster's combination rule as follows:

$$m_{1-n}(A) = \begin{cases} \frac{\sum_{\cap_{k=1}^n b_k = A} \prod_{i=1}^n m_i(b_i)}{1-K}, & \forall A \subseteq \Omega, A \neq \emptyset, b_k \in R_k(\Omega) \\ 0 & \text{if } A = \emptyset \end{cases} \quad (7.7)$$

where

$$K = \sum_{\cap_{k=1}^n b_k = \emptyset} \prod_{i=1}^n m_i(b_i) \quad (7.8)$$

$K$  is the degree of conflict between the combined mass functions, assumed to be strictly smaller than one. This rule is commutative and associative.

### 7.2.3 Decision Making

After the combination step, several decision rules can be used to make a decision. We choose to use the plausibility function to make an optimistic decision since it can be viewed as a loose upper limit to the uncertainty which is not the case of the credibility function that represents a pessimistic decision. Let  $\hat{Y}$  be the predicted label set for an instance  $x$ . To decide whether to classify a given image into a class, we compute the degree of plausibility  $Pl(w_j)$  that the true label set  $Y$  contains the label  $w_j$ , and the degree of plausibility  $Pl(\bar{w}_j)$  that it does not contain the label  $w_j$ . We then define  $\hat{Y}$  as:

$$\hat{Y} = \{w_j \in \Omega | Pl(w_j) \geq Pl(\bar{w}_j)\} \quad (7.9)$$

## 7.3 Experimental Evaluation

To evaluate the effectiveness and the robustness of the proposed methods on user-provided noisy/missing tags, we employ the real-world social images with human annotated tags. Specifically, two publicly available datasets are used for the experiments. We refer the reader back to Section 2.6.2 for dataset statistic details (number of images, number of labels, number of tags...).

First, the pipeline used to compare our system to other ones is detailed. Then we present results of the proposed approach over the two considered benchmarks. Finally, we discuss the effectiveness and robustness of the proposed method.

### 7.3.1 Experimental Setup

The number of classifiers  $Q$  introduced in Section 7.2.1 is equal to three: we used two textual descriptors and one visual descriptor to learn three classifiers using Linear SVM. The number of annotation concepts  $k$  is equal to 93 (respectively 99) for the mageClef'10 (respectively ImageClef'11) dataset. The textual descriptor used for the experiments is the Soft-Bag-of-Concepts (soft-BoC) presented in Chapter 3. Each feature vector is of size 93 (respectively 99) for the ImageClef'10 (respectively ImageClef'11) dataset. For the visual signature, images are described using five various global features, including color and edge features:

- **Color:** We use two color histograms in the RGB space. The first one is quantified on three levels (size  $4^3 = 64$ ) and the second one on five levels (size  $5^3 = 125$ ). A third histogram is computed in the HSV space and quantified on three levels (size  $5^3 = 125$ ). Another color descriptor taking into account the spatial coherence is considered [Stehling et al., 2002].

- **Texture:** We use the local edge pattern (LEP) [Cheng and Chen, 2003] leading to an histogram of size 512.

The visual signature is the concatenation of these five descriptors. The resulting feature vector is of size 890. Each feature vector was used to train a classifier using the Fast Shared Boosting algorithm [Le Borgne and Honnorat, 2010]. In the rest of this chapter, we call Contextual Classifier, the classifier learned using Flickr similarity and Semantic Classifier learned using WordNet similarity detailed in Chapter 3.

## 7.3.2 Experimental Results

We present results of two experiments. In the first experiment, we evaluate the performance of individual classifiers on the ImageClef'11 dataset. Then, we compare the performance of the proposed method to several rules of combination for classifier fusion. In the second experiment, we perform the same experiment on the ImageClef'10 dataset.

### 7.3.2.1 Experiments on the ImageClef'11 dataset

Table 7.1 displays the performances of individual classifiers in terms of mean Average Precision (mAP) for the ImageClef'11 dataset. These results show that individual classifiers exhibit similar performances with a small superiority to the contextual classifier.

We use the Majority voting, Average, Maximum, Minimum and Product rules as baselines for comparison. By comparing results presented in Table 7.2, we can see that the combination of classifiers for both Dempster's rule and average rule

TABLE 7.1: Comparative Performance of individual classifiers in terms of mAP for the ImageClef'11 dataset.

Classifier	Visual Classifier	Contextual Classifier	Semantic Classifier
<i>mAP</i>	29.86	32.13	29.24

TABLE 7.2: Comparative Performance of different combination strategies in terms of mean Average Precision (mAP) for the ImageClef'11 dataset. The best results are marked in bold.

Strategy	Maximum rule	Minimum rule	Product rule	Majority voting	Dempster's rule	Average rule
<i>mAP</i>	36.39	31.93	33.59	40.01	<u>39.05</u>	<b>40.21</b>

TABLE 7.3: Comparative Performance of individual classifiers, Dempster, Average and the ImageClef 2011 Winner [Binder et al., 2011] for some challenging classes in terms of mean Average Precision (mAP).

Classes	Visual classifier	Context classifier	Semantic classifier	Dempster rule	Average rule	[Binder et al., 2011]
Travel	18.85	14.78	17.55	<b>22.12</b>	14.57	16.72
Technical	08.19	06.37	04.52	<b>12.85</b>	07.24	08.51
Boring	07.28	07.78	07.63	<b>15.88</b>	08.79	09.94
Bird	17.55	51.71	56.08	<b>61.52</b>	58.77	58.71
Insect	14.26	47.84	46.44	<b>58.08</b>	53.12	45.21
Airplane	05.36	44.36	42.53	<b>61.66</b>	59.32	22.93
Skate	00.27	10.29	21.54	<b>28.42</b>	11.46	00.56
Scary	18.46	08.31	14.10	<b>19.02</b>	11.29	16.39

gives better results than the best individual classifier. We obtain a gain of  $\approx 10\%$  in terms of classification accuracy (mAP). For this dataset, we observe that the average rule achieves slightly better performances than the Dempster’s rule. These results may be explained by the performance of the individual classifiers which exhibit both identical performances and correlations between estimation errors. In addition, we train individual classifiers with unbalanced data over classes which can generate unreliable confidences (*e.g.* caused by a small training set or by over training).

The average rule is hardly ever theoretically optimal, but performs sometimes surprisingly good except for some classes as shown in Table 7.3. For these challenging classes, Dempster’s rule performs much better than the average rule especially when considering ensembles of “good” and “bad” classifiers, then using the average rule to combine the classification results will not be a good choice. We compare Dempster’s rule to the ImageClef 2011 Winner [Binder et al., 2011] for these challenging classes. The proposed method outperforms the-state-of-art [Binder et al., 2011] for such type of classes. These results are confirmed by the evaluation of the proposed method for the ImageClef’10 dataset in the next section.

### 7.3.2.2 Experiments on the ImageClef’10 dataset

Table 7.4 displays the performances of individual classifiers in terms of mean Average Precision (mAP) on the ImageClef’10 dataset. By comparing results presented in Table 7.5, we can observe that the Dempster combination rule performs better than the average rule. This can be explained by the nature of the classifier to be combined which represent a larger diversity in performances. This diversity may cause uncertainty and conflict between base classifiers.

We can notice that the Belief theory seems to offer a significant advantage to such situations. It is particularly interesting to handle the uncertainty and the conflict that can exist between different classifiers.

In order to illustrate how the proposed method based on the Dempster Theory achieves better predictions, we included example images in Figure 7.2. We compared here the Dempster’s rule only with the Average rule and listed the concepts which the two methods returned. The correct ones are marked with green checks, while the wrong ones are indicated by red crosses.

TABLE 7.4: Comparative Performance of individual classifiers in terms of mean Average Precision (mAP) for the ImageClef’10 dataset.

Classifier	Visual Classifier	Contextual Classifier	Semantic Classifier
$mAP$	31.45	38.58	31.97

Examples	Visual classifier	Contextual classifier	Semantic classifier	Dempster's rule	Average rule	Ground truth labels
Tags : sports dodgers baseball russell martin troy tulowitzki 2ndbase 	✓ Outdoor ✓ Adult ✓ Neutral_Illumination ✓ No_Blur ✗ No_Persons ✗ No_Visual_Season ✓ Male ✗ Visual_Arts	✓ Outdoor ✓ Day ✓ Neutral_Illumination ✓ No_Blur ✗ No_Persons ✗ No_Visual_Season ✗ Visual_Arts	✓ Outdoor ✓ Day ✓ Neutral_Illumination ✓ No_Blur ✗ natural ✗ No_Visual_Season ✗ Visual_Arts	✓ Outdoor ✓ Day ✓ Neutral_Illumination ✓ No_Blur ✗ No_Persons ✓ Male ✓ Adult ✗ Partly_Blurred ✓ Small_Group ✓ Cute	✗ Winter ✗ Indoor ✗ Mountains ✗ Underexposed ✗ Partly_Blurred ✗ Big_Group ✗ artificial	Outdoor Day Neutral_Illumination No_Blur Male Adult Small_Group Cute Sports
Tags : Dal Cinna moroll 	✓ Outdoor ✓ Neutral_Illumination ✓ No_Blur ✗ Female ✗ Adult	✓ Neutral_Illumination ✓ No_Blur ✗ No_Persons ✗ Natural	✓ Neutral_Illumination ✓ No_Persons ✗ Visual_Arts	✓ Outdoor ✓ Neutral_Illumination ✓ No_Blur ✓ Day ✓ Still_Life ✓ No_Persons ✓ Fancy ✓ Cute	✓ Neutral_Illumination ✓ No_Blur ✓ Cute ✗ No_visual_season ✗ No_visual_time	Outdoor Neutral_Illumination No_Blur No_Persons Summer Outdoor Day Still_Life Fancy cute
No tags 	✓ Neutral_Illumination ✓ No_Persons ✓ Natural ✓ No_Visual_Season ✗ Sky ✗ No_Visual_Place ✗ No_Visual_Time	✓ Neutral_Illumination ✓ No_Persons ✓ Natural ✓ No_Visual_Season ✗ Partly_Blurred	✓ Neutral_Illumination ✓ No_Persons ✓ Natural ✓ No_Visual_Season ✗ Partly_Blurred	✓ Neutral_Illumination ✓ No_Persons ✓ Natural ✓ No_Visual_Season ✗ Water ✓ Aesthetic_Impression ✓ Overall_Quality ✓ No_blur ✓ Cute	✓ Neutral_Illumination ✓ No_Persons ✓ Natural ✓ No_Visual_Season ✗ Outdoor ✓ No_Blur	Neutral_Illumination No_Persons Natural No_Blur Partly_Blurred Single_Person Aesthetic_Impression Overall_Quality No_Visual_Season cute

FIGURE 7.2: A qualitative comparison between individual classifiers, the proposed method and the average rule. Predicted labels are shown in each column. The correct ones are marked with green checks, while the wrong ones are indicated by red crosses.

TABLE 7.5: Comparative Performance of different combination strategies in terms of mean Average Precision (mAP) for the ImageClef’10 dataset. The best results are printed in bold.

Strategy	Maximum rule	Minimum rule	Product rule	Majority voting	Dempster’s rule	Average rule
<i>mAP</i>	32.80	33.02	33.59	20.58	<b>43.19</b>	<u>36.58</u>

It can be seen that in most of the cases, the proposed method performs better than the Average rule. The Average rule is not able to predict the whole set of labels and most of the predicted ones are wrong. There are many missing labels with this rule. Whereas, the proposed method removes most of wrong labels predicted by individual classifiers and predict some other correct labels .

## 7.4 Conclusion and Discussions

In this paper, we presented a system for combining classifiers using Belief theory for large-scale multi-label image classification. The exponential complexity of operations in the theory of belief functions has long been seen as a shortcoming of this approach, and has prevented its application to very large frames of discernment. We have shown in this chapter that the complexity of the Dempster-Shafer calculus can be drastically reduced from  $(2^{2^{|\Omega|}})$  to  $|\Omega|$ , while retaining sufficient expressive power, if belief functions are defined over a suitable subset of the power set. The major difference between our work and the state-of-the-art approaches [Liu et al., 2011; Younes et al., 2009] is that we address the problem of combination in a multi-label classification task for a large problem: to the best of our knowledge, this is the first attempt to apply Dempster-Shafer theory for a multimodal multi-label image classification for a large dataset ( $\approx 18k$  images) and a large variety of categories simultaneously (scene, event, objects, image quality and emotions  $\approx 100$  concepts). When individual classifiers present similar performances, experimental results have shown that using simple rules such as averaging can be a good choice. While, for conflicting classifiers, the Belief theory seems to be an interesting framework to handle the uncertainty and the conflict that can exist between different classifiers.

# Chapter 8

## Conclusions and Future Research Directions

In this dissertation, we address the problem of *imperfections in multimodal image annotation in the context of social media*. We distinguish two levels of imperfections: **Representation** and **Decision**. The first level is related to tag imperfections. In fact, we consider as imperfect, tags that are noisy and not related to the image semantic visual content. Our first goal is to identify and define these imperfection aspects at the representation level. Thereafter, we focus on handling such imperfections in order to enhance multimodal image annotation performances. At the decision level, we are interested in handling imperfections that can exist while combining classifiers from different modalities and learned on imperfect data (tags).

### 8.1 Contributions

- To deal with tag imperfections at the representation level, we started by identifying and defining clearly these imperfections in the context of image annotation. Thereafter, we have introduced two novel textual signatures for tag-based image annotation in the context of social media. We reported extensive experimental results on five datasets. From these results, we conclude that both signatures give similar or better results than the state-of-the-art approaches on the five considered datasets on the tag-based image annotation task. Unlike classic BOW models, our both signatures permitted to handle a part of imperfection aspects of tags.
- To deal with the problem of tag incompleteness, we introduced a novel approach for tag suggestion based on local soft coding and Belief theory. First, a list of “candidate tags” is created from the visual neighbors of the untagged

image, using both local soft coding and two consecutive pooling steps. Then, these tag-signatures are used as pieces of evidence to be combined to provide the final list of predicted tags. This fusion is based on the Dempster's rule of combination, in accordance with the Evidential  $k$ NN framework. Hence, both steps support a scheme to tackle with imprecision and uncertainty that are inherent to this type of information in a social media context. The experiments that we carried out for image classification on two publicly available datasets show that we obtain comparable or better results than the state-of-the-art methods. For tag suggestion, we manually annotated 241 queries to propose a new benchmark<sup>1</sup> to the community. For that application as well, we obtained competitive results, with a score two points better than the best recent state-of-the-art method.

- We proposed a new BOW based signature, called Bag-of-Multimedia Words (BOMW), that results from a combination of textual and visual information. It is based on *multimedia codewords* that allow on the one hand cross-coding textual tag-words over visual-words extracted from a document; and on the other hand designing BOMW signatures. Experiments have been conducted on two well-known challenging benchmarks: PASCAL VOC'07 and Image-Clef'12. Obtained results show the competitive performances of the BOMW, ensuring a trade-off between classification accuracy and computation cost. In opposition to classic BoW signatures, classification results remain stable, with a very low fluctuation, when changing sizes of the visual or multimedia codebooks. This is an interesting property useful to reduce the complexity of the classification system in both training and test, which is obtained at the cost of a small pre-processing step for signature design (building tag-coding matrix and clustering it). The performance gain is due to the fact that the proposed multimedia signature lies on a structured space, well appropriate to describe multimedia documents. Therefore, BOMW are probably much more class-discriminative than other types of BOW. The proposed framework is generic and, thus it is possible to exploit it in other application domains (video classification, robotics etc.), with data that include other modes than textual and visual ones.
- To handle imperfections at the decision level, we introduced a novel multimedia feature generation framework which makes use of different modalities in order to obtain efficient image classification. Feature generation is performed in an unified manner and the framework is easy to extend to other potentially useful image representations. The combination of both tag and visual features is performed automatically via the Stack Generalization algorithm. Stacking represents a scheme for minimizing the generalization error rate of one or more models. Stacking classifiers is done by collecting the

---

<sup>1</sup><http://perso.ecp.fr/~znaidiaa/dataset.html>

outputs of so called level-0 classifiers into a new dataset and to train one or more level-1 classifiers on the outputs of the level-0 classifiers to improve learning generalization capacities. The experiments that we carried on four publicly available datasets show that we obtain comparable or better results than the state-of-the-art methods while decreasing the complexity of image representations and concept learning.

- We presented a multimodal framework for combining classifiers using Belief theory for large-scale multi-label image classification. When individual classifiers present similar performances, results have shown that using simple rules such as averaging can be a good choice. While, for conflicting classifiers, the Belief theory seems to be an interesting framework to handle the uncertainty and the conflict that can exist between different classifiers. Our approach ensured that the fusion scheme remained robust in the presence of noise arising from poor classification results due either to the classifiers themselves or to the nature of data (tags) fed to the classifiers. The computational complexity, which limits the use of Belief theory, is reduced by considering only a subset of the frame of discernment.

## 8.2 Perspectives for future research

Many contributions were proposed in this dissertation to deal with imperfections in multimodal image annotation problem in the context of social media. These contributions are in no way complete solutions, and could be improved in several manners. In the following, we propose potential directions that can be explored further.

- To deal with tag imperfections, we relied in the proposed models on two external resources: WordNet and Flickr. In the future, a promising research direction would be to include other knowledge resources such as Wikipedia for tag refinement and noise removing. For example, a similarity based on Wikipedia can be used in our LSTC signature to derive a new tag-based signature. This latter can be combined with the two proposed tag signatures based on Flickr and WordNet.
- Since the proposed BOMW signature shows to be more compact and discriminative than classic BOW representations (visual or tag), it seems interesting to build multimodal pLSA models with BOMW, while being simple and efficient for parameter learning. Instead of considering each modality separately, we argue that combining visual and tag modalities in a same signature, which is more appropriate for multimedia documents, represents an interesting step to apply pLSA. Another possible direction is to extend

recent works addressing the visual Fisher vector [Perronnin and Dance, 2007; Perronnin et al., 2010] and its approximate variant [Jégou et al., 2012] to the proposed BOMW, taking advantages of merging generative models with discriminative classifiers to put in place competitive recognition systems.

- To deal with tag completion, we proposed a method based on visual neighbors which were obtained from an image database containing 1.2 million images extracted from Flickr. This resource is crucial to obtain good raw results for the considered application. Even if our method obtains better results than other recent ones, all of them would benefit from an improved resource. A first direction to improve it is to use a potentially better visual signature to get the neighbors. However, we must keep a certain efficiency in practice to avoid prohibitive time responses to find neighbors. For this, we may search them into a compressed domain that allows to fit large databases into memory [Jégou et al., 2012]. A more difficult direction of research will be the improvement of the annotation of the resource itself. As we explained, a lot of the current annotations are far from being perfect (it is one of the reason we re-annotated the queries to evaluate the work). Hence, this work can naturally be continued into the process of cleaning large multimedia resources.
- To improve the classifier combination based on Belief theory, one direction for future research is to take into account the classifier reliability while combining output scores from different classifiers. In the proposed approaches, classifiers are considered as equal and their reliability are not taken into account. Consequently, an estimation of the classifiers reliability seems to be interesting to improve classifier combination performances. Another interesting direction is to measure the conflict degree between different classifiers and to exploit this measure in classifier combination. An additional direction is to construct mass functions directly in the classifiers.
- One direction is to use tag completion and local soft tag coding approaches for a personalized and interactive tag recommendation system in social media. While a user enters new tags for a particular photo, the system suggests related tags to her, based on the local soft tag coding approach. If no tag is entered, the tag completion approach can be used to suggest a set of tags and give free choice to the user to keep the most relevant ones.
- Based on the BOMW approach, one perspective is to develop an automatic illustration system supported by multimedia information retrieval, that analyzes text and presents a list of candidate images to illustrate it, called "story picturing". This system can be used to illustrate children stories or to describe a newspaper article in order to provide enhanced visual comprehension.

# Appendices

# Appendix A

## Evaluation measures

To assess the performance of the performances of the proposed approaches in this thesis, we use the following evaluation measures:

### Mean Average Precision (mAP)

Among evaluation measures, mAP has been shown to have especially good discrimination and stability. This evaluation measure first ranks the images by their confidence scores, from high to low, for each concept separately. This produces a result vector  $V = \langle v_1, \dots, v_n \rangle$ , where  $s(v_i) \geq s(v_{i+1})$  for all  $1 \leq i \leq n$ . We note by  $rel(v_i)$  the relevance of an image;  $rel(v_i)$  is equal to 1 if  $v_i$  is relevant to this concept and 0 otherwise. The images are inspected one by one and each time a relevant image is encountered the precision and recall values are computed as follows.

- **Precision** The precision measure is based on the observation that users of an Information Retrieval system tend to examine only the first  $k$  results of a search. It measures what fraction of these  $k$  results is relevant to the query on average. The value  $k$  is commonly called the document cut-off value. The precision at  $k$  is defined as:

$$P(@)k(V) = \frac{1}{k} \sum_{i=1}^k rel(v_i) \quad (\text{A.1})$$

- **Recall** The recall measure quantifies what fraction of all the relevant results was ranked to fall within the first  $k$  documents. The recall at  $k$  is defined as:

$$R(@)k(V) = \frac{\sum_{i=1}^k rel(v_i)}{\sum_{i=1}^n rel(v_i)} \quad (\text{A.2})$$

In case of ties we consider all the images with the same confidence score together at once and produce only a single precision and recall value for them using a tie-aware ranking approach [McSherry and Najork, 2008]. We then interpolate the values so that the recall measurements range from 0.0 to 1.0 with steps of 0.1; the precisions at these recall levels are obtained by taking the maximum precision obtained at any non-interpolated recall level equal or greater to the interpolated recall step level under consideration. Formally, the interpolated precision  $P_{interp}$  at a certain recall level  $R$  is defined as the highest precision found for any recall level  $R' \geq R$ :

$$P_{interp}(R) = \max_{R' \geq R} P(R') \quad (\text{A.3})$$

- **Average Interpolated Precision**

The average interpolated precision, called also 11-point average precision, is defined as follows:

$$AP_{interp} = \frac{1}{11} \sum_{R=0.1}^1 P_{interp}(R) \quad (\text{A.4})$$

where  $R \in [0.1, 1.0]$  with steps of 0.1.

To obtain the overall interpolated mAP (mAP), called also 11-point Mean Interpolated Average Precision, we average the average interpolated precisions over all concepts as follows:

$$mAP = \frac{1}{C} \sum_{i=1}^C AP_{interp} \quad (\text{A.5})$$

where  $C$  is the number of annotation concepts.

# Appendix B

## Publications of the Author

### International Journals with Peer-review

Znaidia, A., Shabou, A., Le Borgne, H., and Hudelot, C., Popescu, A., and Paragios, N. . *Handling Tag Imperfections for Image Annotation in Social Media*, IEEE Transaction On Multimedia (Submitted), 2013.

### National Journals with Peer-review

Znaidia, A., Shabou, A., Le Borgne, H., and Hudelot, C. . *Codage des modèles de tags*. Revue d'Intelligence Artificielle, volume 27, number 1, pages 39-63, 2013.

### International Conferences with Peer-review

Znaidia, A., Le Borgne, H., and Hudelot, C. . *Belief Theory for Large- Scale Multi-label Image Classification*. In Belief Functions: Theory and Applications. Advances in Intelligent and Soft Computing, volume 164, pages 205-212, Compiègne, France, 2012.

Znaidia, A., Shabou, A., Popescu, A., Le Borgne, H., and Hudelot, C. . *Multimodal feature generation framework for semantic image classification*. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR'12, pages 38:1-38:8, New York, NY, USA. ACM, 2012.

Znaidia, A., Shabou, A., Le Borgne, H., Hudelot, C., and Paragios, N. . *Bag-of-multimedia-words for image classification*. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR), pages 1509-1512, 2012.

Znaidia, A., Le Borgne, H., and Hudelot, C. . *Tag completion based on belief theory and neighbor voting*. In Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, ICMR, pages 49-56, New York, NY, USA. ACM, 2013.

## National Conferences with Peer-review

Znaidia, A., Le Borgne, H., Hudelot, C., and Popescu, A. . *Prise en compte de l'imperfection des tags pour la classification smantique d'images*. In Actes de la conférence RFIA, Lyon, France, 2012.

## Evaluation Campaign Participations

Znaidia, A. Le Borgne, H., Popescu, A. *CEA LIST's Participation to Visual Concept Detection Task of ImageCLEF 2011*. CLEF (Notebook Papers/Labs/-Workshop), Amsterdam, The Netherlands, 2011.

Znaidia, A. , Shabou, A., Le Borgne, H., Popescu, A. *CEA LIST's Participation to the Concept Annotation Task of ImageCLEF 2012*. CLEF (Online Working Notes/Labs/Workshop), Rome, Italy, 2012.

Borgne, H.L., Popescu, A., Znaidia, A.: *CEA LIST@imageCLEF 2013: Scalable Concept Image Annotation*. CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain, 2013.

## Patent

Znaidia, A., Shabou, A., Le Borgne, H. . *Procédé de classification d'un objet multimodale*, France, Brevet number 1259769, Filing date: le 12-12-2012. (in progress)

# Bibliography

- Al-Ani, A. and Deriche, M. (2002). A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. *Journal of Artificial Intelligence Research*, 17:333–361.
- Altintas, E., Karşligil, E., and Coskun, V. (2005). A new semantic similarity measure evaluated in word sense disambiguation. In *the Proceedings of 15th NODALIBA conference, Joensuu*, pages 8–11.
- Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI*, pages 971–980, New York, NY, USA. ACM.
- Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379.
- Avila, S., Thome, N., Cord, M., Valle, E., and De A. Araújo, A. (2013). Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465.
- Ayache, S. and Quénot, G. (2007). Evaluation of active learning strategies for video indexing. *Journal of Image Communication*, 22(7-8):692–704.
- Ballan, L., Bertini, M., Uricchio, T., and Del Bimbo, A. (2013). Social media annotation. In *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, pages 229–235. IEEE.
- Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence, IJCAI'03*, pages 805–810, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135.
- Bellenger, A. (2013). *Semantic Decision Support for Information Fusion Applications*. These, INSA de Rouen.

- Bellman, R. E. (1961). *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A.
- Binder, A., Samek, W., Kloft, M., Müller, C., Müller, K.-R., and Kawanabe, M. (2011). The Joint Submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task. In *CLEF (Notebook Papers/Labs/Workshop)*.
- Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, pages 127–134, New York, NY, USA. ACM.
- Bloch, I. (2001). Fusion of image information under imprecision and uncertainty: numerical methods. *Courses and lectures-International Centre For Mechanical Sciences*, pages 135–170.
- Bloch, I. (2003). *Fusion d'informations en traitement du signal et des images*. Lavoisier.
- Bloch, I. (2008). *Information Fusion in Signal and Image Processing: Major Probabilistic and Non-Probabilistic Numerical Approaches*. Digital signal and image processing series. Wiley.
- Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web, WWW '07*, pages 757–766, New York, NY, USA. ACM.
- Borotschnig, H., Paletta, L., and Pinz, A. (1999). A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Computing*, 62(4):293–319.
- Bostrom, H. and al. (2007). On the definition of information fusion as a field of research. *Technical report, University of Skovde, School of Humanities and Informatics, Skovde, Sweden*.
- Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). Learning mid-level features for recognition. In *CVPR*, pages 2559–2566.
- Braun, J. J. (2000). Dempster-shafer theory and bayesian reasoning in multisensor data fusion. In *AeroSense 2000*, pages 255–266. International Society for Optics and Photonics.
- Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32(1):13–47.

- Cantador, I., Konstas, I., and Jose, J. M. (2011). Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):1 – 15.
- Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410.
- Carneiro, G. and Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 163–168 vol. 2.
- Chandrika, P. and Jawahar, C. V. (2010). Multi modal semantic indexing for image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10*, pages 342–349, New York, NY, USA. ACM.
- Cheng, Y.-C. and Chen, S.-Y. (2003). Image classification using color, texture and regions. *Image Vision Computing*.
- Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. (2009). Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09*, pages 48:1–48:9, New York, NY, USA. ACM.
- Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.
- Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11*, pages 44:1–44:8, New York, NY, USA. ACM.
- Clough, P., Müller, H., and Sanderson, M. (2010). Seven Years of Image Retrieval Evaluation. In Croft, W. B., Müller, H., Clough, P., Deselaers, T., and Caputo, B., editors, *ImageCLEF*, volume 32 of *The Information Retrieval Series*, chapter 1, pages 3–18. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Coates, A. and Ng, A. Y. (2011). The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization . In *ACM International Conference on Machine Learning (ICML)*, pages 921–928.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. In *Machine Learning*, pages 273–297.
- Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision (ECCV)*, pages 1–22.

- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339.
- Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Transaction on systems, man and cybernetics*, 25:804–813.
- Destercke, S., Dubois, D., and Chojnacki, E. (2009). Possibilistic information fusion using maximal coherent subsets. *IEEE T. Fuzzy Systems*, 17(1):79–92.
- Dork, G. and Schmid, C. (2005). Object class recognition using discriminative local features. Rapport de recherche RR-5497, INRIA.
- Dubois, D. (2007). Uncertainty theories: a unified view. In *Cybernetic Systems Conference*, pages 4–9, Dublin (Ireland). IEEE.
- Dubois, D. and Prade, H. (1988). *Possibility theory*. Plenum Press, New-York.
- Dubois, D. and Prade, H. (1992). When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74. DP164.
- Dubois, D. and Prade, H. (1994). Possibility theory and data fusion in poorly informed environments. *Control Eng. Practice*, 2:811–823.
- Dubois, D., Prade, H., and Smets, P. (2001). New semantics for quantitative possibility theory. In *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ECSQARU '01, pages 410–421, London, UK. Springer-Verlag.
- Duin, R. P. W. (2002). The combining classifier: To train or not to train? In *ICPR (2)*, pages 765–770.
- Durand, T., Thome, N., Cord, M., and Avila, S. (2013). Image classification using object detectors. International Conference on Image Processing, ICIP '13.
- Duygulu, P., Barnard, K., Freitas, J. F. G. d., and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112, London, UK, UK. Springer-Verlag.
- Escalante, H. J., Hérnandez, C. A., Sucar, L. E., and Montes, M. (2008). Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 172–179, New York, NY, USA. ACM.
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

- F. E. White, J. (1987). Data fusion lexicon. *Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego.*
- Fakhar, K., Aroussi, M. E., Saidi, M. N., and Aboutajdine, D. (2012). Score fusion in multibiometric identification based on fuzzy set theory. In Elmoataz, A., Mammass, D., Lezoray, O., Nouboud, F., and Aboutajdine, D., editors, *ICISP*, volume 7340 of *Lecture Notes in Computer Science*, pages 261–268. Springer.
- Fauvel, M., Chanussot, J., and Benediktsson, J. A. (2006). Decision fusion for the classification of urban remote sensing images. *IEEE T. Geoscience and Remote Sensing*, 44(10-1):2828–2838.
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Feng, S., Lang, C., and Xu, D. (2010). Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10*, pages 288–295, New York, NY, USA. ACM.
- Ferret, O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*. European Language Resources Association.
- Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.
- Gao, S., Chia, L.-T., and Cheng, X. (2010). Web image concept annotation with better understanding of tags and visual features. *J. Vis. Comun. Image Represent.*, 21(8):806–814.
- Gao, S., Tsang, I., Chia, L., and Zhao, P. (2011). Local features are not lonely - Laplacian sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3555–3561.
- Garla, V. N. and Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics*, 13(1):261.
- Girolami, M. and Kabán, A. (2003). On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434. ACM.
- Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208.

- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.
- Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. ICCV'09, pages 309 – 316, Kyoto, Japon. IEEE Computer society.
- Guillaumin, M., Verbeek, J., and Schmid, C. (2010). Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 902 – 909.
- Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.
- He, X., Zemel, R., and Carreira-Perpinan, M. (2004). Multiscale conditional random fields for image labeling. In *In CVPR*, pages 695–702.
- Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008). Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In *LREC*. European Language Resources Association.
- Hirst, G. and St Onge, D. (1998). *Lexical Chains as representation of context for the detection and correction malapropisms*. The MIT Press.
- Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, pages 289–296.
- Hofmann, T. (1999b). Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, pages 289–296.
- Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R. (1997). Image indexing using color correlograms. CVPR '97, Washington, DC, USA. IEEE Computer Society.
- Huang, Y., Huang, K., Yu, Y., and Tan, T. (2011). Salient coding for image classification. In *CVPR*.
- Huang, Y., Wu, Z., Wang, L., and Tan, T. (2013). Feature coding in image classification: A comprehensive study. Number 99.
- Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., and Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy.
- Jin, Y., Khan, L., Wang, L., and Awad, M. (2005). Image annotations by combining multiple evidence & wordnet. In *ACM Multimedia*, pages 706–715.

- Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 604–610, Washington, DC, USA. IEEE Computer Society.
- Kawanabe, M., Binder, A., Muller, C., and Wojcikiewicz, W. (2011). Multi-modal visual concept classification of images via markov random walk over tags. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV)*, WACV '11, pages 396–401, Washington, DC, USA. IEEE Computer Society.
- Kennedy, L., Slaney, M., and Weinberger, K. (2009). Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In *Proceedings of the 1st workshop on Web-scale multimedia corpus*, WSMC '09, pages 17–24, New York, NY, USA. ACM.
- Kennedy, L. S., fu Chang, S., and Kozintsev, I. V. (2006). To search or to label?: predicting the performance of search-based automatic image classifiers. In *MIR '06*, pages 249–258.
- Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion*, 14(1):28–44.
- Kludas, J. and Marchand-Maillet, S. (2011). Effective multimodal information fusion by structure learning. In *14th International Conference on Information Fusion (FUSION 2011)*, Chicago, IL.
- Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.
- Landauer, T. K. and Dutnais, S. T. (1997). A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, pages 2169–2178.
- Le Borgne, H. and Honnorat, N. (2010). Fast shared boosting for large-scale concept detection. *Multimedia Tools and Applications*, pages 1–14.
- Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In Fellbaum, C., editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.

- Li, L.-J., Su, H., Xing, E. P., and Fei-Fei, L. (2010a). Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada.
- Li, M., Tang, J., Li, H., and Zhao, C. (2012). Tag ranking by propagating relevance over tag and image graphs. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service, ICIMCS '12*, pages 153–156, New York, NY, USA. ACM.
- Li, W., Min, J., and Jones, G. J. F. (2010b). A text-based approach to the imageCLEF 2010 photo annotation task. In Braschler, M., Harman, D., and Pianta, E., editors, *CLEF (Notebook Papers/LABs/Workshops)*.
- Li, X., Snoek, C. G. M., and Worring, M. (2009a). Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322.
- Li, Y., Crandall, D. J., and Huttenlocher, D. P. (2009b). Landmark classification in large-scale image collections. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1957–1964. IEEE.
- Lienhart, R., Romberg, S., and Hörster, E. (2009). Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09*, pages 9:1–9:8, New York, NY, USA. ACM.
- Lin, D. (1998). An information-theoretic definition of similarity. In Shavlik, J. W., editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann Publishers.
- Lindsey, R., Veksler, V. D., Grintsavay, A., and Gray, W. D. (2007). Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness. In *Proceedings of ICCM 2007: Eighth International Conference on Cognitive Modeling*, pages 279–284, Oxford, UK. Taylor & Francis/Psychology Press.
- Liu, D., Hua, X.-S., Wang, M., and Zhang, H.-J. (2010). Image retagging. In *Proceedings of the international conference on Multimedia, MM '10*, pages 491–500, New York, NY, USA. ACM.
- Liu, D., Hua, X.-S., Yang, L., Wang, M., and Zhang, H.-J. (2009a). Tag ranking. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 351–360, New York, USA. ACM.
- Liu, D., Hua, X.-S., and Zhang, H.-J. (2011a). Content-based tag processing for internet social images. *Multimedia Tools Appl.*, 51(2):723–738.
- Liu, D., Wang, M., Yang, L., Hua, X.-S., and Zhang, H. (2009b). Tag quality improvement for social images. In *ICME*, pages 350–353. IEEE.

- Liu, L., Wang, L., and Liu, X. (2011b). In Defense of Soft-assignment Coding. In *IEEE International Conference on Computer Vision (ICCV)*.
- Liu, N., Dellandra, E., Chen, L., Trus, A., Zhu, C., Zhang, Y., Bichot, C.-E., Bres, S., and Tellez, B. (2012). Liris-imagine at imageclef 2012 photo annotation task. In Forner, P., Karlgren, J., and Womser-Hacker, C., editors, *CLEF (Online Working Notes/Labs/Workshop)*.
- Liu, N., Dellandra, E., Chen, L., Zhu, C., Zhang, Y., Bichot, C.-E., Bres, S., and Tellez, B. (2013). Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. *Computer Vision and Image Understanding*, 117(5):493 – 512.
- Liu, N., Dellandra, E., Tellez, B., and Chen, L. (2011). Associating textual features with visual ones to improve affective image classification. In *International Conference on Affective Computing and Intelligent Interaction (ACII2011)*.
- Liu, N., Zhang, Y., Dellandréa, E., Bres, S., and Chen, L. (2011). Liris-imagine at imageclef 2011 photo annotation task. In *CLEF (Notebook Papers/Labs/Workshop)*.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision (IJCV)*, 60(2):91–110.
- Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208.
- Makadia, A., Pavlovic, V., and Kumar, S. (2008). A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision: Part III, ECCV '08*, pages 316–329, Berlin, Heidelberg. Springer-Verlag.
- Marchand-Maillet, S., Morrison, D., Szekely, E., and Bruno, E. (2010). *Interactive Representations of Multimodal Databases*. Academic Press.
- Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia, HYPERTEXT '06*, pages 31–40, New York, NY, USA. ACM.
- Martin, A., Osswald, C., Dezert, J., and Smarandache, F. (2008). General combination rules for qualitative and quantitative beliefs. *Journal of Advances in Information Fusion*, 3(2):67–89.
- McSherry, F. and Najork, M. (2008). Computing information retrieval performance measures efficiently in the presence of tied scores. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R. W., editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 414–421. Springer.

- Meng, L., Huang, R., and Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12.
- Meyer, G. F., Mulligan, J. B., and Wuerger, S. M. (2004). Continuous audio-visual digit recognition using n-best decision fusion. *Information Fusion*, 5(2):91–101.
- Monay, F. and Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pages 275–278, New York, NY, USA. ACM.
- Monay, F. and Gatica-Perez, D. (2004). pLSA-based image auto-annotation: constraining the latent space. In *ACM Multimedia*, pages 348–351.
- Moser, G. and Serpico, S. B. (2013). Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *IEEE T. Geoscience and Remote Sensing*, 51(5-1):2734–2752.
- Nagel, K., Nowak, S., Kühhirt, U., and Wolter, K. (2011). The fraunhofer idmt at imageclef 2011 photo annotation task. In *CLEF (Notebook Papers/Labs/Workshop)*.
- Niaz, U. and Mérialdo, B. (2013). Fusion methods for multi-modal indexing of web data. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pages 1–4.
- Nikolopoulos, S., Zafeiriou, S., Patras, I., and Kompatsiaris, I. (2013). High order pls-a for indexing tagged images. *Signal Process.*, 93(8):2212–2228.
- Nov, O. and Ye, C. (2010). Why do people tag?: motivations for photo tagging. *Commun. ACM*, 53(7):128–131.
- Nowak, S. and Huiskes, M. J. (2010). New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.
- Nowak, S., Nagel, K., and Liebetrau, J. (2011). The clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*.
- Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145–175.
- Oussalah, M., Maaref, H., and Barret, C. (2001). New fusion methodology approach and application to mobile robotics: investigation in the framework of possibility theory. *Information Fusion*, 2(1):31–48.
- Panchenko, A. (2013). *Similarity Measures for Semantic Relation Extraction*. PhD thesis, Université catholique de Louvain & Bauman Moscow State Technical University.

- Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 17(3):423–435.
- Patwardhan, S. (2003). Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master’s thesis, University of Minnesota.
- Pawlak, Z., Grzymala-Busse, J., Slowinski, R., and Ziarko, W. (1995). Rough sets. *Commun. ACM*, 38(11):88–95.
- Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet: : Similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025.
- Perronnin, F. and Dance, C. R. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society.
- Perronnin, F., Liu, Y., Sánchez, J., and Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3384–3391. IEEE.
- Popescu, A. and Grefenstette, G. (2011). Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 33:1–33:8.
- Quintarelli, E. (2005). Folksonomies: power to the people.
- Quost, B., Masson, M.-H., and Denoeux, T. (2011). Classifier fusion in the dempster–shafer framework using optimized t-norm based combination rules. *Int. J. Approx. Reasoning*, 52:353–374.
- Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. In *IEEE Transactions on Systems, Man and Cybernetics*, pages 17–30.
- Rakotomamonjy, A., Rouen, U. D., Bach, F., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research* 9.
- Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009). Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, ECML PKDD ’09*, pages 254–269.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

- Romberg, S., Lienhart, R., and Hörster, E. (2012). Multimodal image retrieval. *International Journal of Multimedia Information Retrieval*, 1(1):31–44.
- Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.
- Schutze, H. (1993). Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.
- Schwab, D., Lafourcade, M., and Prince, V. (2002). Antonymy and conceptual vectors. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1, COLING '02*, pages 1–7. Association for Computational Linguistics.
- Sebti, A. and Barfroush, A. A. (2008). A new word sense similarity measure in wordnet. In *IMCSIT*, pages 369–373. IEEE.
- Seco, N., Veale, T., and Hayes, J. (2004a). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *ECAI'2004, the 16th European Conference on Artificial Intelligence*.
- Seco, N., Veale, T., and Hayes, J. (2004b). An intrinsic information content metric for semantic similarity in WordNet. *Proc. of ECAI*, 4:1089–1090.
- Semenovich, D. and Sowmya, A. (2010). Geometry aware local kernels for object recognition. In *ACCV (1)*, pages 490–503.
- Sen, S., Vig, J., and Riedl, J. (2009). Tagommenders: connecting users to items through tags. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 671–680, New York, NY, USA. ACM.
- Shabou, A. and Le Borgne, H. (2012). Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, pages 3618–3625. IEEE.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.
- Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web, WWW '08*, pages 327–336, New York, NY, USA. ACM.
- Sivic, J. and Zisserman, A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1470–1477.
- Smarandache, F. and Dezert, J., editors (2009). *Advances and Applications of DSmT for Information Fusion : Collected works. volume 3*. American Research Press, Rehoboth.

- Smets, P. (1989). Constructing the pignistic probability function in a context of uncertainty. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., editors, *UAI*, pages 29–40. North-Holland.
- Smets, P. (1999). Practical uses of belief functions. In Laskey, K. B. and Prade, H., editors, *UAI*, pages 612–621.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22:1349–1380.
- Snoek, C. G. M. (2005). Early versus late fusion in semantic video analysis. In *ACM Multimedia*, pages 399–402.
- Souvannavong, F., Mérialdo, B., and Huet, B. (2005). Multi-modal classifier fusion for video shot content retrieval. In *WIAMIS 2005, 6th International Workshop on Image Analysis for Multimedia Interactive Services, April 13-15, 2005, Montreux, Switzerland, Montreux, SUISSE*.
- Stehling, R. O., Nascimento, M. A., and Falcao, A. X. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 102–109, McLean, Virginia, USA.
- Strohmaier, M., Krner, C., and Kern, R. (2012). Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17(0):1 – 11.
- Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st national conference on Artificial intelligence*, pages 1419–1424. AAAI Press.
- Suchanek, F. M., Vojnovic, M., and Gunawardena, D. (2008). Social tags: meaning and suggestions. In *Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08*, pages 223–232, New York, NY, USA. ACM.
- Sun, A. and Bhowmick, S. S. (2009). Image tag clarity: in search of visual-representative tags for social images. In *Proceedings of the first SIGMM workshop on Social media, WSM '09*, pages 19–26, New York, NY, USA. ACM.
- Sun, F., Li, H., Zhao, Y., Wang, X., and Wang, D. (2013). Towards tags ranking for social images. *Neurocomputing*, 120(0):434 – 440.
- Tang, J., Yan, S., Hong, R., Qi, G.-J., and Chua, T.-S. (2009). Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the 17th ACM international conference on Multimedia, MM '09*, pages 223–232, New York, NY, USA. ACM.

- Tapscott, D. and Williams, A. D. (2006). *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio Hardcover, New York, NY.
- Thomee, B. and Popescu, A. (2012). Overview of the imageclef 2012 flickr photo annotation and retrieval task. In *CLEF (Online Working Notes/Labs/Workshop)'12*, pages –1–1.
- Ting, K. M. and Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289.
- Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970.
- Torresani, L., Szummer, M., and Fitzgibbon, A. (2010). Efficient object category recognition using classemes. In *Proceedings of the 11th European Conference on Computer Vision: Part I, ECCV'10*, pages 776–789, Berlin, Heidelberg. Springer-Verlag.
- Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning, EMCL '01*, pages 491–502, London, UK, UK. Springer-Verlag.
- Tversky, A. (1977). Features of Similarity. In *Psychological Review*, volume 84, pages 327–352.
- van Gemert, J., Veenman, C., Smeulders, A., and Geusebroek, J. (2009). Visual word ambiguity. *PAMI*, pages 1271–1283.
- Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall.
- Wang, C., Jing, F., Zhang, L., and Zhang, H.-J. (2008). Scalable search-based image annotation. *Multimedia Syst.*, 14(4):205–220.
- Wang, G., Chua, T.-S., Ngo, C.-W., and Wang, Y. (2010a). Automatic generation of semantic fields for annotating web images. In *COLING (Posters)*, pages 1301–1309. Chinese Information Processing Society of China.
- Wang, G., Hoiem, D., and Forsyth, D. A. (2009a). Building text features for object image classification. In *CVPR*, pages 1367–1374.
- Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010b). Locality-constrained linear coding for image classification. In *CVPR*.
- Wang, M., Yang, K., Hua, X.-S., and Zhang, H.-J. (2009b). Visual tag dictionary: interpreting tags with visual words. In *Proceedings of the 1st workshop on Web-scale multimedia corpus*, pages 1–8. ACM.

- Wang, Z., Feng, J., Zhang, C., and Yan, S. (2010c). Learning to rank tags. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10*, pages 42–49, New York, NY, USA. ACM.
- Weinberger, K. Q., Slaney, M., and Van Zwol, R. (2008). Resolving tag ambiguity. In *Proceeding of the 16th ACM international conference on Multimedia, MM '08*, pages 111–120. ACM.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.
- Wu, L., Jin, R., and Jain, A. K. (2012). Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints).
- Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Annual Meeting of the Association for Computational Linguistics*, pages 133 –138.
- Xioufis, E. S., Sechidis, K., Tsoumakas, G., and Vlahavas, I. P. (2011). Mlkd’s participation at the clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*.
- Xu, H. and Chua, T.-S. (2006). Fusion of av features and external information sources for event detection in team sports video. *TOMCCAP*, 2(1):44–67.
- Xu, H., Wang, J., Hua, X.-S., and Li, S. (2009). Tag refinement by regularized lda. In *Proceedings of the 17th ACM international conference on Multimedia, MM '09*, pages 573–576, New York, NY, USA. ACM.
- Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801.
- Yang, K., Hua, X.-S., Wang, M., and Zhang, H.-J. (2011). Tag tagging: Towards more descriptive keywords of image content. *Multimedia, IEEE Transactions on*, 13(4):662–673.
- Younes, Z., Abdallah, F., and Dencœux, T. (2009). An Evidence-Theoretic k-Nearest Neighbor Rule for Multi-label Classification. pages 297–308.
- Yu, H., Li, M., Zhang, H.-J., and Feng, J. (2003). Color texture moments for content-based image retrieval. *ICIP '10*, pages 24–28.
- Yu, K., Zhang, T., and Gong, Y. (2009). Nonlinear learning using local coordinate coding. *NIPS*, 22:2223–2231.
- Yuan, J., Wu, Y., and Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.
- Zabih, R. and Kolmogorov, V. (2004). Spatially coherent clustering using graph cuts. In *In CVPR (2)*, pages 437–444.

- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.
- Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28.
- Zhang, D., Islam, M. M., and Lu, G. (2012a). A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346 – 362.
- Zhang, Y., Bres, S., and Chen, L. (2012b). Semantic bag-of-words models for visual concept detection and annotation. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pages 289–295.
- Zhao, W.-L., Jiang, Y.-G., and Ngo, C.-W. (2006). Keyframe Retrieval by Key-points: Can Point-to-Point Matching Help? In *CIVR*.
- Zhou, X., Yu, K., Zhang, T., Huang, T. S., and Huang, T. S. (2010). Image classification using super-vector coding of local image descriptors. In *ECCV (5)*, pages 141–154.
- Zhou, Z., Wang, Y., and Gu, J. (2008). New model of semantic similarity measuring in wordnet. In *3rd International Conference on Intelligent System and Knowledge Engineering*, volume 1, pages 256–261.
- Zhu, G., Yan, S., and Ma, Y. (2010). Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the international conference on Multimedia, MM '10*, pages 461–470, New York, NY, USA. ACM.
- Zhuang, J. and Hoi, S. C. (2011). A two-view learning approach for image tag ranking. In *Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11*, pages 625–634, New York, NY, USA. ACM.
- Znaidia, A., Le Borgne, H., and Hudelot, C. (2012a). Belief Theory for Large-Scale Multi-label Image Classification. In *Belief Functions: Theory and Applications. Advances in Intelligent and Soft Computing*, volume 164, pages 205–212, Compiègne, France.
- Znaidia, A., Le Borgne, H., and Hudelot, C. (2013a). Tag completion based on belief theory and neighbor voting. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, ICMR '13*, pages 49–56, New York, NY, USA. ACM.
- Znaidia, A., Le Borgne, H., Hudelot, C., and Popescu, A. (2012b). Prise en compte de l'imperfection des tags pour la classification sémantique d'images. In *Actes de la conférence RFIA 2012*, Lyon, France.
- Znaidia, A., Shabou, A., Le Borgne, H., and Hudelot, C. (2013b). Codage des modèles de tags. *Revue d'Intelligence Artificielle*, 27(1):39–63.

- Znaidia, A., Shabou, A., Le Borgne, H., Hudelot, C., and Paragios, N. (2012c). Bag-of-multimedia-words for image classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 1509–1512.
- Znaidia, A., Shabou, A., Popescu, A., Le Borgne, H., and Hudelot, C. (2012d). Multimodal feature generation framework for semantic image classification. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12*, pages 38:1–38:8, New York, NY, USA. ACM.