Semantic multimedia analysis using knowledge and context

Nikolopoulos, Spyridon

Semantic multimedia analysis using knowledge and context

Spyridon Nikolopoulos

Department of Electronic Engineering

Queen Mary University of London

Co-supervisors:
Dr. Ioannis (Yiannis) Patras & Dr. Ioannis (Yiannis) Kompatsiaris

PhD Thesis

July 10, 2012

Abstract

The difficulty of semantic multimedia analysis can be attributed to the extended diversity in form and appearance exhibited by the majority of semantic concepts and the difficulty to express them using a finite number of patterns. In meeting this challenge there has been a scientific debate on whether the problem should be addressed from the perspective of using overwhelming amounts of training data to capture all possible instantiations of a concept, or from the perspective of using explicit knowledge about the concepts' relations to infer their presence. In this thesis we address three problems of pattern recognition and propose solutions that combine the knowledge extracted implicitly from training data with the knowledge provided explicitly in structured form. First, we propose a BNs modeling approach that defines a conceptual space where both domain related evidence and evidence derived from content analysis can be jointly considered to support or disprove a hypothesis. The use of this space leads to significant gains in performance compared to analysis methods that can not handle combined knowledge. Then, we present an unsupervised method that exploits the collective nature of social media to automatically obtain large amounts of annotated image regions. By proving that the quality of the obtained samples can be almost as good as manually annotated images when working with large datasets, we significantly contribute towards scalable object detection. Finally, we introduce a method that treats images, visual features and tags as the three observable variables of an aspect model and extracts a set of latent topics that incorporates the semantics of both visual and tag information space. By showing that the cross-modal dependencies of tagged images can be exploited to increase the semantic capacity of the resulting space, we advocate the use of all existing information facets in the semantic analysis of social media.

To my parents ...

Acknowledgements

I would like to acknowledge the opportunity that has been given me by the Information Technologies Institute to work in a stimulating environment and collaborate with many respectful researchers around Europe helping me to identify my research interests, as well as the Queen Mary University of London that helped me to organize my research efforts and work towards my PhD thesis. I would particularly like to thank Dr. Ioannis Kompatsiaris and Dr. Ioannis Patras that took the initiative to establish a radical new form of collaboration between these institutes, allowing me to get the best out of both worlds.

During my thesis I have received significant help from a number of people.

First of all I would like to acknowledge the help received from my two supervisors Dr. Ioannis Patras and Dr. Ioannis Kompatsiaris that contributed with their ideas, guidance, stimulating debates and critical view on my research outcomes. The quality of this work would have been compromised without their help.

Secondly, I need to acknowledge the help received from my fellow workers Elisavet Chatzilari and Christina Lakka that have undertaken a significant part of the technical and scientific work included in this thesis. Many of the demanding experiments presented would have been impossible without their help.

Finally, I must also acknowledge the contribution of my fellow researchers working in the same research team that have turned my working environment into a continuous source of inspiration.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Glossary

AP	Average Precision
ASR	Automatic Speech Recognition
BN(s)	Bayesian Network(s)
CPT(s)	Conditional Probability Table(s)
CRF	Conditional Random Field
DAG	Directed Acyclic Graph
DL	Description Logics
DTSBN(s)	Dynamic Tree-Structure Belief Network(s)
EM	Expectation Maximization
FN	False Negative
FoA	Focus of Attention
FP	False Positive
HMM	Hidden Markov Models
HOSVD	Higher Order Singular Value Decomposition
IPFP	Iterative Proportional Fitting Procedure
JPD	Joint Probability Distribution
KL	Kullback-Leibler
LDA	Latent Dirichlet Allocation
MAP	Mean Average Precision
MFHMM	Multistream Fused Hidden Markov Models
MPEG	Moving Picture Experts Group
MRF	Markov Random Fields
NMI	Normalized Mutual Information
NN	Nearest Neighbor
OWL	Ontology Web Language
pLSA	probabilistic Latent Semantic Analysis
SEMSOC	Semantic, Social and Content-based clustering
SIFT	Scale Invariant Feature Transformation
SVM	Support Vector Machine
TRECVID	TREC Video Retrieval Evaluation

Chapter 1

Introduction

1.1 Theme of the thesis

In the new era of Information and Telecommunication Technologies (ICT) vast amounts of digital content are being generated at a constantly growing pace. The recent advances of Web technologies have effectively turned ordinary people into active members of the Web that generate, share, contribute and exchange considerable amounts of digital information. However, the limitations of machine understanding makes it difficult for automated systems to index all this content in a manner coherent with human visual perception. This fact has turned the discovery of intelligent ways for information consumption into one of the main challenges in computer science [2]. With respect to multimedia, the difficulty of mapping a set of low-level visual features into semantic concepts, generally addressed as bridging the "Semantic Gap" [3], has brought the mechanisms of learning and pattern recognition into the forefront of related scientific research.

As humans we learn to recognize materials, objects and scenes from very few examples and without much effort. A 3-year old child is capable of building models for a substantial number of concepts and recognizing them using these models. By the age of six, humans recognizes more than 10^4 categories of objects [4] and keep learning more throughout their life. Can a computer program learn how to recognize semantic concepts in multimedia content the way a human does? This is the general question addressed by the scientists in computer vision and many other disciplines. But what is exactly the process of building perceptual models? is there indeed a mechanism that

allows humans to initially require many examples to learn, as performed by little babies, and after they have learned how to learn they can learn from just a few examples? and most importantly what is the role of the teacher in this process?

Semantic object detection is one of the most typical examples of perceptual learning and can be considered as one of the most useful operations performed by the human visual system. Many researchers in the field have focused on trying to discover an efficient (in terms of precision and recall), scalable (in terms of the number of concepts) and effortless (in terms of the necessary annotation) way to teach the machine how to recognize visual objects the way a human does. In order to tackle this problem divergent approaches have been proposed, relying either on the use of general knowledge or the abundant availability of data.

The authors of [5] make the hypothesis that once a few visual categories have been learned with significant cost, some information may be abstracted from the process to make learning further categories more efficient. Based on this hypothesis, when learning new visual categories, they take advantage of the general knowledge extracted from previously learned categories by using it in the form of a prior probability density function in the space of model parameters. Similarly in [6] when images of new concepts are added to the visual analysis model, the computer only needs to learn from the new images. What has been learned about previous concepts is stored in the form of profiling models and the computer needs no re-training. In a similar direction online learning algorithms have been investigated by many researchers aiming to exploit their comparatively low computational requirements, since they only need to store and process a single example at a time [7].

On a different perspective, the authors of [8] claim that with the availability of overwhelming amounts of data many problems can be solved without the need for sophisticated algorithms. The authors mention the example of Google's "Did you mean" tool, which corrects errors in search queries by memorizing billions of query-answer pairs and suggesting the one closest to the user query. In their work, the authors present a visual analog to this tool using a large dataset of 79 million images and a non-parametric approach for image annotation that is based on nearest neighbor matching. Additionally, the authors of [9] employ multiple instance learning [10] to learn models from images labeled as containing the semantic concept of interest, but without indication of which image regions are observations of that concept. Similarly

in [11] object recognition is viewed as machine translation that uses expectation maximization in order to learn how to map visual objects (blobs) to concept labels. The approaches relying on human computation such as Google Image Labeler[1] [12] and Peekaboom [13] for image global and regional annotation respectively, also belong to the category of methods that aim at scalable and effortless learning. These are just a few indicative examples that demonstrate the scientific debate around the mechanism of building perceptual models, and the discussion on how much of the knowledge should come in a structured, explicit form and how much can be obtained implicitly from the available training samples.

If we consider that concept detection in multimedia content is the result of a continuous process where the learner interacts with a set of examples and his teacher to gradually develop his system of visual perception, we may identify the following interrelations. The grounding of concepts is primarily achieved through indicative examples that are followed by the description of the teacher (i.e. annotations). Based on these samples the learner uses his senses to build models that are able to ground the annotated concepts, either by relying on the discriminative power of the received stimuli (i.e. discriminative models), or by shaping a model that could potentially generate these stimuli (i.e. generative models). However, these models are typically weak in generalization, at least at their early stages of development. This fact prevents them from successfully recognizing new, un-seen instantiations of the modeled concepts that are likely to differ in form and appearance (i.e. semantic gap). This is where the teacher once again comes into play to provide the learner with a set of logic based rules, or probabilistic dependencies that will offer him an additional path to visual perception through inference. These rules and dependencies are essentially filters that can be applied to reduce the un-certainty of the stimuli-based models, or to generate higher forms of knowledge through reasoning. Finally, when this knowledge accumulates over time it takes the form of experience, which is a kind of information that can be sometimes transferred directly from the teacher to the learner and help him to make rough approximations of the required models.

By no means does this rather simplified description claim to consistently cover all different processes that are required to simulate the system of human visual perception. It is just an effort to put into the same context some of the issues that are intensively

[1] http://images.google.com/imagelabeler/

researched in the field of multimedia understanding. Our thesis is that multimedia analysis requires all aforementioned ingredients to be performed in an efficient manner. Although moving towards the one or the other extreme of the debate may still produce non-trivial recognition models, higher levels of efficiency can only be achieved if explicit and implicit knowledge are effectively combined. In this thesis we verify the aforementioned statement through the development of three different approaches that succeed in boosting the effectiveness of multimedia understanding by: (a) incorporating explicitly provided knowledge and implicitly extracted evidence into a common inference framework, (b) leveraging social media to crowdsource the necessary annotations, and (c) transferring the experience gained from legacy data to produce media representations that are more sensitive to semantics.

1.2 Background on pattern recognition for multimedia analysis and focus of this thesis

Teaching the machine to recognize concepts from visual stimuli has been a great challenge for scientists since the very first steps of computer vision. Throughout the decades there have been remarkable achievements that drastically enhanced the capabilities of the machines both from the perspective of infrastructure (i.e., computer networks, processing power, storage capabilities), as well as from the perspective of processing and understanding of the data. Based on the assumption that humans classify images through models that are built using examples for every single semantic concept, the researchers have been trying to simulate human visual system using pattern recognition.

It is natural that we seek to design and build machines that can recognize patterns. From automated speech recognition, fingerprinting identification, optical character recognition and much more, it is clear that reliable and accurate pattern recognition would offer a great deal of perceptual capabilities. In the following we will use a trivial example to briefly describe some of the most important notions in pattern recognition for multimedia analysis, following the argumentation of [14]. Suppose that we have two different types of objects that appear in mixed order and need to be sorted according to their class using optical sensing. We install a camera, take some sample images and begin to note some physical differences between the two types of objects, e.g. length, lightness, width, shape, etc. These physical differences suggest **features**

that can be explored by the classifier to achieve the aforementioned task. Given that there are truly differences between the population of the two objects, we view them as having different **models**, i.e. different descriptions that are typically mathematical in form. The goal of pattern recognition in this case is to hypothesize the class of these models and for any sensed pattern choose the model that corresponds best. All other techniques employed to achieve this aim should be considered parts of the engineer's conceptual toolbox.

First the camera captures an image of the object. Then, the camera's signal is pre-processed to simplify subsequent operation without loosing relevant information. For instance, one very frequent form of pre-processing is **segmentation**, where the images of different objects are somehow isolated from the background, or from the other objects depicted in the image. The information from a single object is then sent to the **feature extractor** whose purpose is to reduce the data by measuring certain "features" or "properties". The values of these features are then passed to a **classifier** that evaluates the evidence presented and makes a final decision as to the class of the object. Now let us consider how the feature extractor and classifier might be designed. Suppose someone gives us the information that object α is generally longer than object β. This information can be used to create a tentative model for the objects: object α has some typical length, and this is greater than that of object β. Thus, length becomes an obvious feature and we may attempt to classify the object merely by checking whether or not the length of the object exceeds some critical value. To choose this critical value we could obtain some **training samples** of the different objects, take some length measurements and inspect the results.

However, it is very likely that we reach the disappointing conclusion that although object α is somewhat longer than object β, on average, there is a non-trivial number of cases where we can not reliably separate the object α from object β based solely on their length. To improve the situation we may try another feature (e.g. the object shape) and check whether this new feature is more efficient at separating the two object classes. If we need to further improve the resulting performance and there is no other feature available that performs better than the object's shape, we must resort to the use of more than one feature at a time. Using both features our problem now is to partition the two-dimensional space into regions, such that all patterns in one region to be classified as object α and all patterns in the other to be classified as object β.

By measuring the features in a set of training samples and obtaining a scattering plot of their points, we may come up with a **decision boundary** in the two dimensional space that succeeds in separating the objects. Thus, using two features appears to be beneficial for separating our samples and suggests that by incorporating yet more features would probably lead to better results.

However, with such a "solution", our satisfaction would be limited because the central aim of designing a classifier is to make predictions when presented with novel patterns, i.e. object instantiations that are not yet seen. This is the issue of **generalization**. It is unlikely that the aforementioned decision boundary will provide good generalization, since it will be probably "tuned" to the particular training samples, rather than some underlying characteristics or the true model of all type α and β objects. One natural approach to tackle this problem would be to get mode training samples for obtaining a better estimate of the true underlying characteristics. However, in most pattern recognition problems the amount of training data that we can obtain is quite limited. A solution to this problem is discussed in Chapter 4, where we focus on using the collective knowledge aggregated in social sites to automatically determine a set of image regions that can be associated with a certain object. Due to the common background that most users share, the majority of them tend to contribute relevant tags when faced with a similar type of visual content. Thus, by relying on the assumption that the most frequently appearing "term" in both tag and visual information space will converge into the same object, we manage to leverage social media and effortlessly extract a set of training samples that are readily compatible with an object detection classifier.

On the other hand, when we have insufficient training data, a central technique is to incorporate **domain knowledge**. Indeed the less the training data the more important is such knowledge, for instance how the patterns themselves were produced. Many real world pattern recognition problems seek to incorporate at least some knowledge about the method of patterns' production (i.e. incorporate knowledge on how each pattern is generated), or their functional use (e.g. a bicycle is a moving object that stands on two wheels). Then, we may try to deduce some functional properties from the image (e.g. detect two round objects close to each other) and validate the presence of a bicycle in the image. Additionally, we might be able to use **context** (i.e. input-independent

information other than the pattern itself). For instance, we might have the information that our objects are presented to the classifier in sets and if we are examining a sequence of type α objects, it is highly likely that the next object will be of type α. Thus, if after a long series of objects α the classifier detects an ambiguous pattern, it may nevertheless be best to categorize it as object α. However, context can be highly complex and abstract and the appropriate mechanisms for incorporating such information still remain a challenging task. Chapter 3 introduces an approach for addressing this challenge by using ontologies and Bayesian Networks (BNs) to incorporate into a common inference framework both domain knowledge and applications context. A modeling approach is proposed for the BNs that defines a conceptual space suitable for incorporating the evidence derived from content analysis. The evidence derived from the statistical processing of media features is injected into the common inference framework and invoke an evidence-driven probabilistic inference process. The goal of this process is essentially to verify or reject a hypothesis made about the semantic content of the examined media item.

Finally, we are also concerned with the problem of efficient **media representation**. In ideal media representation the structural relationships among the components are naturally revealed and the true (unknown) model of the patterns can be easily expressed. Thus, what we look for is a representation in which the patterns that lead to the same action are somehow "close" to one another, yet far from those that demand a different action. The extent to which we create or learn a proper representation and how we quantify distance will determine the success of our pattern classifier. Towards this objective, Chapter 5 proposes an approach for media representation with a focus on extending the currently used aspect models to higher order, so as to become applicable for more than two observable variables. Using these models we incorporate different information spaces into the analysis process and benefit from the cross-modal dependencies that are likely to exist among them. In this way, we succeed in devising a feature extraction scheme that is more sensitive to semantics, since the co-existence of two information items that are known from experience to appear together rather frequently is more important in defining the resulting feature space, than the co-existence of two information items that rarely appear together and are likely to be the result of noise.

1.3 Contributions of this thesis

1.3.1 Model BNs to incorporate implicit and explicit knowledge into a common inference framework

Towards the objective of making context and explicitly provided knowledge an integral part of multimedia analysis, our contribution can be summarized in the following:

- A generic approach for modeling BNs that defines a conceptual space where both explicitly provided evidence and evidence derived from content analysis can be jointly considered to support or disprove a hypothesis.
- A methodology for transforming ontologies into semantically enhanced BNs such that the process of probabilistic inference is influenced by a set of human-defined, logic-based rules.

Building on those principles we have developed a framework for multimedia analysis that was applied and evaluated in two different case studies. In the first case our goal has been to perform concept detection using image local and global information. Some of the novel aspects that were particularly introduced for this case include:

1. A data-oriented learning strategy for estimating the prior and conditional probabilities required by the BN.
2. A focus-of-attention mechanism capable of exploiting the mutual information between concepts in order to significantly reduce the computational cost of visual inference and still achieve comparable results with the exhaustive case.

In the second case study our goal has been to apply the developed framework for analyzing compound documents by jointly considering the evidence extracted across media. The novel aspects that were particularly introduced for this case include:

1. A concrete problem instantiation where cross-media analysis proves beneficial for the task of concept detection in compound documents.
2. A coherent methodology on how compound documents can be disassembled to their constituent parts and how single media analyzers can be applied on these parts to extract the content-based evidence.

1.3.2 Verify that social media exhibit noise reduction properties and exploit them to facilitate scalable object detection

Towards the objective of effortlessly obtaining large amounts of training data, our contribution can be summarized in the following:

- We show that the collective nature of social media results in noise reduction properties that can be practically exploited to obtain large amounts of annotated image regions.
- We introduce a totally un-supervised method that associates image regions with tags by performing clustering on the visual and tag information space and matching the most populated clusters.
- We study theoretically and experimentally the conditions under which the aforementioned approach is expected to result in valid training samples and derive some intuitive relation between the size of the processed dataset, the amount of visual analysis error and the success probability of our approach.

1.3.3 Extend current models for incorporating the cross-modal dependencies of multi-modal data

Towards the objective of semantics sensitive media representation, our contribution can be summarized in the following:

- We introduce high order probabilistic Latent Semantic Analysis (pLSA) for treating images, visual features and tags as the three observable variables of an aspect model and extract a set of latent topics that incorporate the semantics of both visual and tag information space.
- We introduce the concept of profile to be the occurrence distribution of an information item within a large corpus of images and use the vector distance between two profiles to measure the dependency between the information items of two different modalities.
- We integrate the cross-modal dependencies into the update rules of high order pLSA in order for the co-existence of two information items that are known from

experience to appear together rather frequently, to be more important in defining the topics of the resulting latent space, than the co-existence of two information items that rarely appear together and are likely to be the result of noise.

1.4 Thesis's Structure

In Chapter 2 we review the related literature in the fields that are relevant with the topics addressed in this thesis. More specifically, in Section 2.1 we are mostly concerned with the research efforts aiming at combining implicit and explicit knowledge under a common inference framework. We distinguish between the body of works that employ this combination for the purpose of semantic image annotation and the body of works that use similar techniques to perform semantic analysis in a cross media setting. In Section 2.2 we summarize the research works that aim at learning from weakly annotated and noisy training data by exploiting the arbitrary large amount of available samples. Section 2.3 focuses on the works that seek to exploit the hidden relations of multi-modal data and devise a media representation scheme with increased semantic capacity.

Chapter 3 describes our proposal for performing evidence-driven probabilistic inference in the grounds of knowledge and context. It provides some background information on Ontologies and Bayesian Networks and explains how to model a Bayesian Network so as to smoothly incorporate the conceptual information obtained from content analysis. Finally, two case studies are presented that demonstrate how the proposed approach can be used to perform concept detection using image local and global information, as well as how compound documents can be analyzed using information across media.

Chapter 4 describes the approach proposed for leveraging social media to crowdsource the necessary annotations and achieve scalable object detection. In this chapter we study theoretically and experimentally when the prevailing trends (in terms of appearance frequency) in visual and tag information space converge into the same object, and how this convergence is influenced by the number of utilized images and the accuracy of the visual analysis algorithms. In this way we allow the reader to derive some intuitive conclusions about the success probability of the proposed approach and as a consequence the resulting performance of the object detection models. Thorough

experiments are performed to provide some indicative measures of the performance loss that we suffer when compared to manually trained models.

Chapter 5 introduces the proposed feature extraction scheme that jointly considers visual and tag information to obtain a semantics sensitive feature space. We explain how the currently used aspect models can be extended to handle more than two observable variables and how this property can be exploited to efficiently index a set of tagged images obtained from flickr. The efficiency of the resulting feature space is evaluated for the tasks of image retrieval and clustering, showing how the proposed scheme can be used to devise a semantics sensitive feature space.

Finally Chapter 6 concludes our thesis by discussing the conclusions we have drawn from our theoretical and experimental studies and outlines three trade-offs that we have encountered. Our plans for future work are also included in this Chapter.

Chapter 2

Literature Review

In this chapter we review some of the most indicative works in the literature of semantic multimedia analysis and learning. Our goal is to provide a comprehensive overview of the research activity that has taken place in the fields that are pertinent to the focus of this thesis. We review the related literature in three main sections aiming to highlight in each case the weaknesses of the existing solutions and justify how the approaches proposed in this thesis succeeds in progressing beyond state-of-the-art.

2.1 Combining implicit and explicit knowledge

The combination of implicit and explicit knowledge under a common inference framework has been extensively studied. In the following, we overview the body of works that were considered most relevant with the approach of Chapter 3. We distinguish between the works aiming at semantic image annotation and the ones that perform semantic analysis in a cross media setting

2.1.1 Semantic image annotation

Interpreting images in terms of their semantic content has been primarily addressed by devising methods that map low-level image visual characteristics (i.e., color, shape, texture) to high-level descriptions (i.e., semantic concepts), without making any use of domain knowledge and application context. Some indicative works that have been presented in the literature include [15] where the authors are based on scene-centered rather than object-centered primitives and use the mean of global image features to

represent the gist of a scene, [16] where scene classification is performed using bayesian classifiers that operate on representations determined using a codebook of region types, and [17] where the authors introduce a visual shape alphabet representation with the aim to enable models for new categories to benefit from the detectors build previously for other categories. In this category of solutions we can also classify the methods that make combined use of global and local classification and treat images at a finer level of granularity, usually by taking advantage of image segmentation techniques. In [18] it is demonstrated through several applications how segmentation and object-based methods improve on pixel-based image analysis/classification methods, while in [19] a region-based binary tree representation incorporating adaptive processing of data structures is proposed to address the problem of image classification. Similarly, based on the combined use of local and global classification, [20] proposes a multi-level approach to annotate the semantics of natural scenes by using both the dominant image components (salient objects) and the relevant semantic objects, [21] employs Multiple-Instance-Learning to learn the correspondence between image regions and keywords and uses a bayesian framework for performing classification, while [22] presents a method where a new object is explained solely in terms of a small set of exemplar objects (represented as image regions). For each exemplar object a separate distance function is learned which captures the relative importance of shape, color, texture and position features. However, the inadequacy of the solutions relying solely on visual information to achieve efficient image interpretation has motivated the exploitation of context as a valuable source of information.

Context was defined in [23] as an extra source of information for both object detection and scene classification. Among the methods that make use of such information, we can identify the class of methods that develop models for spatial context-aware object detection, such as [24] that describes one generic outdoor-scene model, [25] that presents a model specific to individual archetypical scene types (e.g., beach, sunset, mountain, or urban), and [26] where multiple class object-based segmentation is achieved through the integration of mean-shift patches. Another class of methods that make use of such extra information includes the ones that exploit temporal context, as this can be derived from the surrounding images of an image collection (i.e., images drawn during a festival). In [27] the authors developed a general probabilistic temporal

context model in which the first-order Markov property is used to integrate content-based and temporal context cues. Temporal context has been also used for active object recognition [28], as well as for identifying temporally related events [29]. Imaging context (i.e., camera metadata tags about scene capture properties, such as exposure time and subject distance) has been also used for aiding in a number of multimedia analysis tasks, including indoor-outdoor classification and event detection [30]. Other works that aim at improving the performance of individual detectors using contextual information are the ones that model the relationships between objects, such as [31] where contextual features are incorporated into a probabilistic framework which combines the outputs of several components, [32] where the authors present a two-layer hierarchical formulation to exploit the different levels of contextual information, and [33] where the authors propose a region-based model which combines appearance and scene geometry to automatically decompose a scene into semantically meaningful regions.

There is also a number of works that exploit conceptual context by developing techniques that are able to handle uncertainty and take advantage of domain knowledge. The authors of [34] introduce "Multijects" as a way to map time sequence of multi-modal, low-level features to higher level semantics using probabilistic rules. "Multinets" are also proposed for representing higher-level probabilistic dependencies between "Mutlijects". In [35] "Multinets" are elaborated by introducing BNs for modeling the interaction between concepts and using this contextual information to perform semantic indexing of video content. A drawback of these approaches lies on the fact that the structure of "Multinets" is customly defined by experts and no methodology is suggested for explicitly incorporating the semantic constraints originating from the domain into the analysis process. In the same lines, [36] proposes a framework for semantic image understanding based on belief networks. The authors use three different image analysis tasks to demonstrate the improvement in performance introduced by extracting and integrating in the same knowledge-based inference framework (based on BNs), both low-level and semantic features. Once again, no systematic methodology is presented on how to seamlessly integrate domain knowledge, expressed with a standard knowledge representation language, into the probabilistic inference process. [37] describes an integrated approach of visual thesaurus analysis and visual context that exploits both conceptual and topological context. Another approach that attempts to

model uncertainty and take advantage of knowledge and context for the task of multimedia analysis is [38]. This work uses low-level features and a BN to perform indoor versus outdoor scene categorization. In [39] a BN is utilized as an inference mechanism for facilitating a classification method based on feature space segmentation. Similarly, [40] propose a generative-model framework, namely dynamic tree-structure belief networks (DTSBNs), and formulates object detection and recognition as an inference process on a DTSBN. Domain knowledge is also used in [41], in order to tackle the problem that when training data is incomplete or sparse, learning parameters in BNs becomes extremely difficult. In their work the authors present a learning algorithm that incorporates domain knowledge into the learning process in order to regularize the otherwise ill-posed problem. Still, the absence of a methodology for integrating ontological knowledge into the inference process is what differentiates these works from our approach.

Works that utilize ontologies as a means to encode domain knowledge are also present in the literature. [42] presents a method for combining ontologies and BNs in an effort to introduce uncertainty in ontology reasoning and mapping. The Ontology Web Language (OWL) is augmented to allow additional probabilistic markups and a set of structural translation rules convert an OWL ontology into a directed acyclic graph of a BN. The conditional probability tables of the nodes are then calculated taking into consideration the ontology semantics. Probabilistic rules are used to cope with uncertainty and ontologies combined with belief networks are employed to express and migrate into a computationally enabled framework, the semantics originating from the domain. The proposed inference approach is validated using a synthetic example and no attempt is made to adjust the scheme for image analysis. [43] proposes a knowledge assisted image analysis scheme that combines local and global information for the task of image categorization and region labeling. In this case, a sophisticated decision mechanism that takes into account visual information, the concepts' frequency of appearance and their spatial relations is used to analyze images. [44] describes a scheme that is intended to enhance traditional image segmentation algorithms by incorporating semantic information. In this case, fuzzy theory and fuzzy algebra are used to handle uncertainty while a graph of concepts carrying degrees of relationship on its edges is employed to capture visual context. In [45] the authors build a concept ontology using

both semantic and visual similarity in an effort to exploit the inter-concept correlations and to organize the image concepts hierarchically. In this process, the authors try to effectively tackle the problem of intra-concept visual diversity by using multiple kernels. However, none of [43], [44], [45] attempt to couple ontology-based approaches with probabilistic inference algorithms for combining concept detectors, context and knowledge. On the other hand, [46] uses ontologies as a structural prior for deciding on the structure of a BN, but in this work ontologies are mostly treated as hierarchies that do not incorporate any explicitly provided semantic constraints.

Finally, we should also note that none of these works is concerned with computational efficiency and the fact that in a real world inference system the number of plausible hypotheses could suffer from a combinatorial explosion. In Chapter 3 we provide a solution to this problem by proposing a focus of attention mechanism that is based on the mutual information between concepts.

2.1.2 Combining information across media

In the research field of multimedia analysis, indexing and retrieval, various methods have been proposed for fusing the evidence extracted from different media sources. Statistical methods are widely used for multimodal integration [47], where the query object is classified based on the distribution of patterns in the space spanned by pattern features. The most frequently encountered methods are Bayesian Networks that assign a pattern to the class which has the maximum estimated posterior probability, and Hidden Markov Models (HMM) that assign a pattern to a class based on a sequential model of state and transition probabilities.

In this context, our study can be considered to share similar objectives with various works in this field. Within the scope of probabilistic inference, Hospedales and Vijayakumar [48] implement a multisensory detection, verification and tracking mechanism by inferring the association between observations. More specifically, in order to solve the who-said-what problem they present a principle probabilistic approach, where Bayesian inference is used for combining multiple sensing modalities. The proposed model is claimed to be sufficient for robust multitarget tracking and data association in audiovisual sequences. In [49] Choi et al. present three classifier fusion methods and evaluate their efficacies on raw data sets. They use class-specific Bayesian fusion, joint optimization of the fusion process and individual classifiers, and employ dynamic fusion

for combining the posterior probabilities from individual classifiers. The results of the proposed approaches are generally better than the majority voting and the naive Bayes fusion approaches, and significantly reduce the overall diagnostic error in automotive systems. Compared to Bayesian Networks, Hidden Markov Models are capable not only to integrate multimodal features but also to include sequential features. In [50] the MFHMM (Multistream Fused Hidden Markov Model) is presented as a generalization of a two-stream fused HMM [51] for integrating coupled audio and visual features. MFHMM is used for linking the multiple HMMs and is claimed by the authors to be an optimal solution according to the maximum entropy principle and the maximum mutual information criterion. In [52] the authors rely on SVMs and present a late fusion scheme where the unimodal features are initially used to learn separate concept classifiers. Then the output of these classifiers are concatenated to determine a new feature space and learn an SVM-based integrated concept classifier.

Recently, semi-supervised graph-based methods have also attracted the interest of researchers for narrowing the semantic gap between the low- and high-level features. Hoi et al. [53] present multi-modal fusion through graphs in addition with a multi-level graph-based ranking scheme for content-based video retrieval. They present the semi-supervised ranking (SSR) method to exploit both labeled and unlabeled data effectively and further explore a multilevel ranking solution to solve the scalability problem of SSR. The proposed multilevel ranking scheme achieves good performance for large scale applications and also provides a solution to the overfitting problem. In the same direction, Wang et al.[54] present the OMG-SSL method, optimized multigraph-based semi-supervised learning, as an efficient video annotation scheme. The proposed approach is equivalent to fusing multiple graphs and then conducting semi-supervised learning on the fused graph. According to the results, the OMG-SSL method improves the learning performance and can be easily extended through utilizing more graphs. The work in [55] proposes a fusion framework in which classification models are build for each data source independently. Then, using a hierarchical taxonomy of concepts, a Conditional Random Field (CRF) based fusion strategy is designed. According to the fusion scheme described in this work, a graph is defined over the hierarchical taxonomy (i.e., a tree over categories) where its node represents a category. The scores from different unimodal classifiers referring to the same category are concatenated in a feature vector, which serves as the observation of the corresponding node. This work is very

similar with our approach from the perspective of integrating explicit knowledge into the analysis process. However, in this case the scores obtained from the unimodal classifiers are concatenated to form the observation vector for each node. The advantage of our approach over this work is that we use the space of likelihood estimates as a "lingua franca" between the heterogeneous types of information, removing the need to homogenize the output of unimodal classifiers. A semi-supervised approach is employed in [56] where the authors propose to facilitate the learning process by integrating both visual and linguistic information, as well as unlabeled multi-modal data. Their approach is based on co-training which is a semi-supervised learning algorithm that requires two distinct "views" of the training data. Co-training first learns a separate classifier for each view using labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data. Compared to our approach the aforementioned solution is unable to exploit the prior information derived from the co-occurrence of concepts, as well as the knowledge derived from the domain.

ClassView [57] is the method presented by Fan et al. for performing video indexing and retrieval. The authors use a hierarchical, semantics sensitive classifier for bridging the semantic gap between low- and high-level features, while the expectation maximization algorithm is used to determine the feature subspace and the classification rule. The domain-dependent concept hierarchy of video contents in the database, similar to our approach, determines the hierarchical structure of the semantics-sensitive video classifier. The proposed scheme turns out to be effective and closer to the human-level video retrieval. Weiet et al. [58] fuse multimodal cues hierarchically via a cross-reference (CR) method. The authors present CR-Reranking for inferring the most relevant (in a semantic sense) shots, achieving high accuracy. First the initial search results are clustered in diverse feature spaces, then the clusters are ranked by their relevance to the query and finally all the clusters are hierarchically fused via the cross-reference strategy. Finally, Lim et al. [59] combine generative with discriminative models in a sequential manner. Generative models that incorporate explicit knowledge are constructed using a small set of training samples. Subsequently, these generative models are used to classify new samples and augment the existing set with new training samples. In this way the authors manage to generate a set of training samples, sufficiently large to learn a robust discriminative classifier. Thus, the incorporation of explicit knowledge is not

so much intended to facilitate the classification process by enforcing certain rules, but to indirectly improve the classification performance of the discriminative classifier by offering more training samples. Compared to [59] the advantage of our approach is that explicit knowledge is made part of the inference process and directly influences the classification performance.

2.2 Leveraging social media to facilitate multimedia analysis

With the rapid evolution of social media considerable interest has been placed on weakly labeled data and their potential to serve as the training samples for various multimedia analysis tasks. The common objective of these approaches is to compensate for the loss in learning from weakly annotated and noisy training data, by exploiting the arbitrary large amount of available samples.

The approach presented in Chapter 4 can be considered to relate with various works in the literature in different aspects. From the perspective of exploring the trade-offs between analysis efficiency and the characteristics of the dataset we find similarities with [60], [61]. In [60] the authors explore the trade-offs in acquiring training data for image classification models through automated web search as opposed to human annotation. The authors try to learn a model that operates on prediction features (i.e. cross-domain similarity, model generalization, concept frequency, within-training-set model quality) and provide quantitative measures in order to estimate when the cheaply obtained data is of sufficient quality for training robust object detectors. In [61] the authors investigate both theoretically and empirically when effective learning is possible from ambiguously labeled images. They formulate the learning problem as partially-supervised multiclass classification and provide intuitive assumptions under which they expect learning to succeed. This is done by using convex formulation and showing how to extend a general multiclass loss function to handle ambiguity.

There are also works [62], [63], [64] that rely on the same principle assumption with our approach, stating that users tend to contribute similar tags when faced with similar type of visual content. In [62] the authors rely on social data to introduce the concept of flickr distance. Flickr distance is a measure of the semantic relation between two concepts using their visual characteristics. The authors rely on the assumption

that images about the same concept share similar appearance features and use images obtained from flickr to represent a concept. Although different in purpose from our approach the authors present some very interesting results demonstrating that social media like flickr can be used to facilitate various multimedia analysis tasks. In [63] the authors make the assumption that semantically related images usually include one or several common regions (objects) with similar visual features. Based on this assumption they build classifiers using as positive examples the regions clustered in a cluster that is decided to be representative of the concept. They use multiple region-clusters per concept and eventually they construct an ensemble of classifiers. They are not concerned with object detection but rather with concept detection modeled as a mixture/constellation of different object detectors. In the same lines, the work presented in [64] investigates inexpensive ways to generate annotated training samples for building concept classifiers. The authors utilize clickthrough data logged by retrieval systems that consist of the queries submitted by the users, together with the images from the retrieved results, that these users selected to click on in response to their queries. The method is evaluated using global concept detectors and the conclusion that can be drawn from the experimental study is that although the automatically generated data cannot surpass the performance of the manually produced ones, combining both automatically and manually generated data consistently gives the best results.

The employment of unsupervised methods (e.g. clustering) for mining images depicting certain objects, is the attribute that relates our approach with [65], [66]. In [65] the authors make use of community contributed collections and demonstrate a location-tag-vision-based approach for retrieving images of geography-related landmarks. They use clustering for detecting representative tags for landmarks, based on their location and time information. Subsequently, they combine this information with vision-assisted process for presenting the user with a representative set of images. Eventually, the goal is to sample the formulated clusters with the most representative images for the selected landmark. In [66] the authors are concerned with images that are found in community photo collections and depict objects (such as touristic sights). The presented approach is based on geotagged photos and the task is to mine images containing objects in a fully unsupervised manner. The retrieved photos are clustered according to different modalities (including visual content and text labels) and Frequent Itemset Mining is applied

on the tags associated with each cluster in order to assign cluster labels. Eventually, the formulated clusters are used to automatically label and geo-locate new photos.

Finally our approach bares also similarities with works like [67], [68] that operate on segmented images with associated text and perform annotation using the joint distribution of image regions and words. In [67] the problem of object recognition is viewed as a process of translating image regions to words, much as one might translate from one language to another. The authors develop a number of models for the joint distribution of image regions and words, using weak annotations. In [68] the authors propose a fully automatic learning framework that learns models from noisy data such as images and user tags from flickr. Specifically, using a hierarchical generative model the proposed framework learns the joint distribution of a scene class, objects, regions, image patches, annotation tags as well as all the latent variables. Based on this distribution the authors support the task of image classification, annotation and semantic segmentation by integrating out of the joint distribution the corresponding variables.

The main factor that differentiates our approach of Chapter 4 from current state-of-the-art is that none of the aforementioned works employ a thorough and systematic analysis of the noise reduction properties that social media are expected to exhibit, due to their collaborative nature of creation. We fill this gap by theoretically and experimentally studying the conditions under which the collective intelligence that is aggregated in social networks can become practically useful in multimedia analysis.

2.3 Exploiting the hidden relations of multi-modal data

The basic motivation for our approach introduced in Chapter 5 it to exploit the multi-modal aspect that is intrinsic in social media. Driven by the same motivation many researchers in the field have investigated specialized methods for the multi-modal analysis of social media.

Among the related works we identify the ones relying on the use of aspect or topic models [69] and the definition of a latent semantic space. For instance the authors of [1; 70] use a pLSA-based model to support multi-modal image retrieval in flickr, using both visual content and tags. They propose to extent the standard single-layer pLSA model to multiple layers by introducing not just a single layer of topics, but a hierarchy of topics. In this way they manage to effectively combine the heterogeneous

information carried by the different modalities of an image. Similarly, pLSA is also the model adopted by the approach presented in [71] for multi-modal image retrieval. However in this case the authors propose an approach to capture 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. The use of aspect models is also the approach followed in [72] for performing tag ranking and image retrieval. The authors extend the model of Latent Dirichlet Allocation (LDA) [73] to a new topic model called regularized LDA, which models the interrelations between images and exploits both the statistics of tags and visual affinities. In this way, they enforce visually similar images to pick similar distributions over topics. In a similar fashion the authors of [74] propose an approach for the multi-modal characterization of social media by combining text features (e.g. tags) with spatial knowledge (e.g. geotags). The proposed approach is based on multi-modal Bayesian models which allow to integrate spatial semantics of social media in well-formed, probabilistic manner.

Improving the retrieval performance of tagged images has been also encountered as a problem of tag relevance learning, with the visual content serving as the driver of the learning process. In this direction the authors of [75; 76] rely 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. Then, based on this intuition, they propose a neighbor voting algorithm for learning tag relevance by propagating common tags through the visual links introduced by visual similarity. Similarly, the work presented in [77] proposes the use of a multi-edge graph for discovering the tags associated with the underlying semantic regions in the image. Each vertex in the graph is characterized with a unique image and the multiple edges between two vertices are defined by thresholding the pairwise similarities between the individual regions of the corresponding images. Then, based on the assumption that any two images with the same tag will be linked at least by the edge connecting the two regions corresponding to the concerned tag, the repetition of such pairwise connections in a fraction of the labeled images is used to infer a common "visual prototype". Tag relevance learning is also the problem addressed in [78], which aims at leaning an optimal combination of the

multi-modality correlations and generate a ranking function for tag recommendation. In order to do this, the authors use each modality to generate a ranking feature, and then apply the Rankboost [79] algorithm to learn an optimal combination of these features.

Recently, there has been also an increasing interest in extending the aspect models to higher order through the use of Tensors [80]. Under this line of works we can mention the tag recommendation system presented in [81] that proposes a unified framework to model the three types of entities that exist in a social tagging system: users, items and tags. These data are represented by a 3-order tensor, on which latent semantic analysis and dimensionality reduction is performed using the Higher Order Singular Value Decomposition (HOSVD) technique [82]. The HOSVD decomposition is used also by the authors of [71] in order to decompose a 3-order tensor in which the first dimension is images, the second is visual words and the third is the text words. By applying the HOSVD decomposition on this 3-order tensor the authors aim to detect the underlying and latent structure of the images by mapping the original data into a lower dimensional space. Finally, a 3-order tensor is used also by the authors of [83] that propose an approach to capture the latent semantics of Web data. In order to do that the authors apply the PARAFAC decomposition [84] which can be considered as a multi-dimensional correspondent to the singular value decomposition of a matrix. In this case the extracted latent topics are used for the task of relevance ranking and producing fine-grained descriptions of Web data.

Compared to our approach presented in Chapter 5 what is missing from the aforementioned works is that they fail to benefit from the fact that being a different representation of the same abstract meaning, there is a certain amount of dependency between the tag and visual information items that appear together very frequently. We rely on legacy data to capture these dependencies and propose a mechanism for incorporating such knowledge into the multimedia analysis pipeline.

Chapter 3

Combining implicit and explicit knowledge for media interpretation

In this chapter we present a general framework for media interpretation that jointly considers implicit and explicit pieces of knowledge that are treated as evidence. As evidence we define the information that (when coupled with the principles of inference) can be used to support or disprove a hypothesis. Our framework implements a generative method for modeling the layer of evidence so as to effectively combine the low-level stimuli carried by a media item, the application context (approximated using the frequency information implicit in the data) and the domain knowledge (provided explicitly by domain experts). Using this framework we manage to drive a probabilistic inference process that verifies or rejects a hypothesis made about the semantic content of the media item. More specifically, we statistically analyze the low-level stimuli to obtain conceptual information about the content, we represent domain knowledge using ontologies and we extract the application context by estimating the conditional probabilistic dependencies between the existing concepts. Then, we combine everything in a bayesian network (BN) that is able to perform inference based on soft evidence. In this way, we provide the means to handle aspects like causality (between evidence and hypotheses), uncertainty (of the extracted evidence) and prior knowledge and hence, imitate some of human's basic perceptual operations. In the following we elaborate on how BNs can be used to model this layer of evidence, as well as on how ontologies can be

employed to migrate domain knowledge into the resulting inference framework. Finally, we present two case studies that experimentally verify the effectiveness of our framework in performing concept detection using: a) the image local and global information, b) the information found across media in compound documents.

3.1 Modeling the Bayesian Network

In this section we describe one of the main contributions of this thesis, which is a generic approach for modeling BNs. The goal of this modeling approach is to define a conceptual space suitable for incorporating the evidence derived from content analysis. We start by providing some background information on BNs and proceed with the proposed modeling approach.

3.1.1 Background on Bayesian Networks

In probability theory, the Bayes' theorem relates the conditional and prior (or marginal) probability distributions of random variables and therefore can be used to update or revise beliefs in light of new evidence. More specifically, a BN relies on the Bayes' theorem to answer probabilistic queries of the form, find out updated knowledge on the state of a subset of variables when other variables (the evidence variables) are observed. The process of computing the posterior distribution of variables given the observed evidence is called probabilistic inference [85]. This process can be used to collect and evaluate pieces of evidence that are meant to be consistent or inconsistent with a given hypothesis [86]. The likelihood of this hypothesis changes as evidence accumulates and provided that enough pieces of evidence are available, the hypothesis belief will become very high or very low. Hypotheses with a very high belief can be accepted as true and those with very low belief can be rejected. The engineering process of a BN can be separated in two phases that need to be carried out before applying probabilistic inference.

3.1.1.1 Network Structure

Bayesian networks are directed acyclic graphs connecting random variables. A graph is composed of nodes and edges and can be represented as $G = (N, E)$, where N is a set whose elements are called nodes and E is a set of unordered pairs between distinct

nodes called edges. A directed graph $G = (N, A)$ is a specialization of a graph where the edges are substituted by arcs i.e. in this case A is a set of ordered pairs of nodes called arcs or directed edges. Finally the cycle free property indicates that no cycles exist between sets of nodes, which means that for a node n_i, there is no non-empty directed path that starts and ends on n_i.

3.1.1.2 Network Parameters

While network structure encodes the qualitative characteristics of causality, i.e. which nodes affect whom, network parameters are used to quantify it, i.e. how much is a node influenced by its neighbors. Conditional probability tables (CPTs) are used to capture the amount of this influence and make it available for inferencing. A conditional probability table is calculated for each node in the network and incorporates two types of probabilities. The prior probability that indicates the probability of the hypothesis attached to this node being true, without considering any evidence, and the conditional probability, which is the probability of the aforementioned hypothesis being true conditioned on the probabilities of its parent nodes. The conditional probability tables can either be defined by experts or learned from observed data.

3.1.1.3 Probabilistic Inference

After completing the aforementioned stages probabilistic inference can take place in the constructed BN. The basic principle adopted by probabilistic inference algorithms can be considered as an interpretation of the Markov property. This property states that the conditional probability distribution of future states, given the present and all past states, depends only upon the present state and not on any of the past states. In a similar fashion, a bayesian network node n_i is considered to be influenced only by its direct parents and not by all network nodes. Hence, once the values of its parents became stable, n_i is shield from the influence of all other predecessors [87]. This entails that in order to estimate the influence of the entire network on node n_i, it suffices to calculate a $CPT(n_i)$ that contains only the conditional probabilities on its direct parents. Although this principle makes feasible the development of a computational framework for probabilistic inference, it leaves open the mechanism by which information flows over the network before the parent nodes reach a stable state.

To address this issue, Pearl [87] introduced a message passing mechanism where messages are exchanged between father and child nodes carrying the information required to update their beliefs. The message propagation mechanism can be intuitively described as a two directional traversal of the graph (i.e. top-down and bottom up) starting from the node that is activated to update its belief. Although intuitively consistent the message passing algorithm proposed by Pearl suffers from scalability issues due to the excessive number of messages that need to be exchanged over the network. In order to overcome this deficiency, Lauritzen and Spiegelhalter [88] proposed to exploit a range of local representations for the network joint probability distribution in order to acquire more computationally efficient methods. The authors made use of some topological changes of the original network in order to facilitate rapid absorption and propagation of evidence. The rough idea of their work is to employ two schemes for modifying topology, termed "marrying" and "filling-in", in order to produce a triangulated version of the network. The maximal cliques are identified in the triangulated graph and are subsequently organized in a junction tree [89]. Conditional probabilities between adjacent cliques are then calculated based on the conditional probabilities of the original network. Eventually, the propagation of probabilities is performed over the formulated cliques using the ordinary message passing approach [87]. To the best of our knowledge junction tree is the most efficient and scalable belief propagation algorithm and was selected for conducting all experiments involving probabilistic inference on BNs.

3.1.2 Proposed modeling approach

After presenting some background information on BNs, in this section, we present our modeling approach for incorporating different types of information into the layer of evidence. More specifically, the proposed modeling approach aims to handle the following types of knowledge: a) information extracted from content analysis that encodes the support received from the analyzed low-level features in favor of some concept, b) conceptual information shared amongst most individuals that determines the logical relations between concepts, such as sub-class, union, intersection, disjoint, etc (i.e. domain knowledge), and c) information that qualitatively evaluates the co-existence of concepts, encoding for example how likely it is for one concept to be present when an-

other concept is verified (i.e. application context). Our approach relies on probabilities and probabilistic inference to define this unified conceptual space.

More specifically, the explicitly provided domain knowledge is used to determine the structure of the BN and in this way enforce the logic rules of the domain during inference. The application context is approximated by the co-occurrence frequency between domain concepts, information that can be extracted using a sample of the population that is being modeled. The application context is encoded into the Conditional Probability Tables (CPTs) of the BN nodes, which influence the inference process when belief propagation takes place. However, the most critical point is how to incorporate the information received from content analysis. In order to do this, we treat the outcome of content analysis as soft evidence that is used to instantiate the nodes of a BN, operating on a conceptual true-false space. The reason for selecting these states (i.e. true, false) to be the only possible states of all network nodes, was to establish a "lingua franca" between the different types of information and facilitate the incorporation of domain knowledge in decision making. In this way the constructed BN does not operate on the low-level features of the content, which would constitute a typical application of the BN theory. Instead, it operates on the space determined by the probability estimates (that we call conceptual true-false space), obtained through the application of pattern recognition on low-level features. In the following we describe how the proposed modeling approach can be used to incorporate information from two different modalities, however the same approach can be seamlessly applied to handle an arbitrary number of modalities, irrespective of their nature.

Let us consider a set of media items D. For the sake of notation simplicity and without loss of generality we will consider the two different modalities to be the visual and textual part of a media item. Thus, for every media item D_i we have:

$$D_i = [T_i, V_i] \tag{3.1}$$

Let also t_i and v_i to be the features extracted from T_i and V_i, respectively. Then, we consider the modality analyzers to be the functions $f_{c_j}(\cdot)$ and $g_{c_j}(\cdot)$ that output the probability of a given concept c_j being valid for a media item, either based on its

textual or visual low-level features, respectively:

$$\begin{array}{ll} f_{c_j}(T_i) = P(c_j = true|t_i), & \text{for the textual part of } D_i \\ g_{c_j}(V_i) = P(c_j = true|v_i), & \text{for the visual part of } D_i \end{array} \tag{3.2}$$

Thus, if we have a modality analyzer that is trained to detect all domain concepts $\forall c_j \in C$, it produces $|C|$ probabilities when applied on the media item D_i. In order to construct a BN that operates on a conceptual true-false space, for every concept c_j we create a node with two states $r_z = \{true, false\}$. Then, we link these nodes based on their logical relations (as explained in Section 3.2.2.1) and learn the CPTs by applying the Expectation Maximization (EM) algorithm on sample data (as detailed in Section 3.2.2.2). We consider the output of each modality analyzer to formulate a new feature space y, determined from the probability estimates. We refer to this new feature space as conceptual true-false space. By applying the Bayes rule in feature space y we have for each concept c_j:

$$P_{c_j}(r_z|y) = \frac{P_{c_j}(y|r_z)P_{c_j}(r_z)}{P_{c_j}(y)}, \quad \forall c_j \in C \tag{3.3}$$

$P_{c_j}(r_z)$ represents our prior knowledge about c_j and in the conceptual true-false space we accept that $P_{c_j}(r_z = true)$ is equal to the frequency of appearance of c_j in the domain (i.e. how often it appears in the training set, divided by the total members of the dataset). Respectively, we accept that $P_{c_j}(r_z = false) = 1 - P_{c_j}(r_z = true)$. $P_{c_j}(y)$ is a scale factor that guarantees that the posterior probabilities sum to one and equals:

$$P_{c_j}(y) = \sum_{r_z \in \{true, false\}} P_{c_j}(y|r_z)P_{c_j}(r_z) \tag{3.4}$$

$P_{c_j}(y|r_z)$ is the likelihood (or class conditional probability) of r_z with respect to y and $P_{c_j}(r_z|y)$ is the posterior probability of r_z after considering the analysis outcome and taking into consideration prior knowledge. In order to facilitate the analysis process we need to calculate the posterior probabilities for each independent piece of conceptual information (i.e. $\forall c_j \in C$), thus we need to estimate $P_{c_j}(r_z = true|y)$. It is clear from eqs. (3.3) and (3.4) that in order to estimate this value, what we are missing is $P_{c_j}(y|r_z = true)$ and $P_{c_j}(y|r_z = false)$. However, recalling that $f_{c_j}(\cdot)$ and $g_{c_j}(\cdot)$ provides us with a probability expressing how much support c_j receives from the textual

or visual low-level features of the media item, we incorporate the content analysis outcome into the decision process by instantiating the nodes of the BN as follows:

$$P_{c_j}(y|r_z = true) = \begin{cases} f_{c_j}(T_i), & \text{for textual evidence} \\ g_{c_j}(V_i), & \text{for visual evidence} \end{cases} \tag{3.5}$$

$$P_{c_j}(y|r_z = false) = \begin{cases} 1 - f_{c_j}(T_i), & \text{for textual evidence} \\ 1 - g_{c_j}(V_i), & \text{for visual evidence} \end{cases} \tag{3.6}$$

Thus, during the analysis process we inject, as explained above, the output of modality analyzers into the BN and perform probabilistic inference by propagating evidence beliefs. Eventually, the resulting posterior probability for the "true" state of the node corresponding to the concept that we want to detect, is considered to be the updated confidence degree for this concept.

What is evident from the above is that the one and only requirement of the proposed modeling approach, is for the modality analyzers to generate a probabilistic output when applied on the low-level features of media items. This fact makes our approach generic enough to facilitate various different analysis tasks, as verified by the two case studies presented later in this section.

3.2 Mapping ontologies to Bayesian Networks

As already mentioned in the previous section, domain knowledge is used to enforce the logic rules of the domain during inference. However, in order to incorporate these rules into our inference framework domain knowledge will have to be elucidated and represented in machine understandable format. In other words, the logical relations between the concepts should be represented in a format manageable by the analysis module. To fulfil this objective, ontologies have emerged as a very powerful tool able to express knowledge in different levels of granularity and incorporate from abstract notions such as general rules governing time and space, to more tangible concepts such as domain specific material entities [90]. In the following we provide some background information on ontologies and present the adopted approach for constructing a BN out of the knowledge expressed in an ontology.

3.2.1 Background on Ontologies

Let C be the set of unary predicate symbols that are used to denote domain concepts (e.g. *seaside*, *sea*, *boat*, *wave*, *sand*, etc) R to be the set of binary predicates that are used to denote relations between concepts (e.g. *sea* is sub-class of *seaside*) and O the algebra defining the allowable operators for these sets (e.g. *sea* is disjoint with *desert*). The part of experience that relates to the general knowledge of a specific domain can be represented using C, R, O. This conceptualization does not claim to formally describe all different variations of knowledge modeling, it mostly adheres to the syntax of "Description Logics" [91] and serves in our work to highlight the knowledge capacities of our framework. Balancing between expressivity and decidability is the critical tradeoff that most knowledge representation languages have to cope with. OWL–DL [92] is a syntactic variant of the SHOIN(D) DL [93] that was constructed to provide the maximum possible expressivity while being decidable and was deemed the most appropriate for serving the purposes of our framework. Thus, the general knowledge about a specific domain D can be expressed by a structure K_D that associates the domain concepts and relations using the allowable operands:

$$K_D = S(C^D, R^D, O), \quad O \in OWL - DL \tag{3.7}$$

DL stands for "Description Logics" [91] and constitutes a specific set of constructs such as intersection, union, complement, equivalent, disjoint, etc. For instance such constructors can be used to express that two concepts are disjoint with each other and can not be depicted in the same image simultaneously. Our goal is to use these constructors for explicitly imposing semantic constraints in the process of image interpretation, which can not be captured by typical machine learning techniques. Loosely speaking, we use the knowledge structure to obtain three different types of information: a) what are the semantic restrictions that apply in the examined domain, b) which of the domain concepts should be considered as evidence and therefore used to trigger the probabilistic inference process, and c) which evidence is expected to support one hypothesis or another. In this sense, the knowledge structure sets the tracks to which evidence belief is allowed to propagate by determining the structure of the BN.

Apart from ontologies, other representation structures capable of encoding explicit knowledge also exist (e.g. conceptual graphs). However, the use of ontologies was ad-

vocated by their wide acceptance and appeal in the area of knowledge engineering [90]. It is true that ontologies have been widely established as the main tool for encoding explicit knowledge in machine understandable format. This is witnessed by the fact that in many disciplines considerable effort has been already allocated on engineering ontologies that encode domain concepts and relations. Therefore, enabling our framework to automatically handle ontologies makes it directly applicable in these domains.

3.2.2 Ontology to BN mapping

In this section we describe the methodology adopted in our work for transforming an ontology into a BN. As mentioned previously, a BN is a directed acyclic graph $G = (N, A)$ whose nodes $n \in N$ represent variables and whose arcs $a \in A$ encode the conditional dependencies between them. Given that the network structure is capable of encoding the qualitative characteristics of causality (i.e. which nodes affect which), and the CPTs can be used to quantify the causality relations between concepts (i.e. how much is a node influenced by its connected nodes), the resulting BN will be able to facilitate three different operations: a) Provide the means to store and utilize domain knowledge K_D; this is achieved by mapping K_D to the network structure. b) Organize and make accessible information coming from the application context; this is achieved through the CPTs attached to the network nodes. c) Allow the propagation of evidence belief in a mathematically coherent manner; this is performed using the message passing belief propagation algorithms. In [42] Ding et al. introduce a probabilistic extension to OWL ontology based on BNs and define a set of structural translation rules to convert this ontology into a directed acyclic graph. Moreover, the authors describe how the parameters of the network can be estimated based on the information received by an expert. Here, we propose an adaptation of this method that learns the network parameters from sample data.

3.2.2.1 Mapping the network structure

Intuitively, deciding on the structure of a BN based on an ontology can be seen as mapping ontological elements (i.e. concepts and relations) to graph elements (i.e. nodes

and arcs):

$$\begin{array}{c} S(C,R,O) \rightarrow G(N,A), \quad O \in DL \\ where \quad C \rightarrow N, \quad R \rightarrow A, \quad O \rightarrow (N,A) \end{array} \tag{3.8}$$

$O \rightarrow (N,A)$ indicates that in order to migrate a DL constructor into the network structure both nodes and arcs will have to be employed.

The structural transformation process adopted in our framework is similar to the one proposed in [42] and takes place in two stages. In the first stage, the BN incorporates the hierarchical information of the ontology. In order to do so, all ontology concepts are transformed into network nodes with two states (i.e. true and false). These nodes are called concept nodes n_{cn}. Then, an arc is drawn between two concept nodes in the network, if and only if they are connected with a superclass-subclass relation in K_D and with the direction from the superclass to the subclass. The adoption of this principle was motivated by the fact that when an instance belongs to a certain class it is automatically subsumed that it can also belong to one of its subclasses, thus imposing a kind of causality. At the second stage, the BN incorporates the semantic constraints between concepts that are expressed in the ontology using DL constructors. This is done by creating a control node n_{cl} for each DL constructor. This node is connected to the concept nodes that correspond to the concepts associated with the DL constructor. The way in which the connection is made depends on the type of the DL constructor and results in a different sub-network structure. The DL constructors that can be handled by the adopted methodology are owl:intersectionOf, owl:unionOf, owl:complementOf, owl:equivalentClass and owl:disjointWith. The structural translation rules can be found in [42], however for the sake of completeness we also include a brief description below:

(1) Every concept $c \in C$ is mapped into a two-state (i.e. true-false) variable node in the BN.

(2) There is always an arc from a parent superclass node to a subclass node.

(3) A concept c defined by the set intersection operation (owl:intersectionOf) of concepts $c_i, (i = 1, \ldots, m)$ is mapped into a sub-network (Fig. 3.1) of the resulting BN with one arc from each c_i to c, and one arc from c and c_i to a control node called "Intersection":

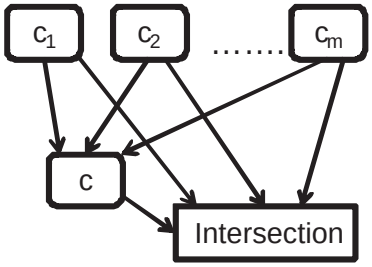


Figure 3.1: Sub-network mapping the owl:intersectionOf constructor.

(4) A concept c defined by the set union operation (owl:unionOf) of concepts $c_i, (i = 1\ldots, m)$ is mapped into a sub-network (Fig. 3.2) of the resulting BN with one arc from c to each c_i, and one arc from c and each c_i to a control node called "Union".

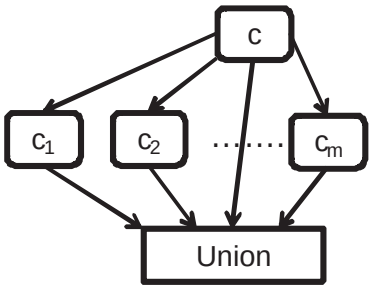
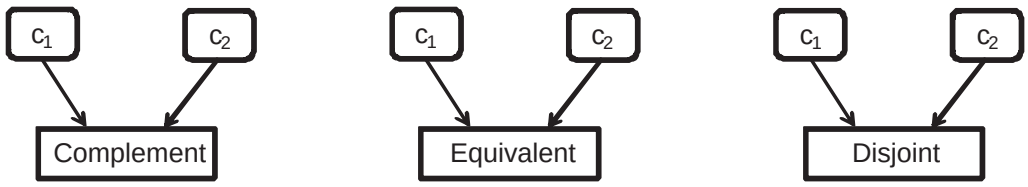


Figure 3.2: Sub-network mapping the owl:unionOf constructor.

(5) If two concepts c_1 and c_2 are related with the owl:complementOf, owl:equivalentClass, or the owl:disjointWith constructor, then a control node (named "Complement", "Equivalent" and "Disjoint", respectively) is added to the resulting BN, with two arcs pointing from c_1 and c_2 to this node.

3.2.2.2 Learning the network parameters

Once the network structure is fixed, each concept node n_{cn} needs to be assigned a prior probability if it is a root node, or a conditional probability table if it is a child of one or more nodes. In [42] these probabilities are set manually (i.e. by domain experts) and formulate the original probability distribution of the network. In order to learn the probability distribution model of a network that is enhanced with the semantic



(a) (b) (c)

Figure 3.3: Sub-network mapping the owl:complementOf, owl:equivalentClass and owl:disjointWith constructors.

constraints of the domain, the authors developed an algorithm called D-IPFP. This algorithm is based on the "iterative proportional fitting procedure" (IPFP), which is a procedure that modifies a given distribution to meet a set of constraints (i.e. the semantic constraints of the domain in our case), while minimizing the *KL-divergence* (Kullback-Leibler divergence) [94] to a target distribution (i.e. the original probability distribution of the network in our case). One important drawback of the aforementioned approach is that the target probability distribution needs to be explicitly provided by an expert. Such an approach, apart from the fact that it requires human intervention when switching to a different domain, is likely to introduce bias in the initial conditions of the BN.

To overcome these limitations we propose a variation of the aforementioned methodology where the original probability distribution is learned from sample data instead of being explicitly provided by humans. In our case, the sample data are concept labels that have been used to annotate the images. Given a sufficiently large amount of annotated images, the original probability distribution of the network can be approximated using the frequency information implicit in the data. This kind of solution has attracted considerable attention in the field of computer vision and particularly in cases where graph-based probabilistic networks are used. For instance, in contrast to [42], both [95] and [96] use a sample portion of the data that is being modeled in order to learn the necessary conditional probabilities.

In our work the conditional probabilities are learned by employing the Expectation Maximization (EM) [97] algorithm, using as training data the media items annotated with concept labels. Initially, we apply the EM algorithm to a BN that incorporates only the hierarchical information of the ontology. Then, we add the control nodes to model the semantic constraints and we once again apply the EM algorithm to the

modified BN. In this case, since no annotated samples are available for the control nodes, these nodes are treated as latent variables with two states (i.e. true and false). The last step is to manually set the CPTs of all control nodes n_{cl} as described in [42] (also depicted in Figs. 3.4 and 3.5) and fix their state to "true" (i.e. set the belief of the true state equal to 100%). This is done in order to enforce the semantic constraints into the probabilistic inference process.

C1	C2	True	False
True	True	0.00	100.00
True	False	100.00	0.00
False	True	100.00	0.00
False	False	0.00	100.00

(a)

C1	C2	True	False
True	True	100.00	0.00
True	False	0.00	100.00
False	True	0.00	100.00
False	False	100.00	0.00

(b)

C1	C2	True	False
True	True	0.00	100.00
True	False	100.00	0.00
False	True	100.00	0.00
False	False	100.00	0.00

(c)

Figure 3.4: CPTs for the control nodes corresponding to owl:complementOf, owl:equivalentClass and owl:disjointWith constructors: a) When its state is set to "true" c_1 and c_2 are complement of each other, b) When its state is set to "true" c_1 and c_2 are equivalent with each other, and c) When its state is set to "true" c_1 and c_2 are disjoint with each other.

C1	C2	C	True	False
True	True	True	100.00	0.00
True	True	False	0.00	100.00
True	False	True	0.00	100.00
True	False	False	100.00	0.00
False	True	True	0.00	100.00
False	True	False	100.00	0.00
False	False	True	0.00	100.00
False	False	False	100.00	0.00

(a)

C1	C2	C	True	False
True	True	True	100.00	0.00
True	True	False	0.00	100.00
True	False	True	100.00	0.00
True	False	False	0.00	100.00
False	True	True	100.00	0.00
False	True	False	0.00	100.00
False	False	True	0.00	100.00
False	False	False	100.00	0.00

(b)

Figure 3.5: CPTs for the control nodes corresponding to owl:intersectionOf and owl:unionOf constructors: a) When its state is set to "true" c is the intersection of c_1 and c_2, b) When its state is set to "true" c is the union of c_1 and c_2.

3.3 Case study on concept detection using image local and global information

In this section we describe how the proposed framework can be effectively used to boost the performance of concept detection classifiers by combining image local and global information, with domain knowledge and application context [98], [99], [100]. In the method developed to facilitate this case study (depicted in Fig. 3.6), the application context and the domain knowledge are considered to be the a priori/fixed information. On the contrary, the visual stimulus depends on the examined image and is considered to be the observed/dynamic information. Based on the proposed generative approach for modeling the layer of evidence (see Section 3.1) we manage to effectively combine and exploit both a priori and observed information and evaluate our method using content from three different domains.

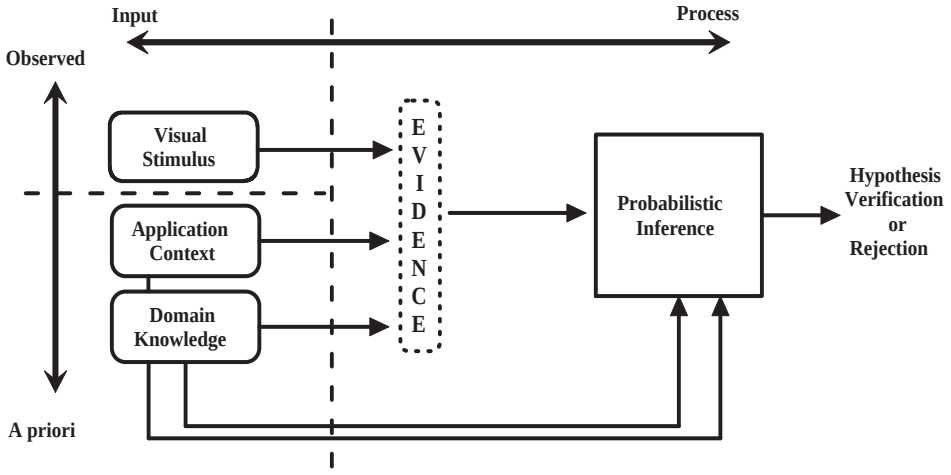


Figure 3.6: Functional relations between the different components of the developed method.

The main outcome of this case study is to show how global and local evidence, as obtained from the application of concept classifiers on global and local image data respectively, can be probabilistically combined within a BN that incorporates domain knowledge and application context. We demonstrate that combining information in this way leads to statistically significant improvements for the tasks of image categorization, localized region labeling and weak annotation of video shot key-frames. Moreover, we introduce a mechanism that exploits the mutual information between concepts, in

order to significantly reduce the computational cost of visual inference and still achieve results comparable to the exhaustive case. In the following we present the individual components of our method, we detail the functional settings of the employed tasks and describe our experimental findings.

3.3.1 Components Description

3.3.1.1 Extracting conceptual information from visual stimulus

Machine learning methods are widely established as a powerful tool for processing visual stimulus and extracting conceptual information. Here, we consider the supervised learning paradigm where a classifier is trained to identify an object category, provided that a sufficiently large number of examples are available. We denote by C the set of domain concepts and by I_q a visual representation that refers to the piece of content to be analyzed. Depending on the circumstances, I_q can be an image region, the whole image, a video shot, etc. A concept detector can then be implemented using a classifier F_c that is trained to recognize instances of the concept $c \in C$. We denote by $F_c(I_q)$ the output of F_c applied to image I_q. When F_c is a probabilistic classifier we have $F_c(I_q) = Pr(c|I_q)$. These probabilities $Pr(c|I_q)$ are essentially the soft evidence that are provided to the BN for triggering probabilistic inference.

3.3.1.2 Domain Knowledge

As already mentioned the general knowledge about a specific domain can be represented using C, R and O. Following our general guidelines OWL–DL was employed to construct the structure $K_D = S(C, R, O)$, describing how the domain concepts are related to each other using the allowable operators, $O \in DL$. More specifically, for each domain of discourse that was examined in this case study, a separate ontology was manually constructed by the domain experts, as described in Section 3.3.3.

3.3.1.3 Application context

The role of K_D is to capture information about the domain of discourse in general, but not to deliver information concerning the context of the analysis process at hand. No information is provided to the framework in terms of where within the analyzed content the anticipated evidence are likely to reside. For instance, in the image categorization

task, this type of information could suggest the analysis mechanism to look for evidence in specific image regions. Moreover, information on how to quantitatively evaluate the existence of the extracted evidence (i.e. how much each hypothesis is affected by the existence of one evidence or another) is also missing from K_D. We consider this type of information to be part of the application context. Let *app* denote the type of application specific information used to guide the analysis mechanism in searching for evidence (Section 3.3.2 details this information for each of the analysis tasks), and $W = [W_{i,j}]$ the matrix whose elements $W_{i,j}$ quantify the effect of concept c_i on c_j. Then we consider the application context $X = S(app, W)$ to be the information consisting of both *app* and W. W_{ij} is approximated by the frequency of co-occurrence between concepts c_i and c_j in the training set, as discussed in Section 3.2.2.2. This information, that is implicitly extracted from the training data, is encoded into the Conditional Probability Tables (CPTs) of the BN nodes and influences the probabilistic inference process when belief propagation takes place.

3.3.1.4 Evidence-driven probabilistic inference

Having defined K_D and X, the following steps are applied to achieve semantic image interpretation: a) we use K_D to decide which of the domain concepts should constitute the evidence set c^E, b) we use *app* to decide where to physically look for this evidence, c) we apply the probabilistic classifiers F_c on I_q to obtain the degrees of confidence for the concepts in c^E, d) we use *app* and K_D to decide which of the domain concepts should constitute the hypotheses set c^H, e) we provide the degrees of confidence for the concepts in c^E to the BN and trigger probabilistic inference by using these degrees as soft evidence, f) we propagate evidence beliefs using the network's inference tracks and the corresponding causality quantification functions W_{ij}, g) we calculate the posterior probabilities for all concepts in c^H and decide which of the hypotheses should be verified or rejected.

Let us assume that the degree of confidence that the analyzed image I_q depicts a concept c_i, is estimated by a classifier as described in Section 3.3.1.1. We denote with $h(I_q, c_i) = Pr(c_i|I_q)$ the function estimating the degree of confidence that concept c_i appears in image I_q. We also denote with $H(I_q) = \{h(I_q, c_i) : c_i \in c^H\}$ the set of confidence degrees that the concepts belonging to the hypotheses set are depicted in image I_q and with $E(I_q) = \{h(I_q, c_i) : c_i \in c^E\}$ the set of confidence degrees that

the concepts belonging to the evidence set are depicted in the image I_q (Section 3.3.3 clarifies the hypothesis and the evidence sets for each of the analysis tasks). Then, we provide $H(I_q)$ and $E(I_q)$ to the BN and using probabilistic inference we calculate the posterior probabilities of the network nodes using information coming from knowledge R, O and context W_{ij}. If we denote with $\acute{h}(I_q, c_i) = Pr(c_i \mid H(I_q), E(I_q), R, O, W_{ij})$ the function that calculates the posterior probabilities of the network nodes, the set of posterior probabilities of the concepts belonging to the hypotheses set can be represented as $\acute{H}(I_q) = \{\acute{h}(I_q, c_i) : c_i \in c^H\}$ and the formula used to achieve semantic image interpretation can be expressed as follows:

$$c = \underset{c_i \in c^H}{\arg \otimes} (\acute{h}(I_q, c_i)) \tag{3.9}$$

$\otimes$ is an operator (e.g. max) that depends on the specifications of the analysis task (Section 3.3.3 describes the functionality of this operator for each of the analysis tasks). Table 3.1 summarizes the notation of all introduced terms.

3.3.1.5 Computational efficiency

Another aspect that is usually important in semantic image interpretation is related to the computational efficiency of the employed methods. Our evidence-driven probabilistic inference framework is essentially a method that connects a symbol (visual stimulus in our case) to real-world objects/concepts to which the symbol is associated. However, in the real-world the number of plausible hypotheses could suffer from a combinatorial explosion, rendering testing for them intractable. This problem is usually addressed using exclusion principles determined by the faculties of attention and perception [101]. In our case, the exclusion principles are derived from the domain ontology which determines the set of plausible hypotheses for each task.

Still, the computational cost for gathering the necessary evidence is often so expensive that it can be prohibitive in highly complex domains. For this purpose we introduce a Focus of Attention (FoA) mechanism that improves the computational efficiency of the proposed framework. In particular, we apply an iterative process that initially examines the hypothesis and evidence that are more likely, in statistical terms, to be valid. If the hypothesis is verified the process is terminated, otherwise the next most

3. COMBINING IMPLICIT AND EXPLICIT KNOWLEDGE FOR MEDIA INTERPRETATION

Table 3.1: Legend of Introduced Terms

Term	**Symbol**	**Role**
Trained Classifier	F_c	- Estimate the degree of confidence that the visual representation I_q depicts concept c.
Domain Knowledge	$K_D = S(C, R, O)$	- Determine which concepts belong to the evidence set and which to the hypothesis set. - Specify qualitatively relations between evidence and hypotheses (i.e. which evidence support a certain hypothesis).
Application Context	$X = S(app, W)$	- Determine where to "physically" look for evidence, expressed with *app* (i.e. application specific information). - Specify quantitative relations (causality) between evidence and hypotheses, expressed with W.
Hypotheses	$h(I_q, c_i) = Pr(c_i \mid I_q)$ and $H(I_q) = \{h(I_q, c_i) \quad : \quad c_i \in c^H\}$	- Constitute the initial degrees of confidence for the concepts belonging to the hypotheses set c^H, as determined by $N_C \in K_D$ and $app \in X$, obtained by applying the classifiers F_c.
Evidence	$E(I_q) = \{h(I_q, c_i) \quad : \quad c_i \in c^E\}$	- Constitute the degrees of confidence for the concepts belonging to the evidence set c^E, as determined by $N_C \in K_D$ and $app \in X$, obtained by applying the classifiers F_c.
Evidence driven probabilistic inference	$\acute{h}(I_q, c_i) = Pr(c_i \mid H(I_q), E(I_q), R, O, W_{ij})$ and $\acute{H}(I_q) = \{\acute{h}(I_q, c_i) \quad : \quad c_i \in c^H\}$	- Perform inference using $\acute{h}(I_q, c_i)$ and estimate the posterior probabilities $\acute{H}(I_q))$, using $E(I_q)$ as trigger, $R, O \in K_D$ as belief propagation tracks and $W \in X$ as causality quantification functions.
Semantic image interpretation	$c = \arg \otimes_{c_i \in c^H} (\acute{h}(I_q, c_i))$	Achieve semantic image interpretation based on the operator $\otimes$ that depends on the analysis task.

likely hypothesis is examined. More specifically, instead of examining the complete hypotheses set $H(I_q) = \{h(I_q, c_i) : c_i \in c^H\}$, we initially examine the hypothesis with the maximum confidence degree corresponding to c_k, such that $k = arg\max_i(h(I_q, c_i))$ and $c_i \in c^H$. This is performed by inserting this value to the corresponding network node and comparing the node's posterior probability against a pre-defined belief threshold. If the posterior probability exceeds the belief threshold the process is terminated. Otherwise, a ranked list of the evidence concepts (i.e. $\forall c_i \in c^E$), that would have caused maximum impact on the hypothesis if were observed, is formed. This is performed by calculating the mutual information between the node corresponding to the concept c_k and all other nodes corresponding to the concepts of c^E. The mutual information between two discrete random variables is the expected reduction in entropy of one variable (measured in bits) due to a finding in the other variable. The mutual information between c_k and c_i, $\forall c_i \in c^E$ is calculated according the following equation:

$$I(c_k; c_i) = \sum_{\{true,false\}} \sum_{\{true,false\}} Pr(c_k, c_i) \log_2 \frac{Pr(c_k, c_i)}{Pr(c_k)Pr(c_i)}, \tag{3.10}$$

where $Pr(c_k, c_i)$ is the joint and $Pr(c_k)$, $Pr(c_i)$ the marginal probability distributions of c_k and c_i. The efficient calculation of $P_r(c_k, c_i)$ is performed using the junction tree [89]. Subsequently, the nodes are ranked in descending order based on their mutual information with c_k and the confidence degrees of the concepts corresponding to the most highly ranked nodes are extracted. The resulting degrees are inserted into the BN causing belief propagation to take place. If the posterior probability of the examined hypothesis still fails to exceed the pre-defined belief threshold, the hypothesis is rejected and the process is repeated for the hypothesis with the next highest confidence value in $H(I_q)$. If none of the hypotheses overcomes the belief threshold the image is categorized based on the maximum confidence degree of $H(I_q)$. One disadvantage of this approach lies in the difficulty of estimating an optimal belief threshold adapted to the statistical characteristics of each hypothesis. However, the fact that only a small portion of the available classifiers is required to reach a decision makes this approach attractive for complex domains.

3.3.2 Functional Settings

3.3.2.1 Image analysis tasks

This section describes how the proposed framework can be adapted to three different image analysis tasks. For each of these tasks we clarify the task specific contextual information $app \in X$ (i.e. where to physically look for evidence) as well as the way that the hypotheses $H(I_q)$ and evidence $E(I_q)$ sets are determined.

Image categorization is the task of selecting the category concept c_i that best describes an image I_q as a whole. In this case, a hypothesis is formulated for each of the category concepts, that is $H(I_q) = \{Pr(c_i|I_q) : i = 1, \ldots, n\}$ where n is the number of category concepts in K_D. Global classifiers (i.e. models trained using global image information) are applied to estimate the initial probability for each hypothesis. For this task, the application context *app* determines that evidence should be taken from the image local information, and specifically from the regions extracted using a segmentation algorithm. For instance, knowing that a specific region depicts *road* is a type of contextual information that the algorithm can exploit when trying to decide whether the image depicts a *Seaside* or a *Roadside* scene. Local classifiers (i.e. models trained using regional image information) are applied to the pre-segmented image regions $I_q^{s_j}$, in order to generate a set of confidence values that constitute the evidence $E(I_q) = \{Pr(\acute{c}_i|I_q^{s_j}) : i = 1, \ldots, k \quad \& \quad j = 1, \ldots, m\}$, where k is the number of regional concepts in K_D and m is the number of identified segments. In this case, the category concepts c_i constitute the hypothesis set c^H and the regional concepts $\acute{c}_i$ comprise the evidence set c^E.

Localized region labeling is the task of assigning labels to pre-segmented image regions, with one of the available regional concepts $\acute{c}_i$. In this case, a hypothesis is formulated for each of the available regional concepts and for each of the image segments. That is $H(I_q) = \{Pr(\acute{c}_i|I_q^{s_j}) : i = 1, \ldots, k \quad \& \quad j = 1, \ldots, m\}$, where k is the number of regional concepts and m is the number of identified segments. Local classifiers are used to estimate the initial probability for each of the formulated hypotheses. In this task, the contextual information *app* is considered to be the image as a whole. For example, knowing that an image depicts a *Roadside* scene can be considered the application context and facilitate the algorithm to decide whether a specific region depicts *sea* or *road*. The degrees of confidence for each of the category

concepts c_i, obtained by applying the global classifiers to I_q, constitute the evidence of this task. That is $E(I_q) = \{Pr(c_i|I_q) : i = 1, \ldots, n\}$, where n is the number of category concepts. In this case, the regional concepts $\acute{c}_i$ constitute the hypothesis set c^H and the category concepts c_i comprise the evidence set c^E.

In practice, our framework can be used to improve region labeling when there is a conflict between the decisions suggested by the global and local classifiers. A conflict occurs when the concept suggested by the local classifiers does not belong to the set of child nodes of the concept suggested by the global classifiers. Since there is no reason to trust one suggestion over another we make two different hypotheses. The first one assumes that the suggestion of the global classifiers is correct. The regional concept corresponding to the maximum confidence degree, among the child nodes of the category concept, is selected and the overall impact on the posterior probability of the regional concept is measured. The second approach considers that the suggestion of the local classifiers is correct. The category concept corresponding to the maximum confidence degree, among the parent nodes of the regional concept suggested by the local classifiers, is selected and the overall impact on the posterior probability of the regional concept is measured. Among the two cases, the regional concept with the maximum positive impact on its posterior probability is selected to label the examined region.

Weak annotation of video shot key-frames is the task of detecting all concepts depicted in an image, but without having to associate them with specific image regions. Thus, there is no distinction between category and regional concepts and more than one label can be assigned to the image. A hypotheses set is formulated, $H(I_q) = \{Pr(c_i|I_q) : \quad i = 1, \ldots, n\}$ where n is the number of all available concepts in the domain. All classifiers are employed to extract the initial probability for all formulated hypotheses. The application context *app* determines that evidence should be searched for in the global image information. For instance, if an image is being examined for the presence of the concept *sports*, it would be helpful for the algorithm to know that the concept *soccer-player* is also depicted in the image. Thus, the evidence are considered to be the confidence values of all other concepts except the one examined by the current hypothesis. That means that when we examine the hypothesis $H(c_k|I_q)$, the evidence are $E(I_q) = \{Pr(c_i|I_q) : \forall i \in [1, n] \backslash \{k\}\}$.

3.3.2.2 Low-level image processing

The low level processing of visual stimulus consists in visual features extraction, segmentation and learning the concept detection models. Four different visual descriptors proposed by the MPEG-7 standard [102] namely Scalable Color, Homogeneous Texture, Region Shape, Edge Histogram, were employed as described in [43] to formulate the visual feature space. Region Shape descriptor was used only at the region level. Segmentation was performed using an extension of the Recursive Shortest Spanning Tree algorithm [103] that produces a segmentation mask $S = \{s_i : \quad i = 1, \ldots, m\}$ for each image, with s_i representing the identified segment. Support Vector Machines (SVMs) were employed for learning the concept detection models (represented as F_c in Table 3.1), using a gaussian radial basis as the kernel function. Global and local classifiers were created off-line using manually annotated images as training samples and for all concepts included in K_D.

For the task of weakly annotating video shot key-frames we have utilized the detectors released by Columbia University [104]. In this case, individual SVMs are trained at global level independently over each feature space and a simple late fusion mechanism is subsequently applied to produce the average score. Three types of features were used, namely grid color moments, edge histogram direction and texture [104]. In all cases the SVM-based models were constructed using the libsvm library [105] and their soft output (i.e. confidence degree) was calculated based on the distance between the decision boundary and the classified feature vector in the kernel space. More specifically, for extracting the soft output of the classifiers we calculate the distance d of the test image I_q feature vector, from the separating hyperplane that has been learned by the SVM model [106]. This distance is positive in case of affirmative recognition and negative otherwise. Then, a sigmoid function [107] is employed to compute the respective degree of confidence for concept c, as follows:

$$Pr(c|I_q) = \frac{1}{1 + e^{-td}} \tag{3.11}$$

where t is a scale factor.

3.3.3 Experimental Study

To experimentally evaluate our method we have employed two datasets with different domain complexity and volume, namely the "Personal Collection" (PS) and "News" (NW). PS was assembled internally in our lab by merging various photo albums while NW was taken from the TRECVID 2005 competition. Using these datasets we demonstrate the improvement in performance achieved by exploiting context and knowledge, compared to baseline detectors that rely solely on low-level visual information. Additionally, we evaluate a FoA mechanism that is based on the mutual information between concepts. We show that we can significantly reduce the computational cost of visual inference and still achieve comparable performance with the exhaustive case. All experiments were conducted using the Netica[1] software provided by Norsys for handling BNs and the Protégé[2] ontology editor for constructing the ontologies.

A collection of 648 images I^{PS} comprised the dataset for the PS domain. Let us denote by C_G={*Countryside_buildings, Seaside, Rockyside, Forest, Tennis, Roadside*} the set of category concepts and by C_L={*Building, Roof, Tree, Stone, Grass, Ground, Dried-plant, Trunk, Vegetation, Rock, Sky, Person, Boat, Sand, Sea, Wave, Road, Road-line, Car, Court, Court-line, Board, Gradin, Racket*} the set of regional concepts. Each image was manually annotated at global and region level using concepts from C_G and C_L. For the NW domain 374 semantic concepts were defined by the Columbia University [104] to characterize its content. For this domain the TRECVID2005 development dataset [108] containing 137 annotated video clips was used. The annotations were provided at the level of subshots, extracted using temporal criteria (see [104] for details). By extracting a key-frame from each subshot a dataset consisting of 61600 still images I^{NW} annotated at global level was constructed.

In both cases, OWL–DL was utilized to represent domain knowledge using manually constructed ontologies. The ontology for the PS domain is depicted in Fig. 3.7 and the automatically derived BN is depicted in Fig. 3.8. For the NW domain, the ontology was constructed using the guidelines of Naphade et. al. in [109]. More specifically, the concepts were associated on the basis of program categories N_G={*politics, finance/bussiness, science/technology, entertainment, weather, commercial/advertisement*}

[1] http://www.norsys.com/
[2] http://protege.stanford.edu/

that were placed at the top of the hierarchy, having the rest of the concepts N_L as subclasses. Subsequently, the methodology of Section 3.2 was applied to construct the corresponding BN. Both the ontology and the BN of the NW can be accessed through the web page [1].

The total set of images in I^{PS} was split in half to formulate the test I^{PS}_{test} and train I^{PS}_{train} sets, each one containing 324 images. I^{PS}_{train} was used both for training the classifiers F_c and learning the parameters of the BN. In a similar fashion out of 137 video clips for the NW domain, the key-frames included in the first 100 (i.e. 45276 still images) I^{N}_{train} were selected for learning the parameters of the BN. The key-frames of the remaining 37 video clips (i.e. 16624 still images) I^{N}_{test} were used as ground truth for testing. As for the classifiers the baseline detectors released by Columbia University [104] for all 372 concepts were adopted.

[1] http://mklab.iti.gr/content/evidence-driven-image-interpretation-using-ontologies-bayesian-networks

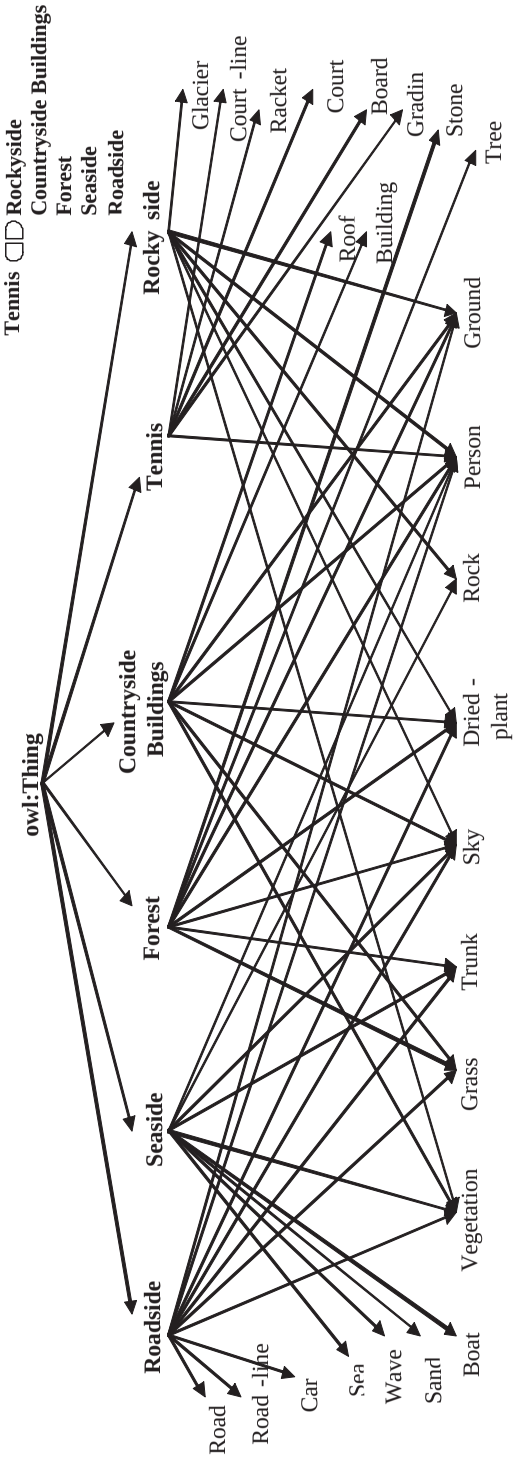


Figure 3.7: Ontology encoding the domain knowledge about the "Personal Collection" domain.

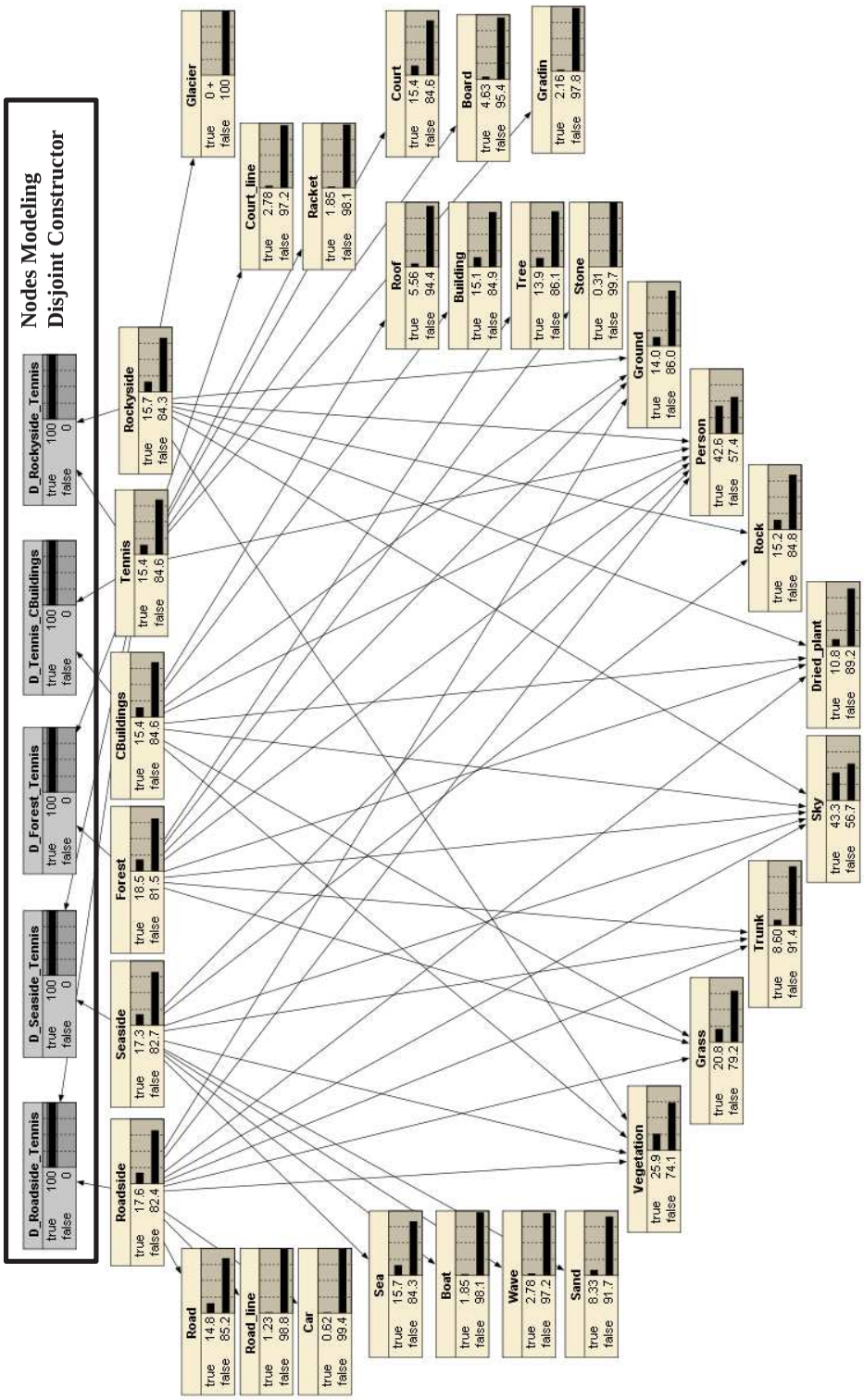


Figure 3.8: Bayesian network derived from the ontology of Fig. 3.7 modeling the "Personal Collection" domain. The nodes in the black frame are control nodes that are used to model the disjointness between the concept *Tennis* and all other category concepts.

3.3.3.1 Image Categorization

In this experiment we measure the efficiency of categorizing the images of I_{test}^{PS} to one of the categories in C_G using three configurations. These configurations vary in the amount of utilized context and knowledge. In the baseline configuration $CON1$ we assess the performance of image categorization based solely on visual stimulus. Images are categorized based on the maximum value of the global concept classifiers. The second configuration $CON2$ uses context (i.e. $X = S(app, W)$) and knowledge (i.e. $K_D = S(N_C, R, O)$) in order to extract the existing evidence and facilitate the process of evidence driven probabilistic inference. In this case, information from the image regions is incorporated into the analysis process but no semantic constraints are taken into account. The BN employed by this configuration is the one depicted in Fig. 3.8 without the nodes enclosed by the black frame. Since the joint probability distribution (JPD) of the random variables included in a BN is equal to the product of the conditional probability distributions of these variables, given the variables corresponding to the parent nodes of the former [110], the JPD defined by this BN is:

$$Pr(C_G^1, \ldots, C_G^{|G|}, C_L^1, \ldots, C_L^{|L|}) = \prod_{i=1}^{|G|} Pr(C_G^i) \prod_{j=1}^{|L|} Pr(C_L^j | Parent(C_L^j)) \qquad (3.12)$$

where $Parent(C_L^j)$ is the set of parent nodes of C_L^j according to the BN. The fact that none of the category concepts C_G has parent nodes (as shown in Fig. 3.8) allows us to include in the expression of the JPD, the first product on the right hand side of eq. (3.12). This expression represents the product of the marginal probabilities of the category concepts.

The third configuration $CON3$ takes into account the semantic constraints of the domain using the methodology presented in Section 3.2 to construct the BN. In this case the BN used for performing probabilistic inference is extended with the addition of the control nodes (i.e. the set of nodes enclosed by the black frame of Fig. 3.8) that are used for modeling the disjointness between *Tennis* and all other category concepts of the ontology. In this case if we define C_D to be the set of control nodes the JPD defined by the utilized BN is:

$$Pr(C_G^1,\ldots,C_G^{|G|},C_L^1,\ldots,C_L^{|L|},C_D^1,\ldots,C_D^{|D|}) = \prod_{i=1}^{|G|} Pr(C_G^i) \prod_{j=1}^{|L|} Pr(C_L^j|Parent(C_L^j)) \prod_{k=1}^{|D|} Pr(C_D^k|C_G^k,C_G^{Tennis}) \tag{3.13}$$

The use of the common superscript k in both C_D and C_G in the last product of eq. (3.13), indicates that every node of the subnetwork that is used to model the disjointness between each category concept and *Tennis*, is conditioned on the node of the corresponding category concept and the node corresponding to *Tennis*.

The reason for treating $CON2$ and $CON3$ as two different configurations was to examine how much of the overall improvement comes from the use of regional evidence and concept hierarchy information ($CON2$), and how much comes from the enforcement of semantic constraints in the analysis process ($CON3$). In both $CON2$ and $CON3$ configurations the analysis process runs as follows. Initially, we formulate the hypotheses set using all category concepts. Then, we look for the presence of all possible regional concepts determined in K_D (i.e. $\forall c_j \in C_L$) before deciding which of them should be used as evidence. This approach requires applying all available classifiers, global and local, and producing one set of confidence values for the image as a whole, $LK_{global} = \{Pr(c_i|I_q) : \forall c_i \in C_G\}$ (see Fig. 3.10, table with title "Global Classifiers") and one set per identified image region, $LK_{local} = \{Pr(c_j|I_q^{s_k}) : \forall c_j \in C_L \quad \& \quad \forall s_k \in S\}$. The latter is a matrix where its columns correspond to the regions identified by the segmentation algorithm of Section 3.3.2.2 and its rows correspond to the confidence degrees of the regional concepts determined in K_D (see Fig. 3.10, table with title "Local Classifiers"). All values of LK_{global} and the maximum per column values of LK_{local} are inserted as soft evidence into the corresponding nodes of the BN. Then, the network is updated to propagate evidence impact and the concept corresponding to the node with the highest resulting posterior probability, among the nodes representing category concepts, is selected to categorize the image (i.e. in this case $\otimes \equiv \max$, see Table 3.1). Fig. 3.9 summarizes the obtained results for the three different configurations. It is clear that the performance obtained using the $CON2$ is superior to the one obtained using $CON1$, since an average improvement of $\approx 5\%$ units is observed for the F-measure [111] metric.

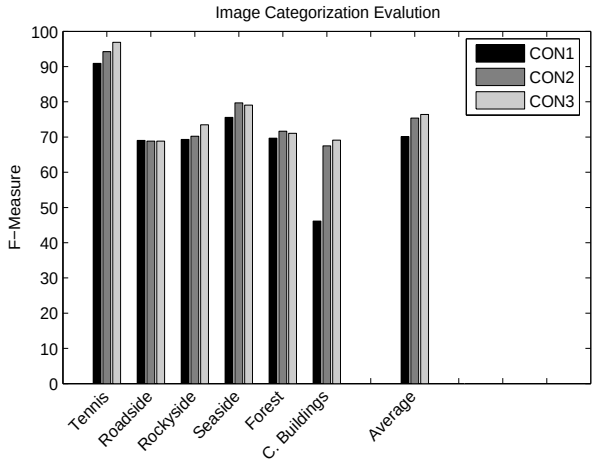
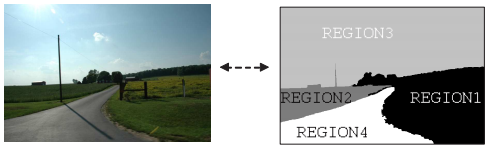


Figure 3.9: F-Measure scores for the task of image categorization using CON1: the output of the global concept classifiers is used to categorize the image, CON2: uses also knowledge and application context for categorizing the image, CON3: takes also into account the semantic constraints expressed in an ontology.

The running example of Fig. 3.10 demonstrates how evidence collected using regional information ($CON2$) can correct a decision erroneously taken by a global classifier that relies solely on visual stimulus ($CON1$). In Fig. 3.10 the Table "Global Classifiers" depicts the probabilities $Pr(c_i|I_q)$ that are obtained after the global classifiers are applied on image I_q. Using only this information the image is categorized as *Seaside* (i.e. this is the result of $CON1$). *Seaside* is the chosen category even after inserting the values $Pr(c_i|I_q)$ into the network and performing inference (i.e. second row of table with title "*Belief Evolution*" in Fig. 3.10). However, as the pieces of regional evidence (i.e. the maximum value from each column of the "*Local Classifiers*" table), are consecutively inserted into the BN, belief propagation causes the posterior probabilities of the category concepts to change. The last four rows of "*Belief Evolution*" table illustrate how the posterior probabilities of each category concept evolve in the light of new evidence. Eventually the correct category, that is *Roadside*, emerges as the one with the highest posterior probability. What is interesting in this example is the fact that only two out of four local classifiers (the ones corresponding to regions 1 and 3) predicted correctly the regional concept. Nevertheless, this information was sufficient for our framework to infer the correct prediction, since the relation between

the concepts *grass* (identified in region 1) and *Roadside*, was strong enough to raise the inferred posterior probability of this category above the corresponding value of *Seaside*. This is a reasonable result since the *Seaside* category receives no support from the evidence *grass*, as shown in Fig. 3.7.



	Global Classifiers
Tennis	45,97
Roadside	54,21
Rockyside	47,07
Seaside	**56,31**
Forest	52,46
C.Buildings	56,00

Bayesian Network of Fig. 3.8 (CON2)

	Local Classifiers			
	Region1	Region2	Region3	Region4
Board	48,51	49,51	50,62	46,84
Court	49,72	52,53	51,83	54,00
Person	52,56	50,16	51,34	52,93
Gradin	49,07	53,01	51,40	51,67
Court line	50,13	49,25	51,85	51,93
Racket	48,93	50,99	50,26	48,34
Sky	47,03	**60,69**	**74,54**	**56,03**
Dried plant	47,14	49,63	48,09	47,97
Grass	**56,36**	53,29	47,66	50,31
Road	52,35	48,47	49,45	54,97
Vegetation	49,17	47,18	46,85	53,74
Ground	50,55	48,56	50,70	49,51
Road line	48,55	49,93	49,89	50,10
Car	47,83	49,51	47,60	47,92
Trunk	48,82	47,68	49,00	48,41
Rock	49,13	47,40	47,79	48,00
Glacier	50,00	50,00	50,00	50,00
Sea	48,73	51,32	47,87	44,08
Sand	49,62	47,86	49,66	47,41
Wave	52,85	46,54	47,96	48,29
Boat	49,76	49,74	47,63	48,77
Tree	50,47	47,61	48,41	48,25
Stone	49,58	49,50	49,39	49,55
Building	44,62	47,71	46,14	43,97
Roof	48,60	52,30	49,47	49,90

	Belief Evolution					
%	Tennis	Roadside	Rockyside	Seaside	Forest	C,Buildings
Prior Probabilities	15,4	17,6	15,7	17,3	**18,5**	15,4
Global	13,4	20,2	14,2	**21,2**	20,1	18,8
Evidence-Region1	13,4	**21,7**	14,2	21,2	20,9	20,6
Evidence-Region2	13,4	**23,5**	15,2	22,9	21,5	22,6
Evidence-Region3	13,4	**27,2**	17,3	26,3	22,6	26,8
Evidence-Region4	13,4	**27,2**	17,6	26,9	22,8	27,4

Figure 3.10: Example of image categorization using the framework's $CON2$ configuration where local information helps to correct a misclassification error about the image category.

The lower of cells in Table 3.2 depict the confusion matrix that corresponds to the output of $CON2$. By looking at the relations between regional and category concepts in Fig. 3.7 in conjunction with Table 3.2, it is clear that our framework tends to confuse categories that share many regional evidence. This is the case for *Rockyside* and *Forest* or *Countryside Buildings* and *Roadside*. Another interesting observation (Fig. 3.7) is the small amount of regional evidence that *Tennis* shares with the rest of image categories. This can be practicaly considered as domain information (i.e. semantic constraint) that can be used to aid image analysis. In order to do so, we associate the *Tennis* concept and all other concepts in C_G with the *"owl:disjointWith"* DL-constructor. Then, we reconstruct the BN using the enhanced ontology. The nodes of the BN that are enclosed

Table 3.2: Confusion Matrix for Image Categorization - $CON2$ lower of the cells - $CON3$ upper of the cells

%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Buildings
Tennis	98.00 94.00	0.00 0.00	0.00 2.00	2.00 4.00	0.00 0.00	0.00 0.00
Roadside	1.75 0.00	73.68 73.68	0.00 0.00	8.77 8.77	10.53 12.28	5.26 5.26
Rockyside	5.88 0.00	3.92 3.92	64.71 70.58	5.88 5.88	19.61 19.61	0.00 0.00
Seaside	0.00 0.00	5.36 5.36	3.57 3.57	91.07 91.07	0.00 0.00	0.00 0.00
Forest	0.00 0.00	10.00 10.00	8.33 8.33	10.00 10.00	71.67 71.67	0.00 0.00
C. Buildings	2.00 0.00	24.00 24.00	6.00 6.00	12.00 12.00	2.00 2.00	54.00 56.00

by the frame in Fig. 3.8 are used to model the disjointness between *Tennis* and all other category concepts. We can see from Fig. 3.9, that using the semantic constrains (CON3) the performance of image analysis is further increased with an average improvement of $\approx 6.5\%$ units from the baseline configuration ($CON1$). By inspecting the upper of the cells in Table 3.2 where the confusion matrix for the $CON3$ is depicted, we can see that the improvement comes basically from the correction of the test samples that were mis-categorized as *Tennis*.

In order to examine the statistical significance of this improvement we apply the McNemar test [112] on the output of $CON1$ and $CON3$ configurations. We selected this test since it is a non-parametric method that can be applied on qualitative variables, such as the output of our different configurations. McNemar test is basically affected by the number of times a transition of the type (success $\rightarrow$ failure) or (failure$\rightarrow$ success) is observed. For a two-tailed test the null hypothesis states that there is equal probability of going from failure to success and vice versa and that this probability is no better than the totally random case. That is H_0 : P(failure $\rightarrow$ success) = P(success $\rightarrow$ failure) = 1/2. The alternative hypothesis states that there is significant difference in statistical terms between the results generated by the two prediction schemes, that is H_1 : P(failure $\rightarrow$ success) $\neq$ P(success $\rightarrow$ failure). The goal is to reject the null hypothesis in favor of the alternative and verify that the difference in

performance observed in the evaluation metrics is statistically significant. The 2×2 contingency table summarizing the transitions observed between $CON1$ and $CON3$ configurations is depicted in Table 3.3.

Table 3.3: Contingency Matrix - Image Categorization

		before		
		+	-	**Total**
after	+	218	30	**248**
	-	15	61	**76**
	Total	**233**	**91**	**324**

Since the number of discordant pairs ($30 + 15$ in our case) is more than 25, the chi-squared approximation with Yates' correction and 1 degree of freedom is calculated to be 4.536. Thus, the $p - value$ calculated by the McNemar's test equals 0.0369. By adopting the conventional criteria used for deciding on statistical significance that considers the significance level α to be 0.05, we have $p - value < \alpha$. Thus, it is safe to conclude that the difference in performance introduced by $CON3$ configuration is statistically significant.

3.3.3.2 Image categorization using a Focus of Attention mechanism

In order to assess the benefit of using the proposed Focus of Attention (FoA) mechanism, we measure the gain in computational cost in terms of two quantities. The number of classifiers (#Classifiers) that need to be applied and the number of inferences (#Inferences) that need to be performed. #Inferences is the number of times a confidence degree is inserted into one of the BN nodes and as a result triggers an inference process. When the FoA mechanism is not employed, the #Inferences that need to be performed for analyzing a single image is equal to the number of confidence values estimated for the global concepts of the image (i.e. the 6 values of LK_{global} in our experiments) plus the number of regions identified in the image (i.e. maximum per column values of LK_{local}). Thus, the total #Inferences for the complete set of 324 test images is $324 * 6$ plus the number of regions identified in all 324 test images, which was calculated to be 2010. Table 3.4 shows the #Classifiers and #Inferences for the exhaustive case of Section 3.3.3.1 (i.e. $CON3$). These values will serve as the baseline reference when estimating the computational gain of the FoA mechanism.

Table 3.4: Computational Cost Quantities - $CON3$ Configuration

	324 (# Test Images) * 6 (# Global Classifiers) + 2010 (# Total Regions) * 25 (# Local Classifiers)
# Classifiers	**52194**
	324(# Test Images) * 6 (# Global Classifiers) + 2010 (max of local classifiers per region)
# Inferences	**3954**

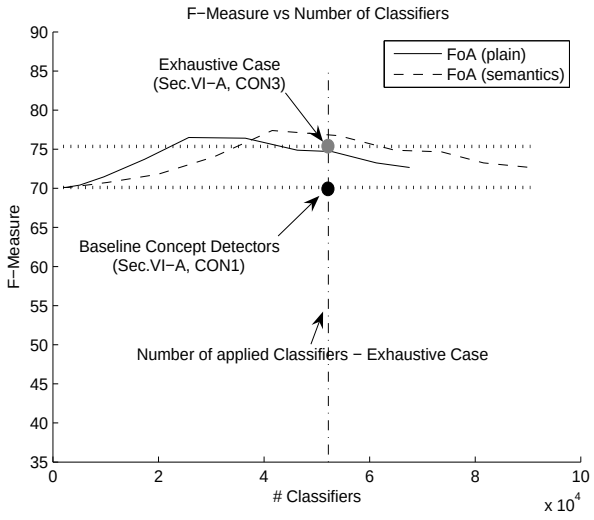
In our experimental setting the belief threshold receives one of the following discrete values $\{0.1, 0.2, \ldots, 1.0\}$. Using each of these values as a common belief threshold for all formulated hypotheses, we obtain 10 different F-Measure scores. Given that the belief threshold affects also the #Classifiers and the #Inferences, we practically obtain 10 pairs of values for {F-Measure, #Classifiers} and 10 pairs of values for {F-Measure, #Inferences}. These pairs are used to draw the curves depicted in Figs. 3.11(a) and 3.11(b). In both diagrams we demonstrate the performance of: a) the baseline concept detectors (i.e. $CON1$ of Section 3.3.3.1) (black dot), b) the probabilistic inference using exhaustive search (i.e. $CON3$ of Section 3.3.3.1) (gray dot), c) the plain FoA mechanism (solid curve), and d) the FoA mechanism using also the methodology of Section 3.2 for incorporating semantic constraints (dashed curve). The baseline figures of Table 3.4 are also displayed in Figs. 3.11(a) and 3.11(b) using the vertical lines. The horizontal dotted lines are drawn for allowing comparisons with the performance of the baseline configurations. It is clear that the proposed FoA mechanism manages to achieve (for the optimal value of the belief threshold, F-Measure $= 76,40$) performance comparable to the one obtained by the best of the configurations in Section 3.3.3.1, using a remarkably smaller number of classifiers. On the other hand, for the same optimal threshold value, the number of inferences that need to be performed increases, see Fig. 3.11(b). More specifically, the number of classifiers reduces from 52194 to 25753 (# classifiers corresponding to the peak of the solid curve in Fig. 3.11(a)), while the number of inferences increases from 3954 to 4538 (# inferences corresponding to the peak of the solid curve in Fig. 3.11(b)). For the case where the FoA mechanism incorporates semantic constraints (dashed curve), the number of applied classifiers reduces from 52194 to 41560 (# classifiers corresponding to the peak of the dashed curve in Fig. 3.11(a)), while the number of inferences increases from 3954 to 6860 (# inferences corresponding to the peak of the dashed curve in Fig. 3.11(b)).

In order to estimate these numbers in terms of time we have calculated the average time per classifier and per inference to be $0,12$ (sec) and $0,69 * 10^{-3}$ (sec), respectively. Thus. the gain in computational time is approximately 3172 (sec) using the plain FoA mechanism and 1274 (sec) using the FoA with semantic constraints, which can be considered as a significant reduction of the overall computational cost. Finally, let us note that in both approaches for image categorization (Section 3.3.3.1 and Section 3.3.3.2) the configuration incorporating semantic constraints outperforms the other configurations, which is an additional argument for the effectiveness of the framework presented in Sections 3.1 and 3.2.

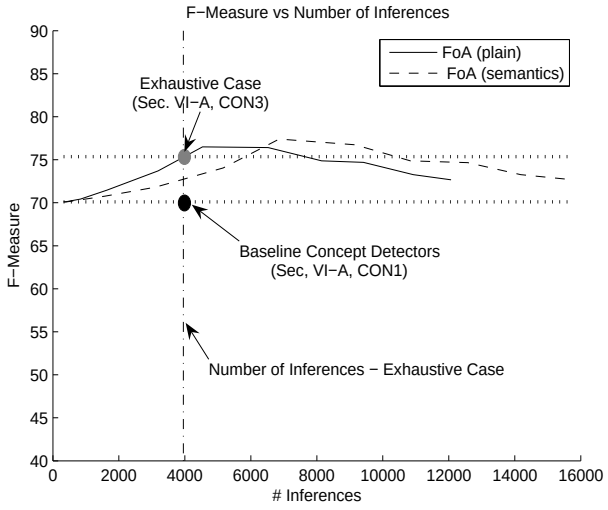
3.3.3.3 Localized Region Labeling

In order to evaluate the performance of our framework for the task of assigning labels to pre-segmented regions, we have used the BN of Fig. 3.8 (without the nodes enclosed by the black frame) and the JPD of eq. (3.12). As mentioned in Section 3.3.2.1, our framework can reinforce region labeling when there is a conflict between the decisions suggested by the global and local classifiers. Let $Child(c_k) = \{c_j : k \rightarrow_{parent} j\}$ be the subset of C_L corresponding to the child nodes of $c_k \in C_G$. Let also $LK_{global} = \{Pr(c_i|I_q) : \forall c_i \in C_G\}$ be the set of confidence values obtained from the global classifiers applied to image I_q and $LK^{s_w}_{local} = \{Pr(c_j|I^{s_w}_q) : \forall c_j \in C_L\}$ be the set of confidence values obtained from the local classifiers applied to a region $I^{s_w}_q$ of the image. A conflict occurs when $c_l \notin Child(c_g)$ with $g = \arg\max_i(LK_{global})$ and $l = \arg\max_j(LK^{s_w}_{local})$.

In the first case we follow the suggestion of the global classifiers and select the concept c_g. Then, the local concept c_l is selected such that $l = \arg\max_j(LK^{s_w}_{local})$ and $c_l \in Child(c_g)$. The confidence values corresponding to c_g and c_l are inserted into the BN as evidence and the overall impact on the posterior probability of the hypothesis stating that the region under examination $I^{s_w}_q$ depicts c_l is measured. In the second case, we follow the suggestion of the local classifiers and select $c_{\acute{l}}$, such that $\acute{l} = \arg\max_j(LK^{s_w}_{local})$. The confidence values of the global classifiers are examined and the $c_{\acute{g}}$ with $\acute{g} = \arg\max_i(LK_{global})$ and $c_{\acute{g}} \in F(c_{\acute{l}})$ is selected. As in the previous case, the confidence values corresponding to $c_{\acute{l}}$ and $c_{\acute{g}}$ are inserted into the network and the overall impact on the posterior probability of the hypothesis stating that the examined region $I^{s_w}_q$ depicts $c_{\acute{l}}$ is measured. Eventually, the values representing the impact on the posterior probabilities of the two different cases are compared and depending on the



(a)



(b)

Figure 3.11: F-Measure scores using the Focus of Attention mechanism against: a) # Classifiers, b) # Inferences. Each point in a curve corresponds to a belief threshold that receives one of the following discrete values $\{0.1, 0.2, \ldots, 1.0\}$.

largest value, c_l or $c_{\bar{l}}$ is chosen to label the region in question (i.e. this is the functionality of $\otimes$ operator depicted in Table 3.1, for this task). If no conflict occurs, the concept corresponding to the local classifier with maximum confidence is selected. Fig. 3.12 shows that when using the proposed framework an average increase of approximately

4.5% units is accomplished.

Table 3.5: Contingency Matrix - Localized Region Labeling

		before		
		+	-	**Total**
after	+	1035	61	**1096**
	-	22	892	**914**
	Total	**1057**	**953**	**2010**

In order to apply the McNemar's test for this case we calculate the 2×2 contingency matrix depicted in Table 3.5. The $p - value$ estimated by the McNemar's test is found to be less than 0.0001 showing that the improvement is statistically very significant, since $p - value << \alpha$.

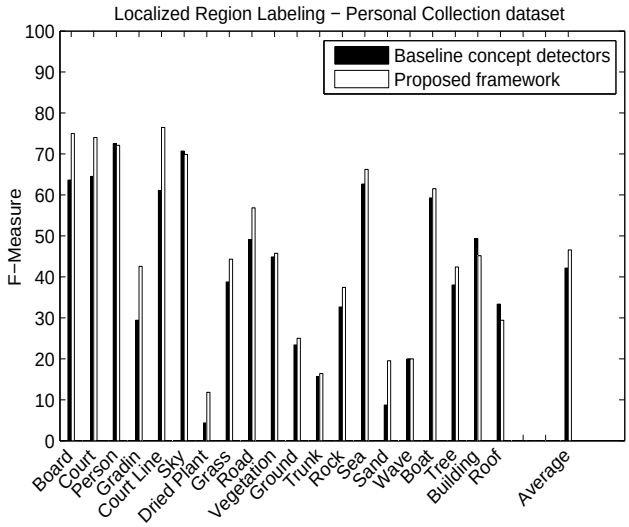


Figure 3.12: F-Measure scores for the localized region labeling task applied on the Personal Collection dataset. Scores are reported for the baseline case, where decisions are based solely on the output of the classifiers, and for the case where knowledge and context are employed to improve image analysis.

3.3.3.4 Weakly annotating video shot key-frames

In this experiment, the task was to weakly annotate (i.e. identify the presence of a concept but not localize it within the image) the key-frames of video sub-shots. In

contrast to the previous cases this task does not require the existence of region-level annotations, which are very difficult to obtain, and therefore allows us to perform tests on a much larger set of semantic concepts. The TRECVID 2005 dataset was used for this purpose. Recalling that N_G denotes the set of category concepts that were placed at the top of the hierarchy and N_L the rest of domain concepts that were used as subclasses of N_G, the JPD defined by the utilized BN is:

$$Pr(N_G^1, .., N_G^{|G|}, N_L^1, .., N_L^{|L|}) = \prod_{i=1}^{|G|} Pr(N_G^i) \prod_{j=1}^{|L|} Pr(N_L^j | Parent(N_L^j)) \tag{3.14}$$

The benefit of using such a large dataset is the existence of semantic relations between the available concepts. These relations are necessary for assessing the effectiveness of our framework, since our goal is to exploit domain knowledge for improving the efficiency of image interpretation. On the other hand, many of the concepts appear rarely in the training set; a fact that makes difficult approximating the conditional probabilities using frequency information. In order to assess the efficiency of our framework we compare its performance against the performance of baseline concept detectors that make no use of domain knowledge and application context. In the first case we use the fused output of the global detectors released by the Columbia University [104]. The concepts corresponding to the K maximum confidence values produced by the global detectors are selected to weakly annotate the key-frames. In the second case, the fused detection confidence values of all classifiers are provided as evidence to the BN. Belief propagation is performed and the resulting posteriors are recorded for all concepts. Finally, the K concepts that exhibit maximum positive impact on their posteriors are selected as the analysis outcome (i.e. this is the functionality of $\otimes$ operator depicted in Table 3.1, for this task). For both cases, K was determined by varying its value between 2 and 20 and selecting the one that yields the optimal average F-Measure score.

In order to examine the relation between a concept's Appearance Frequency (AF) in the training set and the efficiency of the proposed framework, we report the F-Measure scores sorted based on the AF of the concepts. By inspecting Fig. 3.13(a) we observe that for the concepts with $AF \geq 10\%$ our framework outperforms the baseline in almost all cases. In Fig. 3.13(b), where the concepts with $10\% > AF \geq 5\%$ are depicted, we observe a similar behavior, but with the average improvement to be inferior from that

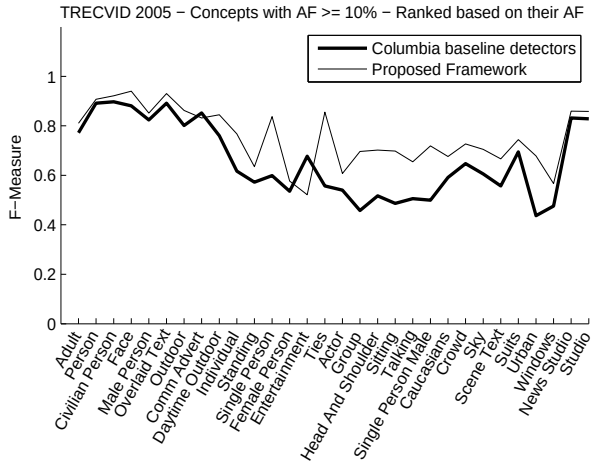
of Fig. 3.13(a). Finally, Fig. 3.14 verifies that when the AF of a concept is relatively small (Fig. 3.14 depicts concepts with $5\% > AF \geq 2\%$) our framework does not deliver any improvements. Similar conclusions can be drawn when $AF < 2\%$. It is evident that the availability of realistic prior and conditional probabilities is important for the efficiency of our framework and learning them from data is feasible only when there are enough training samples to learn from.

3.3.3.5 Comparison with existing methods

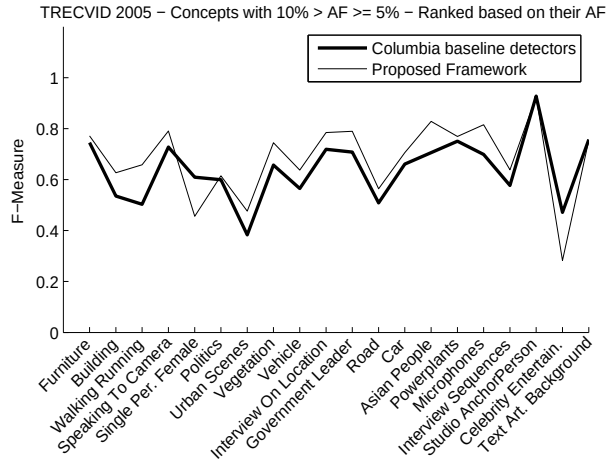
In order to compare our work with other methods in the literature we apply the localized region labeling task on the 591 images of the MSRC dataset[1] I^{MSRC}. In order to do so we categorized all 591 images into 6 categories (i.e. global concepts) namely, *Cityscape*, *Countryside*, *Forest*, *Indoors*, *ManMade* and *Waterside*. As regional concepts we used 21 out of the 23 semantic classes provided by MSRC, treating as void the *horse* and *mountain* classes that appear very rarely. An ontology was created to represent the relations between the aforementioned global and regional concepts and a BN was derived from it using the methodology presented in Section 3.2. Both the ontology and the BN for the MSRC dataset can be accessed through our web page[2]. All images of I^{MSRC} were segmented by the segmentation algorithm described in Section 3.3.2.2 and the ground truth label of each segment was taken to be the label of the hand-labeled region (hand-segmented and hand-labeled regions are provided by the Microsoft Research Cambridge team for all 591 images) that overlapped with the segment by more than the 2/3 of the segment's area. In any other case the segment is labeled as void. We should note that although we could use directly the hand-segmented images included in the MSRC dataset, such an approach would not be realistic since we cannot reasonably expect segmentation information for an unknown image. The overlap rule has been used by many works in the literature that utilize automatic image segmentation and require for labeling the automatically extracted segments, based on a set of manually generated (ground truth) segments. For instance in [95] the authors use 20×20 image patches whose labels are taken to be the most frequent ground truth pixel label within the block. Similarly in [113] in order to find the best possible combination

[1] http://research.microsoft.com/vision/cambridge/recognition

[2] http://mklab.iti.gr/content/evidence-driven-image-interpretation-using-ontologies-bayesian-networks



(a)



(b)

Figure 3.13: F-Measure scores for the concepts of TRECVID 2005 dataset ranked based on their appearance frequency (AF) in the training set: a) $AF \geq 10\%$ and b) $10\% > AF > 5\%$

of automatically extracted segments for constructing a figure-ground segmentation, the authors include an automatically identified segment into the foreground if it has more than 50% overlap with the ground-truth foreground in terms of the segment's area. The I^{MSRC} was split randomly in 295 training I^{MSRC}_{train} and 296 testing I^{MSRC}_{test} images,

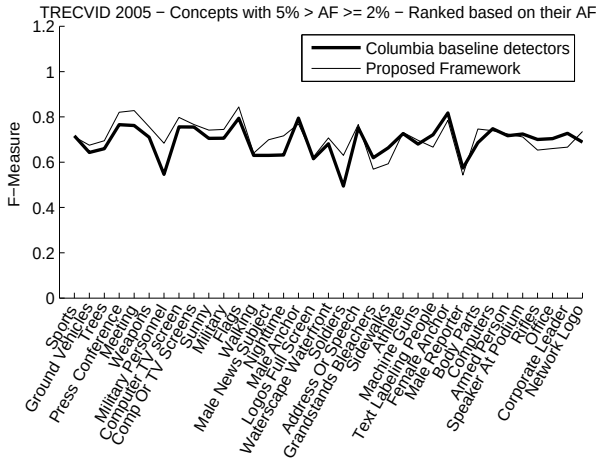


Figure 3.14: F-Measure scores for the concepts of TRECVID 2005 dataset ranked based on their appearance frequency (AF) in the training set with $5\% > AF > 2\%$.

ensuring approximately proportional presence of each class in both sets. I_{train}^{MSRC} was used to train the concept classifiers and learn the parameters of the BN. Fig. 3.15 reports the performance for both the baseline concept classifiers as well as the proposed framework configured as described in Section 3.3.3.3, on the I_{test}^{MSRC}. We can see that performance is increased in 14 out of the 21 regional concepts giving an average improvement of approximately 4.5% units in terms of the F-Measure metric. However, there are concepts like *sky*, *chair*, and *cat* that exhibit performance lower from the baseline. This can be attributed to the fact that our framework operates on top of the classifiers' outcomes, which usually come with a high number of erroneous predictions. Intuitively, the framework compensates for the misleading predictions by favoring the co-occurrence of evidence that is known from experience to usually co-exist and constitute the analysis context. It does so by adjusting the final output so as to comply with the extracted collection of evidence. Therefore, provided that an adequate amount of evidence are accurate, the framework is expected to make the correct decision by absorbing any misleading cues produced by the erroneous visual analysis. However, there can also be cases, like the ones mentioned above, where the evidence extracted from context are misleading, causing our framework to change the correct prediction of the local classifier.

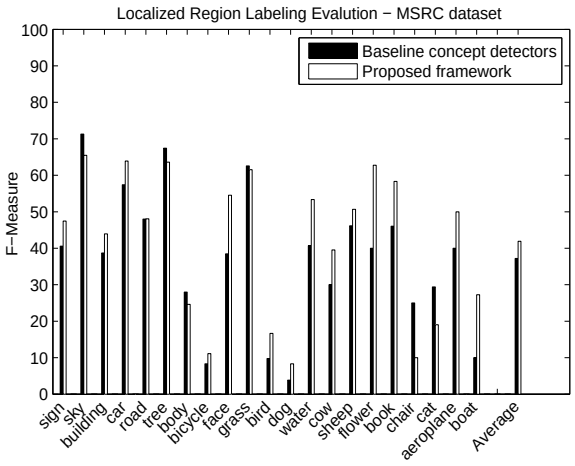


Figure 3.15: F-Measure scores for the localized region labeling task applied on the Microsoft Research Cambridge dataset. Scores are reported for the baseline case, where decisions are based solely on the output of the classifiers, and for the case where knowledge and context are employed to improve image analysis.

Table 3.6: Comparing with existing methods in object recognition

	Buildings	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	Average
Textonboost [96]	**62**	**98**	86	58	50	83	60	53	**74**	63	**75**	63	35	**19**	**92**	15	86	**54**	19	**62**	7	58
PLSA-MRF/P [95]	52	87	68	**73**	**84**	94	**88**	**73**	70	68	74	**89**	33	19	78	**34**	**89**	46	**49**	54	**31**	**64**
Prop. Framework	32	55	**87**	40	73	**96**	57	56	50	**76**	8	64	**38**	12	46	5	51	12	8	29	18	44

In order to present results on the same dataset with the Textonboost system [96] and the Markov field aspect models of [95], we calculated the classification rate (i.e. number of correctly classified cases divided by the total number of correct cases) achieved by our framework for each of the 21 object classes in MSRC dataset. We hereby note that the reported classification results are not directly comparable since the results are reported at different level. In [96] at pixel level, in [95] at the level of 20x20 image patches, and in our case at the level of arbitrary shaped segments which are extracted by an automatic segmentation algorithm. In addition, the methods are not relying on the same set of visual features, and the training/test split is likely to be different. Table 4.6 summarizes the classification rates per class.

It is clear that none of the three systems manages to outperform the others for a

significant portion of the 21 classes. Moreover, error rates are often quite different on individual classes showing that while there are some classes that can be modeled very efficiently using the visual features and the model proposed by one method, there are other classes that are best modeled using a different set of visual features and model. For instance while the visual features employed by our method perform very poorly in recognizing the class *grass*, they are pretty efficient in recognizing the class *car* or *sky*.

3.4 Case study on compound document analysis using information across media

Existing methods for the semantic analysis of multimedia, although effective for single-medium scenarios, are inherently flawed in cases where knowledge is spread over different media types. The group of research approaches that seek to enhance semantic metadata extraction by exploiting information across media are usually referred as cross media analysis. Practically, the aim of such methods is to combine the evidence extracted from different media types and accumulate their effect in favor or against a certain hypothesis. These pieces of evidence can belong to different levels of granularity and used differently by the analysis mechanism. For instance, we can consider cross media analysis to be a general fusion problem that is carried out at different levels of abstraction, namely result-level [114], [115], [116], extraction-level [117], [118], [119] and feature-level [120], [121], [122].

The goal of this case study is to demonstrate how the framework proposed in Sections 3.1 and 3.2 can be used to boost the efficiency of cross media analysis by exploiting the knowledge explicitly provided by domain experts (i.e. domain knowledge). Towards this direction, the framework developed for this case study operates on the result-level of abstraction and allows domain knowledge to become part of the inference process [123], [124]. More specifically, our framework combines the soft evidence collected from different media types, to support or disprove a certain hypothesis made about the semantic content of the analyzed resource. Soft evidence are obtained by applying single-medium analyzers on the low-level features of the different media types. Subsequently, these pieces of evidence are used to drive a probabilistic inference process that updates the observable variables of the BN and verify or reject the examined hypothesis based on the posteriori probability of the remaining variables. Fig. 3.16 demonstrates

the functional relations between the components of the proposed cross media analysis approach.

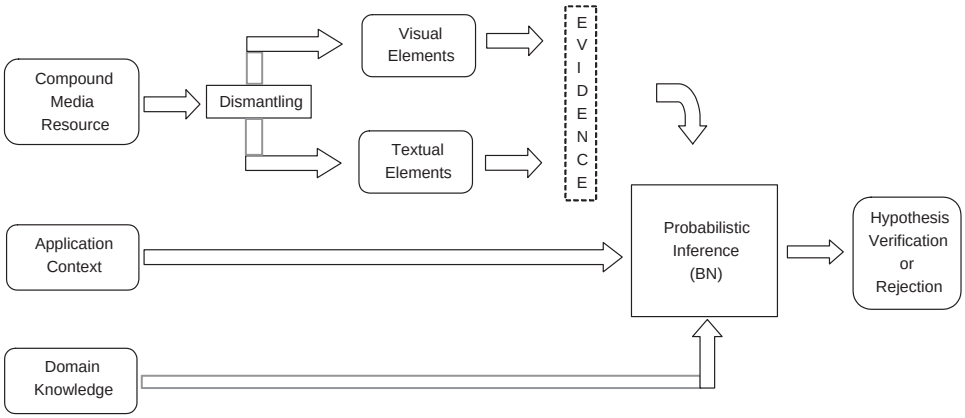


Figure 3.16: Cross media analysis scheme

In this case study we use content from a real world application taken from the car manufacturing industry [125], as well as content from the TRECVID2010 competition, to verify that performing cross media analysis using the proposed approach leads to significant improvements compared to the cases where single-medium analyzers act separately. In the following we describe the implemented cross media analysis approach including details for the utilized single-medium analyzers.

3.4.1 Cross media analysis approach

To demonstrate the ability of the proposed evidence-driven probabilistic inference framework to efficiently handle evidence across media, we have chosen to implement a cross media analysis approach that uses probabilistic inference to detect high-level concepts in compound documents. High-level concept detection is usually the output of knowledge-related tasks and typically requires the synergy of information scattered in different places. The more the available information, the more easy it is for the knowledge worker to infer the presence of a high-level concept. Independently of whether these pieces of information act cumulatively or in a complementary way, they have an impact (i.e. positive or negative) on the confidence of the fact that a certain high-level concept is valid for the analyzed resource. In order to model this process we rely on the framework presented in Sections 3.1 and 3.2 to implement a generative classifier based

on BNs. The role of this classifier is to: i) fuse the information extracted from different media types on the grounds of knowledge and context, ii) produce a confidence degree about the validity of a high-level concept in the analyzed resource, and iii) make a decision by applying a fixed threshold on this confidence degree. Since cross media analysis is mostly about simultaneously evaluating the appropriate evidence extracted across different media types, an important issue for making the aforementioned approach suitable for such purposes is the strategy by which evidence (and as a consequence their source modalities) are considered to be co-related.

In the following we elaborate on the components that are used to implement the cross media analysis approach for compound documents, which are: a) a dismantling mechanism and a modality synchronization strategy for handling the compound media resources, b) the single-medium analysis techniques for extracting evidence using low level features, and c) the actions required to adjust the techniques presented in Sections 3.1 and 3.2 for performing cross media analysis.

3.4.1.1 Compound documents dismantling & modality synchronization

Compound documents are multimedia documents that incorporate more that one media type in the same digital resource. OpenDocument, Microsoft Office's documents, PDF and web pages are indicative representation formats of such documents where visual and textual elements co-exist. A compound document may contain evidence for a concept to be extracted across different media. However, it is not straightforward to know which media elements refer to the same concept. Moreover, these documents carry additional information such as cross references or layout features (e.g. spatial proximity between a caption and an image frame) that have a major effect on the content essence. These features, although very important for human perception, are difficult for knowledge extraction algorithms to encode and exploit.

Document processing literature discusses several approaches to extract layout information from PDF, HTML and other structured documents, see [126] for an overview. Most of these approaches [127], [128] are based on manual or semi-automatically extracted templates that characterize each part of the document. However, the variety of layouts that a document editor is likely to use for expressing the intended meaning, makes it difficult for automated systems to model and make this information available

for analysis. This process is further hindered by the absence of a uniform document representation standard that could reduce the diversity of existing formats.

All the above, makes the employment of a dismantling and synchronization mechanism an important module of cross media analysis. This mechanism will be able to disassemble a compound document to its constituent parts and decide which of these parts should be considered simultaneously by the fusion process. For the purposes of our work, assuming a certain layout for the analyzed documents, we accept that a different topic is covered in each document page and disregard cases where more that one topics exist in the same page or a single topic extents to many pages. Thus, all media elements of the same document page are considered to be conceptually related. Given this assumption, we analyze a document on a per page basis by fusing the output of single-medium analyzers that are independently applied on the media elements residing on the same page. Although such an assumption may seem inconsistent with a non-negligible number of cases, in this work we basically focus on how to effectively fuse cross media evidence on the grounds of knowledge and context, while existing approaches can be employed in cases where this assumption does not hold.

3.4.1.2 Single-medium analysis techniques

In this section we detail the techniques we have used to analyze the low-level features of a document and produce confidence degrees for the related concepts.

Visual analysis: Visual evidence is extracted by applying concept detectors on the images contained in a document page. The method adopted for implementing the concept detectors is based on the Viola and Jones detection approach [129]. The functionality of this approach can be characterized by three key aspects, a) a scheme for image representation called *integral image*, that allows for very fast feature extraction, b) a method for constructing a classifier by selecting a small number of important features using AdaBoost [130], and c) a method for combining successively more complex classifiers in a cascade structure, which dramatically increases the speed of the detector by focusing attention on promising regions of the image.

In more detail, the visual information contained in an image is described by Haar-like features, introduced in [131] and depicted in Fig. 3.17. The values of these features are the differences between the sums of the white and black rectangular regions. In

order to compute these sums efficiently, Viola and Jones make use of integral images. An integral image is an array corresponding to an image that contains in position (x, y) the sum of the intensity values of all pixels above and to the left of (x, y). For the Haar features that are rotated by 45^o, a rotated integral image is used, which accumulates the values inside a triangle starting from point (x, y) and ending at the top of the image. The construction of an integral image requires a linear scan through the actual image and results in computing the feature responses in constant time. The efficient computation of the feature responses is essential, since all of them are computed at all positions and scales in an image, resulting in a very dense representation of approximately 100,000 feature responses for an image of size 20x20 pixels.

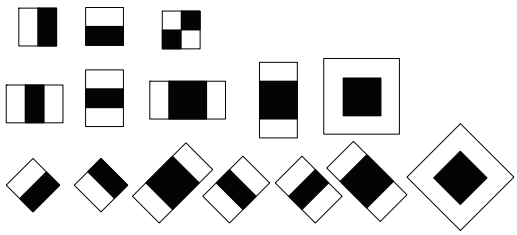


Figure 3.17: Haar-like features. The values of these features are the differences between the sums of the white and black rectangular regions.

Then, the AdaBoost algorithm is used in order to train a classifier for an object category. AdaBoost creates a degenerate decision tree based on the responses of m Haar features that best describe the depicted concept. Classification time is reduced by using several low precision, fast classifiers connected in a cascade, instead of one high precision and slow classifier. In order to classify a sub-window of an image as positive (depicting the object), the sub-window has to be classified as positive by all the classifiers in the cascade, also called stages. If a sub-window is classified as negative (not depicting the object) by any single classifier, then it is rejected and not processed by the following stages, as depicted in Fig. 3.18. The detection task of finding the precise position and scale of the object is performed in a sliding window manner, checking every possible position and scale.

The output of the local concept detector is the exact position and scale at which a concept c_j was found in the analyzed image, as well as a confidence degree associated to every detection result. The confidence degree is extracted from the detectors inner structure, as depicted in Fig. 3.18. More specifically, the output of each classifier

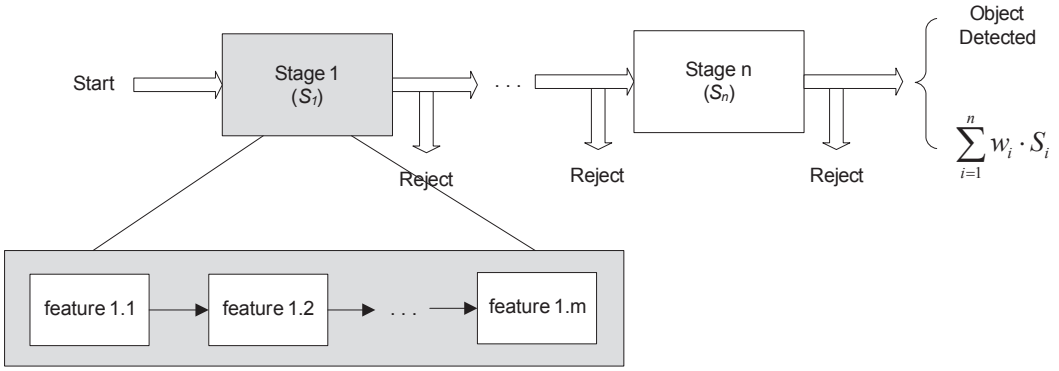


Figure 3.18: Confidence value derived from the cascade of classifiers.

in the cascade is associated with a confidence degree S_i derived by a combination of the calculated decision thresholds. These thresholds are related to the partial or full detection of the concept of interest. The values extracted from all classifiers of the cascade, are then combined in a weighted sum to provide a confidence value for each examined sub-window. The weights w_i applied to each stage output, emphasize the response of the last stages which are more discriminative than the initial low precision ones. The confidence value is then normalized in [0 1], based on the training set used to create the detector. For the purposes of our work we filter out cases with a very small confidence degree and we select the case with maximum confidence degree when multiple instances of the same concept are found on the same image.

Textual analysis: For obtaining textual evidence we need to estimate the semantic relatedness of a concept with the linguistic information contained in a document page. In order to do so, we should be able to measure the semantic relatedness between any two individual concepts and apply a page oriented summarization strategy, as detailed later in this section. Approximating human judgement and measuring the semantic relatedness between concepts has been a challenging task for many researchers. Most works in the literature make use of the WordNet lexical database [132] for achieving this objective.

WordNet can be viewed as a large graph where each node represents a real world concept and each link between nodes represents a relationship between the corresponding concepts. Every node consists of a set of words (synset), that linguistically describe

the real world concept associated with the node, as well as a short description of this concept (gloss). Using the above, WordNet encodes a significantly large amount of knowledge and is able to facilitate a great number of methods determining the semantic relatedness between concepts. Methods existing in the literature can be divided to the ones that use only the structure and content of WordNet to measure semantic relatedness [133], while others achieve this by also exploiting statistical data from large corpora, [134], [135], [136], [137], [138]. Another important characteristic of such methods is whether they are able to operate on all parts of speech [138], [136] or nouns only [133], [134], [135], [137]. For the purposes of our work we decided to employ a semantic relatedness measure that is based on context vectors and was originally presented by Patwardhan in [139]. The method introduced in this work relies on a different representation for WordNet glosses that is based on multidimensional vectors of co-occurrence counts. Its main advantage derives from its ability to combine the benefits of methods that use the knowledge from a large data corpus and the ones that rely solely on the strict definitions of WordNet (glosses).

In order to describe the method in more detail we need to determine the meaning of *word vectors* and *context vectors*. Every word in the word space has a corresponding word vector. The word vector corresponding to a given word is calculated as a vector of integers. The integers are the frequencies of occurrence of each word from the word space in the context. The context of a word is considered to be the words that appear close in the text with this word. Thus, each word in the word space represents a dimension of the vector of integers. Once the word vectors for all words in the word space are calculated, they are used to calculate the context vectors for every instance of a word. This is done by adding the word vectors of all words that appear in the context of this word.

In order to measure the semantic relatedness between two concepts the method of [139] represents each concept in WordNet by a *gloss vector*. A *gloss vector* is essentially a context vector formed by considering a WordNet gloss as the context. More specifically, having created the word vectors for all words in the word space, the gloss vector for a WordNet concept is created by adding the word vectors of all words contained in its gloss. For example, consider the gloss of *lamp - an artificial source of visible illumination*. The gloss vector for lamp would be formed by adding the word vectors

of *artificial*, *source*, *visible* and *illumination*. Eventually, the semantic relatedness between two concepts is defined as the cosine of the angle between the corresponding normalized gloss vectors:

$$SemanticRelatedness(c_1, c_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|v_1||v_2|} \tag{3.15}$$

where c_1, c_2 are the two given concepts, and $\overrightarrow{v_1}$, $\overrightarrow{v_2}$ are the gloss vectors corresponding to the concepts. The motive behind our choice of relying on context vectors over the other existing measures for semantic relatedness is threefold. Context vectors are able to: i) exploit information both from a large data corpora as well as from the WordNet descriptions (glosses), ii) handle all different parts of speech placing no limitations on the amount of linguistic information contained in a document page that can be used to derive an overall degree of semantic relatedness with the query concept, iii) produce values normalized to [0,1], which is crucial for our analysis given the probabilistic standpoint of our framework.

After having defined a method for measuring the semantic relatedness between any two individual concepts, we need a methodology for extracting the overall semantic relatedness between a concept and the linguistic information contained in a document page. In order to do so, we use the previously described approach to measure the semantic relatedness between the word expressing the concept of interest and all words contained in a document page. In this way we get as many semantic relatedness values as the number of words contained in the document page. Subsequently, we only consider the words with semantic relatedness above the 64% of the maximum semantic relatedness value of all words in this page. This percentage was found to yield optimal performance in a series of preliminary experiments. By averaging between the selected values we get a number between [0,1] that indicates the semantic relatedness of the query word with the linguistic information contained in a document page. This number is used as the confidence degree of this concept for the examined document page.

3.4.1.3 Adjusting our framework to perform cross media analysis

Having described the methods for extracting conceptual information out of the low level stimuli of compound documents, the techniques described in Sections 3.1 and

3.2 can be directly applied to support knowledge-assisted analysis of compound documents. More specifically, the modeling approach of Section 3.1 is used to introduce the conceptual true/false space and link the visual and textual analyzers with the resulting BN. Then using the domain ontology as a reference, the methodology of Section 3.2 is employed to map the explicitly provided knowledge into the structure of a BN. As in the previous case study, all concepts are directly translated into network nodes, with an arc being drawn between two nodes if the corresponding concepts are related by an "rdfs:subClassOf" relation in the domain ontology. In contrast to the previous case study, more complex relations of the form owl:disjointWith, owl:unionOf, owl:intersectionOf, owl:unionOf, owl:complementOf and owl:equivalentClass are not exploited, since no such need arises from the domain ontology. For learning the network parameters the knowledge implicit in the data is captured and translated into the prior and conditional probabilities associated with each node in the BN, as explained in Section 3.2.2.2. The CPTs of all network nodes are learned by applying the Expectation Maximization (EM) [97] algorithm on a set of compound documents annotated with concept labels. Finally, evidence-driven probabilistic inference is performed using the junction tree algorithm [89], as explained in Section 3.1.1.3.

3.4.2 Experimental Study

Our goal in this section is to experimentally examine the performance of the developed framework in three different aspects: i) how much improvement is achieved by the employment of the proposed cross-media analysis scheme compared to single-medium solutions, ii) whether the choice of a generative over a discriminative model is more suited for fusing evidence coming from heterogeneous sources, and iii) whether the additional cost of engineering an ontology for expressing domain knowledge, actually pays off in terms of efficiency when compared with less costly approaches like using a simplified BN or learning its structure from data using the K2 algorithm [140]. Finally, we also evaluate the performance of our framework in the context of a video shot classification scheme.

3.4.2.1 Testbed

The domain selected for performing our experimental study concerns forecasting the launch of competitors' models, as defined in cooperation with Centro Ricerche Fiat

(CRF)[1]. The goal of a competitor analysis department is to constantly monitor the existent competitors' products, understand market trends and try to anticipate customer needs. The information needed to achieve that, is scattered throughout the Internet (i.e. blogs and forums), and covered by a long tail of international and national automotive magazines. In a typical scenario the main role is played by the person responsible for data acquisition that has the responsibility of daily inspecting a number of resources such as WWW pages, car exhibitions, car magazines, etc, that are likely to publish material of potential interest. The collected information is subsequently used in the *set-up* stage of new vehicles (i.e. the development stage where a first assessment of the future vehicle's features is carried out). This process is of great value to many companies because it contributes to keeping new product designs up to date. One of the tasks defined by the experts was to be able to automatically evaluate a document with respect to its interest for the *car components ergonomic design*. The fact that most of the collected documents use both visual and textual descriptions, motivated the construction of a cross media classifier recognizing compound resources that are valid for the high-level concept *car components ergonomic design*.

For the purposes of our evaluation a dataset of 162 pdf documents (containing 1453 pages) was collected, that are primarily advertising brochures describing the characteristics of new car models. Each pdf document was dismantled into its visual and textual constituent parts using the xpdf library[2]. All media elements extracted from the same page were kept together so as not to lose any conceptual relations originating from the document's layout. The linguistic information was gathered in a single text file while the visual representations were extracted to independent image files as depicted in Fig. 3.19.

Two different manual annotation efforts were carried out for the purposes of our work. Since we have decided to consider the pdf documents on a per page basis, the first annotation effort was to manually inspect each of the 1453 document pages and record in an annotation file whether they are valid for the high level concept *car components ergonomic design*. The second annotation effort involved going through all 1453 document pages and marking for each page which of the ontology concepts are present or not. The result of this annotation process was a set of concept labels for each

[1] http://www.crf.it/
[2] http://www.foolabs.com/xpdf/

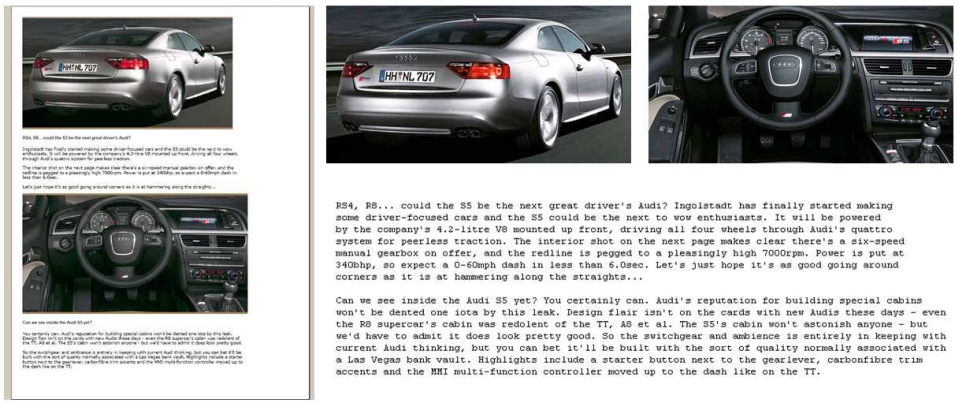


Figure 3.19: Dismantling a pdf document to its constituent parts

of the 1453 document pages, suitable for measuring the co-occurrence between any two concepts of the domain. These sets of concept labels were used to learn the CPTs of the BN nodes. More specifically, out of the 162 documents, 149 (928 pages) I^{train} were used for learning the BN parameters and 13 (525 pages) were used for testing I^{test}.

3.4.2.2 High level concept detection using the cross media analysis scheme

For conducting our experiments we have engineered three ontologies (one for each of the evaluated cases: textual-only, visual-only and cross media) that are mostly concerned with concepts related to the ergonomic design of car components. All three ontologies were engineered based on the knowledge acquired by going through a sufficient number of related documents and getting acquaintance with the domain details. These ontologies were used to determine the structure of three different BNs (one for each evaluation case). In all cases, the node modeling the high-level concept *car components ergonomic design* was placed at the root node of the constructed BN. For learning the CPTs of the BN nodes, the Expectation Maximization algorithm was applied on I^{train}. Depending on the concepts included in the employed ontology, only the annotations referring to these concepts were included in the corresponding training set.

After constructing the BNs the analysis process runs as follows. Depending on the examined case (textual-only, visual-only, or cross media) the single-medium analyzers are applied on the constituent parts of a document page. Their probabilistic output is injected into the BN nodes as described in Section 3.1. This triggers an inference process

that progressively modifies the posterior probabilities of all connected nodes in the network using message passing belief propagation. When the process is completed the posterior probability of the root node modeling the high-level concept *car components ergonomic design* (represented with the *CA_ED* symbol in all figures), is compared against a fixed threshold. If the threshold is exceeded the detector decides positively, otherwise the document page is considered as not being relevant with the ergonomic design of car components. An illustration of this procedure for the cross-media case is depicted in Fig. 3.20. For measuring the efficiency of the high-level concept detector we have used precision versus recall curves. The threshold value of Fig. 3.20 is uniformly scaled between [0,1] for conducting the experiments in all cases.

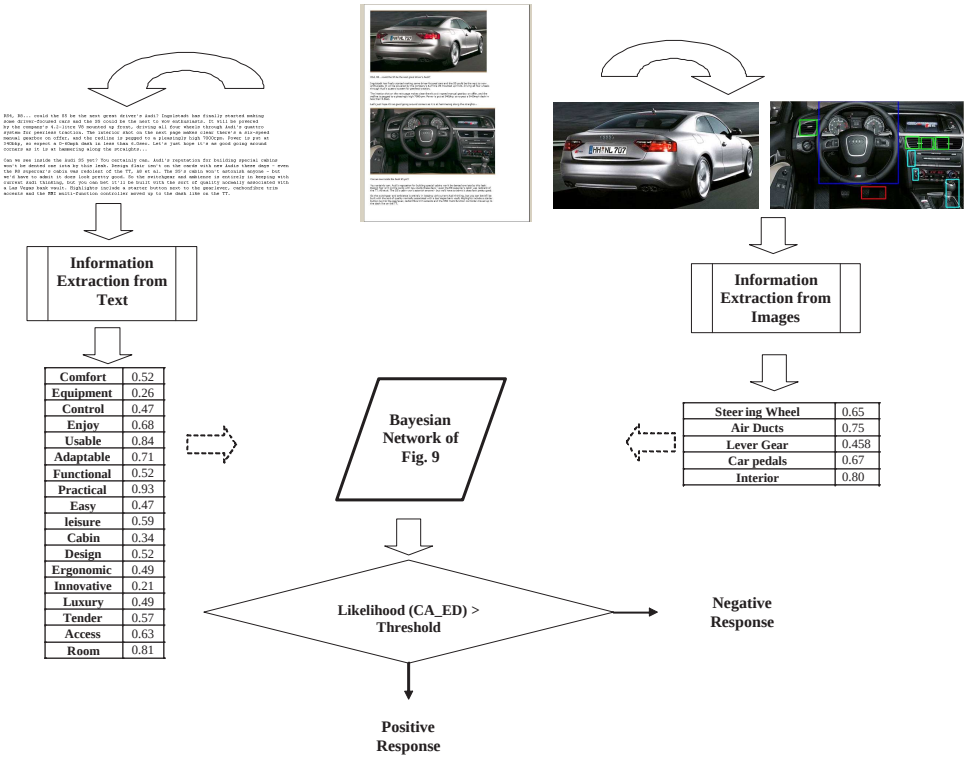
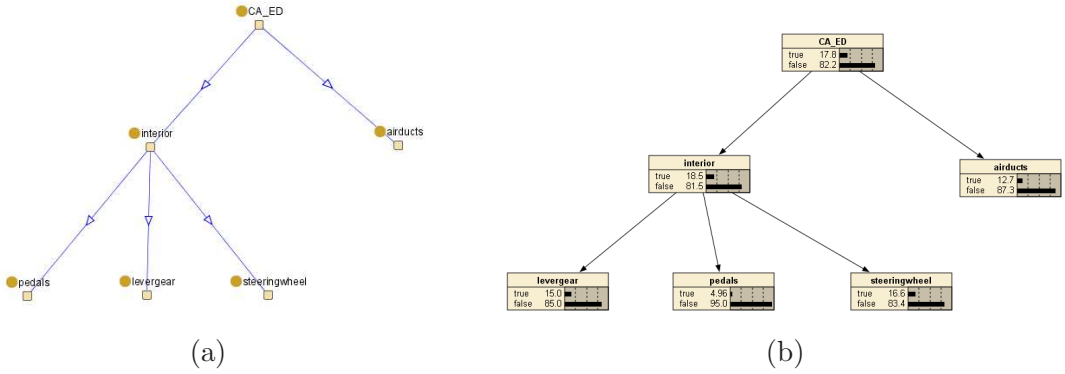


Figure 3.20: Inference process illustration for the cross media setting

Single vs Cross media analysis: In the case of visual-only analysis, the general knowledge about the specific domain was expressed by the ontology depicted in Fig. 3.21(a). This ontology associates five visual concepts, namely *air ducts, steering*

wheels, gear levers, car pedals and interior with the high-level concept *car components ergonomic design*. The trained BN used for this setting is depicted in Fig. 3.21(b). Five detectors trained to identify the five concepts of the domain ontology were implemented using the method of Section 3.4.1.2. These detectors were trained using an independent dataset of 3230 images depicting car interiors that was strongly annotated at region-detail. Each of these detectors was attached to the corresponding BN node of Fig. 3.21(b) and was used to trigger the process of probabilistic inference. By applying these five detectors on every image contained in a document page and using their output to instantiate the network nodes, we are able to decide about the existence of the high-level concept *car components ergonomic design* in a document page, based solely on the information depicted on the images of this page. The obtained results are depicted in Fig 3.26.



(a) (b)

Figure 3.21: Experimental setting using only visual concepts, a) Domain ontology for document analysis using only visual evidence, b) Bayesian Network for visual analysis

In the case of textual-only analysis, we used eighteen different concepts, namely *access, cabin, design, leisure, comfort, easy, enjoy, luxury, room, tender, ergonomic, equipment, innovative, usable, practical, functional, adaptable and control* for obtaining the textual evidence. Using these eighteen concepts we constructed the ontology of Fig. 3.22 that encodes the associations between the textual concepts and the high-level concept of *car component ergonomic design*. The trained BN used in this setting is depicted in Fig. 3.23. The confidence degrees that are used to instantiate the BN nodes are obtained by applying the textual analysis method described in Section 3.4.1.2 for

each document page and using the above linguistic descriptions as query words. As in the previous case this setting allows us to decide about the existence of the high-level concept *car components ergonomic design* in a document page, based solely on the information included in the textual descriptions of this page. The precision versus recall curve obtained from textual-only analysis is depicted in Fig 3.26.

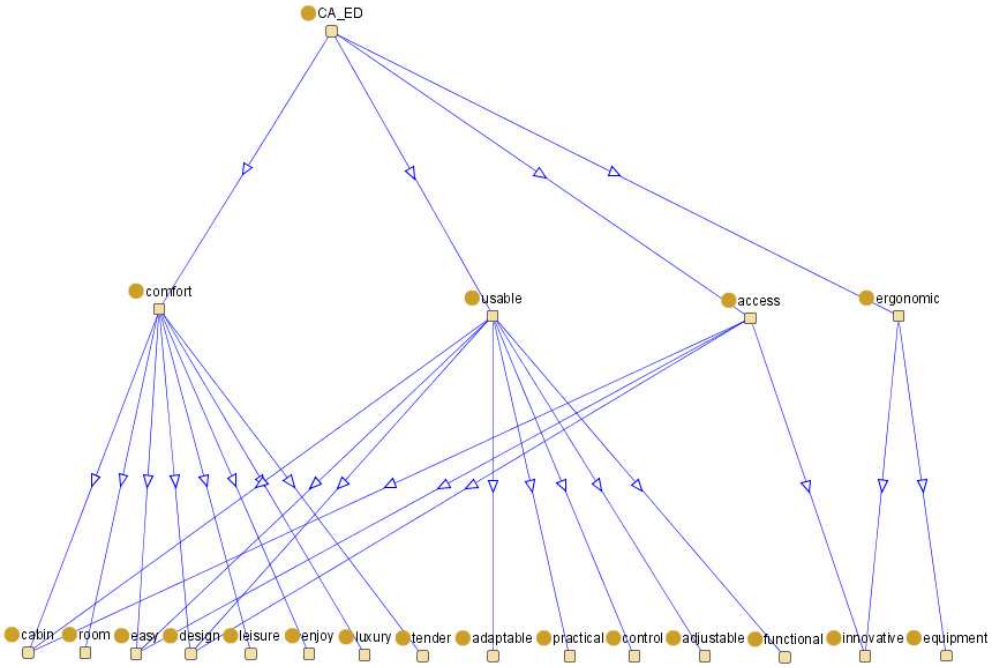


Figure 3.22: Domain ontology for document analysis using only textual concepts

For the case of cross-media analysis, both textual and visual concepts were used for the construction of the ontology depicted in Fig. 3.24. This ontology expresses the domain knowledge across media and reflects the cross-relations between textual and visual concepts. The trained BN that was used for performing inference in this setting is depicted in Fig. 3.25. The confidence degrees obtained by applying the aforementioned textual and visual single-medium analyzers on the constituent parts of a document page, are used to instantiate the BN nodes and perform inference using evidence across media. The results achieved by the high level concept detector in this setting are depicted in Fig. 3.26.

It is clear from the comparative diagram of Fig. 3.26 that the configuration of the

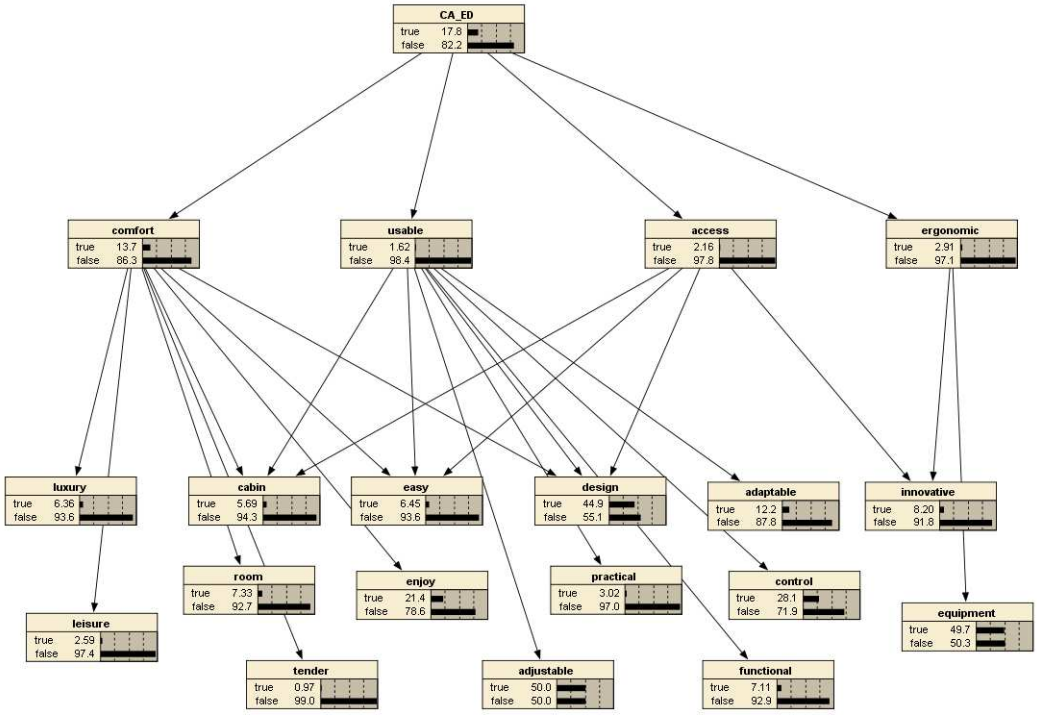


Figure 3.23: Bayesian Network for textual-only analysis

framework using evidence across media, outperforms the cases where evidence originates exclusively from one media type. We can see that textual analysis performs significantly better than visual analysis mainly due to the increased amount of evidence that has been used in this setting. However, when the textual and visual evidence are combined in the cross media setting, the high-level concept detector manages to further improve its efficiency for most of the applied threshold values. This outcome verifies that there are many cases where the evidence existing across different media types carry complementary information, which can only be translated into facts when considered in a synergetic fashion.

Generative vs Discriminative model: The second goal of our experimental study was to investigate the superiority of generative models like BNs over discriminative models like Support Vector Machines (SVMs) [141], to more efficiently incorporate and benefit from explicit knowledge. The motive behind using BNs in our work was their ability to smoothly incorporate explicit knowledge through their parameters and

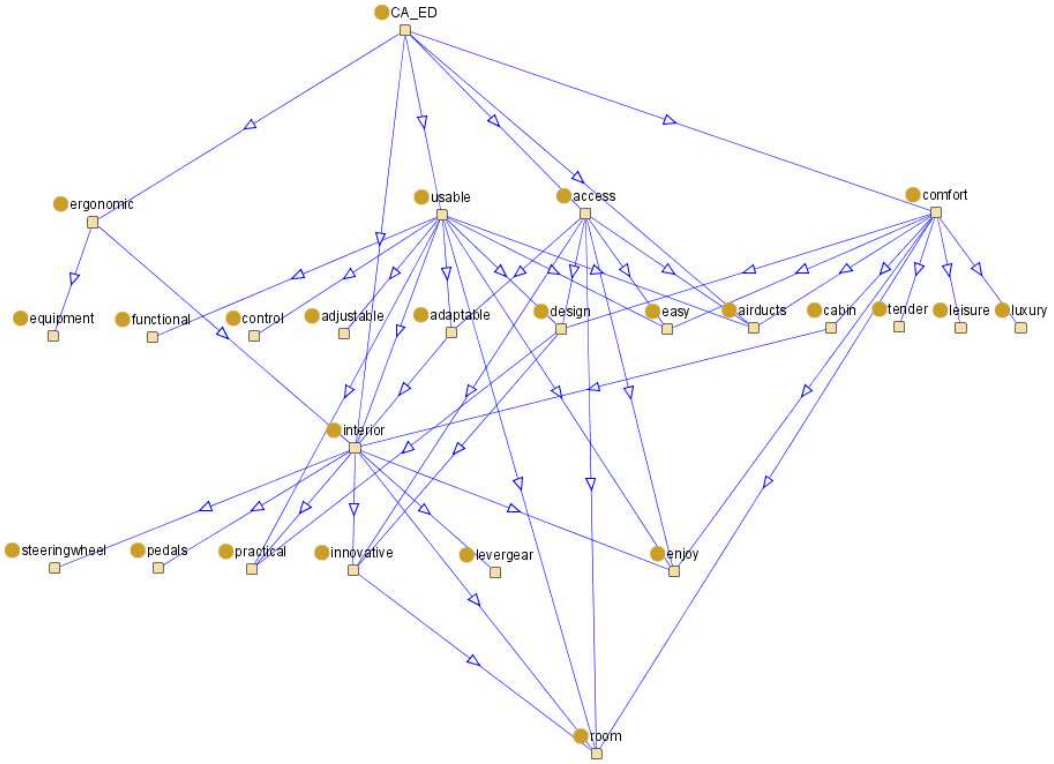


Figure 3.24: Domain ontology for document analysis using both visual and textual concepts

structure, as well as to learn efficient models from small training sets. This is in contrast to the approaches based on SVMs, since there is no straightforward way to incorporate explicit knowledge in these cases, as it can only be done at the level of the kernel. Moreover, when relying on SVMs, robust models can only be learned when there is a significant number of training samples available.

In order to verify the above, we compared our generative classifier based on BNs with a discriminative classifier implemented using SVMs. The feature space for training the SVM models was determined by concatenating the confidence degrees generated from the single-medium analyzers, resulting in a 23-dimensional feature vector for each document page. The SVMlight library [142] was employed for learning an one-class classifier recognizing the concept *car components ergonomic design*, using the same train/test split as in the case of BNs. A polynomial kernel function was used for learning the SVM models. Since the one class SVM models are known to be rather

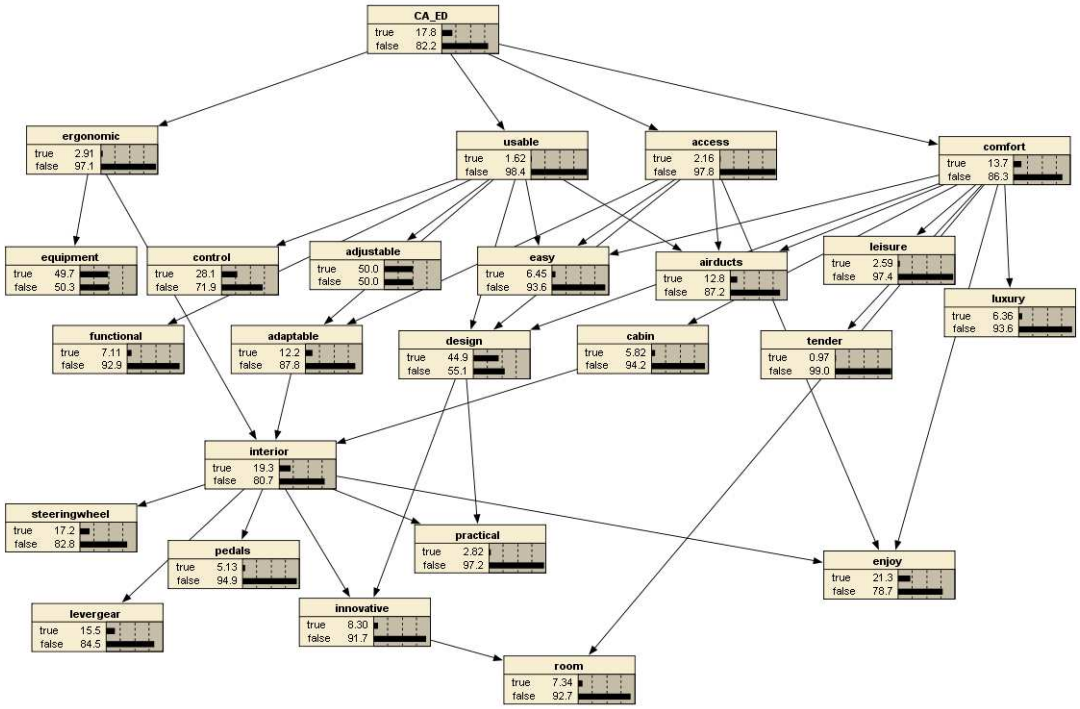


Figure 3.25: Bayesian Network for cross media analysis

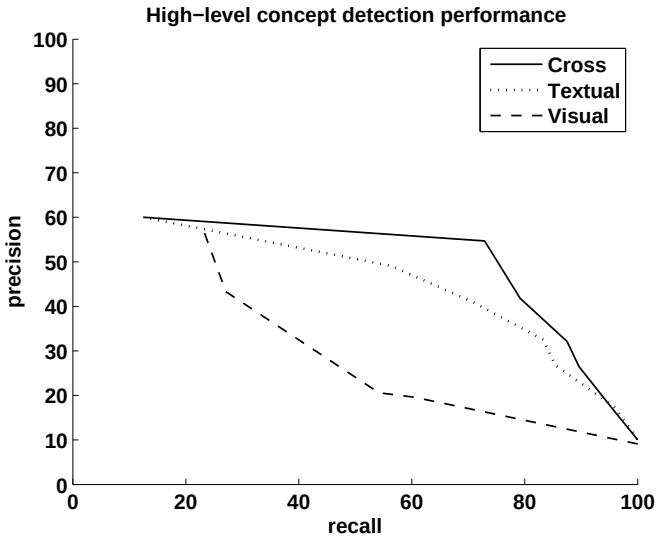
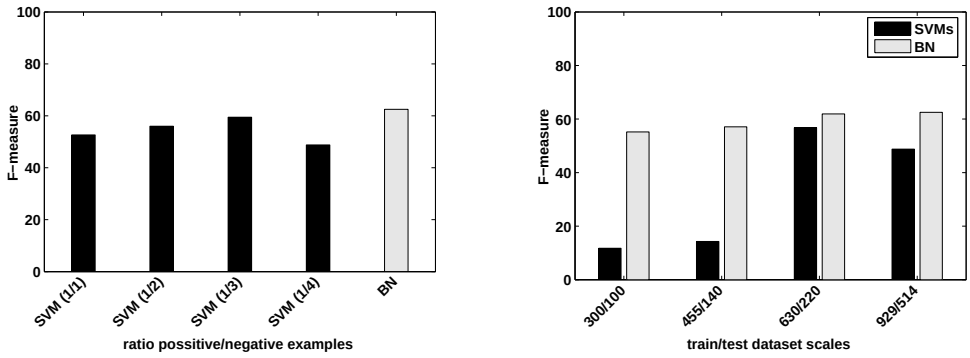


Figure 3.26: Cross vs single media analysis performance

sensitive on the ratio between positive and negative examples, we have tried 4 different ratios (i.e. 1/1, 1/2, 1/3 and 1/4) in order to optimally tune the classifier. Using the full train set the positive/negative ratio is approximately 1/4. The bar diagrams of Fig. 3.27(a) shows the F-measure scores achieved by the SVM-based classifiers using all four positive/negative ratios, as well as the score achieved by the BN classifier for the optimal threshold value. We can see that the BN classifier outperforms all SVM-based classifiers with the smallest improvement being $\approx 3\%$ (1/3 case) and the largest being $\approx 12\%$ (1/4 case).

Moreover, in order to verify that, in contrast to SVMs, BNs are able to learn efficient models even from just a few examples, we performed several experiments by reducing the number of samples included in the train/test datasets. Fig.3.27(b) shows the F-measure scores achieved using both approaches for four different scales of the train/test datasets. For this experiment the SVM-based classifiers were trained using all positive and negative samples included in each of the different dataset scales. It is clear that the models learned using BNs manage to deliver good performance even when trained with a particularly small number of samples. This is not the case for the models learned using SVMs, where the number of training samples needs to grow approximately 600 in order to deliver good results. All findings of these experiments verify the superiority of generative models in more efficiently handling prior knowledge and learning from a few examples. This attribute is particularly useful in cross media analysis since the cost of manual annotation in a cross media fashion is even higher from the single-medium cases.

Cases with missing or noisy domain knowledge: It is evident that our framework benefits from the existence of knowledge about the domain. However, there can be cases where such knowledge is either noisy or missing (i.e. the list of domain concepts is known but the relations between them are not). In such cases, our framework can be applied using either a trivial structure for the BN, or using a BN the structure of which is determined from sample data. In order to evaluate the performance of our framework when domain knowledge is noisy, we have considered the following two approaches for determining the structure of the BN. The first approach assumes the most trivial structure for the BN and initiates our framework using a naive BN. The naive BN is the simplest classifier based on Bayes' rule, it assumes that all variables



(a) (b)

Figure 3.27: a) Comparing generative with discriminative models using different ratios for the positive/negative examples b) Comparing generative with discriminative models using different scales for the train/test datasets

are independent from each other and all nodes are directly connected to the root node. The second approach is based on methods that are able to derive the structure of the BN from sample data. One such method is [143] where prior knowledge, provided in the form of a temporal BN called *prior network*, is combined with sample data in order to learn one or more BNs that are much closer to the actual structure of the domain than the initial *prior network*. A similar method is the well-established, score-based Cooper's K2 algorithm [140] which attempts to recover the underlying distribution of nodes in the form of a Directed Acyclic Graph (DAG), without making any assumptions about their structure. For the purposes of our work we have decided to employ the K2 algorithm in order to evaluate the performance of a BN, the structure of which is determined without using any prior information about the relations between the domain concepts.

More specifically, the K2 algorithm takes as input the number and ordering of nodes ($n = 24$ in our case), an upper bound for the parents of its node and a set of training data, which in our case correspond to the concept label annotations described in Section 3.4.2.1. The set of nodes includes the 23 visual and textual concepts as well as the high level concept *car components ergonomic design*. The ordering of the nodes was determined based on the frequency of appearance (in descending order)

of the corresponding concepts in the training data. In order to avoid networks with high complexity we have set the upper bound of parent nodes to be four. The BN generated using the K2 algorithm is depicted in Fig. 3.28. In Fig. 3.29, we compare the performance achieved by a BN constructed based on the cross media domain ontology of Fig 3.24, against the performance of a naive BN and the performance of a BN, the structure of which is determined using the K2 algorithm. In all cases the curves were drawn by modifying the threshold value between [0,1].

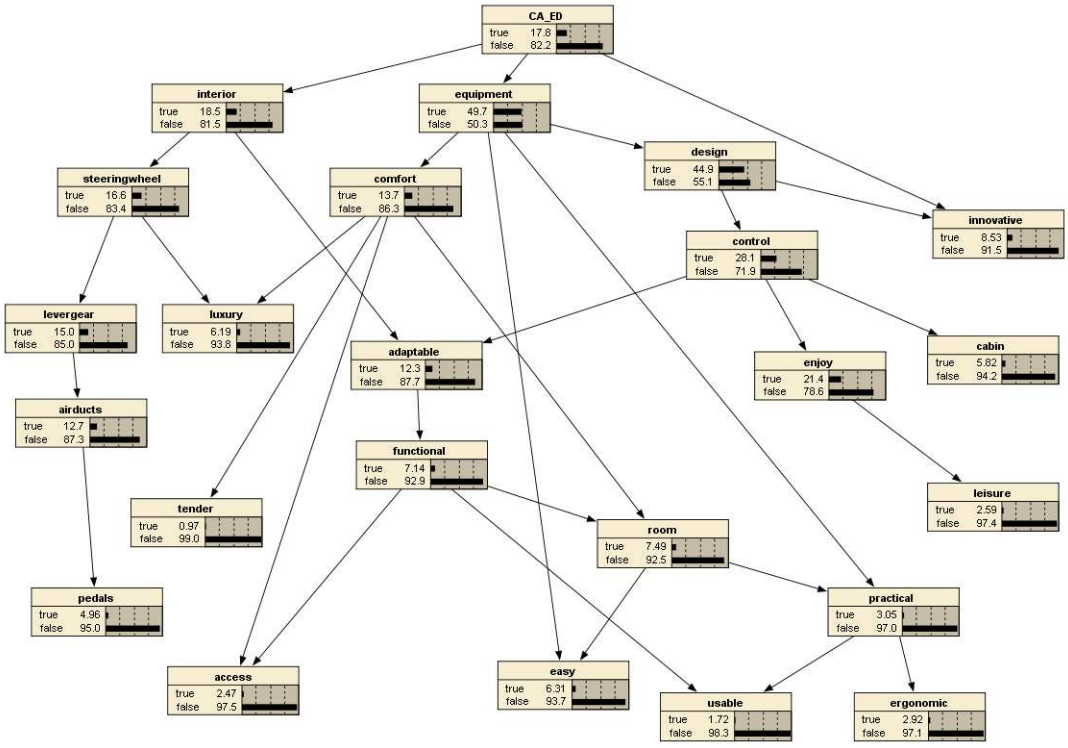


Figure 3.28: Bayesian Network derived from sample data using the K2 algorithm

It is clear from the results that the incorporation of explicit knowledge is particularly useful when combining information from heterogenous sources. We can see that the BN using the ontology, clearly outperforms the naive and K2 algorithm-based approaches. This is attributed to the fact that the domain ontology manages to capture the underlying cross-modal relations and boost the classification performance. Moreover, the fact that the naive BN approach achieves better results from K2, further advocates the need for incorporating explicit knowledge (even as a simple two level hierarchy) when

combining information from heterogeneous sources.

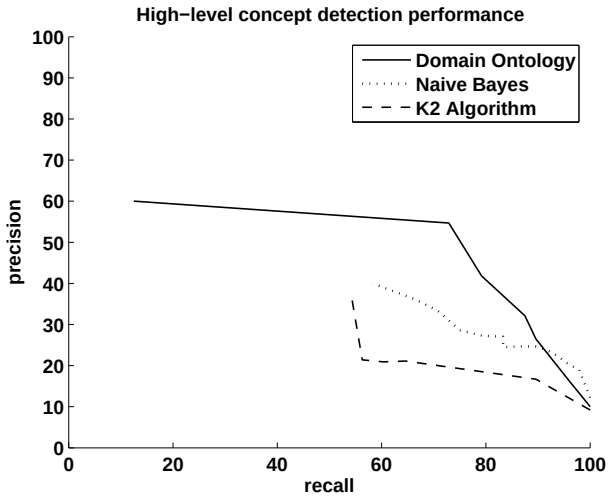


Figure 3.29: Comparative diagram for the different methods used to determine the BN structure

3.4.2.3 Video shot classification

In order to verify the efficiency of our framework to handle more general applications, we have developed an ontology-based classifier for video shots. For building and evaluating this classifier we have relied on the TRECVID2010 development dataset IACC.1.tv10.training[1], that has been provided by TRECVID organizers to facilitate training in various different tasks of 2010 competition. The dataset is composed of 118581 shots annotated with 130 concepts[2]. The reason for choosing this dataset over the datasets used in the previous years, was that 2010 was the first year where the organizers provided an ontology with the relations between 104 of the 130 available concepts. The availability of such an ontology is an important motivation for employing our approach, since the incorporation of domain knowledge in the analysis process is one of its great advantages. In order to facilitate training and testing we have split the 118581 shots to 59291 training T^{train} and 59290 testing T^{test} shots.

Engineering the ontology and building the BN: By examining the ontology relations provided with the dataset, we observed that there were 9 concepts, namely *Person,*

[1] http://www-nlpir.nist.gov/projects/tv2010/tv2010.html#IACC.1.tv10.training
[2] http://www-nlpir.nist.gov/projects/tv2010/TV10-concepts-130_UPDATED.xlsx

Outdoor, Indoor, Vegetation, Vehicle, Politics, Animal, Sports, Science_Technology, that acted as super-classes of all other concepts in the ontology. Based on this fact, and given that the goal of our approach is to infer the presence of a high-level concept by accumulating the effect of the existing evidence, we consider these 9 concepts to be the root concepts or our ontologies. Out of the remaining 95 concepts, 45 were chosen as textual concepts based on the availability of Automatic Speech Recognition (ASR) transcripts for a relatively high number of the shots annotated with these concepts. This selection strategy was motivated by the need to ensure that there will be sufficient textual information to extract evidence for the textual concepts. The remaining 50 concepts were considered as visual. When considering the textual-only or visual-only analysis case, the root concepts are only supported by the 45 textual or the 50 visual concepts, respectively. In the cross media analysis case all available concepts are used. The output of the multi-class video-shot classifier is a confidence degree for each of the 9 root concepts. Crisp decisions can be taken by applying a threshold on these confidence degrees. Having engineered the ontologies for the three analysis cases (i.e. textual-only, visual-only and cross-media), we used the methodology described in Sections 3.1 and 3.2 to construct the corresponding BNs. The CPTs were learned by applying the EM algorithm on the concept labels of the shots included in T^{train} and evidence-driven probabilistic inference was performed as described in Section 3.4.1.3.

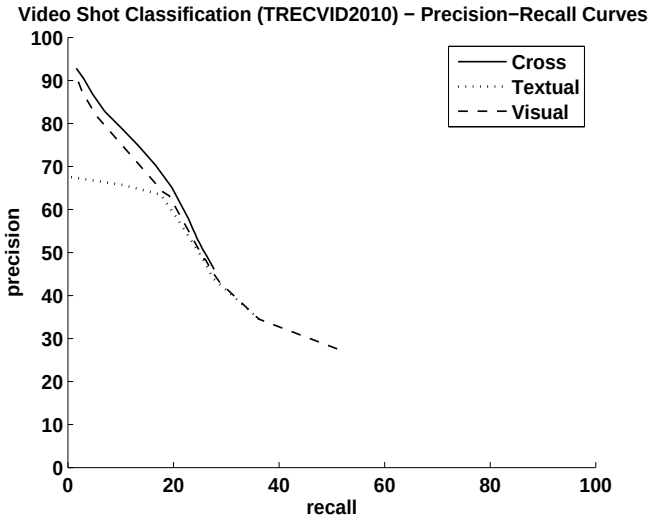
Modality synchronization: Each of the shots included in the TRECVID2010 development datasaet consists of its key-frame (i.e. an image) and the ASR transcripts of the spoken dialogs within the shot time-frame. In this case we consider that a conceptual relation exists between the key-frame and the ASR transcript of a shot. Thus, classification is performed for every shot by combining the visual and textual evidence extracted from the corresponding key-frame and ASR transcript, respectively.

Single-medium analysis: For extracting the likelihood estimates of the textual concepts we have employed the textual analysis approach described in Section 3.4.1.2. In this case, the values of semantic relatedness are estimated between the textual concept and every word included in the ASR transcript of the analyzed shot. By averaging the semantic relatedness values as described in Section 3.4.1.2 we obtain a likelihood estimate per textual concept, for each shot.

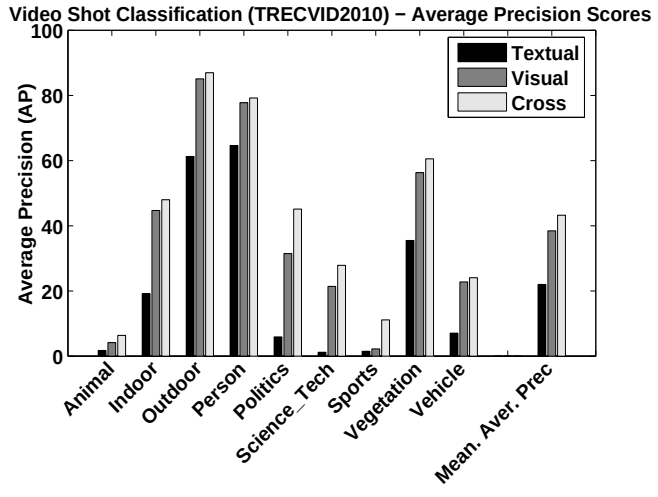
Due to the fact that the annotations provided by TRECVID are at the global level of the image and not at the level of regions, as required by the technique of Section 3.4.1.2, we have employed a different method for visual analysis. In this case, the visual representation of the images was extracted by applying the feature extraction technique described in [144]. More specifically, a set of interest points was detected in every image by applying the Harris-Laplace point detector on intensity channel [145]. For each of the identified interest points a 128-dimensional SIFT descriptor was computed using the version described by Lowe [146]. Then, a Visual Word Vocabulary (Codebook) [147] was created by using the K-Means algorithm to cluster in 500 clusters, approximately 3 million SIFT descriptors that were sub-sampled from a total amount of ≈200 million SIFT descriptors, extracted from ≈120 thousand training images. The Codebook allows the SIFT descriptors of all interest points to be vector quantized against the set of Visual Words and create a histogram of 500 dimensions. Finally, additional histograms were extracted from specific parts of the image. Using a 2x2 subdivision of the image, one histogram was extracted for each image quarter. Similarly, using a 1x3 subdivision consisting of three horizontal bars, one histogram was extracted for each bar. In the end all histograms were concatenated to form a 4000-dimensional visual representation of the image. After obtaining the visual representation of the images, Support Vector Machines (SVMs) [141] were used for generating the concept detection models. The 59291 key-frames included in T^{train} were used for training the concept detection models. Tuning arguments included the selection of Gaussian radial basis kernel and the use of cross validation for selecting the kernel parameters.

Video-shot classification results: The performance of our video-shot classifier was evaluated on T^{test}, for the cases of visual-only, textual-only and cross media analysis. In Fig. 3.30 we report results for the 9 root concepts. Fig. 3.30(a) depicts the precision-recall curves achieved by each analysis case. The curves are obtained by uniformly scaling the decision threshold between [0,1] and averaging between all root concepts. As expected the video-shot classifier incorporating evidence across media outperforms the classifiers that incorporate only textual or only visual information. In contrast to the analysis results on compound documents reported in Fig. 3.26, in this case the video-shot classifier based on visual analysis performs better than the classifier relying

on textual analysis. This can be attributed to the low quality of ASR transcripts or the complete absence of transcripts for a non-negligible amount of shots.



(a)



(b)

Figure 3.30: Cross vs single media analysis performance using TRECVID2010 dataset a) Precision-recall curves obtained by uniformly scaling the decision threshold between [0,1] and averaging between all root concepts, b) Average precision scores for the 9 root concepts

In Fig. 3.30(b) we report the Average Precision (AP) scores for the 9 root concepts, since this is the metric used by the TRECVID organizers. We can see that the improvement in performance achieved by the cross-media classifier is consistent across all root concepts and in certain cases by a significant amount, as in the case of *Sports*. Our experimental results show that the superiority of the cross media classifier over its single-medium counterparts is evident in all experimental settings, advocating the efficiency of our approach.

In order to compare our work with existing state-of-the art methods, we have relied on the evaluation results released by the organizers of TRECVID2010 for the task of *Semantic Indexing*. In the context of this task all submissions were evaluated for a set of 30 concepts[1], subset of the total set of 130 concepts. In order to facilitate the comparison of our work with the methods participated in the competition, we have employed a modified version of our video shot classifier. This version works in a similar way with the previous case, with the additional functionality that likelihood estimates are also given for the root nodes of the BN, providing useful evidence for the existence of their child nodes. In this way we manage to obtain inferred confidence degrees for 26 of the concepts that have been used for evaluation. No confidence degrees were obtained for the concepts *Doorway*, *Explosion Fire*, *Hand*, *Telephones*, since they were not included in the ontology provided by the organizers. Fig. 3.31 compares the Average Precision achieved by our framework against the top-scoring run and the average performance among all 101 runs, submitted for the *Semantic Indexing* task [148].

It is important to note that the performance figures depicted in Fig. 3.31 are not directly comparable due to the following reasons. The dataset used for training and testing are not identical, since we have trained our classifier using half portion of the development dataset and evaluated its performance using the other half. On the contrary, the methods submitted for the *Semantic Indexing* competition used the full development dataset for training and evaluated their performance using an independent test set. Moreover, the performance scores provided by the organizers refer to the

[1] Airplane flying, Animal, Asian_People, Bicycling, Boat-ship, Bus, Car_Racing, Cheering, Cityscape, Classroom, Dancing, Dark-skinned_People, Demo or protest, Doorway, Explosion_Fire, Female-Human-Face-Closeup, Flowers, Ground_Vehicles Hand, Mountain, Nighttime, Old_People, Running, Singing, Sitting_Down, Swimming, Telephones, Throwing, Vehicle, Walking

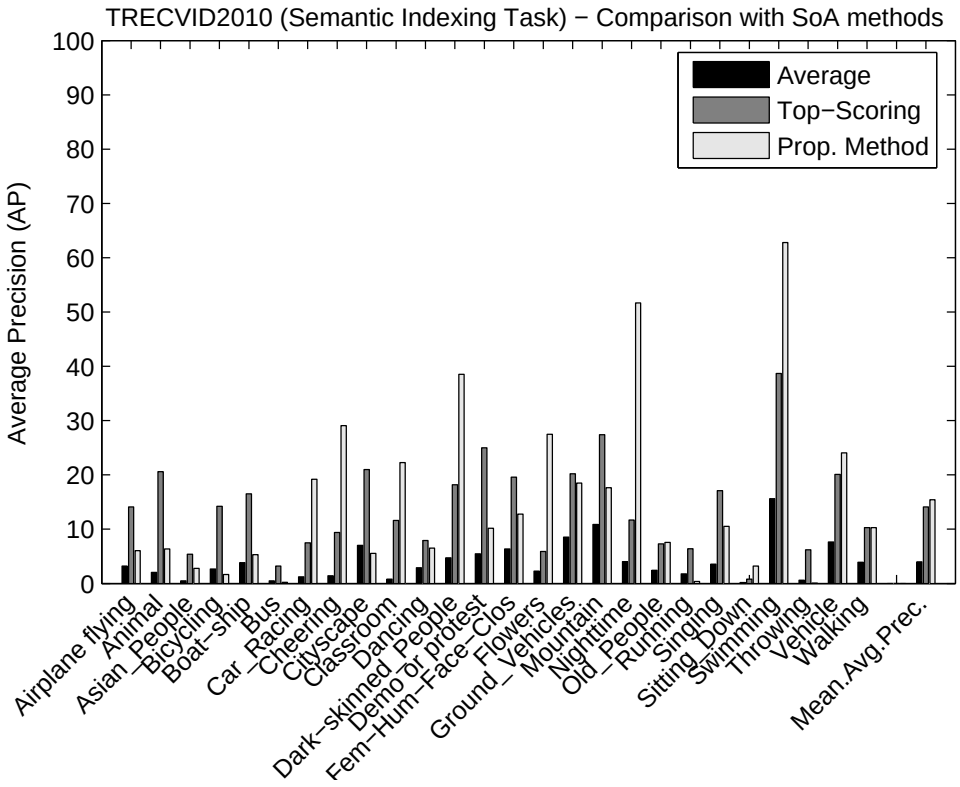


Figure 3.31: Comparison of our framework for 26 concepts against the top-scoring method and the average performance among all 101 runs, submitted for TRECVID2010 *Semantic Indexing* task.

Inferred Average Precision [149] which is an approximation of Average Precision when the available annotations are incomplete. The figures provided for our framework refer to Average Precision since we had complete annotations for our test set. Despite the above, it is clear that our method compares favorably with the performance achieved by the state-of-the-art methods. Among the 26 evaluated concepts our method outperforms the top-scoring methods in 11 and surpass the average performance score in 21 cases. The Mean Average Precision achieved by our framework (15.4%) is improved by 1.3% units compared to the Mean Average Precision of the top-scoring methods (14.1%) and by 11.4% units compared to the average performance scores (4%).

3.5 Discussion of our experimental findings

Both case studies have verified the effectiveness of the proposed probabilistic inference framework in improving the performance of concept detectors by using their output as evidence. We have seen how domain knowledge and application context act beneficially in media interpretation by favoring the co-occurrence of evidence that are known from experience to co-exist. Given that in all examined cases the improvement in performance derives mainly from the incorporation of knowledge and context to the analysis process, we may rightfully claim that the proposed framework can be used to improve the performance of any set of concept detectors that produce a probabilistic output.

Going a bit deeper in Section 3.3 particularly interesting have been the results of Section 3.3.3.1, which led us to the conclusion that the amount and nature of the semantic information that can be used to enhance image interpretation depends largely on the special characteristics of the domain. More specifically, although using the information from the knowledge structure K_D and the causality relations $W_{ij} \in X$ obtained from context was proven to be useful in all cases, the semantic constraints originating from the domain were only able to facilitate image interpretation when the imposed rules were sufficiently concrete. For instance, the disjointness between "Tennis" and all other category concepts of the *PS* domain expresses a rather strict distinction that is suggested by knowledge. On the contrary, attempts to incorporate semantic constraints that, although valid from the point of logic, were less strict from the visual inference point of view didn't result in performance improvements. Another interesting experimental finding was observed in Section 3.3.3.4, showing that a sufficiently large amount of training data is required for approximating the prior and conditional probabilities using frequency information. Indeed, it was evident that the availability of realistic prior and conditional probabilities for the BN nodes is particularly important for the efficiency of our framework. Learning them from data was only possible when there were enough training samples to learn from. However, given that the manual annotation of images is a cumbersome procedure, especially at region level, this option is not always feasible. In such cases a potential solution to the problem could be to mine the necessary annotations from social sites like Flickr that are being populated with hundreds of user tagged images on a daily basis. This was actually our basic motivation for developing the framework for scalable object detection, presented in Chapter 4.

Similar conclusions have also been drawn from the experimental study presented in Section 3.4, that was conducted with the goal to examine whether the proposed probabilistic inference framework can be used to effectively combine evidence extracted across media. Our experimental findings in Section 3.4.2.2 have indeed verified that there are many cases where the high-level concepts contained in a multi-modal resource can only be extracted if evidence is considered across media. Moreover, it has been shown in Section 3.4.2.2 that the information coming from the domain knowledge is particularly useful when dealing with heterogeneous types of content, even if provided in a very simplistic and rough form. Interesting were also the results of Section 3.4.2.2 showing that when performing cross media analysis, the generative models are more suitable for incorporating explicit knowledge and outperform the discriminative models that lack a straightforward way to benefit from such knowledge. Finally, a drawback related to the amount of annotation effort was also revealed in this case study posing the requirement to deeply model the analysis context, both in terms of engineering the domain ontology and producing the necessary cross media annotations. This is a critical requirement that makes our framework appropriate for cases where this effort is justified by the added value in the application, or in cases where social media can be used to mine the necessary annotations.

3. COMBINING IMPLICIT AND EXPLICIT KNOWLEDGE FOR MEDIA INTERPRETATION

Chapter 4

Scalable object detection by leveraging social media

In this chapter we present an approach that leverages social media for the effortless learning of object detectors [150]. We are motivated by the fact that the increased training cost of methods demanding manual annotation, limits their ability to easily scale in different types of objects and domains. At the same time, the rapidly growing social media applications have made available a tremendous volume of tagged images, which could serve as a solution for this problem [151]. However, the nature of annotations (i.e. global level) and the noise existing in the associated information (due to lack of structure, ambiguity, redundancy and emotional tagging), prevents them from being readily compatible (i.e. accurate region level annotations) with the existing methods for training object detectors. We overcome this deficiency by using the collective knowledge aggregated in social sites to automatically determine a set of image regions that can be associated with a certain object [152], [153], [154].

4.1 Description of the proposed approach

Machine learning algorithms for object detection fall within two main categories that are characterized by the annotation granularity of their learning samples. The algorithms that are designed to learn from strongly annotated samples [155], [156], [157] (i.e. samples in which we know the exact location of an object within an image) and the algorithms that learn from weakly annotated samples [158], [11], [9], [66] (i.e. samples in

which we known that an object is depicted in the image, but its location is unknown). In the first case, the goal is to learn a mapping from visual features f_i to semantic labels c_i (e.g. a face [155], [157] or a car [156]) given a training set made of pairs (f_i, c_i). New images are annotated by using the learned mapping to derive the semantic labels that correspond to the visual features of the new image. On the other hand, in the case of weakly annotated training samples the goal is to estimate the joint probability distribution between the visual features f_i and the semantic labels c_i given a training set made of pairs between sets $\{(f_1, \ldots, f_n), (c_1, \ldots, c_m)\}$. New images are annotated by choosing the semantic labels that maximize the learned joint probability distribution given the visual features of the new image. Some indicative works that fall within the weakly supervised framework include the ones relying on aspect models like probabilistic Latent Semantic Analysis (pLSA) [158], [159] and Latent Dirichlet Allocation (LDA) [160], [161], which are typically used for estimating the necessary joint probability distribution.

While model parameters can be estimated more efficiently from strongly annotated samples, such samples are very expensive to obtain raising scalability problems. On the contrary, weakly annotated samples can be easily obtained in large quantities from social networks but the estimation of model parameters is far more difficult. Motivated by this fact our work aims at combining the advantages of both strongly supervised (learn model parameters more efficiently) and weakly supervised (learn from samples obtained at low cost) methods, by allowing the strongly supervised methods to learn from training samples that can be mined from collaborative tagging environments. The problem we consider is essentially a multiple-instance learning problem in noisy context, where we try to exploit the noise reduction properties that characterize massive user contributions, given that they encode the collective knowledge of multiple users. Indeed, Flickr hosts a series of implicit links between images that can be mined using criteria such as geo-location information, temporal proximity between the image timestamps, or images associated with the same event. The goal of this work is to exploit the social aspect of the contributed content at the level of tags. More specifically, given that in social tagging environments the generated annotations may be considered to be the result of the collaboration among individuals, we can reasonably expect that tag assignments are filtered by the collaborative effort of the users, yielding more consistent annotations. In this context, drawing from a large pool of weakly annotated images,

our goal is to benefit from the knowledge aggregated in social tagging systems in order to automatically determine a set of image regions that can be associated with a certain object.

In order to achieve this goal, we consider that if the set of weakly annotated images is properly selected, the most populated tag-"term" and the most populated visual-"term" will be two different representations (i.e. textual and visual) of the same object. We define tag-"terms" to be sets of tag instances grouped based on their semantic affinity (e.g. synonyms, derivatives, etc.). Respectively, we define visual-"terms" to be sets of region instances grouped based on their visual similarity (e.g. clustering using the regions' visual features). The most populated tag-"term" (i.e. the most frequently appearing tag, counting also its synonyms, derivatives, etc.) is used to provide the semantic label of the object that the developed classifier is trained to recognize, while the most populated visual-"term" (i.e. the most populated cluster of image regions) is used to provide the set of positive samples for training the classifier in a strongly supervised manner. Our method relies on the fact that due to the common background that most users share, the majority of them tend to contribute relevant tags when faced with similar types of visual content [162]. Given this fact, it is expected that as the pool of the weakly annotated images grows, the most frequently appearing "term" in both tag and visual information space will converge into the same object.

In the following we describe the general architecture of our approach and provide technical details for the independent analysis components. Subsequently, we provide some theoretical insight on the convergence properties of our approach and present the experimental findings that are used to support our claims.

4.2 Architecture and Components Description

4.2.1 General Architecture

The approach we propose for leveraging social media to train object detection models is depicted in Fig. 4.1. The analysis components that we can identify are: a) construction of an appropriate image set, b) image segmentation, c) extraction of visual features from image regions, d) clustering of regions using their visual features, and e) supervised learning of object recognition models using strongly annotated samples.

More specifically, given an object c_k that we wish to train a detector for (e.g. *sky* in Fig. 4.1), our approach starts from a large collection of user tagged images and performs the following actions. Images are appropriately selected so as to formulate a set of images that emphasizes on object c_k. By emphasizing we refer to the case where the majority of the images within the image set depict a certain object and that the linguistic description of that object can be obtained from the most frequently appearing tag (see Section 4.2.2.1 for more details). Subsequently, clustering is performed on all regions extracted from the images of the image set, that have been pre-segmented using an automatic segmentation algorithm. During region clustering the image regions are represented by their visual features and each of the generated clusters typically contains visually similar regions. Since the majority of the images within the selected image set depicts instances of the desired object c_k, we anticipate that the majority of regions representing the object of interest will be gathered in the most populated cluster, pushing all irrelevant regions to the other clusters. Eventually, we use as positive samples the visual features extracted from the regions belonging to the most populated cluster, to train in a supervised manner an SVM-based binary classifier for recognizing instances of c_k. After training the classifier, object detection is performed on unseen images by using the automatic segmentation algorithm to extract their regions and apply the classifier to decide whether these regions depict c_k.

4.2.2 Analysis Components

We use the notation of Table 4.1 to provide technical details, formalize the functionality and describe the links between the components employed by our framework.

4.2.2.1 Construction of an appropriate image set

In this section we refer to the techniques that we use in order to construct a set of images emphasizing on object c_k, based on the associated textual information (i.e. annotations). If we define $ling(c_k)$ to be the linguistic description of c_k (e.g. the words "sky", "heaven", "atmosphere" for the object sky), a function describing the functionality of this component takes as input a large set of images and $ling(c_k)$, and returns a set of images S^{c_k}, subset of the initial set, that emphasizes on object c_k.

$$imageSet(S, ling(c_k)) = S^{c_k} \subset S \tag{4.1}$$

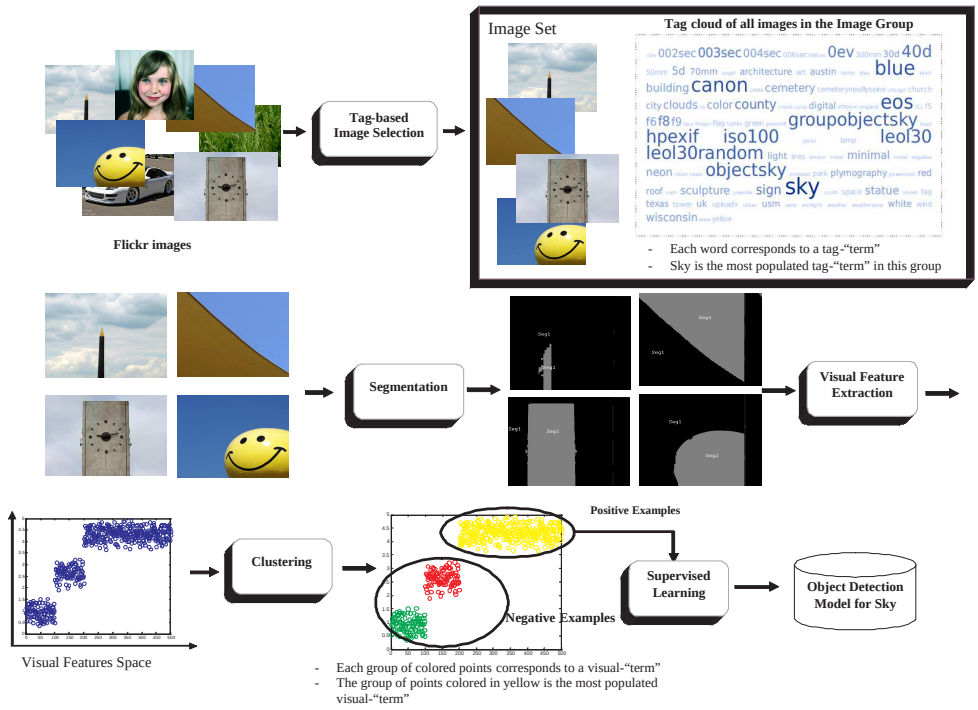


Figure 4.1: Proposed framework for leveraging a set of user tagged images to train a model for detecting the object *sky*.

For the purposes of our work we use three different implementations of this function based on the type of associated annotations.

Keyword-based selection: This approach is used for selecting images from strongly annotated datasets. These datasets are hand-labeled at region detail and the labels provided by the annotators can be considered to be mostly accurate and free of ambiguity. Thus, in order to create S^{c_k} we only need to select the images where at least one of its regions is labeled with $ling(c_k)$.

Flickr groups: *Flickr groups*[1] are virtual places hosted in collaborative tagging environments that allow social users to share content on a certain topic, which can be also an object. Although managing *flickr groups* still involves some type of human an-

[1] http://www.flickr.com/groups/

Table 4.1: Legend of used notation

Symbol	Definition
S	The complete social media dataset
N	The number of images in S
S^{c_k}	An image set, subset of S that emphasizes on object c_k
n	The number of images in S^{c_k}
I	An image from S
$R = \{r_i, i = 1, \ldots, m\}$	Complete set of regions identified in all images of S^{c_k} by an automatic segmentation algorithm
$T = \{t_i, i = 1, \ldots, n\}$	Complete set of tags contributed for all images of S^{c_k} by web users
$F = \{f_(r_i), i = 1, \ldots, m\}$	Complete set of visual features extracted from all regions in R
$C = \{c_i, i = 1, \ldots, t\}$	Set of distinct objects that appear in the image set S^{c_k}
$\mathbf{R} = \{\mathbf{r_i}, i = 1, \ldots, o\}$	Set of clusters created by performing clustering on the regions extracted from all images of S^{c_k} based on their visual similarity (i.e. visual-terms)
$\mathbf{T} = \{\mathbf{t_j}, j = 1, \ldots, d\}$	Set of clusters created by clustering together the tags contributed for all images in S^{c_k}, based on their semantic affinity (i.e. tag-terms)
p_{c_i}	Probability that tag-based image selection draws from S an image depicting c_i
TC_i	Number of regions depicting object c_i in S^{c_k}

*we use normal letters (e.g. z) to indicate individuals of some population and bold face letters (e.g. $\mathbf{z}$) to indicate clusters of individuals of the same population

notation (i.e. a human assigns an image to a specific *flickr group*) it can be considered weaker than the previous case since this type of annotation does not provide any information about the boundaries of the object depicted in the image. From here on we will refer to the images obtained from *flickr groups* as roughly-annotated images. In this case, S^{c_k} is created by taking a predefined number of images from a *flickr group* that is titled with $ling(c_k)$. Here, the tags of the images are not used as selection criteria. One drawback of *flickr groups* derives from the fact that since they are essentially virtual places they are not guaranteed to constantly increase their size and therefore cater for datasets of arbitrary scale. Indeed, the total number of positive samples that can be

extracted from the images of a *flickr group* has an upper limit on the total number of images that have been included in this group by the users, which is typically much smaller than the total number of flickr images that actually depict this object. This is the reason that we also investigate the following selection technique that operates on image tags and is therefore capable of producing considerably larger sets of images emphasizing on a certain object.

SEMSOC: SEMSOC stands for SEmantic, SOcial and Content-based clustering and is applied on weakly annotated images in order to create sets of images emphasizing on different topics. SEMSOC was introduced by Giannakidou et. al. in [163] and is an un-supervised model for the efficient and scalable mining of multimedia social-related data that jointly considers social and semantic features. Given the tendency of social tagging systems to formulate knowledge patterns that reflect the way content is perceived by the web users [162], SEMSOC aims at identifying these patterns and creating an image set emphasizing on c_k. The reason for adopting this approach is to overcome the limitations that characterize collaborative tagging systems such as tag spamming, tag ambiguity, tag synonymy and granularity variation (i.e. different description level). The outcome of applying SEMSOC on a large set of images S, is a number of image sets $S^{c_i} \subset S, \quad i = 1, \ldots, m$, where m is the number of created sets. This number is determined empirically, as described in [163]. Then in order to obtain the image set S^{c_k} that emphasizes on object c_k, we select the SEMSOC-generated set S^{c_i} where its most frequent tag closely relates with $ling(c_k)$. Although the image sets generated by SEMSOC are not of the same quality as those obtained from *flickr groups*, they can be significantly larger favoring the convergence between the most populated visual- and tag-"term". In this case, the total number of positive samples that can be obtained is only limited by the total number of images that have been uploaded on the entire flickr repository and depict the object of interest. Moreover, since SEMSOC considers also the social and semantic features of tags when creating the sets of images, the resulting sets are expected to be of higher semantic coherence than the sets created using for instance, a straightforward tag-based search. Fig. 4.2 shows four examples of image clusters generated by SEMSOC along with the corresponding most frequent tag.

(a) Vegetation (b) Sky

(c) Sea (d) Person

Figure 4.2: Examples of image sets generated using SEMSOC (in caption the corresponding most frequent tag). It is clear that the majority of images in each set include instances of the object that is linguistically described by the most frequent tag. The image is best viewed in color and with magnification.

4.2.2.2 Segmentation

Segmentation is applied on all images in S^{c_k} with the aim to extract the spatial masks of visually meaningful regions. In our work, we have used a K-means with connectivity constraint algorithm as described in [164]. The output of this algorithm, when applied to a single image, is a set of segments which roughly correspond to meaningful objects, as shown in Fig. 4.1. Thus, the segmentation analysis component takes as input the full set of images that are included in S^{c_k} and generates an extensive set of independent image regions:

$$segm(S^{c_k}) = \{r_i \in R : \forall I \in S^{c_k}\} \tag{4.2}$$

4.2.2.3 Visual Descriptors

In order to visually describe the segmented regions we have employed an approach similar to the one described in [144], with the important difference that in our case descriptors are extracted to represent each of the identified image regions, rather than the whole image. More specifically, for detecting interest points we have applied the Harris-Laplace point detector on intensity channel, which has shown good performance for object recognition [165]. In addition, we have also applied a dense-sampling approach where interest points are taken every 6^{th} pixel in the image. For each interest point (identified both using the Harris-Laplace and dense sampling) the 128-dimensional SIFT descriptor is computed using the version described by Lowe [166]. Then, a Visual Word Vocabulary (Codebook) is created by using the K-Means algorithm to cluster in 300 clusters, approximately 1 million SIFT descriptors that were sub-sampled from a total amount of 28 million SIFT descriptors extracted from 5 thousand training images. The Codebook allows the SIFT descriptors of all interest points enclosed by an image region, to be vector quantized against the set of Visual Words and create a histogram. Thus, a 300-dimensional feature vector $f(r_i)$ is extracted $\forall r_i \in R$, which contains information about the presence or absence of the Visual Words included in the Codebook. Then, all feature vectors are normalized so that the sum of all elements of each feature vector is equal to 1. Thus, the visual descriptors component takes as input the full

set of independent image regions R extracted from all images in S^{c_k} and generates an equivalent number of feature vectors.

$$vis(R) = \{f(r_i) \in F : \forall r_i \in R\} \tag{4.3}$$

4.2.2.4 Clustering

For performing feature-based region clustering we applied the affinity propagation clustering algorithm on all extracted feature vectors F. Affinity propagation was proposed by Frey and Dueck [167] and selected for our work due to the following reasons:

a) The requirements of our framework imply that in order to learn an efficient object detection model, clustering will have to be performed on a considerably large number of regions, making computational efficiency an important issue. In contrast to common clustering algorithms that start with an initial set of randomly selected centers and iteratively refine this set so as to decrease the sum of squared errors, affinity propagation simultaneously considers all data points as potential centers. By viewing each data point as a node in a network, affinity propagation recursively transmits real-valued messages along the edges of the network until a good set of centers and corresponding clusters emerges. In this way, it removes the need to re-run the algorithm with different initializations, which is very beneficial in terms of computational efficiency.

b) The fact that the number of objects depicted in the full set of images can not be known in advance, poses the requirement for the clustering procedure to automatically determine the appropriate number of clusters based on the analyzed data. Affinity propagation, rather than requiring that the number of clusters is pre-specified, takes as input a real number for each data point, called "preference". These "preference" values influence the number of identified clusters, which also emerges from the message-passing procedure. If a priori, all data points are equally suitable as centers (as in our case) the preferences should be set to a common value. This value can be varied to produce different numbers of clusters and taken for example to be the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters). The minimum value has been used in our experiments.

Thus, the clustering component takes as input the full set of feature vectors extracted by the visual descriptors component and generates clusters of feature vectors based on a similarity distance between those vectors. These clusters of feature vectors

can be directly translated to clusters of regions since there is one to one correspondence between regions and feature vectors. Thus, the functionality of the clustering component can be described as follows:

$$clust(F) = \{\mathbf{r}_i \in \mathbf{R}\} \tag{4.4}$$

Out of the generated clusters of regions we select the most populated $\mathbf{r}_v$, as described in detail in Section 4.3, and we use the regions included in this cluster to learn the parameters of a model recognizing c_k.

4.2.2.5 Learning Model Parameters

Support Vector Machines (SVMs) [141] were chosen for generating the object detection models due to their ability in smoothly generalizing and coping efficiently with high-dimensionality pattern recognition problems. All feature vectors corresponding to the regions assigned to the most populated $\mathbf{r}_v$ of the generated clusters, are used as positive samples for training a binary classifier. Negative examples are chosen arbitrarily from the remaining dataset. Tuning arguments include the selection of Gaussian radial basis kernel and the use of cross validation for selecting the kernel parameters. Thus, the functionality of the model learning component (m_{c_k}) can be described by the following function:

$$svm(vis(\mathbf{r}_v), c_k) = m_{c_k} \tag{4.5}$$

4.3 Theoretical grounding & intuitive analysis

4.3.1 Problem Formulation

The goal of our framework is to train an SVM-based binary classifier in order to recognize whether a region r_i of an un-seen image I depicts a certain object c_k. In order to do that, we need to provide the classifier with a set of positive and a set of negative samples (i.e. image regions) for c_k. Given that negative samples can be chosen arbitrarily from a random population, our main problem is to find a set of image regions depicting the object c_k, $(\mathbf{r}^+, c_k)$. The $+$ superscript indicate positive training samples. However, the annotations found in social networks are in the form of tagged images

$\{I, (t_1, t_2, \ldots, t_n)\}$, which can be transformed to $\{(r_1, r_2, \ldots, r_m)^I, (t_1, t_2, \ldots, t_n)^I\}$ after segmenting I into regions. Ideally, the tagged images could be used to extract the positive samples for c_k if we could perfectly cluster the visual and tag information space. More specifically, If we take R and T to be the total set of regions and tags extracted from all images in S respectively, by performing clustering based on the *similarity* between the individuals of the same population (i.e. visual similarity for image regions and semantic affinity for contributed tags), we are able to generate clusters of individuals in each population as shown below:

$$\begin{array}{rll} visualCluster(R) = \mathbf{r}_i, & \mathbf{r}_i \subseteq R & \text{visual-terms} \\ tagCluster(T) = \mathbf{t}_j, & \mathbf{t}_j \subseteq T & \text{tag-terms} \end{array} \tag{4.6}$$

Now, given a large set of tagged images $I \in S$ this process would produce for each object c_l depicted by the images of S, a triplet of the form $(\mathbf{r}_i, \mathbf{t}_j, c_l)$. Ideally in each triplet, $\mathbf{r}_i$ is the set of regions extracted from all images in S that depict c_l, and $\mathbf{t}_j$ is the set of tags from all images in S that were contributed to linguistically describe c_l. We consider that an object c_l may have many different instantiations in both visual (e.g. different angle, illumination, etc.) and tag (e.g. synonyms or derivatives of the words expressing the object; for instance the object sea can be linguistically described using many different words such as "sea", "seaside", "ocean", etc.) information space. Thus, $\mathbf{r}_i$ can be used to provide the positive samples required to train the SVM-based classifier, while $\mathbf{t}_i$ can be used to provide the linguistic description of the object that the classifier is trained to recognize. However, the aforementioned process can only be made feasible in the ideal case where the image analysis works perfectly and there is no noise in the contributed tags. This is highly unlikely due to the following reasons. From the perspective of visual analysis, in case of over or under segmentation, or in case the visual descriptors are inadequate to perfectly discriminate between different semantic objects, it is very likely that the clustering algorithm will create a different number of clusters than the actual number of semantic objects depicted by the images of S, or even mix regions depicting different objects into the same cluster. From the perspective of tag-analysis the well known problems of social networks (i.e. lack of structure, ambiguity, redundancy and emotional tagging) hinders the process of clustering together the tags contributed to refer to the same object.

For this reason, in our work, we relax the constraints of the aforementioned problem and instead of requiring that one triplet is extracted for every object c_l depicted by the

images of S, we only aim at extracting the triplet corresponding to the object c_k, which is the object emphasized by the processed image set. Thus, the first step is to create an appropriate set of images S^{c_k} that emphasizes on object c_k. Then, based on the assumption that there will be a connection between what is depicted by the majority of the images in S^{c_k} and what is described by the majority of the contributed tags, we investigate the level of semantic consistency (i.e. the level of which the majority of regions included in $\mathbf{r}_v$ depict c_k and the majority of tags included in $\mathbf{t}_g$ are linguistically related with c_k) of the triplet $(\mathbf{r}_v, \mathbf{t}_g, c_k)$, if v and g are selected as follows. Since both $\mathbf{r}_i$ and $\mathbf{t}_j$ are clusters (of images regions and tags, respectively), we can apply the $Pop(\cdot)$ function on them, that calculates the population of a cluster (i.e. the number of instances included in the cluster). Then v and g are selected such as the corresponding clusters are the most populated from all clusters generated by the clustering functions of eq. (4.6), that is $v = \arg\max_i(Pop(\mathbf{r}_i))$ and $g = \arg\max_j(Pop(\mathbf{t}_j))$.

Although the errors generated from imperfect visual analysis may have different causes (e.g. segmentation error, imperfect discrimination between objects), they all hinder the creation of semantically consistent region clusters. Therefore, in our work, we consider that the error generated from the inaccurate clustering of image regions with respect to the existing objects ($error_{cl-obj}$), incorporates all other types of visual analysis error. Similarly, although the contributed tags may incorporate different types of noise (i.e. ambiguity, redundancy, granularity variation, etc.) they all hinder the process of associating a tag with the objects that are depicted in the image, and thus is reflected on the level of emphasis that is given on object c_k when collecting S^{c_k}. Eventually, the problem addressed in this work is what should be the characteristics of S^{c_k} and $error_{cl-obj}$ so as the triplet $(\mathbf{r}_v, \mathbf{t}_g, c_k)$ determined as described above, to satisfy our objective (i.e. that the majority of regions included in $\mathbf{r}_v$ depicts c_k and the majority of tags included in $\mathbf{t}_g$ are linguistically related with c_k).

4.3.2 Image set construction

In order to investigate how the characteristics of the constructed image set S^c impact the success probability of our approach, we need to analytically express the association between the number of images included in S^c with the expected number of appearances of any object depicted by those images. Using image tag information to construct an image set that emphasizes on a certain object (e.g. c_1), can be viewed as the process

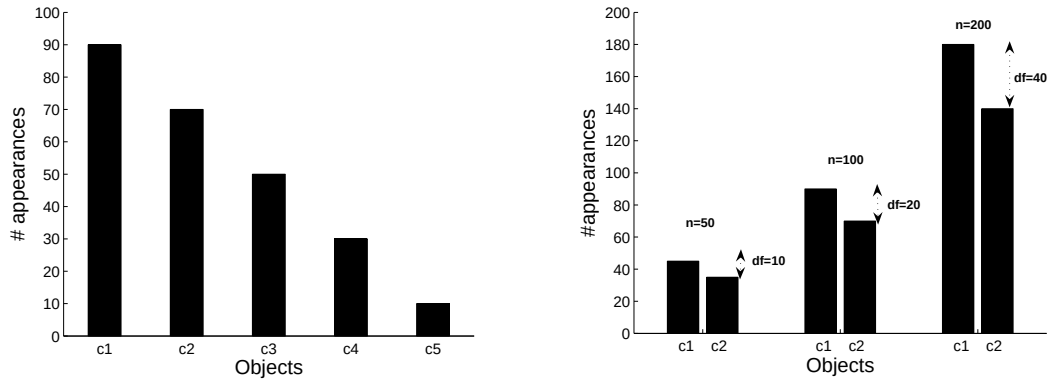
of selecting images from a large pool of weakly annotated images using as argument $ling(c_1)$ (i.e. the linguistic description of c_1 along with possible synonyms, derivatives, etc). Although misleading and ambiguous tags will hinder this process, the expectation is that as the number of selected images grows, there will be a connection between what is depicted in the majority of the selected images and what is described by the majority of the contributed tags. This can be formalized as follows. When one picks an image from a pool of weakly annotated images using $ling(c_1)$ as an argument, the probability that the selected image depicts c_1 is greater than the probability that the image depicts any other object.

Let us assume that we construct an image set $S^{c_1} \subset S$ that emphasizes on object c_1. What we are interested in is the frequency distribution of objects $c_i \in C$ appearing in S^{c_1} based on their frequency rank. We can view the process of constructing S^{c_1} as the act of populating an image set with images selected from a large dataset S using certain criteria. In this case, the number of times an image depicting object c_i appears in S^{c_1}, can be considered to be equal with the number of successes in a sequence of n independent success/failure trials, each one yielding success with probability p_{c_i}. Given that S is sufficiently large, drawing an image from this dataset can be considered as an independent trial. Thus, the number of images in S^{c_1} that depict object $c_i \in C$ can be expressed by a random variable K following the binomial distribution with probability p_{c_i}. Eq. (4.7) shows the probability mass function of a random variable following the binomial distribution:

$$Pr_{c_i}(K = k) = \binom{n}{k} p_{c_i}^k (1 - p_{c_i})^{n-k} \tag{4.7}$$

Given the above, we can use the expected value $E(K)$ of a random variable following the binomial distribution to estimate the expected number of images in S^{c_1} that depict object $c_i \in C$, if they are drawn from the initial dataset S with probability p_{c_i}. This is actually the value of k maximizing the corresponding probability mass function, which is:

$$E_{c_i}(K) = np_{c_i} \tag{4.8}$$



(a) (b)

Figure 4.3: a) Distribution of *#appearances* $\forall c_i \in C$ based on their frequency rank, for n=100 and p_{c_1}=0.9, $p_{c_2} = 0.7$, $p_{c_3} = 0.5$, $p_{c_4} = 0.3$, $p_{c_5} = 0.1$. b) Difference of *#appearances* between c_1, c_2, using fixed values for $p_{c_1} = 0.8$ and $p_{c_2} = 0.6$ and different values for n.

If we consider γ to be the average number of times an object appears in an image, then the number of appearances (*#appearances*) of an object in S^{c_1} is:

$$TC_i = \gamma n p_{c_i} \tag{4.9}$$

Moreover, based on the assumption mentioned earlier in this section, we accept that there will be an object c_1 that is drawn (i.e. appears in the selected image) with probability p_{c_1} higher than p_{c_2}, which is the probability that an image depicting c_2 is drawn, and so forth for the remaining $c_i \in C$. This assumption is experimentally verified in Section 4.4.1 where the frequency distribution of objects for different image sets are measured in a manually annotated dataset. Finally, using eq. (4.9) we can estimate the expected number of appearances (*#appearances*) of an object in S^{c_1}, $\forall c_i \in C$. Fig. 4.3(a) shows the *#appearances* $\forall c_i \in C$ against their frequency rank, given some example values for p_{c_i} with $p_{c_1} > p_{c_2} > \ldots$. It is clear from eq. (4.9) that if we consider the probabilities p_{c_i} to be fixed, the expected difference, in absolute terms, on the *#appearances* between the first and the second most highly ranked objects c_1 and c_2 increases as a linear function of n (see Fig. 4.3(b) for some examples). Based on this observation and given the fact that as N increases n will also increase, we examine how the population of the generated region clusters relates with $error_{cl-obj}$ and n.

4.3.3 Intuitive analysis

The purpose of this section is to help the reader derive some intuitive conclusions about the impact of the dataset size and the error introduced by the visual analysis algorithms $error_{cl-obj}$, on the success probability of our approach. In order to do this we examine clustering from the perspective of how much a possible solution deviates from the perfect case. This allows us to approximate $error_{cl-obj}$ with a measurable quantity and derive an analytical form of the association between the visual analysis error, the size of the dataset and an indicator of the success probability of our approach.

Given an image set S^{c_1} that emphasizes on object c_1 the goal of region clustering is to group together regions representing the same object. If perfect grouping is accomplished in a semantic sense, the distribution of clusters' population based on their population rank coincides with the distribution of objects' $\#appearances$ based on their frequency rank. In this case, the most populated cluster contains all regions depicting the most frequently appearing object. However, as the visual analysis techniques are expected to introduce error, we are interested on the connection between the $error_{cl-obj}$ and the population of the resulting clusters. Since there is no way to explicitly measure the $error_{cl-obj}$, we use the notation of Table 4.2 to approximate its effect on the population of the generated clusters.

Table 4.2: Notations for Clustering

Symbol	Definition
Pop_j	Population of cluster $\mathbf{r}_j$
$FP_{i,j}$	False positives of $\mathbf{r}_j$ with respect to c_i
$FN_{i,j}$	False negatives of $\mathbf{r}_j$ with respect to c_i
$DR_{i,j} = FP_{i,j} - FN_{i,j}$	Displacement of $\mathbf{r}_j$, with respect to c_i

Without loss of generality we work under the assumption that due to the $error_{cl-obj}$, it is more likely for the cluster corresponding to the second most frequently appearing object to become more populated than the cluster corresponding to the first most frequently appearing object, than any other cluster. A cluster that corresponds to an object c_i is considered to be the cluster that exhibits the highest F-measure (F_1) score, with respect to that object, among all generated clusters. Thus, the cluster

corresponding to object c_i is found using function Z, which is defined as:

$$Z(c_i, \mathbf{R}) = \mathbf{r}_\kappa, \quad \kappa = \arg\max_j(F_1(c_i, \mathbf{r}_j)) \tag{4.10}$$

where F_1 is the harmonic mean of precision (prec) and recall (rec) and is calculated using the following equation:

$$\begin{gathered} F_1(c_i, \mathbf{r}_j) = \frac{2prec_{i,j}rec_{i,j}}{prec_{i,j}+rec_{i,j}} \quad with \\ rec_{i,j} = \frac{TC_i - FN_{i,j}}{TC_i}, \quad prec_{i,j} = \frac{TC_i - FN_{i,j}}{Pop_j} \end{gathered} \tag{4.11}$$

Then, given that $\mathbf{r}_\kappa$ has been decided to be the corresponding cluster of c_i, the population Pop_κ of the cluster $\mathbf{r}_\kappa$ is equal to the number of regions TC_i depicting c_i, adding the number of false positives $FP_{i,\kappa}$ and removing the number of false negatives $FN_{i,\kappa}$ that have been generated from the $error_{cl-obj}$. Thus, we have:

$$\begin{gathered} Pop_\kappa = TC_i + FP_{i,\kappa} - FN_{i,\kappa} \Rightarrow \\ Pop_\kappa = TC_i + DR_{i,\kappa} \end{gathered} \tag{4.12}$$

$DR_{i,\kappa}$ is defined to be the displacement of $\mathbf{r}_k$ with respect to c_i and is an indicator of how much the content of $\mathbf{r}_k$ deviates from the perfect solution. $DR_{i,\kappa}$ shows how the Pop_κ of cluster $\mathbf{r}_\kappa$ is modified according to the $error_{cl-obj}$ introduced by the visual analysis algorithms. Positive values of $DR_{i,\kappa}$ indicate inflows in $\mathbf{r}_\kappa$ population, while negative values indicate leakages. In the typical case where the clustering result does not exhibit high values for $FP_{i,\kappa}$ and $FN_{i,\kappa}$ simultaneously (see Section 4.4.2), $DR_{i,\kappa}$ is also an indicator of result's quality since it shows how much the content of a cluster has been changed with respect to the perfect case. Let us denote $\mathbf{r}_\alpha = Z(c_1, \mathbf{R})$ and $\mathbf{r}_\beta = Z(c_2, \mathbf{R})$ the clusters corresponding to c_1 (i.e. the most frequently appearing object in S^{c_1}) and c_2 (i.e. the second most frequently appearing object in S^{c_1}), respectively. We are interested in the relation connecting Pop_α and Pop_β given $DR_{1,\alpha}$, $DR_{2,\beta}$. Thus we have:

$$\begin{gathered} Pop_\alpha - Pop_\beta = TC_1 + DR_{1,\alpha} - TC_2 - DR_{2,\beta} \Rightarrow \\ Pop_\alpha - Pop_\beta = (TC_1 - TC_2) + (DR_{1,\alpha} - DR_{2,\beta}) \end{gathered} \tag{4.13}$$

We know about the first parenthesis on the right hand side of the equation that since S^{c_1} emphasizes on c_1 this object will appear more frequently than any other object in

S^{c_1}, thus $TC_1 - TC_2 > 0$. In the case where the second parenthesis on the right hand side of the equation is also positive (i.e. $DR_{1,\alpha} - DR_{2,\beta} > 0$), the value $Pop_\alpha - Pop_\beta$ will be greater than zero since it is the sum of two positive numbers. This indicates that despite the $error_{cl-obj}$, cluster $\mathbf{r}_\alpha$ remains the most populated of the generated clusters and continues to be the most appropriate (i.e. in terms of the maximum F_1 criterion) cluster for training a model detecting object c_1. When $DR_{1,\alpha} - DR_{2,\beta} > 0$ we can distinguish between the three qualitative cases for clustering that are described in Table 4.3. The superscripts are used to indicate the sign (i.e. positive or negative) of the corresponding displacement in each case.

Table 4.3: Qualitative cases for clustering

$DR_{1,\alpha} - DR_{2,\beta} > 0$	$DR^+_{1,\alpha} > DR^+_{2,\beta}$	Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ increase their population but the inflows of $\mathbf{r}_\alpha$ are greater than the inflows of $\mathbf{r}_\beta$.
	$DR^+_{1,\alpha} \quad DR^-_{2,\beta}$	$\mathbf{r}_\alpha$ increases its population while $\mathbf{r}_\beta$ reduces its own.
	$DR^-_{1,\alpha} > DR^-_{2,\beta}$	Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ reduce their population but the leakages of $\mathbf{r}_\alpha$ are lesser than the leakages of $\mathbf{r}_\beta$.
$DR_{1,\alpha} - DR_{2,\beta} < 0$	$DR^+_{1,\alpha} < DR^+_{2,\beta}$	Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ increase their population but the inflows of $\mathbf{r}_\alpha$ are lesser than the inflows of $\mathbf{r}_\beta$.
	$DR^-_{1,\alpha} \quad DR^+_{2,\beta}$	$\mathbf{r}_\alpha$ reduces its population while $\mathbf{r}_\beta$ increases its own.
	$DR^-_{1,\alpha} < DR^-_{2,\beta}$	Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ reduce their population but the leakages of $\mathbf{r}_\alpha$ are greater than the leakages of $\mathbf{r}_\beta$.

*the superscripts indicate the sign (i.e. positive or negative) of the corresponding displacement

If $DR_{1,\alpha} - DR_{2,\beta} < 0$, the two parentheses of the right hand side of the eq. (4.13) have different signs and the sign of the value $Pop_\alpha - Pop_\beta$ depends on the difference between the absolute values of $|TC_1 - TC_2|$ and $|DR_{1,\alpha} - DR_{2,\beta}|$. In this case one of the

factors controlling whether the most populated cluster $\mathbf{r}_\alpha$ will be the most appropriate cluster for training a model detecting c_1, is the absolute difference between TC_1 and TC_2, which according to our analysis in Section 4.3.2 depends largely on the number of images n in S^{c_1}. The three qualitative cases for clustering that we can identify when $DR_{1,\alpha} - DR_{2,\beta} < 0$ are shown in Table 4.3.

In order to get an intuitive view of the relation between n and the probability of selecting the most appropriate cluster when $DR_{1,\alpha} - DR_{2,\beta} < 0$, we approximate the effect of $error_{cl-obj}$ on the distribution of the generated clusters' population by measuring how much a certain clustering solution deviates from the perfect solution. In order to do this, we view clustering as a recursive process with starting point the perfect solution. Then, the deviation of some clustering solution $t+1$ from the perfect solution depends on the deviation of the previous solution t from the perfect solution. Respectively, the population of a cluster in solution $t+1$ is equal to the population of this cluster in the previous solution t, adding the number of false positives and removing the number of false negatives that have been generated from the transition $t \rightarrow t+1$. This can be expressed using the following recursive equation:

$$\begin{gathered} Pop_k^{t+1} = Pop_k^t + FP_{i,k}^{t \rightarrow t+1} - FN_{i,k}^{t \rightarrow t+1} \Rightarrow \\ Pop_k^{t+1} = Pop_k^t + DR_{i,k}^{t \rightarrow t+1} \end{gathered} \tag{4.14}$$

If we take as starting point the perfect solution, we have $Pop_k^0 = TC_i$. If we also consider $DR_{i,k}^{dt}$ to be constant for all transitions, we can find a closed-form solution for the recursive equation:

$$Pop_k^{t+q} = TC_i + qDR_{i,k}^{dt} \tag{4.15}$$

Where q is the number of transitions that have taken place and provides and intuitive measure of how much distance there is between current clustering solution and the perfect solution. However, TC_i is the number of times the object c_i appears in S^c ($\#appearances$) and according to eq. (4.9) we have $TC_i = \gamma n p_{ci}$. By substituting TC_i in eq. (4.15) we have:

$$Pop_k^{t+q} = \gamma n p_{c_i} + qDR_{i,k}^{dt} \tag{4.16}$$

Given that $DR_{1,\alpha} - DR_{2,\beta} < 0$, the population of cluster $\mathbf{r}_\alpha$ is increasing/decreasing with a rate lower/higher from the rate that $\mathbf{r}_\beta$ increases/decreases. So, we are interested in the number of transitions that are needed for causing the population of $\mathbf{r}_\alpha$ to become equal or less than the population of $\mathbf{r}_\beta$. The equality corresponds to the minimum number of transitions.

$$\begin{gathered} Pop_\alpha^{t+q} - Pop_\beta^{t+q} \leq 0 \\ \gamma n p_{c_1} + qDR_{1,\alpha}^{dt} - \gamma n p_{c_2} - qDR_{2,\beta}^{dt} \leq 0 \\ q \geq \frac{\gamma n (p_{c_1} - p_{c_2})}{(DR_{2,\beta}^{dt} - DR_{1,\alpha}^{dt})} \end{gathered} \tag{4.17}$$

In order to derive some conclusions from this equation we need to make the following remarks. Given our basic assumption we have $p_{c_1} > p_{c_2}$. Moreover, given that $DR_{1,\alpha} - DR_{2,\beta} < 0$ we can also accept that $DR_{1,\alpha}^{dt} - DR_{2,\beta}^{dt} < 0$. Thus, all terms on the right hand side of eq. (4.17) are positive. Then, it is clear from eq. (4.17) that the number of transitions q required for causing $\mathbf{r}_\alpha$ not to be the most populated of the generated clusters, increases proportionally to the dataset size n and the difference of probabilities $(p_{c_1} - p_{c_2})$. It is important to note that q does not correspond to any physical value since clustering is not a recursive process, it is just an elegant way to help us derive the intuitive conclusion that as n increases, there is higher probability in $\mathbf{r}_\alpha$ being the most appropriate cluster for learning c_1, due to the increased amount of deviation from the perfect solution that can be tolerated.

4.4 Experimental study

The goal of our study is to use real social data for experimentally validating our expectations on the size of the processed dataset and the error introduced by the visual analysis algorithms. We examine the conditions under which the most populated visual- and tag-"term" converge into the same object and evaluate the efficiency of the object detection models generated by our framework. To this end, in Section 4.4.1 we experimentally verify that the absolute difference between the first and second most frequently appearing objects in a dataset constructed to emphasize on the former, increases as the size of the dataset grows. Section 4.4.2 provides an experimental insight

on the $error_{cl-obj}$ introduced by the visual analysis algorithms and examines whether our expectation on the most populated cluster holds. In Section 4.4.3 we compare the quality of object models trained using flickr images leveraged by the proposed framework, against the models trained using manually provided, strongly annotated samples. Moreover, we also examine how the volume of the initial dataset affects the efficiency of the resulting models. In addition to the above, in Section 4.4.4 we examine the ability of our framework to scale in various types of objects. We close our experimental study in Section 4.4.5 where we compare our work with other existing methods in the literature.

To carry out our experiments we have relied on three different types of datasets. The first type includes the strongly annotated datasets constructed by asking people to provide region detail annotations of images pre-segmented with the automatic segmentation algorithm of Section 4.2.2.2. For this case we have used a collection of 536 images S^B from the *Seaside* domain annotated in our lab[1] and the publicly available MSRC dataset[2] S^M consisting of 591 images. The second type refers to the roughly-annotated datasets like the ones obtained from *flickr groups*. In order to create a dataset of this type S^G, for each object of interest, we have downloaded 500 member images from a *flickr group* that is titled with a name related to the name of the object, resulting in 25 groups of 500 images each (12500 images in total). The third type refers to the weakly annotated datasets like the ones that can be collected freely from collaborative tagging environments. For this case, we have crawled 3000 and 10000 images from flickr which will be referred to as S^{F3K} and S^{F10K} respectively, in order to investigate the impact of the dataset size on the efficiency of the generated models. Depending on the annotation type we use the tag-based selection approaches presented in Section 4.2.2.1 to construct the necessary image sets S^c. Table 4.4 summarizes the information of the datasets used in our experimental study. Note that since our approach is working on the level of regions rather than the level of images, the number of media objects handled by our framework (i.e. feature extraction, clustering, SVM-learning) is much larger than the number of images depicted in Table 4.4, approximately multiplied by 7.

[1] http://mklab.iti.gr/project/scef
[2] http://research.microsoft.com/vision/cambridge/recognition

Table 4.4: Datasets Information

Symbol	Source	Annotation Type	No. of Images	objects	Selection approach
S^B	internal dataset	strongly annotated	536	sky, sea, vegetation, person, sand, rock, boat	keyword based
S^M	MSRC	strongly annotated	591	aeroplane, bicycle, bird, boat, body, book, cat, chair, cow, dog, face, flower, road, sheep, sing, water, car, grass, tree, building, sky	keyword based
S^G	*flickr groups*	roughly-annotated	12500 (500 for each object)	sky, sea, vegetation, person and the 21 MSRC objects	*flickr groups*
S^{F3K}	flickr	weakly annotated	3000	cityscape, seaside, mountain, roadside, landscape, sport-side	SEMSOC
S^{F10K}	flickr	weakly annotated	10000	jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis	SEMSOC

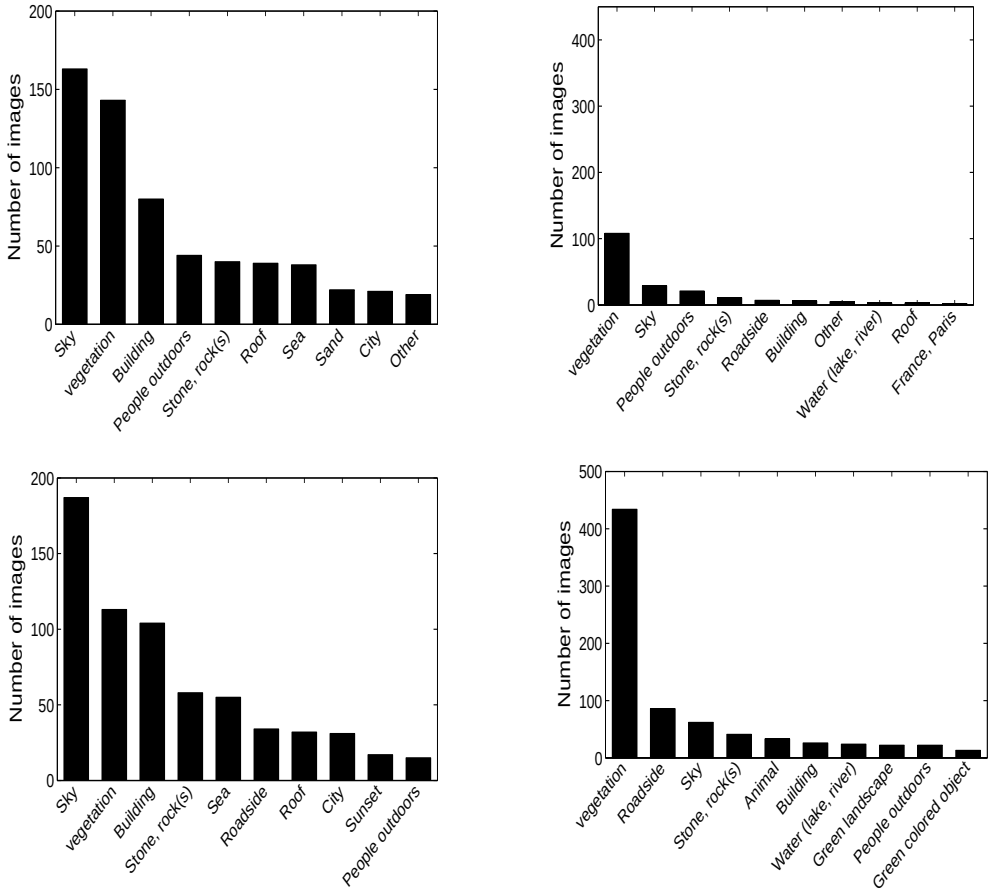
4.4.1 Objects' distribution based on the size of the image set

As claimed in Section 4.3.2, we expect the absolute difference between the number of appearances (*#appearances*) of the first (c_1) and second (c_2) most highly ranked objects within an image set S^{c_1}, to increase as the volume of the dataset increases. This is evident in the case of keyword-based selection since, due to the fact that the annotations are strong, the probability that the selected image depicts the intended object is equal to 1, much greater than the probability of depicting the second most

frequently appearing object. Similarly, in the case of *flickr groups*, since a user has decided to assign an image to the *flickr group* titled with the name of the object, the probability of this image depicting the intended object should be close to 1. On the contrary, for the case of SEMSOC that operates on ambiguous and misleading tags this claim is not evident. For this reason and in order to verify our claim experimentally, we plot the distribution of objects' *#appearances* in four image sets that were constructed to emphasize on objects *sky*, *sea*, *vegetation* and *person*, respectively. These image sets were generated from both S^{F3K} and S^{F10K} using SEMSOC. Each of the bar diagrams depicted in Figs. 4.4 and 4.4 describes the distribution of objects' *#appearances* inside an image set S^c, as evaluated by humans. This annotation effort was carried out in our lab and its goal was to provide weak but noise-free annotations in the form of labels for the content of the images included in both S^{F3K} and S^{F10K}. It is clear that as we move from S^{F3K} to S^{F10K} the difference, in absolute terms, between the number of images depicting c_1 and c_2 increases in all four cases, advocating our claim about the impact of the dataset size on the distribution of objects' *#appearances* when using SEMSOC.

4.4.2 Clustering assessment

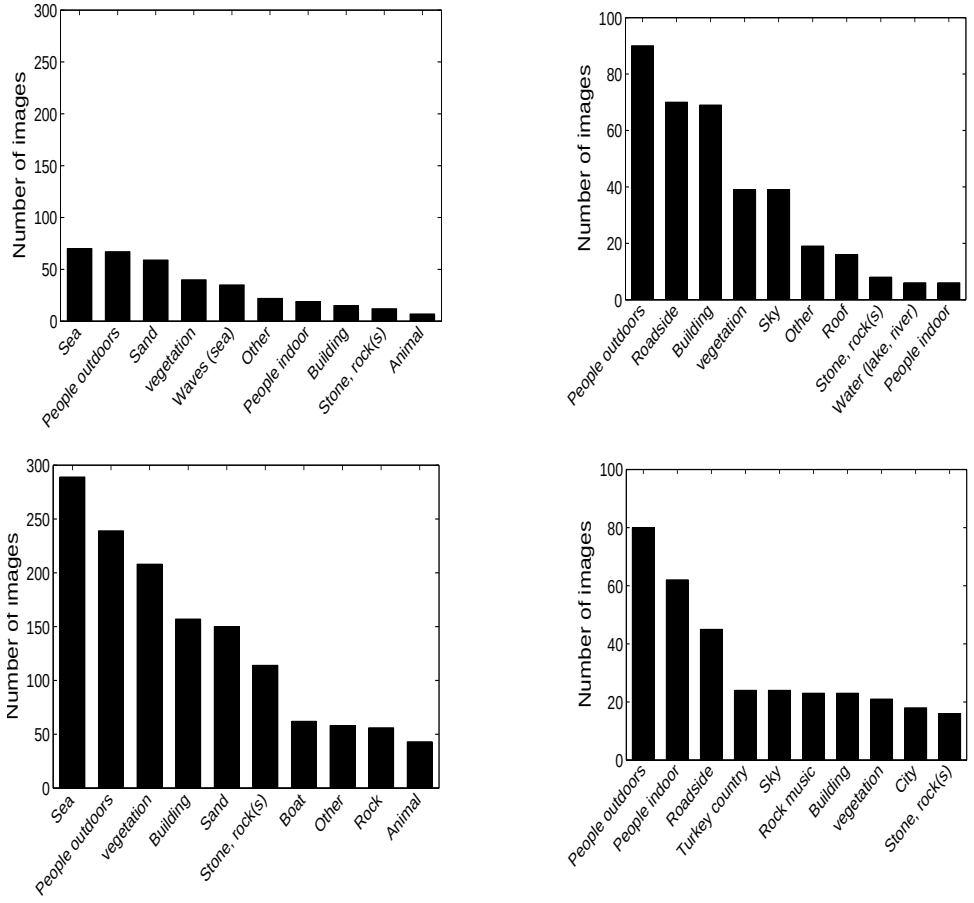
The purpose of this experiment is to provide insight on the validity of our approach in selecting the most populated cluster in order to train a model recognizing the most frequently appearing object. In order to do so we evaluate the content of each of the formulated clusters using the strongly annotated datasets S^B and S^M. More specifically, $\forall c_i$ depicted in S^B or S^M we obtain $S^{c_i} \subset S^B$ or $S^{c_i} \subset S^M$ using keyword based search and apply clustering on the extracted regions. Then, for each S^{c_i} we calculate the values TC_1, $DR_{1,\alpha}$ and Pop_α for the most frequently appearing object c_1 and its corresponding cluster $\mathbf{r}_a$; and TC_2, $DR_{2,\beta}$ and Pop_β for the second most frequently appearing object c_2 and its corresponding cluster $\mathbf{r}_\beta$. Both $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ are determined based on eq. (4.10) of Section 4.3.3. Subsequently, we examine whether $\mathbf{r}_\alpha$ is the most populated among all the clusters generated by the clustering algorithm, not only among $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ (i.e. we examine if $Pop_\alpha = \max(Pop_i)$ for all generated clusters). If this is the case we consider that our framework has succeeded in selecting the most appropriate cluster for training a model to recognize c_1 (a $\surd$ is inserted in the corresponding



(a) Sky (b) Vegetation

Figure 4.4: Distribution of objects' #appearance for objects *sky* and *vegetation* in an image set S^c, generated from S^{F3K} (upper line) and S^{F10K} (bottom line) using SEMSOC

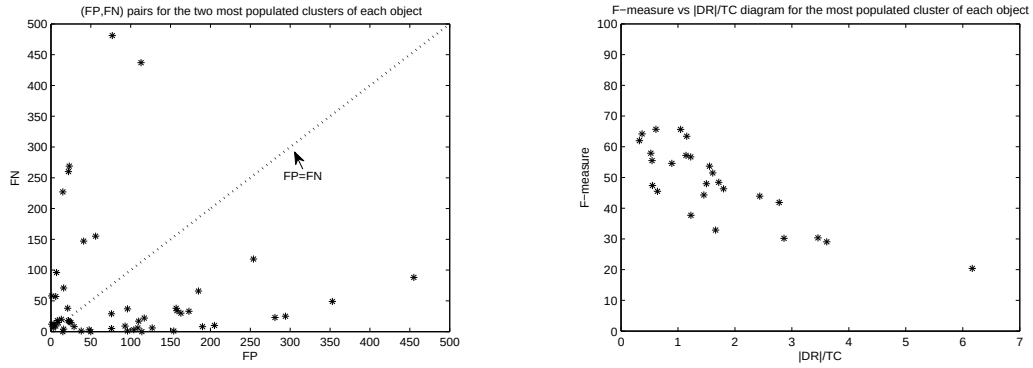
entry of the *Suc* column of Table 4.5). If $\mathbf{r}_\alpha$ is not the most populated cluster, we consider that our framework has failed in selecting the appropriate cluster (a X is inserted in the corresponding entry of the *Suc.* column). Table 4.5 summarizes the results for the 7 objects of S^B and the 19 objects of S^M (the objects bicycle and cat were omitted since there was only one cluster generated). We notice that the appropriate cluster is selected in 21 out of 26 cases advocating our expectation that the $error_{cl-obj}$ introduced by the visual analysis process is usually limited and allows our framework to work efficiently. By examining the figures of Table 4.5 more thoroughly we realize that $DR_{1,\alpha} - DR_{2,\beta} > 0$ for all success cases, with the only exception of object *sky* for S^B. This is in accordance with our analysis in Section 4.3.3 which showed that if



(a) Sea (b) Person

Figure 4.5: Distribution of objects' #appearance for objects *Sea* and *Person* in an image set S^c, generated from S^{F3K} (upper line) and S^{F10K} (bottom line) using SEMSOC

the relative inflow from $\mathbf{r}_\alpha$ to $\mathbf{r}_\beta$ is positive our framework will succeed in selecting the appropriate cluster. In the case of object *sky* our analysis does not hold due to the excessive level of over-segmentation. Indeed, by examining the content of the images belonging to the image set $S^{sky} \subset S^B$ we realize that despite the fact that *sky* is the most frequently appearing object in the image set, after segmenting all images in S^{sky} and manually annotating the extracted regions, the number of regions depicting *sky* $TC_1 = 470$ is less than the number of regions depicting *sea* $TC_2 = 663$. This is a clear indication that the effect of over-segmentation has inverted the objects' distribution making *sea* the most frequently appearing object in S^{sky}. In accordance with our analysis are also the fail cases where the relative inflow from $\mathbf{r}_\alpha$ to $\mathbf{r}_\beta$ is negative (i.e.



(a) (b)

Figure 4.6: a) Diagram showing (FP,FN) scatter plot for $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ clusters of all objects. It is evident that the (FP,FN) pairs produced by the clustering algorithm lay close to the diagonal ($FP = FN$) only when they are close to (0,0). b) Diagram showing the F-Measure scores exhibited for the $\mathbf{r}_\alpha$ cluster of each object, against the observed $|DR_{i,j}|$ value of this cluster normalized with the total number of true positives TC_i. The qualitative aspect of $|DR_{i,j}|$ is advocated by the observation that the F-measure tends to decrease as the ratio $|DR_{i,j}|/TC_i$ increases.

$DR_{1,\alpha} - DR_{2,\beta} < 0$). In none of these 5 cases the difference between $(TC_1 - TC_2)$ was high enough to compensate for the error introduced by the visual analysis process.

Additionally, we have used the experimental observations of Table 4.5 in order to verify the qualitative aspect of $|DR_{i,j}|$ mentioned in Section 4.3.3. More specifically, by producing the (FP,FN) scatter plot for the $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ clusters of the 7 *Seaside* and 19 *MSRC* objects (Fig. 4.6(a)), we verify that no (FP,FN) pairs lay close to the diagonal ($FP = FN$) unless they are close to (0,0). Thus, given that $DR_{i,j} = FP_{i,j} - FN_{i,j}$, there are no cases exhibiting high values for both FP and FN and low values for $|DR_{i,j}|$. This renders $|DR_{i,j}|$ a valid indicator for the quality of the result since a poor quality cluster exhibiting high values for either FP or FN, exhibit also high values for $|DR_{i,j}|$. This qualitative aspect of $|DR_{i,j}|$ is also verified by the diagram of Fig. 4.6(b). In this diagram we plot the F-measure scores for the $\mathbf{r}_\alpha$ cluster of each object (see Section 5.3), against the observed $|DR_{i,j}|$ value of this cluster normalized by the total number of true positives TC_i. It is evident from the diagram that the F-Measure tends to decrease as the ratio $|DR_{i,j}|/TC_i$ increases, showing a clear connection between the $|DR_{i,j}|$ quantity used in our analysis and the quality of the result.

Table 4.5: Clustering Output Insights

S^{c_i}	n	c_1	TC_1	$DR_{1,\alpha}$	Pop_α	c_2	TC_2	$DR_{2,\beta}$	Pop_β	Suc.	$sign(DR_{1,\alpha} - DR_{2,\beta})$
S^B (*Seaside*)											
S^{sea} *	395	sea	732	-404	328	sky	395	-212	183	X	-
S^{sand}	359	sand	422	136	558	sky	337	-103	234	√	+
S^{rock}	53	rock	155	95	250	sea	86	47	133	√	+
S^{boat}	68	boat	96	120	216	sky	69	-57	12	√	+
S^{person}	215	person	435	-238	198	sea	406	-99	307	X	-
$S^{vegetation}$	80	vegetation	157	140	297	sea	114	59	173	√	+
S^{sky}	418	sky	470	-246	224	sea	663	-324	339	X	+
S^M (*MSRC*)											
S^{sign}	27	sign	65	101	166	building	19	-10	9	√	+
S^{sky}	129	sky	139	-89	50	building	115	119	234	X	-
$S^{building}$	88	building	209	304	513	sky	52	-17	35	√	+
S^{car}	6	car	6	37	43	road	7	-3	4	√	+
S^{road}	74	road	94	269	363	sky	32	93	125	√	+
S^{tree}	100	tree	226	258	484	sky	45	124	169	√	+
S^{body}	32	body	54	195	249	face	19	4	23	√	+
S^{face}	21	face	35	121	156	body	17	10	27	√	+
S^{grass}	154	grass	221	367	588	sky	48	133	181	√	+
S^{bird}	29	bird	58	71	129	grass	15	-6	9	√	+
S^{dog}	27	dog	56	84	140	road	11	21	32	√	+
S^{water}	62	water	113	182	295	sky	19	7	26	√	+
S^{cow}	43	cow	109	114	223	grass	57	-51	6	√	+
S^{sheep}	5	sheep	13	15	28	grass	13	-11	2	√	+
S^{flower}	28	flower	60	103	163	grass	8	12	20	√	+
S^{book}	33	book	149	-55	94	face	5	153	158	X	-
S^{chair}	19	chair	39	95	134	road	9	-3	6	√	+
$S^{aeroplane}$	18	aeroplane	12	50	68	sky	12	-8	4	√	+
S^{boat}	15	boat	25	45	70	water	25	-7	18	√	+

∗ although $Pop_\alpha > Pop_\beta$ in this case, the population Pop_γ of the cluster corresponding to the third most frequently appearing object was found to be the highest, which is why we consider this case as a failure

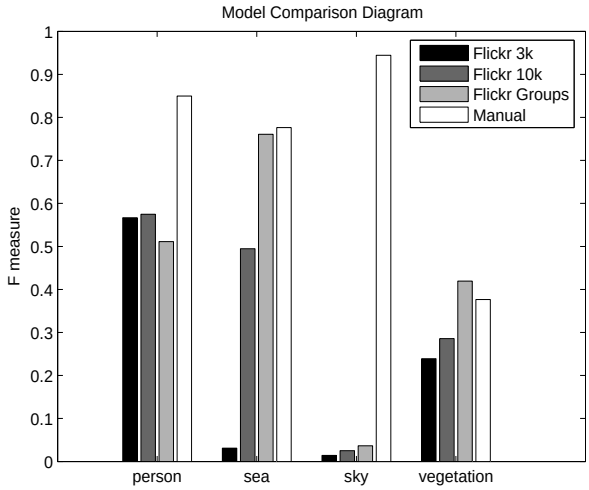


Figure 4.7: Performance comparison between four object recognition models that are learned using images of different annotation quality (i.e. strongly, roughly and weakly)

4.4.3 Comparing object detection models

In order to compare the efficiency of the models generated using training samples of different annotation type (i.e. strongly, roughly, weakly), we need a set of objects that are common in all three types of datasets. For this reason after examining the contents of S^B, reviewing the availability of groups in flickr and applying SEMSOC on S^{F3K} and S^{F10K}, we determined 4 object categories C^{bench}={sky, sea, vegetation, person}. These objects exhibited significant presence in all different datasets and served as benchmarks for comparing the quality of the different models. For each object $c_i \in C^{bench}$ one model was trained using the strong annotations of S^B, one model was trained using the roughly-annotated images contained in S^G and two models were trained using the weak annotations of S^{F3K} and S^{F10K}, respectively. In order to evaluate the performance of these models, we test them using a subset (i.e. 268 images) of the strongly annotated dataset $S^B_{test} \subset S^B$, not used during training. The F_1 metric was used for measuring the efficiency of the models.

By looking at the bar diagram of Fig. 4.7, we derive the following conclusions: a) Model parameters are estimated more efficiently when trained with strongly annotated samples, since in 3 out of 4 cases they outperform the other models and sometimes by a significant amount (e.g. sky, person). b) *Flickr groups* can serve as a less costly

alternative for learning the model parameters, since using the roughly-annotated samples we get comparable and sometimes even better (e.g. vegetation) performance than manually trained models, while requiring considerable less effort to obtain the training samples. c) The models learned from weakly annotated samples are usually inferior from the other cases, especially in cases where the proposed approach for leveraging the data has failed in selecting the appropriate cluster (e.g. *sea* and *sky* for the S^{F3K} dataset). However, the efficiency of the models trained using weakly annotated samples improves when the size of the dataset increases. From the bar diagram of Fig. 4.7, it is clear that when using S^{F10K} the incorporation of a larger number of positive samples into the training set improves the generalization ability of the generated models in all four cases. Moreover, in the case of object *sea* we note also a drastic improvement of the model's efficiency. This is attributed to the fact that the increment of the dataset size compensates, as explained in Section 4.3, for the $error_{cl-obj}$ and allows the proposed method to select the appropriate cluster. On the other hand, in the case of object *sky* it seems that the correct cluster is still missed despite the use of a larger dataset. The correct cluster is also missed for the object *sky* when the weakly annotated samples are obtained from *flickr groups*. This shows that $error_{cl-obj}$ is considerably high for this object and does not allow our framework to select the correct cluster.

4.4.4 Scaling in various types of objects

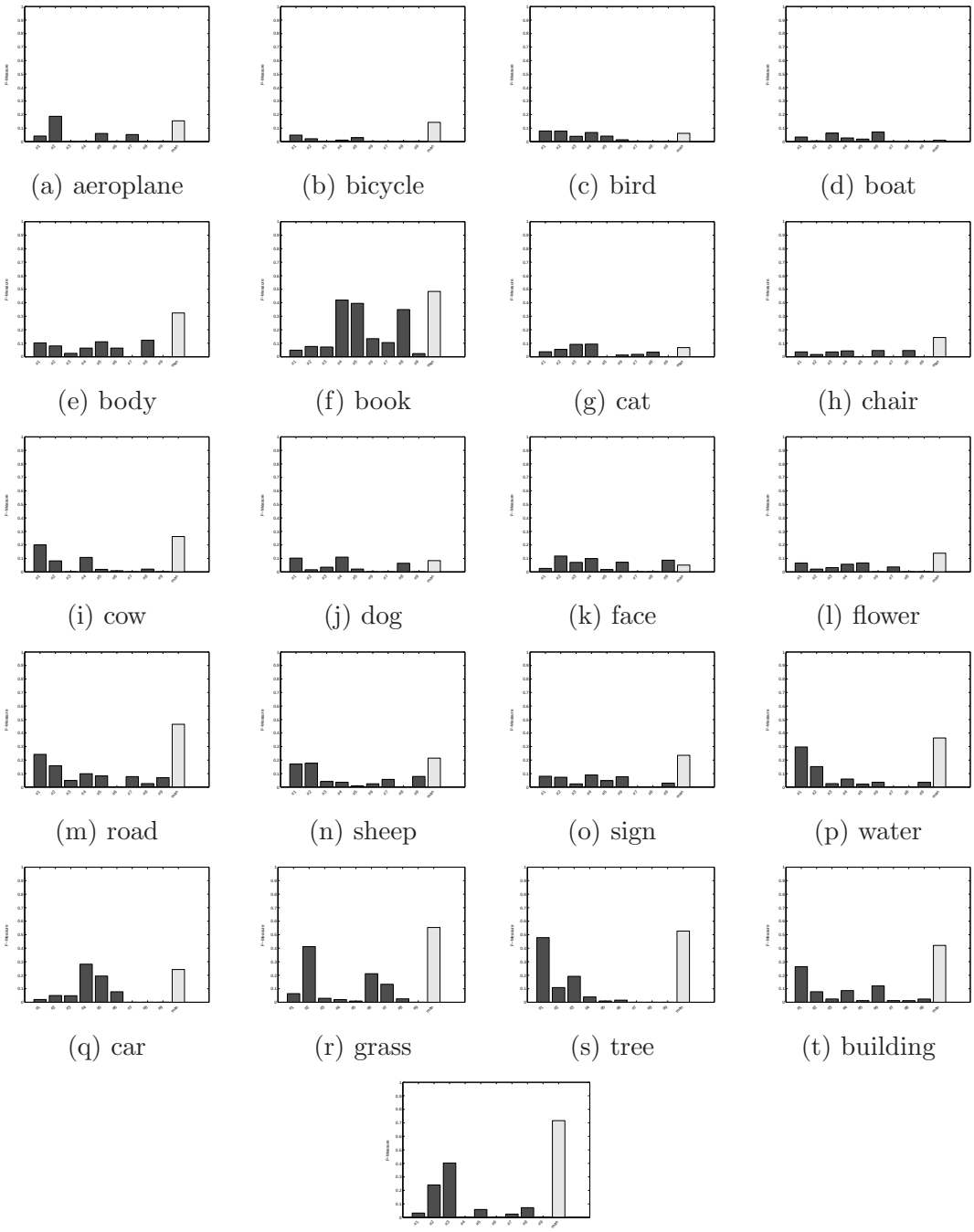
In order to test the ability of our approach in scaling to various types of objects we have performed experiments using the MSRC dataset. MSRC (S^M) is a publicly available dataset that has been widely used to evaluate the performance of many object detection methods. The reason for choosing MSRC over other publicly available benchmarking datasets, such as the the PASCAL VOC challenge [168], was its widespread adoption by many works in the literature allowing us to compare our work with state of the art methods (see Section 4.4.5). MSRC consists of 591 hand-segmented images annotated at region detail for 23 objects. Due to their particularly small number of samples *horse* and *mountain* objects were ignored in our study. In order to test our approach for these objects we have relied on *flickr groups* to obtain 21 image groups, with 500 members each, suitable for training models for the 21 objects of S^M. All images of S^M were segmented by the segmentation algorithm described in Section 4.2.2.2 and the ground truth label of each segment was taken to be the label of the hand-labeled region that

overlapped with the segment by more than the 2/3 of the segment's area. In any other case the segment was labeled as void. The S^M was split randomly in 295 training S^M_{train} and 296 testing S^M_{test} images, ensuring approximately proportional presence of each object in both sets.

In an attempt not only to evaluate the efficiency of the developed models but also to discover whether the root cause for learning a bad model is the selection of an inappropriate set of training samples, or the deficiency of the employed visual feature space to discriminate the examined object, we perform the following. Since we don't have strong annotations for the images obtained from *flickr groups* and it is impossible to assess the quality of the generated clusters as performed in Section 4.4.2, we train as many models as the number of generated clusters (not only using the most populated) and test them using S^M_{test}. Our aim is to assess the quality of the generated clusters indirectly, by looking at the recognition rates of the models trained with the member regions of each cluster. The bar diagrams of Fig. 4.8 show the object recognition rates (measured using the F_1 metric) for the models trained using as positive samples the members of each of the nine most populated (in descending order) clusters. The last bar in each diagram corresponds to the performance of the model trained using the strong annotations of S^M_{train} and tested using S^M_{test}. Moreover, in order to visually inspect the content of the generated clusters we have implemented a viewer that is able to read the clustering output and simultaneously display all regions included in the same cluster. By having an overall view of the regions classified in each cluster we can better understand the distribution of clusters to objects and derive some conclusions on the reasons that make the proposed approach to succeed or fail. By looking at the bar diagrams of Fig. 4.8 we can distinguish between four cases.

In the first case we classify the objects *bird*, *boat*, *cat*, *dog* and *face* that are too diversiform with respect to the employed visual feature space and as a consequence, none of the developed models (not even the one trained using the manual annotations) manage to achieve good recognition rates. In addition to that, the particular small number of relevant regions in the testing set renders most of these objects inappropriate for deriving useful conclusions.

In the second case we classify the objects *bicycle*, *body*, *chair*, *flower* and *sign* that although they seem to be adequately discriminated in the visual feature space (i.e. the model trained using the manually annotated samples performs relatively well), none



(a) aeroplane (b) bicycle (c) bird (d) boat

(e) body (f) book (g) cat (h) chair

(i) cow (j) dog (k) face (l) flower

(m) road (n) sheep (o) sign (p) water

(q) car (r) grass (s) tree (t) building

(u) sky

Figure 4.8: Experiments on the 21 objects of MSRC dataset. In each bar diagram the nine first bars (colored in black) show the object recognition rates (measured using F_1 metric) for the models trained using as positive samples the members of each of the nine most populated (in descending order) clusters. The last bar (colored in gray) in each diagram correspond to the performance of the model trained using strongly annotated samples.

of the models trained using the formulated clusters manages to deliver significantly better recognition rates from the other clusters. Thus, none of the generated clusters contains good training samples which indicates that the images included in the selected *flickr group* are not representative of the examined object, as perceived by the MSRC annotators.

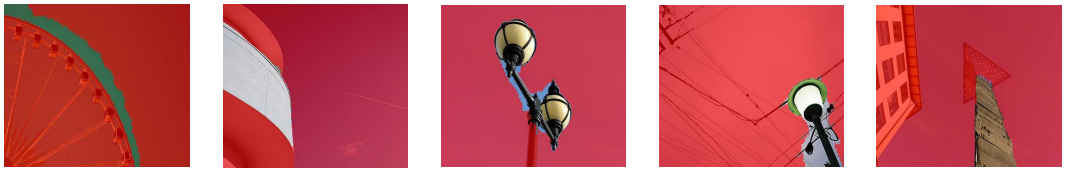
Aeroplane, *book*, *car*, *grass*, *sky*, *sheep* are classified in the third case including the objects that are effectively discriminated in the visual feature space (i.e. the model trained using the manually annotated samples performs relatively well) and there is at least one cluster that delivers performance comparable with the manually trained model. However, the increased $error_{cl-obj}$ has prevented this cluster from being the most populated, since the regions representing the examined object are split in two or more clusters. Indeed, if we take for instance the object *sky* and use the viewer to visually inspect the content of the formulated clusters, we realize that clustering has generated many different clusters containing regions depicting sky. As a result the cluster containing the regions of textured objects has become the most populated. Fig. 4.9 shows indicative images for some of the generated clusters for object *sky*. The clusters' rank (#) refers to their population. We can see that the clusters ranked #2, #3, #6 and #7 contain *sky* regions while the most populated cluster #1 contains the regions primarily depicting statues and buildings. Consistently, we can see in Fig. 4.8 that the performance of the models trained using clusters #2, #3 is much better than the performance of the model trained using cluster #1.

Finally, in the last case we classify the objects *cow*, *road*, *water*, *tree*, *building*, where our proposed approach succeeds in selecting the appropriate cluster and allows the classifier to learn an efficient model. Fig. 4.10 presents some indicative regions for 6 out of the 9 clusters, generated by applying the proposed approach for the object *tree*. For each cluster we present five indicative images in order to show the tendency, in a semantic sense, of the regions aggregated in each cluster. It is interesting to see that most of the formulated clusters tend to include regions of a certain semantic object such as *tree* (#1), *grass* (#2), *sky* (#5), *water* (#9) or noise regions. In these cases where the $error_{cl-obj}$ is limited, it is clear that the regions of the object that appears more frequently in the dataset (*tree* in this case) are gathered in the most populated cluster.

#1 Cluster - architecture (statues, buildings)

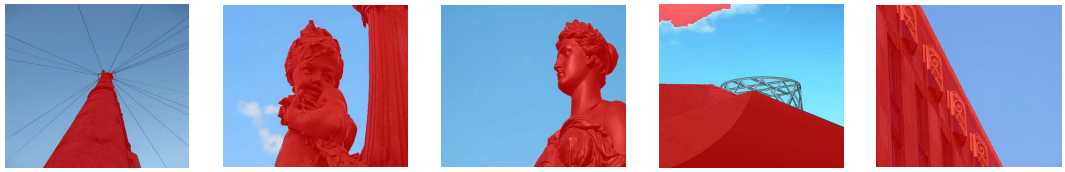
#2 Cluster - sky (but a bit noisy)

#3 Cluster - sky (best performing model)



#5 Cluster - noise

#6 Cluster - sky (mostly dark)



#7 Cluster - sky (mostly light)

Figure 4.9: Indicative regions from the clusters generated by applying our approach for the object *sky*. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

#1 Cluster - trees

#2 Cluster - grass

#3 Cluster - mountain with noise

#4 Cluster - noise

#5 Cluster - cloudy sky

#9 Cluster - water

Figure 4.10: Indicative regions from the clusters generated by applying our approach for the object *tree*. The regions that are not covered in red are the ones that have been assigned to the corresponding cluster.

4.4.5 Comparison with existing methods

Our goal in the previous experiments was to highlight the potential of social media to serve as the source of training samples for object recognition models. Thus, we have focused on the relative loss in performance that results from the use of leveraged rather than manually annotated training samples, and not on the absolute performance values of the developed models. However, in order to provide an indicative measure of the loss in performance that we suffer when compared with other existing works in the literature, we calculate the classification rate (i.e. number of correctly classified cases divided by the total number of correct cases) of our framework for the 21 objects of MSRC. Then, we compare the results with two methods [96], [95] that are known to deliver state of the art performance on this dataset. Textonboost [96] uses conditional random fields to obtain accurate image segmentation and is based on textons, which jointly model shape and texture. The combination of Markov Random Fields (MRF) and aspect models is the approach followed in [95] in order to produce aspect-based spatial field models for object detection. Note that the reported classification rates are not directly comparable since the methods are not relying on the same set of visual features, the training/test split is likely to be different and the results are reported at different level (in [96] at pixel level, in [95] at the level of 20x20 image patches, and in our case at the level of arbitrary shaped segments which are extracted by an automatic segmentation algorithm). However, the comparison of these methods allows us to make some useful conclusions about the trade-off between the annotation cost for training and the efficiency of the developed models. Table 4.6 summarizes the classification rates per object for each method.

On average, the accuracy obtained from our approach (45%) is inferior to the one obtained from PLSA-MRF/I (50%) which is again inferior to the accuracy obtained from Textonboost (58%). It is interesting to see that the performance scores obtained by the three methods are ranked proportionally to the amount of annotation effort required to train their models. Indeed, Textonboost [96] requires strongly annotated images that can only be produced manually, the PLSA-MRF/I algorithmic version of [95], requires weakly but noise-free annotated images the generation of which typically involves light human effort, and our framework operates on weakly but noisy annotated images that can be automatically collected from social sites at no cost.

Table 4.6: Comparing with existing methods in object detection. The reported scores are the classification rates (i.e. number of correctly classified cases divided by the total number of correct cases) per object for each method.

	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	Average
Prop. Framework	87	9	65	45	45	14	29	53	56	12	75	88	27	30	25	50	44	59	71	29	41	45
PLSA-MRF/I [95]	45	64	**71**	75	**74**	**86**	81	47	1	**73**	55	88	6	6	**63**	18	**80**	27	26	**55**	8	50
Prop.Fram./M-F/W	**83**	**72**	69	**91**	70	1	**87**	**53**	**33**	12	**87**	**100**	**47**	**79**	53	**47**	55	**33**	**67**	11	**61**	**57**
Textonboost [96]	62	**98**	**86**	58	50	**83**	**60**	**53**	**74**	**63**	75	63	35	19	**92**	15	**86**	54	19	**62**	7	58
Prop.Fram./M-F/S	**63**	67	76	**73**	**70**	51	27	47	67	17	**94**	**100**	**53**	**47**	59	**47**	68	**92**	**73**	59	**55**	**62**

The costless nature of our approach motivated the execution of two additional experiments that are essentially variations of our original approach, mixing manually labeled data from MSRC and noisy data from flickr. More specifically, the first variation Prop.Fram./M-F/W mixes MSRC and flickr data at the level of images. Initially, the strong region-to-label associations provided by MSRC are relaxed to become weak associations of the form image-to-label(s). Then, these weakly annotated MSRC images are mixed with images from flickr and the proposed framework is applied on the mixed set of images. Finally, the samples used for training the object recognition models consist of the regions belonging to the most populated of the clusters generated from the mixed set. The Prop.Fram./M-F/W variation is directly compared with PLSA-MRF/I [95] since they use the MSRC annotations in the same way. The second variation Prop.Fram./M-F/S mixes MSRC and flickr data at the level of regions. The samples used for training the object recognition models consist of the strongly annotated regions from MSRC plus the regions belonging to the most populated of the clusters generated from flickr data. The Prop.Fram./M-F/S variation is directly compared with Textonboost [96] since they use the MSRC annotations in the same way. Table 4.6 shows that both variations of our approach, mixing MSRC and flickr data, outperform their directly comparable state-of-the art approaches. In the case of Prop.Fram./M-F/W the obtained average accuracy (57%) outperforms PLSA-MRF/I by 7%, while in the case of Prop.Fram./M-F/S the obtained average accuracy (62%) outperforms Textonboost by 4%.

4.5 Discussion of our experimental findings

In this chapter we have shown that the collective knowledge collected in social media can be successfully used to remove the need for close human supervision when training object detectors. The experimental results have demonstrated that although the performance of the detectors trained using leveraged social media is inferior to the one achieved by manually trained detectors, there are cases where the gain in effort compensates for the small loss in performance. In addition, we have seen that by increasing the number of utilized images we manage to improve the performance of the generated detectors, advocating the potential of social media to facilitate the creation of reliable and effective object detectors. The value of social media was also advocated by the experiments showing that when mixing manually labeled and effortlessly obtained flickr data, we manage to outperform the state-of-the-art approaches relying solely on manually labeled samples. Finally, despite the fact that there will always be a strong dependence between the discriminative power of the employed feature space and the efficiency of the proposed approach in selecting the appropriate set of training samples, our analysis has shown that we can maximize the probability of success by using large volumes of user contributed content.

Chapter 5

Tagged image indexing using cross-modal dependencies

In this chapter we present our approach for the efficient indexing of tagged images. Tagged images are a common resource of social networks and occupy a large portion of the social media stream. Their basic characteristic is the co-existence of two heterogeneous information modalities i.e. visual and tag, which refer to the same abstract meaning. This multi-modal nature of tagged images makes their efficient indexing a challenging task that apart from dealing with the heterogeneity of modalities, it needs to also exploit their complementary information capacity [169].

5.1 Description of the proposed approach

The need to obtain a joint, unique representation of tagged images calls for techniques that will manage to handle the very different characteristics exhibited by the visual and tag information. This is true both in terms of the raw features' nature, i.e. sparse, high-dimensional tag co-occurrence vectors extracted from tag descriptions, compared to usually dense and low-dimensional descriptors extracted from visual content, as well as in terms of their semantic capacity, i.e. while abstract concepts like "freedom" are more easily described with text, ambiguous concepts like "rock" are more easily grounded using visual information. Based on the above, one can pursue a solution to the multi-modal indexing problem by defining a joint feature space where the projection of uni-modal features will yield a homogeneous and semantically enriched image

representation.

The most trivial approach in this direction is to define a joint feature space by concatenating the individual uni-modal features extracted from both modalities, also known as early fusion. However, by indiscriminately placing features extracted from different modalities into a common feature vector, the resulting space is likely to be dominated by one of the combined modalities or lose its semantic consistency. This was the reason that researchers turned into the statistical characteristics of the data to overcome these problems. For instance, [120] uses information theory and a maximum entropy model in order to integrate heterogeneous data into a unique feature space, [170] finds statistical independent modalities from raw features and applies super-kernel fusion to determine their optimal combination, while [122] presents several cross-modal association approaches under the linear correlation model.

The most recent approaches rely on the use of probabilistic Latent Semantic Analysis (pLSA) to facilitate the combination of heterogeneous modalities. The pLSA-based aspect or topic model is a method originally proposed in [171] that allows to map a high-dimensional word distribution vector to a lower-dimensional topic vector (also called aspect vector). This model assumes that the content depicted by every image can be expressed as a mixture of multiple topics and that the occurrences of words in this content is a result of the topic mixture. Thus, the latent layer of topics that is introduced between the image and the tag or visual words appearing in its content, acts as a feature space where both types of words can be combined meaningfully. Moreover, given that the goal of pLSA is to learn a set of latent topics that will act as bottleneck variables when predicting words, apart from handling the heterogeneity of multimodal sources, pLSA is also encouraged for discovering the hidden relations between images. Examples of pLSA-based approaches include [172] where pLSA is used to infer which visual patterns describe each concept, as well as [173] where Latent Dirichlet Allocation (LDA) [73] is used to model each image as the mixture of topics/object parts depicted in the image.

However, even if the space of latent topics can be considered to satisfy the requirement of combining the words extracted from heterogeneous modalities without introducing any bias or rendering them meaningless, it still neglects the fact that, being different expressions of the same abstract meaning, there is a certain amount of dependance between the tag and visual words that appear together very frequently.

This additional requirement motivates the employment of methods that will allow the cross-word dependencies to influence the nature of the extracted latent topics. In this context we examine the use of high order pLSA to improve the semantic capacity of the derived latent topics. High order pLSA is essentially the application of pLSA to more than two observable variables allowing the incorporation of different word types into the analysis process. We treat images, visual content and tags as the three observable variables of an aspect model and we manage to extract a set of latent topics that incorporate the semantics of both the visual and tag information space. Moreover, we integrate the cross-word dependencies into the update rules of high order pLSA in order to devise a feature extraction scheme where the co-existence of two words that are known from experience to appear together rather frequently is more important in defining the latent topics, than the co-existence of two words that rarely appear together and are likely to be the result of noise. In the following we formulate image retrieval as a problem of defining a semantics sensitive feature space and describe different approaches for using the information carried by tagged images to define such a space. Moreover, we present our approach on how to apply high order pLSA using cross-word dependencies and present a distributed calculation model for tackling the high computational and memory requirements of this method. Finally, we present our experimental findings for the tasks of image retrieval and clustering.

5.2 Problem formulation

In order to index tagged images based on their semantic meaning we need to define a feature space where the distance between two images is proportional to their semantic affinity. To put this formally, given an image d, the set of concepts depicted by this image $C_d = \{c_1, c_2, \ldots, c_{|C|}\}$, a representation $F_S^d = \{f_{s_1}, f_{s_2}, \ldots, f_{s_{|S|}}\}$ of the image in feature space S, the distance $dist(F_S^{d_i}, F_S^{d_j}) \geq 0$ between the representations of two images in S and a set of D images indexed based on their representations; we need to define the feature space S where $\forall d \in D$ the typical image retrieval process $Q(d_q, D) = rank_r(dist(F_S^{d_q}, F_S^{d_r}))$ returns a ranked list of all images in D such that when $dist(F_S^{d_q}, F_S^{d_i}) \leq dist(F_S^{d_q}, F_S^{d_j})$ it also stands that $|C_{d_q} \cap C_{d_i}| \geq |C_{d_q} \cap C_{d_j}|$. Thus, image retrieval is essentially a problem of defining a semantics sensitive feature

space. In the following we describe different techniques for defining a feature space suitable for indexing tagged images.

5.3 Building a semantics sensitive space for tagged images

5.3.1 Codebook-based representation

One of the most popular approaches for image representation is based on defining a set of representative "words" (i.e. a Codebook $W = \{w_1, w_2, \ldots, w_{|W|}\}$), that are able to span a sufficiently large portion of the information space that they are used to describe. Then, based on this Codebook each image can be represented as an occurrence count histogram of the representative "words" in its content. The critical factor in this process is to define a highly expressive Codebook, so as to cover every potential instantiation of the image content. In the following we describe how the Codebook representation approach can be applied in the case of visual content and tags, as well as how to mix different Codebooks for obtaining a multi-modal image representation.

5.3.1.1 Visual codebook

In order to represent the visual information carried by an image using the aforementioned Codebook-based approach, we need to define the set of visual words that will act as the representative "words" of our information space. For the purposes of our work we have used the scheme adopted in [174] that consists of the following 3 steps: a) the Difference of Gaussian filter is applied on the gray scale version of an image to detect a set of key-points and scales respectively, b) the Scale Invariant Feature Transformation (SIFT) [166] is computed over the local region defined by the key-point and scale, and c) a Visual Word Vocabulary (i.e. Codebook $V = \{v_1, v_2, \ldots, v_{|V|}\}$) [147] is created by applying the k-means algorithm to cluster in 500 clusters, the total amount of SIFT descriptors that have been extracted from all images. Then, using the Codebook V we vector quantize the SIFT descriptor of each interest point against the set of representative visual words. This is done by mapping the SIFT descriptor to its closest visual word and increasing the corresponding word count. By doing this for all key-points found in an image, the resulting 500-dimensional image representation is the occurrence count histogram of the visual "words" in its content, $F_V^d = \{f_{v_1}, f_{v_2}, \ldots, f_{v_{|V|}}\}$.

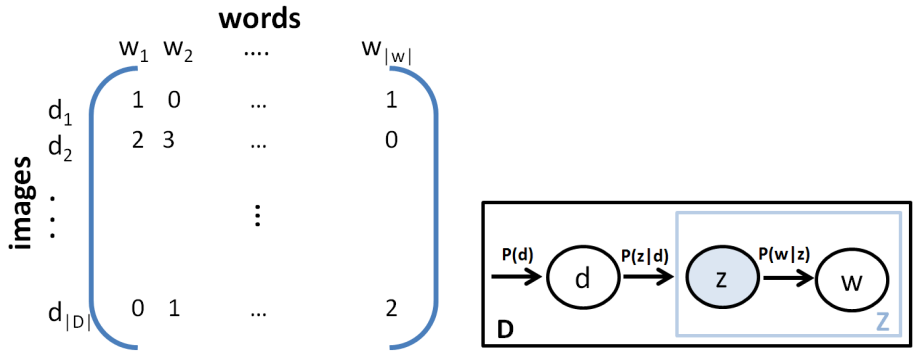
5.3.1.2 Tag codebook

A similar approach has been adopted for representing the tag information that accompanies an image using a tag Codebook. As in the previous case, we need to define the set of representative tag "words" that will manage to span a sufficiently large portion of the tag information space. However, in this case there is no need to employ clustering for determining which words should be included in the Tag Word Vocabulary (i.e. Codebook $T = \{t_1, t_2, \ldots, t_{|T|}\}$). Instead, from a large volume of utilized tags we need to select the ones with minimum level of noise and maximum usage by the users. For the purposes of our work we have used the Codebook constructed in [174] using the following steps. 269,648 images were downloaded from flickr along with their accompanying tags. Among the total set of 425,059 unique tags that have been used by the users, there are 9,325 tags that appear more than 100 times. Many of these unique tags arise from spelling errors, while some of them are names etc, which are meaningless for general image annotation. Thus, all these 9,325 unique tags were checked against the WordNet Lexical Database [132] and after removing the non-existing ones, a list with 5,018 tags was determined. For the purposes of our work, out of the 5,018 tags the first 1,000 that have been used more frequently were selected to form the tag Codebook. Eventually, we use this Codebook to obtain for each image a 1000-dimensional occurrence count histogram of the tag "words" in its content, $F_T^d = \{f_{t_1}, f_{t_2}, \ldots, f_{t_{|T|}}\}$.

5.3.1.3 Combining visual and tag codebooks

The most straightforward approach to produce a multi-modal image representation is to consider a combined Codebook composed by simply extending the list of representative visual-"words" with the list of representative tag-"words" (i.e. $VT = \{v_1, v_2, \ldots, v_{|V|}, t_1, t_2, \ldots, t_{|T|}\}$). In this case the generated image representation is essentially the concatenation of visual- and tag-based representations, which results in a 1500-dimensional occurrence count histogram, $F_{VT}^d = \{f_{vt_1}, f_{vt_2}, \ldots, f_{vt_{|V|+|T|}}\}$.

The major drawback of the codebook combination approach is that concatenation is performed between heterogeneous quantities. This results in an non-uniform feature space that is unable to exploit the complementary effect of different modalities. Motivated by this fact, pLSA has been proposed to create a uniform space for the combination of different modalities.



(a) (b)

Figure 5.1: a) co-occurrence data table $n(d, w)$ for images and words, b) the standard pLSA model.

5.3.2 Mixture of latent topics

pLSA aims at introducing a latent (i.e. unobservable) topic layer between two observable variables (i.e. images and "words" in our case). Let us denote $D = \{d_1, \ldots, d_{|D|}\}$ the set of images and $W = \{w_1, \ldots, w_{|W|}\}$ the set of "words". The key idea is to map high-dimensional word occurrence count vectors, as the ones described in Section 5.3.1, to a lower dimensional representation in a so-called latent semantic space [171]. pLSA is based on a statistical model which has been called aspect model [69]. The aspect model is a latent variable model for co-occurrence data $n(d, w)$ (see Fig 5.1(a) for an example), which associates an unobserved class variable $z \in Z = \{z_1, \ldots, z_{|Z|}\}$ with each observation as shown in Fig. 5.1(b). Then, given that $P(w|d)$ is the conditional probability of "words" given images, its value can be obtained by performing row-wise normalization of $n(d, w)$. Then, using the asymmetric model for pLSA a joint probability model over the set of images D and the set of words W is defined by the mixture:

$$P(d, w) = P(d)P(w|d), \quad P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{5.1}$$

where $P(d)$ denotes the probability of an image to be picked, $P(z|d)$ the probability of a topic given the current image, and $P(w|z)$ the probability of a word given a topic. In this case, if we introduce $R(z, w, d)$ to indicate which hidden topic z is selected to generate w in d such that $\sum_z R(z, w, d) = 1$, the complete likelihood for this model can be formulated as:

$$L = \sum_D \sum_W P(d,w) \sum_Z R(z,w,t)[logP(d) + logP(w|z) + logP(z|d)] \quad (5.2)$$

The function that we need to maximize is:

$$E[L] = \sum_D \sum_W P(d,w) \sum_Z P(z|d,w)[logP(d) + logP(w|z) + logP(z|d)] \quad (5.3)$$

Thus, using Expectation Maximization (EM) [97] the latent topics can be learned by randomly initializing $P(w|z)$, $P(z|d)$ and iterating through the following steps:

E-step:

$$P(z|w,d) = \frac{P(w|z)P(z|d)}{\sum_Z P(w|z)P(z|d)} \quad (5.4)$$

M-step:

$$P(d) = \frac{\sum_W \sum_Z P(d,w)P(z|w,d)}{\sum_D \sum_W \sum_Z P(d,w)P(z|w,d)}$$

$$P(w|z) = \frac{\sum_D P(d,w)P(z|w,d)}{\sum_D \sum_W P(d,w)P(z|w,d)}$$

$$P(z|d) = \frac{\sum_W P(d,w)P(z|w,d)}{\sum_Z \sum_W P(d,w)P(z|w,d)} \quad (5.5)$$

whereas for indexing a new image I_q we just need to repeat the above steps but without updating $P(d)$ and $P(w|z)$ that have been obtained from the learning stage. Once a topic mixture $P(z|d)$ is derived for an image d, we have a high-level representation of this image with less dimensions from the initial representation that is based on the co-occurrence of "words". This is because we commonly choose the number of topics $|Z|$ to be much smaller than the number of words so as to act as bottleneck variables in predicting words. The resulting $|Z|-$dimensional topic vectors can be used directly in an image retrieval or a clustering setting, if we take the distance (e.g. L_1, Euclidean, cosine) between the topic vectors of two images to express their similarity.

5.3.2.1 Visual-based latent topics

In the visual information space, pLSA can be applied by considering the representative visual "words" of the visual codebook to constitute the second observable variable. Then, using the co-occurrence vectors between images and visual words $n(d, v)$, each image of D can be represented in the visual-based latent space ZV using the following joint probability model:

$$P(d, v) = P(d)P(v|d), \quad P(v|d) = \sum_{zv \in ZV} P(w|zv)P(zv|d) \tag{5.6}$$

In the visual-based latent space $P(zv|d)$, the vector elements of each image representation denote the degree to which an image can be expressed using the corresponding visual based latent topics, $F^d_{ZV} = \{f_{zv_1}, f_{zv_2}, \ldots, f_{zv_{|ZV|}}\}$.

5.3.2.2 Tag-based latent topics

Similarly, in the tag information space pLSA can be applied by considering the representative tag "words" of the tag codebook to constitute the second observable variable. Then, using the tag-word co-occurrence vectors between images and tag words $n(d, t)$, each image of D can be represented in the tag-based latent space ZT using the following joint probability model:

$$P(d, t) = P(d)P(t|d), \quad P(t|d) = \sum_{zt \in ZT} P(w|zt)P(zt|d) \tag{5.7}$$

In the tag-based latent space $P(zt|d)$, the vector elements of each image representation denote the degree to which an image can be expressed using the corresponding tag-based latent topics, $F^d_{ZT} = \{f_{zt_1}, f_{zt_2}, \ldots, f_{zt_{|ZT|}}\}$

5.3.2.3 Combining visual and tag based latent space

Motivated by the fact that both topic vectors refer to the so-called latent semantic space and express probabilities (i.e. the degree to which a certain topic exists in the image), we assume that the topics obtained from both modalities are homogeneous and can be indiscriminately considered as the representative "words" of a combined codebook. Based on this assumption, an image representation that combines information from both modalities can be constructed by concatenating into a common multi-modal image

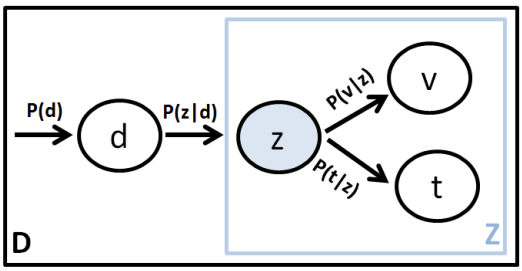


Figure 5.2: Graphical representation of the *highOrder-plsa* model

representation, the two image representations of visual and tag based latent space, $F_Z^d = \{f_{zv_1}, f_{zv_2}, \ldots, f_{zv_{|ZV|}}, f_{zt_1}, f_{zt_2}, \ldots, f_{zt_{|ZT|}}\}$.

However, even if the concatenation is performed between values of similar nature (i.e. latent topics obtained through the application of pLSA), the simple combination of visual and tag based topics completely neglects the dependencies that may exist between the original visual- and tag-"words". Thus, even if we know by experience that the visual word v_i has low dependency with the tag word t_j, there is no way for the aforementioned approach to exploit this knowledge. This shortcoming was our basic motivation for applying high order pLSA as detailed subsequently.

5.3.3 High order pLSA

High order pLSA is the extension of pLSA to more than two observable variables. Using high order pLSA our goal is to apply the previously described aspect model for our three observable variables namely images, visual words and tag words. Using the asymmetric approach for pLSA, the generative model for our three observable variables is graphically represented in Fig 5.2 and can be expressed as follows:

$$P(d, v, t) = P(d) \sum_Z P(v|z)P(t|z)P(z|d)) \tag{5.8}$$

Then, in direct analogy with Section 5.3.2, if we introduce $R(z, v, t, d)$ to indicate which hidden topic z is selected to generate v and t in d such that $\sum_z R(z, v, t, d) = 1$, the complete likelihood can be formulated as:

$$L = \sum_D \sum_V \sum_T P(d,v,t) \sum_Z R(z,d,v,t) \\ [logP(d) + logP(v|z) + logP(t|z) + logP(z|d)] \tag{5.9}$$

and the function that we need to maximize is:

$$E[L] = \sum_D \sum_V \sum_T P(d,v,t) \sum_Z P(z|d,v,t) \\ [logP(d) + logP(v|z) + logP(t|z) + logP(z|d)] \tag{5.10}$$

Thus, using Expectation Maximization (EM) [97] the latent topics can be learned by randomly initializing $P(v|z)$, $P(t|z)$ and $P(z|d)$ and iterating through the following steps:

E-step:

$$P(z|d,v,t) = \frac{P(v|z)P(t|z)P(z|d)}{\sum_Z P(v|z)P(t|z)P(z|d)} \tag{5.11}$$

M-step:

$$P(d) = \frac{\sum_V \sum_T \sum_Z P(d,v,t)P(z|v,t,d)}{\sum_D \sum_V \sum_T \sum_Z P(d,v,t)P(z|v,t,d)}$$

$$P(v|z) = \frac{\sum_D \sum_T P(d,v,t)P(z|v,t,d)}{\sum_D \sum_T \sum_V P(d,v,t)P(z|v,t,d)}$$

$$P(t|z) = \frac{\sum_D \sum_V P(d,v,t)P(z|v,t,d)}{\sum_D \sum_T \sum_V P(d,v,t)P(z|v,t,d)}$$

$$P(z|d) = \frac{\sum_V \sum_T P(d,v,t)P(z|v,t,d)}{\sum_Z \sum_V \sum_T P(d,v,t)P(z|v,t,d)} \tag{5.12}$$

whereas for indexing a new image I_q we just need to repeat the above steps but without updating $P(d)$, $P(v|z)$ and $P(t|z)$ that have been obtained from the learning stage. The iterations stop when the value of eq.(5.10) converge to its maximum (either local or global). In order to guarantee this, we rely on the relative change of $E[L]$ between consecutive iterations, as shown in eq.(5.13). If this relative change is below a predefined threshold the process is terminated, otherwise the EM steps are repeated.

$$\frac{E_{current}[L] - E_{previous}[L]}{abs(E_{previous}[L])} \tag{5.13}$$

In eqs.(5.9-5.12) we have used the joint probability distribution $P(d, v, t)$ of the observable variables (i.e. documents, visual words and tag words), in order to formulate high order pLSA. Due to the normalizing denominators, instead of $P(d, v, t)$ any unnormalized approximation to it can be used. The classical pLSA formulations use the frequency of occurrence $n(d, v, t)$, which is the number of times a visual word v_i appears together with a tag word t_j in a given image d_k. However, in our effort to incorporate prior knowledge into the generation process of the latent topics, we have followed an approach where $P(d, v, t)$ is approximated using the cross-word dependencies. More specifically, we accept that there is a certain degree of dependence on how visual words appear together with tag words, and that this dependence can be learned from data. In order to estimate these dependencies we introduce the concept of *word-profiles*. The *word-profile* is a $|D|$-dimensional binary vector that models the occurrence distribution of a *word* in a set of $|D|$ images, having 1's in the places corresponding to the images where the *word* appears at least once and 0 in all other places. In other words, the *word-profiles* are the column vectors of $n(d, t)$ and $n(d, v)$ after thresholding them with 1. Using the occurrence distribution of each *word* in a corpus of images, we have a natural way to estimate the dependency between *words* of different type (i.e. visual and tag), by measuring their vector distance. For the purposes of our work, given that the values of *word-profiles* cannot be negative, we have used the complement of cosine similarity to calculate the dependency between two words v and t, as shown below:

$$J(v, t) = 1 - \frac{v * t}{\| v \| \| t \|} \tag{5.14}$$

Then, $\forall v \in V$ and $\forall t \in T$ we calculate $J(v,t)$ in order to measure the dependency degree of every possible combination between the visual and tag words. Finally, we incorporate this information during the approximation of $P(d,v,t)$ as follows:

$$P(d,v,t) = \overline{n}(d,v) * \overline{n}(d,t) * J(v,t) \tag{5.15}$$

where $\overline{n}(d,v)$ and $\overline{n}(d,t)$ are the matrices $n(d,v)$, $n(d,t)$ after thresholding. The rationale behind using eq.(5.15) to approximate $P(d,v,t)$ is to penalize or favor the contribution of some pair (v,t) to the sum of eq.(5.12), based on the prior knowledge that we have about the dependency of v with t. In this way, the co-existence of a pair (v,t) with high cross-word dependency is more important in defining the mixture of latent topics, than the co-existence of a pair with low cross-word dependency, which can be the result of noise.

5.4 A distributed model for calculating high-order pLSA

Although flexible for incorporating two or more random variables in a single latent space, high-order pLSA comes at the price of particularly high computational and memory requirements. As illustrated in eqs.(5.11-5.12) the algorithmic implementation of high order pLSA will have to store in memory and traverse one 4-dimensional array for executing the update steps of EM. Given that the dimensionality of the codebook-based representation in both tag and visual space can range from a few hundreds to a few thousands, it is obvious that the resulting 4-dimensional matrix will become difficult to handle when the number of considered images becomes high. Although data sparseness can be used to alleviate this burden, still the high dimensionality of the matrices that need to be processed renders the proposed approach intractable for very large datasets.

Motivated by this fact, we propose a distributed calculation model for high-order pLSA that could benefit from the multi-core facilities offered by modern processors. Drawing from the literature in distributed clustering [175] and in analogy with the approach presented in [176] for distributed pLSA, we divide the full set of images into equally sized nodes. Each of these nodes is able to apply the algorithm locally and periodically communicate with a central super-node in order to synchronize with

the other nodes. More specifically, using the notation of Section 5.3.3, the algorithm proceeds as follows:

1. Initially, the normalized, term-document co-occurrence matrices $P(d,v)$ and $P(d,t)$ are split along the images dimension into equally sized chunks $P^i(d,v)$ and $P^i(d,t)$. Every chunk is then transmitted to one of the K nodes so that each node carries the information for $|D|/K$ images, except the last node that may have less.

2. The super-node initializes with random values the matrices $P(v|z)$, $P(t|z)$, $P(z|d)$ and with equal priors the matrix $P(d)$. A copy of the matrices $P(v|z)$ and $P(t|z)$ is transmitted in all K nodes, while the matrices $P(d)$ and $P(z|d)$ are split along the images dimension into equally sized chunks $P^i(d)$ and $P^i(z|d)$, in order to be transmitted to each of the K nodes.

3. Each node calculates the local joint probability distribution $P^i(z|d,v,t)$ according to eq.(5.11) and estimates the local value $E^i[L]$ according to eq.(5.10). Then, the super-node sums the $E^i[L]$ values collected from all nodes in order to calculate the central $E[L]$ value for this iteration.

4. Each node locally calculates $P^i(d,v,t)$ based on eq.(5.15), by using $P^i(d,v)$ and $P^i(d,t)$, as well as the cross-words dependencies $J(v,t)$ that are common for all nodes.

5. After calculating $P^i(d,v,t)$ each node locally proceeds to the maximization step and produces the local matrices $P^i(d)$, $P^i(v|z)$, $P^i(t|z)$ and $P^i(z|d)$. The only difference from eq.(5.12) is that all 4 matrices are un-normalized (i.e. all denominators in eq.(5.12) are set to 1).

6. The local matrices $P^i(v|z)$ and $P^i(t|z)$ are collected from all nodes. The super-node performs element wise summation across i and normalizes the resulting matrices so that each column sum to 1. In this way the super-node updates the values of the global matrices $P(v|z)$ and $P(t|z)$, which are once again transmitted to all nodes.

7. Using the updated global matrices $P(v|z)$ and $P(t|z)$ and the corresponding $P^i(d)$ and $P^i(z|d)$ each node re-calculates the new local joint probability distribution

$\acute{P}^i(z|d,v,t)$ according to eq.(5.11) and estimates the new local value of $\acute{E}^i[L]$ according to eq.(5.10). As in step 4, the super node sums the $\acute{E}^i[L]$ values collected from all nodes in order to calculate the new central $\acute{E}[L]$ value.

8. Using $\acute{E}[L]$ and $E[L]$ the super-node checks whether the convergence criterion of eq. (5.13) is satisfied. If yes, the local matrices $P^i(z|d)$ from all nodes are collected and concatenated in order to re-assemble the global matrix $P(z|d)$. If not, the process continues with Step 4.

By adopting this model for the distributed calculation of high order pLSA the benefit is twofold. Firstly, the fact that there is no need for communication or concurrent memory access between the nodes, allows them to run in parallel and synchronize only when they need to communicate with the super node. This parallel computation allow us to expect a reduction of the total computational time by a factor that approximates the number of cores offered by the utilized processor. Secondly, the proposed distributed model provides an elegant way for regulating the memory requirements of the algorithm independently of the dataset size. Indeed, given that in a non-parallel mode the minimum amount of data that should be loaded into RAM is bounded by $P^i(z|d,v,t)$ instead of $P(z|d,v,t)$, allow us to implement a version of the model that fits the memory specifications of the utilized computer. This can be done by using more nodes with smaller size or vice versa. In section 5.5.3.4 of our experimental study we measure the gain in computational cost of the distributed calculation model and show how we can regulate our algorithm to process a significantly large set of images.

Finally, we should mention that apart from dealing with computational and memory limitations, the distributed calculation model is also suggested for applications where data sources are distributed over a network and collecting all data at a central location is not a viable option. These applications include privacy-preserving environments where each node is only allowed to share a sub-set or an encoded representation of the local data, as well as sensor networks where each node collects a set of observations and needs to design local processing rules that perform at least as well as global ones, which rely on all observations being centrally available.

5.5 Experimental Study

Our experimental evaluation is primarily focused on comparing the performance achieved by the different feature spaces described in Section 5.3, in an image retrieval setting. Our aim is to verify that by exploiting the multi-modal nature of tagged images and introducing the cross-word dependencies when performing the modality fusion, we succeed in defining a feature space that is more sensitive to semantics. We also verify the efficiency of our approach in handling tasks of varying requirements by evaluating its performance in image clustering. Moreover, we experimentally measure the gain in computational cost achieved by the distributed calculation model and show how we can significantly reduce the memory requirements of our algorithm and run high order pLSA on a significantly large set of images. Finally, we compare our work with two state-of-the-art approaches that are also oriented towards exploiting the multi-modal nature of tagged images for improving the performance of an image retrieval system

5.5.1 Data set

To carry out our evaluation we have used the NUS-WIDE dataset[1] that was created by the NUS's Lab for Media Search [174]. The dataset contains 269,648 images that have been downloaded from flickr together with their tags. For all images the authors released 500-dimensional co-occurrence vectors for visual words (as described in Section 5.3.1.1), as well as 1000-dimensional co-occurrence vectors for tag-words (as described in Section 5.3.1.2). Moreover, the ground-truth for all images with respect to 81 concepts has been provided to facilitate evaluation. The full set of 269,648 images has been split by the authors to 161,789 train and 107,859 test images. For the purposes of our work we have used a sub-sample of 5,000 (I^{train}) images for training and 5,000 (I^{test}) images for testing. The selection was random, however in order to remove the effects of incomplete tagging and noisy annotation, we have used an additional restriction so as to select images with at least one concept present in their annotation info and at least one tag present in their tag-based representation.

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

5.5.2 Evaluation protocol

The adopted evaluation protocol for image retrieval is implemented as follows. Initially a set of training images is utilized to learn all necessary parameters that require training, such as the latent topics of simple and high order pLSA, as well as to calculate the cross-words dependencies between the visual and tag words. Subsequently, using the learned parameters the training images are indexed . Finally, an independent set of testing images is used to query the index and evaluate the system's performance based on the relevance between the query and the retrieved images.

For assessing the relevance between the query and the retrieved images we have used the Average Precision (AP) metric. AP favors the algorithms that are able not only to retrieve the correct images, but to retrieve them as early as possible in a ranked list of results. This is a crucial attribute for an image retrieval systems since users rarely take the time to browse through the results beyond the first pages. Average precision is expressed by the following equation.

$$AP = \frac{\sum_{r=1}^{N} Pr(r) \cdot rel(r)}{\#\text{ relevant images}} \tag{5.16}$$

where r is the current rank, N is the number of retrieved images, $rel()$ is a binary function that determines the relevance of the image at the given rank with the query image. $rel()$ outputs 1 if the image in the given rank is annotated with at least one concept in common with the query image and 0 otherwise. $Pr(r)$ is the precision at rank r and is calculated by:

$$Pr(r) = \frac{\#\text{ relevant retrieved images of rank r or less}}{r} \tag{5.17}$$

AP measures the retrieval performance of the method using one image as query. Finally, in order to facilitate fast image matching the images were indexed using a kd-tree multidimensional indexing structure [177] that supports k-NN (Nearest Neighbor) queries.

Apart from the AP and in order to evaluate the efficiency of our algorithm in a task different from image retrieval, we have also implemented the Normalized Mutual Information (NMI) measure for clustering comparison. NMI belongs to the class of information theoretic based measures that rely on the mutual information shared between two random variables. The mutual information measures how much knowing one

of these variables reduces the uncertainty about the other, which makes it appropriate for measuring the similarity between two clustering solutions. The NMI measure that we have used in our work is a normalized version of the Mutual Information defined as:

$$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{5.18}$$

where $I(\cdot)$ calculates the Mutual Information between the clustering solutions U and V and $H(\cdot)$ calculates the information entropy of each solution. NMI takes the value of 1 when the two clustering solutions are identical and 0 when they are independent. More information about NMI and how this measure can be applied to compare two clustering solutions can be found in [178].

5.5.3 Results

5.5.3.1 Retrieval performance

In order to obtain one global performance score for all query images we employed the Mean Average Precision (MAP) score, which is the mean of AP scores over the full set of query images. In our experiments we have set the value of N to be equal with the total number of indexed images. As baseline we have used the performance scores obtained using the 6 different feature spaces described in Section 5.3 namely *visual-words*, *tag-words*, *visualtag-words*, *plsavisual-words*, *plsatag-words* and *plsavisual_plsatag-words*. The performance score for the proposed approach appears under *highOrder-plsa*. The number of topics in all cases involving aspect models was selected to be 30, except from the *plsavisual_plsatag-words* case where the concatenation of the uni-modal plsa models resulted in a dimensionality of 60 topics. Moreover, in order to counterbalance the effect of initial randomization all experiments involving aspect models were repeated 5 times to obtain an average performance value.

Table 5.1 shows the MAP scores for all evaluated feature spaces. We notice that *visual-words* being a more dense representation of the image content performs better than *tag-words*, which are typically very sparse. As expected, the straightforward combination of both modalities by simply concatenating their word count vectors *visualtag-words*, fails to combine them efficiently and performs slightly better than the best of the uni-modal cases. When moving to the space of pLSA-based latent topics we can see an increase of the retrieval performance for both uni-modal cases, which verifies the

Table 5.1: Performance scores for image retrieval

Feature Space	#dims	MAP (%)
tag-words	1000	29,45
visual-words	500	31,07
visualtag-words	1500	31,08
plsatag-words	30	35,674
plsavisual-words	30	31,728
plsavisual_plsatag-words	60	35,906
highOrder-plsa	**30**	**37,75**

efficiency of aspect models to discover semantic relations between the images. Moreover, it is interesting to note that the relative improvement achieved by *plsatag-words* is considerably higher than the relative improvement of *plsavisual-words*. This can be attributed to the ability of pLSA in more efficiently handling sparse data, since the co-occurrence table of tag-words is much more sparse than the corresponding table of visual words. Additionally, the performance achieved by *plsavisual_plsatag-words* introduces some improvement over the uni-modal cases, in contrast to the behavior of *visualtag-words*. This verifies the ability of the latent space to more efficiently combine the heterogeneous modalities of tagged images, compared to the original space of word counts. Finally, the performance achieved by the proposed method verifies the usefulness of cross-word dependencies in creating a semantics sensitive feature space. Indeed, we can see that *highOrder-plsa* outperforms all other cases that neglect this kind of dependencies, introducing an improvement of approximately 1.8% units over the best performing baseline.

In order to gain more insight into the retrieval performance of our system we have calculated the MAP on a concept basis. In order to do this, for each concept, we have used only the images depicting this concept to query the index. Then, the MAP score of this concept is calculated by averaging the AP scores obtained for each of the issued queries. Fig. 5.3 depicts the MAP scores achieved by the *plsavisual_plsatag-words* and *highOrder-plsa* approaches for the 30 concepts that appear more frequently in the NUS_WIDE dataset. We can see that the proposed *highOrder-plsa* approach outperforms the best performing baseline in 21 out of the 30 considered concepts.

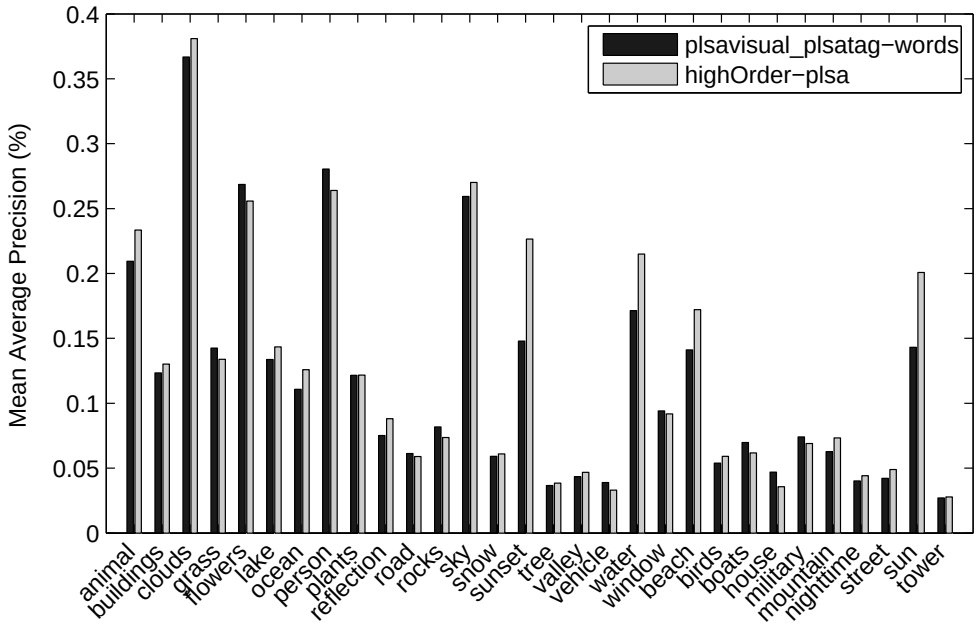


Figure 5.3: Performance scores on a concept-basis

5.5.3.2 Clustering Performance

In order to also evaluate the performance of the proposed approach on a different setting, we have designed an experiment where the task was to mine the conceptual categories characterizing the images included in the NUS_WIDE dataset. More specifically, the authors of [174] provide a concept list where each of the 81 annotation concepts is classified to one of six categories namely *Event/Activities*, *Program*, *Scene/Location*, *People*, *Objects* and *Graphics*. The task was to automatically identify these categories by performing clustering on the images included in our test set I^{test}. In each case one of the aforementioned feature spaces was used for calculating the distance similarity matrix. Then, NMI was employed to compare each of the obtained clustering solutions against the solution derived from the ground truth information. The L1-norm metric was used to calculate the similarity distance between images and the k-means algorithm was employed to perform clustering. In all cases, the number of requested clusters was set to be equal with the number of categories and 100 repetitions were imposed on the clustering process in order to alleviate the sensitivity of k-means to the

Table 5.2: Performance scores for image clustering

Feature Space	#dims	NMI
tag-words	1000	0.0448
visual-words	500	0.0164
visualtag-words	1500	0.0166
plsatag-words	30	0.06591
plsavisual-words	30	0.01977
plsavisual_plsatag-words	60	0.04809
highOrder-plsa	**30**	**0.07979**

initial conditions. The obtained results are depicted in Table 5.2.

It is evident from the NMI scores that the tag information space is more efficient in identifying the existing categories. Indeed, the clustering solutions obtained using *tag-words* and *plsatag-words* are much closer to the optimal solution than using *visual-words* and *plsavisual-words*, respectively. The poor performance of the visual information space is also observed in the cases of *visualtag-words* and *plsavisual_plsatag-words*, where the inclusion of visual-words in a joint space with tags has a negative effect on the clustering efficiency of the resulting space. Nevertheless, the use of cross-word dependencies by *highOrder-plsa* allows the resulting space to filter out the misleading information of visual words and obtain a clustering solution that is closer to the optimal case than all other baselines.

5.5.3.3 Latent space dimensionality and convergence threshold

In this experiment we investigate the impact of the employed latent space dimensionality on the retrieval performance of the pLSA-based methods. Our interest is on roughly estimating the number of dimensions where a performance peak is exhibited by each of the examined cases. Fig. 5.4 plots the MAP scores achieved by each method against the dimensionality of the latent space. We can see that the performance peak for *highOrder-plsa* appears between the range of $15 - 30$ dimensions. A similar kind of behavior is also exhibited by the uni-modal aspect models (i.e. *plsatag-words* and *plsavisual-words*) where the performance peak is located around the 30 dimensions. However, this is not the case for *plsavisual_plsatag-words* where the number of latent

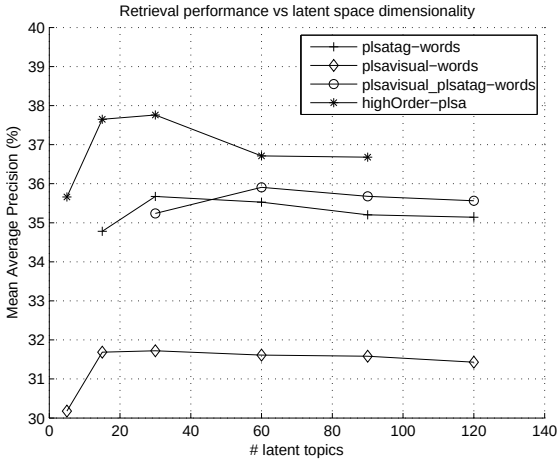


Figure 5.4: Impact of the latent space dimensionality on the retrieval performance

topics will have to reach 60 before achieving the peak of its performance. Thus, the fact that our approach reaches its performance peak using considerably fewer dimensions than the best performing baseline constitutes an additional argument in its favor. This is because the efficiency of the indexing mechanisms, which are typically employed in image retrieval systems, benefit substantially from the low dimensionality of the utilized feature space.

Another interesting aspect of the proposed algorithm is the relation between the convergence threshold employed during the EM procedure and the retrieval performance of the resulting feature space. As already mentioned in Section 5.3.3 the iterations of the EM algorithm stop when the value of eq.(5.10) becomes lower than a predefined threshold. In all experiments so far this threshold was set to 10^{-3}. Here, we evaluate the retrieval performance of the proposed approach using as convergence threshold the values 10^{-4}, 10^{-5} and 10^{-6}. Fig. 5.5 shows the MAP scores for each of the aforementioned values. It is evident that by making the convergence criterion more strict the retrieval performance of the resulting latent space increases. However, for values that are very close to zero (e.g. 10^{-5} and 10^{-6}) the improvement is only marginal and does not compensate for the increased computational overhead.

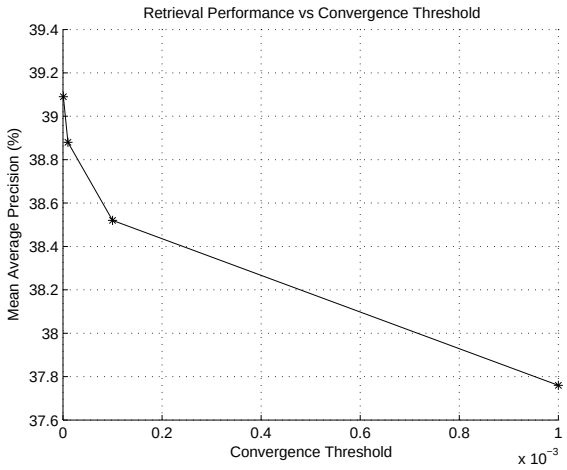


Figure 5.5: Impact of the convergence threshold on the retrieval performance

5.5.3.4 Distributed calculation model

In order to estimate the gain in computational cost achieved by the proposed distributed calculation model we have measured the time required by our algorithm to complete on an i7-950 processor with 4 physical cores and 12GBs of RAM, using the centralized and the distributed calculation model respectively. Moreover, for the distributed case we have considered two different configurations. In the first configuration we consider that the memory facilities of the utilized computer are adequate to load in RAM the 4-dimensional $P(z|d, v, t)$ array that derives from the processed dataset, while in the second case we consider that the memory required to load the $P(z|d, v, t)$ array exceeds the available resources. In this case the hard disk is used by each node to store and load the corresponding chunk $P^i(z|d, v, t)$ in every iteration. Table 5.3 demonstrates our experimental findings. We can see that the time required by our algorithm to complete reduces by a factor of ≈ 4 when employing the distributed calculation model, which is a reasonable outcome given that the whole process has been parallelized in 4 physical cores. On the other hand, when employing the configuration of the algorithm using the hard disk, the computational overhead introduced by read/write operations doubles the execution time but still remains considerably lower than the centralized version.

Finally, by exploiting the ability of the distributed calculation model to regulate its memory requirements, we have managed to apply the proposed high order pLSA algo-

Table 5.3: Execution time for different calculation models

Calculation model	Elapsed Time (sec)	
	Train	Test
Centralized	12288	2502
Distributed (Memory)	3563	349
Distributed (Disk)	6687	549

Table 5.4: Performance scores for image retrieval - Full NUS_WIDE Dataset

Feature Space	#dims	MAP (%)
tag-words	1000	29,90
visual-words	500	30,470
visualtag-words	1500	30,476
plsatag-words	30	35,512
plsavisual-words	30	31,128
plsavisual_plsatag-words	60	35,686
highOrder-plsa	**30**	**38,50**

rithm to the full set of images provided by the NUS_WIDE dataset. More specifically, we have applied high order pLSA on 121,920 train and 81,589 test images, which is the set that constitutes the full NUS_WIDE dataset after removing the images that did not satisfy the restrictions described in Section 5.5.1. Table 5.4 shows the obtained MAP scores. It is interesting to note that the improvements observed when moving from one feature space to another are equivalent to those observed in Table 5.1, showing that the conclusions we have drawn from our previous experiments can be considered valid.

5.5.4 Comparison with existing methods

In order to compare our work with state-of-the-art methods in multi-modal indexing we have generated two additional feature spaces by implementing the methods proposed in [1] and [71]. More specifically, we have implemented one of the variations presented in [1] that treats the visual and tag-based latent topics obtained from the application of the uni-modal pLSA, as the observed words for learning a second level pLSA model. This model (*ml-plsa* [1]) allows the image to be represented as a vector of meta-topics

Table 5.5: Performance scores for image retrieval

Feature Space	#dims	MAP (%)
ml-plsa [1]	30	35,956
mm-plsa [71]	30	34,162
highOrder-plsa	**30**	**37,75**

as illustrated in Fig. 5.6. Similarly, we have also implemented the multi-modal pLSA scheme (*mm-plsa*) presented in [71]. In this work the authors' goal is to exploit the interactions between the different modes when defining the latent space. However, in order to simplify their model, they assume that the pair of random variables representing the visual and tag words are conditional independent, given the respective image d_i. Given this assumption, we have $P(v|t,d) = P(v|d)$ and the joint probability model of text words, visual words and images can be written as:

$$P(d,v,t) = P(d)P(t|d)P(v|t,d) \Rightarrow P(d,v,t) = P(d)P(t|d)P(v|d) \tag{5.19}$$

Given eq. 5.19 we have used the EM-steps of Section 5.3.3 to generate a feature space for the *mm-plsa* model. Table 5.5 compares the performance of the three methods obtained using I^{train} and I^{test}.

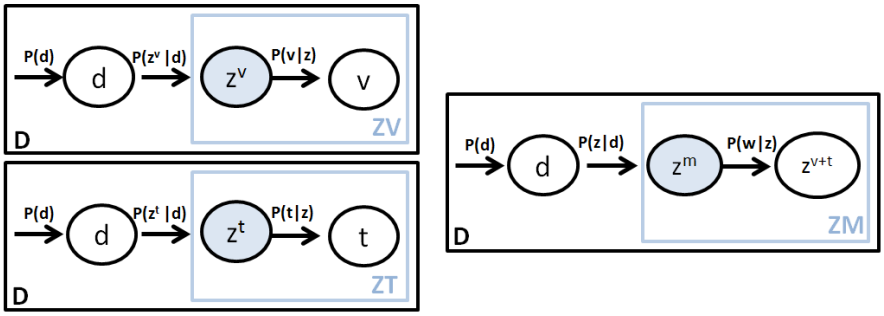


Figure 5.6: Graphical representation of the *ml-plsa* model [1]

The fact that the performance of *mm-plsa* model is lower than two of the baselines presented in Section 5.5.3.1 shows that there is important information neglected under the cross words independence assumption. In addition, the approximation of $P(d,v,t)$

without using the cross-words dependencies is misleading in the generation of a semantics sensitive latent space. On the other hand, the *ml-plsa* model manages to introduce some improvement over the best performing baseline of Section 5.5.3.1. However, the improvement is marginal showing the the second level pLSA has little to offer when applied on dense data (i.e. such as the data produced by the application of the first level of pLSA). Finally, the fact that *highOrder-plsa* outperforms all other methods shows that the use of cross-word dependencies is beneficial for combining the semantics of both visual and tag information space in a semantics sensitive latent space.

5.6 Discussion of our experimental findings

In discussing our results, we should stress the great potential of exploiting the information carried by the different modalities of tagged images when designing a semantics sensitive feature space. We have seen that almost all methods incorporating information from both modalities outperform their uni-modal counterparts. Moreover, the use of aspect models has proven to be an efficient solution for overcoming the heterogeneity of sources and even discover the hidden relations between images, as demonstrated in Sections 5.5.3.1, 5.5.3.2 and 5.5.3.4. Furthermore, the superior performance of *highOrder-pLSA* in all different settings has proven that, being different representations of the same abstract meaning, the visual and tag words appearing in the image content exhibit some cross-word dependencies that can be used to improve the effectiveness of the resulting feature space. In addition, the execution times observed in Section 5.5.3.4 advocate the development of distributed calculation models for high computational and memory demanding algorithms, since the multi-core facilities offered by modern processors are able to speed-up significantly the whole process. Finally, we should note that although the approach presented in this chapter performs fusion between the modalities of visual content and tags, a similar methodology can be used to incorporate additional modalities of social media such as geo-located or user-related information.

Chapter 6

Conclusions and Future Work

6.1 Discussion and Conclusions

In concluding this thesis we would like to provide a kind of walk-through to our motivations, the key choices that we have made in designing the proposed approaches and the conclusions we have reached. The starting point of our work was to tackle the inherent limitations of the example-based learning paradigm by proposing ways to smoothly incorporate into the learning process, knowledge that was provided explicitly in a human understandable format. In this effort BNs seemed like a natural choice due to their ability to incorporate explicit restrictions through their network structure and at the same time accommodate for the evidence that was extracted from the content itself. Two of our major problems were how to integrate the human expressed knowledge into the inference process and how to turn into evidence the support received from the low-level stimuli of multimedia content. Based on the approaches presented in Sections 3.1 and 3.2 we have managed to tackle those problems and develop a multimedia analysis framework based on evidence driven probabilistic inference. However, our experimental observations on the limited benefit of complex domain knowledge and the apparent requirement for plentiful annotated samples, was basically the motive for shifting our interest to social networks and their potential to serve as a rich source of low-cost, annotated samples.

Driven by the abundant availability of user tagged images in social networks, our expectation was that we would be able to obtain the desired annotated samples if we could exploit the noise reduction properties stemming from the collaborative nature of

their creation. Relying on a totally un-supervised method like clustering was favored for fulfilling the objective for low-cost annotated samples. Moreover, the solution that was dictated by our intuition was to associate the prevailing tendencies in both tag and visual information space, expecting to converge into the same object. Our theoretical and experimental study has indeed proven that this kind of convergence is possible especially when processing large volumes of data. However, it was clear from our experiments that although there were many cases where the gain in effort did actually compensate for the small loss in performance, the performance of the detectors trained using leveraged social media was inferior to the one achieved by manually trained detectors. This was one of our motives for seeking more efficient ways to exploit the tag information spaces in the multimedia analysis process.

The heterogeneous nature of visual features and tags was the main reason for putting aspect models at the core of our research efforts. However, the currently used aspect models like pLSA were unable to handle more than two observable variables, preventing us from incorporating the semantics of both visual and tag information space into a joint learning process. Thus, extending the currently used models to higher order became the obvious focus of our research, keeping also in mind to find the appropriate place in the analysis process where aggregated training information could be injected. By proposing high order pLSA and integrating the cross-modal dependencies into the update rules of EM execution, we have succeeded in increasing the semantic capacity of the resulting feature space.

Thus, recalling that the goal of our thesis was to verify that high levels of learning efficiency can only be achieved if explicit and implicit knowledge are combined efficiently, we may rightfully claim that the aforementioned statement has become evident from the approaches presented in Chapters 3, 4 and 5. In all cases the ability to incorporate more knowledge into the learning process has resulted in significant performance improvements. Indeed, in Chapter 3 we have seen the proposed BNs modeling approach to improve the performance of a set of baseline concept classifiers by using their output as evidence. Similarly, in Chapter 4 we have shown how the collective knowledge encoded in social media can be successfully used to boost the scalability of current object detection schemes, by removing the need for close human supervision. Finally, Chapter 5 has verified that the knowledge encoded in the cross modal dependencies of multi-modal data is particularly useful when designing a semantics sensitive

feature space. However, during the development of the proposed approaches we were faced with a series of tradeoffs that we discuss below.

The first trade-off that we encountered is between the advantages of using ontologies and the effort required to encode the relations between all different domain concepts. It is true that in order to generate a valid ontological representation of the examined domain, substantial effort will have to be allocated by domain experts. Similarly, in order to obtain realistic priors for the BN a sufficiently large amount of samples will have to be manually annotated, as verified by the experiment of Section 3.3.3.4. Thus, there is a trade-off between the time required to obtain a deep modeling of the domain and the efficiency of the resulting analysis framework. In searching the best option to balance this trade-off it is interesting to note that in the experiment of Section 3.3.3.2 we have seen that the incorporation of semantic constraints into the inference process was able to help only under certain circumstances. On the other hand the hierarchical information of the ontology was proven useful in all cases, as it became particularly obvious from the experiments of Section 3.4.2.2. Based on the above one general conclusion that we can draw is that expert's provided knowledge, even at a shallow depth, is complementary to the knowledge that can be mined from training samples. Thus, in designing semantic multimedia analysis systems, there should be a clear preference towards spending the available resources towards acquiring simple logic-based rules, rather than spending the resources to obtain more training samples. On the other hand, complex logic-based rules and deep modeling of the domain should only be employed in cases where the additional effort is justified by the added value in the application.

The second trade-off that we came across during the preparation of this thesis is between the available computational resources for handling large amounts of data and the amount of human intervention that can be exploited to drastically improve the visual analysis error. In our effort to crowdsource annotated samples by leveraging social media, we have seen that in order to ensure high probability of success for our approach the size of the processed dataset would have to grow particularly high. On the other hand, we have also observed that the aforementioned probability of success was very sensitive to the amount of error introduced by the visual analysis algorithms, which was roughly related with an invert proportional manner to the dataset size. This observation, although mostly intuitive, can be a very powerful tool in balancing

between the computational and human resources that are needed to achieve optimal performance. For instance, if instead of the mostly un-supervised process that was employed in Chapter 4 we have the luxury to include a human annotator in the analysis loop, we may as well achieve comparable results with a significantly smaller dataset [179]. On the other hand, if human intervention is completely out of the question, more computational resources can be committed to ensure the successful completion of the undertaken task.

Finally, the trade-off that we encountered in the last of the approaches presented in this thesis is also between the availability of computational and memory requirements against the gain in performance efficiency. By incorporating more than two observable variables into the analysis process, although beneficial from the perspective of incorporating the cross modal dependencies into the resulting latent space, we significantly raised the computational and memory requirements of our method. This is a problem that is usually encountered by methods that try to incorporate many sources of information into a joint learning process and becomes a serious obstacle when the complexity increases exponentially for every additional source, as in the case of Chapter 5. Typical solutions to this problem aim at exploiting the sparseness that is likely to characterize the processed data, or at designing distributed calculation models to exploit the multi-core facilities offered by modern systems, as in our case (cf. Section 5.4). However, when dealing with large scale problems like the ones that are usually addressed in social media, special attention will have to be paid on avoiding situations where the scale of the analysis that is necessary for extracting meaningful results, is so large that it renders the application of some methods intractable. Thus, in developing methods that are intended to work in the space of social media, there should be a clear design preference on algorithms with low complexity and even more on solutions that are able to trade computational complexity with small losses in performance. This can be done either by using rough approximations instead of accurate values, or by re-designing the algorithms to rely on pre-computed information that can be calculated once and in an off-line mode.

6.2 Plans for future extensions

Our plans for future research are motivated by the fact that social media are commonly described by a high diversity of features. For instance, an image in flickr is associated with the tags that have been assigned to it, the users that seem to like it and mark it as favorite, the visual features that describe the visual content of the image, and possibly the information that denotes the spatial and temporal context of this image. Even though all these facets of information are not combined naturally with each other, still they carry complementary knowledge about the resource since each facet is essentially the representation of the resource in a different feature space. We consider the efficient exploitation of such information to be an important advancement from traditional multimedia analysis methods, since managing the diversity of all available features and successfully exploiting their complementary information capacity poses new requirements and challenges. Based on the above our plans for future work include the extension of the approach presented in Chapter 5 to further exploit the intrinsic multi-modal nature of social media.

More specifically, given that the proposed high order pLSA method can in principal extend to an arbitrary high number of observable variables, we plan to research and develop a social media analysis framework that will be flexible in handling any combination of the aforementioned information facets. Important problems related to the heterogeneity of information, the computational and memory requirements, as well as the mechanism for correlating the different modalities will have to be addressed for achieving the necessary levels of efficiency. In addition, we plan to experiment further with the amount of information that can be extracted from legacy data and examine whether the incorporation of feature kernels into the update rules of high order pLSA will actually lead to a joined feature space of increased sematic capacity.

Bibliography

[1] Rainer Lienhart, Stefan Romberg, and Eva Hörster. **Multilayer pLSA for multimodal image retrieval**. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM. xi, 22, 155, 156

[2] Shih-Fu Chang. **The Holy Grail of Content-Based Media Analysis**. *IEEE MultiMedia*, **9**(2):6–10, 2002. 1

[3] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. **Content-Based Image Retrieval at the End of the Early Years**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**(12):1349–1380, 2000. 1

[4] Irving Biederman. **Recognition-by-components: A theory of human image understanding**. *Psychological Review*, **94**:115–147, 1987. 1

[5] Fei-Fei Li, Robert Fergus, and Pietro Perona. **One-Shot Learning of Object Categories**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**(4):594–611, 2006. 2

[6] Jia Li and James Z. Wang. **Real-Time Computerized Annotation of Pictures**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**(6):985–1002, 2008. 2

[7] Antoine Bordes. *New Algorithms for Large-Scale Support Vector Machines*. PhD thesis, pour obtenir le Grade de Docteur en Sciences de lUniversit?e Paris VI – Pierre et Marie Curie, 2010. 2

[8] Antonio Torralba, Rob Fergus, and William T. Freeman. **80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**:1958–1970, 2008. 2

[9] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos. **Supervised Learning of Semantic Classes for Image Annotation and Retrieval**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**(3):394–410, 2007. 2, 95

[10] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. **Solving the multiple instance problem with axis-parallel rectangles**. *Artif. Intell.*, **89**(1-2):31–71, 1997. 2

[11] Pinar Duygulu, Kobus Barnard, João F. G. de Freitas, and David A. Forsyth. **Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary**. In *ECCV (4)*, pages 97–112, 2002. 3, 95

[12] Luis von Ahn and Laura Dabbish. **Labeling images with a computer game**. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM. 3

[13] Luis von Ahn, Ruoran Liu, and Manuel Blum. **Peekaboom: a game for locating objects in images**. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64, New York, NY, USA, 2006. ACM. 3

[14] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd edition, 2001. 4

[15] Aude Oliva and Antonio Torralba. **Building the gist of a scene: the role of global image features in recognition**. In *Progress in Brain Research*, page 2006, 2006. 13

[16] Demir Gokalp and Selim Aksoy. **Scene Classification Using Bag-of-Regions Representations**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **0**:1–8, 2007. 14

BIBLIOGRAPHY

[17] Andreas Opelt, Axel Pinz, and Andrew Zisserman. **Incremental learning of object detectors using a visual shape alphabet**. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3–10, Washington, DC, USA, 2006. IEEE Computer Society. 14

[18] T. Blaschke. **Object-based contextual image classification built on image segmentation**. In *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003*, pages 113 – 119, oct. 2003. 14

[19] Zhiyong Wang, Dagan Feng, and Zheru Chi. **Region-based binary tree representation for image classification**. In *Proceedings of the International Conference on Neural Networks and Signal Processing, 2003.*, **1**, pages 232 – 235 Vol.1, dec. 2003. 14

[20] Jianping Fan, Yuli Gao, and Hangzai Luo. **Multi-level annotation of natural scenes using dominant image components and semantic concepts**. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 540–547, New York, NY, USA, 2004. ACM. 14

[21] Changbo Yang, Ming Dong, and Farshad Fotouhi. **Region based image annotation through multiple-instance learning**. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 435–438, New York, NY, USA, 2005. ACM. 14

[22] Tomasz Malisiewicz and Alexei A. Efros. **Recognition by association via learning per-exemplar distances**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **0**:1–8, 2008. 14

[23] P. Murphy, A. Torralba, and W. T. Freeman. **Using the forest to see the trees: a graphical model relating features, objects and scenes**. In *in Advances in Neural Information Processing Systems 16 (NIPS)*, Vancouver, 2003. BC: MIT Press. 14

[24] Amit Singhal, Jiebo Luo, and Weiyu Zhu. **Probabilistic Spatial Context Models for Scene Content Understanding**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **1**:235, 2003. 14

[25] M.R. Boutell, J. Luo, and C.M. Brown. **Improved semantic region labeling based on scene context**. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005.*, page 4 pp., july 2005. 14

[26] Lin Yang, Peter Meer, and David J. Foran. **Multiple Class Segmentation Using A Unified Framework over Mean-Shift Patches**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **0**:1–8, 2007. 14

[27] Matthew Boutell, Jiebo Luo, and Christopher M. Brown. **A generalized temporal context model for classifying image collections**. *Multimedia Syst.*, **11**(1):82–92, 2005. 14

[28] Lucas Paletta, Manfred Prantl, and Axel Pinz. **Learning Temporal Context in Active Object Recognition Using Bayesian Analysis**. *International Conference on Pattern Recognition,*, **1**:1695, 2000. 15

[29] Dan Moldovan, Christine Clark, and Sanda Harabagiu. **Temporal context representation and reasoning**. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1099–1104, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc. 15

[30] M. Boutell and J. Luo. **Bayesian fusion of camera metadata cues in semantic scene classification**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004*, **2**, pages II–623 – II–630 Vol.2, june-2 july 2004. 15

[31] Xuming He, Richard S. Zemel, and Miguel. Carreira-Perpinan. **Multiscale Conditional Random Fields for Image Labeling**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**:695–702, 2004. 15

[32] Sanjiv Kumar and Martial Hebert. **A Hierarchical Field Framework for Unified Context-Based Classification**. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1284–1291, Washington, DC, USA, 2005. IEEE Computer Society. 15

[33] Stephen Gould, Richard Fulton, and Daphne Koller. **Decomposing a Scene into Geometric and Semantically Consistent Regions**. In *Proceedings of the IEEE International Conference on Computer Vision*, Kyoto, Japan, September 2009. 15

[34] Milind R. Naphade, Trausti T. Kristjansson, Brendan J. Frey, and Thomas S. Huang. **Probabalistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems**. In *ICIP (3)*, pages 536–540, 1998. 15

[35] Milind R. Naphade and Thomas S. Huang. **A probabilistic framework for semantic video indexing, filtering, and retrieval**. *IEEE Transactions on Multimedia*, **3**(1):141–151, 2001. 15

[36] Jiebo Luo, Andreas E. Savakis, and Amit Singhal. **A Bayesian network-based framework for semantic image understanding**. *Pattern Recognition*, **38**(6):919–934, 2005. 15

[37] Phivos Mylonas, Evaggelos Spyrou, Yannis Avrithis, and Stefanos Kollias. **Using visual context and region semantics for high-level concept detection**. *IEEE Transanctions on Multimedia*, **11**(2):229–243, 2009. 15

[38] Michael J. Kane and Andreas E. Savakis. **Bayesian Network Structure Learning and Inference in Indoor vs. Outdoor Image Classification**. In *ICPR (2)*, pages 479–482, 2004. 16

[39] Leonardo Nogueira Matos and João Marques de Carvalho. **Combining global and local classifiers with Bayesian network**. In *ICPR (4)*, page 952, 2006. 16

[40] Sinisa Todorovic and Michael C. Nechyba. **Interpretation of complex scenes using dynamic tree-structure Bayesian networks**. *Comput. Vis. Image Underst.*, **106**(1):71–84, 2007. 16

[41] Wenhui Liao and Qiang Ji. **Learning Bayesian network parameters under incomplete data with domain knowledge**. *Pattern Recogn.*, **42**(11):3046–3056, 2009. 16

[42] Zhongli Ding, Yun Peng, and Rong Pan. **A Bayesian Approach to Uncertainty Modeling in OWL Ontology**. In *Proceedings of the International Conference on Advances in Intelligent Systems - Theory and Applications*, November 2004. 16, 33, 34, 35, 36, 37

[43] G. Th. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. **Combining global and local information for knowledge-assisted image analysis and classification**. *EURASIP J. Adv. Sig. Proc.*, **2007**(2):18–18, 2007. 16, 17, 46

[44] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias. **Semantic Image Segmentation and Object Labeling**. *IEEE Transactions on Circuits and Systems for Video Technology*, **17**(3):298–312, March 2007. 16, 17

[45] Jianping Fan, Yuli Gao, and Hangzai Luo. **Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classifier Training for Multilevel Image Annotation**. *IEEE Transactions on Image Processing*, **17**(3):407–426, 2008. 16, 17

[46] Christopher Town. **Ontological inference for image and video analysis**. *Mach. Vis. Appl.*, **17**(2):94–115, 2006. 17

[47] Cees G.M. Snoek and Marcel Worring. **Multimodal Video Indexing: A Review of the State-of-the-art**. *Multimedia Tools and Applications*, **25**:5–35, 2003. 17

[48] Timothy M. Hospedales and Sethu Vijayakumar. **Structure Inference for Bayesian Multisensory Scene Understanding**. *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, **30**(12):2140–2157, 2008. 17

BIBLIOGRAPHY

[49] Kihoon Choi, S. Singh, A. Kodali, K.R. Pattipati, J.W. Sheppard, S.M. Namburu, S. Chigusa, D.V. Prokhorov, and Liu Qiao. **Novel Classifier Fusion Approaches for Fault Diagnosis in Automotive Systems**. *Instrumentation and Measurement, IEEE Transactions on*, **58**(3):602–611, March 2009. 17

[50] Zhihong Zeng, Jilin Tu, B.M. Pianfetti, and T.S. Huang. **AudioVisual Affective Expression Recognition Through Multistream Fused HMM**. *IEEE Transactions on Multimedia*, **10**(4):570–577, June 2008. 18

[51] Hao Pan, S.E. Levinson, T.S. Huang, and Zhi-Pei Liang. **A fused hidden Markov model with application to bimodal speech processing**. *IEEE Transactions on Signal Processing*, **52**(3):573–581, March 2004. 18

[52] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. **Early versus late fusion in semantic video analysis**. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM. 18

[53] Steven C.H. Hoi and Michael R. Lyu. **A Multi-Modal and Multi-Level Ranking Scheme for Large-Scale Video Retrieval**. *IEEE Transactions on Multimedia*, **10**(4):607–619, 2008. 18

[54] Zhiyong Wang, David D. Feng, Zheru Chi, and Tian Xia. **Annotating Image Regions Using Spatial Context**. *Multimedia, International Symposium on*, **0**:55–61, 2006. 18

[55] Zheshen Wang, Ming Zhao, Yang Song, S. Kumar, and Baoxin Li. **YouTubeCat: Learning to categorize wild web videos**. pages 879 –886, jun. 2010. 18

[56] Sonal Gupta, Joohyun Kim, Kristen Grauman, and Raymond J. Mooney. **Watch, Listen & Learn: Co-training on Captioned Images and Videos**. In *ECML/PKDD (1)*, pages 457–472, 2008. 19

[57] Jianping Fan, A.K. Elmagarmid, Xingquan Zhu, W.G. Aref, and Lide Wu. **ClassView: hierarchical video shot classification, indexing, and accessing**. *IEEE Transactions on Multimedia*, **6**(1):70–86, Feb. 2004. 19

[58] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. **Multimodal Fusion for Video Search Reranking**. *IEEE Transactions on Knowledge and Data Engineering*, **99**(PrePrints), 2009. 19

[59] Shiau Hong Lim, Li-Lun Wang, and Gerald DeJong. **Integrating prior domain knowledge into discriminative learning using automatic model construction and phantom examples**. *Pattern Recogn.*, **42**(12):3231–3240, 2009. 19, 20

[60] Lyndon S. Kennedy, Shih-Fu Chang, and Igor Kozintsev. **To search or to label?: predicting the performance of search-based automatic image classifiers**. In *Multimedia Information Retrieval*, pages 249–258, 2006. 20

[61] Timothee Cour, Ben Sapp, Chris Jordan, and Ben Taskar. **Learning from Ambiguously Labeled Images**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09)*, 2009. 20

[62] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. **Flickr distance**. In *ACM Multimedia*, pages 31–40, 2008. 20

[63] Yongqing Sun, Satoshi Shimada, Yukinobu Taniguchi, and Akira Kojima. **A novel region-based approach to visual concept modeling using web images**. In *ACM Multimedia*, pages 635–638, 2008. 20, 21

[64] Theodora Tsikrika, Christos Diou, Arjen P. de Vries, and Anastasios Delopoulos. **Image annotation using clickthrough data**. In *8th ACM International Conference on Image and Video Retrieval*, Santorini, Greece, 8-10 July 2009. 20, 21

[65] Lyndon S. Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. **How flickr helps us make sense of the world: context and content in community-contributed media collections**. In *ACM Multimedia*, pages 631–640, 2007. 21

[66] Till Quack, Bastian Leibe, and Luc J. Van Gool. **World-scale mining of objects and events from community photo collections**. In *CIVR*, pages 47–56, 2008. 21, 95

[67] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. **Matching Words and Pictures**. *Journal of Machine Learning Research*, **3**:1107–1135, 2003. 22

[68] Li-Jia Li, Richard Socher, and Li Fei-Fei. **Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 22

[69] Thomas Hofmann. **Unsupervised Learning from Dyadic Data**. pages 466–472. MIT Press, 1998. 22, 138

[70] Stefan Romberg, Eva Hörster, and Rainer Lienhart. **Multimodal pLSA on visual features and tags**. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, pages 414–417, Piscataway, NJ, USA, 2009. IEEE Press. 22

[71] Chandrika Pulla and C. V. Jawahar. **Multi modal semantic indexing for image retrieval**. In *Conference on Image and Video Retrieval*, pages 342–349, 2010. 23, 24, 155, 156

[72] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. **Tag refinement by regularized LDA**. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 573–576, New York, NY, USA, 2009. ACM. 23

[73] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. **Latent dirichlet allocation**. *J. Mach. Learn. Res.*, **3**:993–1022, March 2003. 23, 134

[74] Sergej Sizov. **GeoFolk: latent spatial semantics in web 2.0 social media**. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 281–290, New York, NY, USA, 2010. ACM. 23

[75] Xirong Li, Cees G.M. Snoek, and Marcel Worring. **Learning tag relevance by neighbor voting for social image retrieval**. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 180–187, New York, NY, USA, 2008. ACM. 23

[76] Xirong Li, Cees G. M. Snoek, and Marcel Worring. **Learning social tag relevance by neighbor voting**. *Trans. Multi.*, **11**:1310–1322, November 2009. 23

[77] Dong Liu, Shuicheng Yan, Yong Rui, and Hong-Jiang Zhang. **Unified tag analysis with multi-edge graph**. In *Proceedings of the international conference on Multimedia*, MM '10, pages 25–34, New York, NY, USA, 2010. ACM. 23

[78] Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. **Learning to tag**. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 361–370, New York, NY, USA, 2009. ACM. 23

[79] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. **An efficient boosting algorithm for combining preferences**. *J. Mach. Learn. Res.*, **4**:933–969, December 2003. 24

[80] Tamara G. Kolda and Brett W. Bader. **Tensor Decompositions and Applications**. *SIAM Review*, **51**(3):455–500, September 2009. 24

[81] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. **Tag recommendations based on tensor dimensionality reduction**. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 43–50, New York, NY, USA, 2008. ACM. 24

[82] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. **A Multilinear Singular Value Decomposition**. *SIAM J. Matrix Anal. Appl.*, **21**(4):1253–1278, 2000. 24

[83] Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. **TripleRank: Ranking Semantic Web Data by Tensor Decomposition**. In *ISWC '09: Proceedings*

of the 8th International Semantic Web Conference, pages 213–228, Berlin, Heidelberg, 2009. Springer-Verlag. 24

[84] Richard A. Harshman and Margaret E. Lundy. **PARAFAC: Parallel factor analysis**. *Computational Statistics & Data Analysis*, **18**(1):39 – 72, 1994. 24

[85] Finn V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001. 26

[86] William M. Bolstad. *Introduction to Bayesian Statistics*. John Wiley, 2004. 26

[87] Pearl. **Fusion, Propagation, and Structuring in Belief Networks**. *Artif. Intell.*, **29**(3):241–288, 1986. 27, 28

[88] S. L. Lauritzen and D. J. Spiegelhalter. **Local computations with probabilities on graphical structures and their application to expert systems**. pages 415–448, 1990. 28

[89] F. V. Jensen and F. Jensen. **Optimal junction trees**. In CA: Morgan Kaufmann, editor, *Proc. of the 10th Conf. on Uncertainty in Artif. Intel.*, San Mateo, 1994. 28, 43, 74

[90] Jorge Cardoso. **The Semantic Web Vision: Where Are We?** *IEEE Intelligent Systems*, **22**(5):84–88, 2007. 31, 33

[91] Ian Horrocks. **Description Logics in Ontology Applications**. In *Automated Reasoning with Analytic Tableaux and Related Methods*, pages 2–13. 2005. 32

[92] Deborah L. McGuinness and Frank van Harmelen. **OWL Web Ontology Language Overview**. W3C recommendation, W3C, February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/. 32

[93] Ian Horrocks and Peter F. Patel-Schneider. **A proposal for an owl rules language**. In *WWW*, pages 723–731, 2004. 32

[94] Solomon Kullback and Richard A. Leibler. **On Information and Sufficiency**. *The Annals of Mathematical Statistics*, **22**(1):79–86, 1951. 36

[95] Jakob Verbeek and Bill Triggs. **Region Classification with Markov Field Aspect Models**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **0**:1–8, 2007. 36, 62, 65, 129, 130

[96] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. *TextonBoost*: **Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation**. In *ECCV (1)*, pages 1–15, 2006. 36, 65, 129, 130

[97] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley and Sons, 2nd edition, 1997. 36, 74, 139, 142

[98] S. Nikolopoulos, G.T. Papadopoulos, I. Kompatsiaris, and I. Patras. **Evidence-Driven Image Interpretation by Combining Implicit and Explicit Knowledge in a Bayesian Network**. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **41**(5):1366 –1381, oct. 2011. 38

[99] Spiros Nikolopoulos, Georgios Th. Papadopoulos, Ioannis Kompatsiaris, and Ioannis Patras. **An Evidence-Driven Probabilistic Inference Framework for Semantic Image Understanding**. In *Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '09, pages 525–539, Berlin, Heidelberg, 2009. Springer-Verlag. 38

[100] Spiros Nikolopoulos, Georgios Th. Papadopoulos, Ioannis Kompatsiaris, and Ioannis Patras. **Image Interpretation by Combining Ontologies and Bayesian Networks**. In Ilias Maglogiannis, Vassilis Plagianakos, and Ioannis Vlahavas, editors, *Artificial Intelligence: Theories and Applications*, **7297** of *Lecture Notes in Computer Science*, pages 307–314. Springer Berlin Heidelberg, 2012. 38

[101] Jozsef A. Toth. *Reasoning agents in a dynamic world: The frame problem.*, **73** of *Artificial Intelligence*. Elsevier, 1995. 41

[102] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada. **Colour and texture descriptors**. *IEEE Trans. Circuits and Systems*

for Video Technology, Special Issue on MPEG-7, **11**(6):703–715, Jun 2001. 46

[103] N. Murphy T. Adamek, N. O'Connor. **Region-based Segmentation of Images Using Syntactic Visual Features**. In *Workshop on Image Analysis for Multimedia Interactive Services, (WIAMIS)*, Montreux, Switzerland, 2005. 46

[104] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. **Columbia Universitys Baseline detectors for 374 LSCOM Semantic Visual Concepts**. Advent technical report 222-2006-8, Columbia University, March 2007. 46, 47, 48, 61

[105] C.-C. Chang and C.-J. Lin. **LIBSVM: a Library for Support Vector Machines**, 2001. 46

[106] Christopher J. C. Burges. **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Min. Knowl. Discov.*, **2**(2):121–167, 1998. 46

[107] D.M.J. Tax and R.P.W. Duin. **Using two-class classifiers for multiclass classification**. In *International Conference on Pattern Recognition*, Quebec, Canada, 2002. 46

[108] NIST. **TREC video retrieval evaluation (TRECVID)**, 2001-2006. 47

[109] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. **A light scale concept ontology for multimedia understanding for TRECVID 2005**. Technical report, IBM, 2005. 47

[110] R. Neapolitan. *Learning bayesian networks*. Prentice Hall Upper Saddle River, NJ, 2003. 51

[111] C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, London,, 1979. 52

[112] Eric D. Taillard, Philippe Waelti, and Jacques Zuber. **Few statistical tests for proportions comparison**. *European Journal of Operational Research*, **185**(3):1336–1350, March 2008. 55

[113] Feng Ge, Song Wang, and Tiecheng Liu. **Image-Segmentation Evaluation From the Perspective of Salient Object Extraction**. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1146–1153, Washington, DC, USA, 2006. IEEE Computer Society. 62

[114] Milind R. Naphade and Thomas S. Huang. **A probabilistic framework for semantic video indexing, filtering, and retrieval**. *IEEE Transactions on Multimedia*, **3**(1):141–151, 2001. 66

[115] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. **Multi-Modal Classifier Fusion for Video Shot Content Retrieval**. In *In Proceedings of the 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, April 2005. 66

[116] V. Mihajlovic, M. Petkovic, W. Jonker, and H.M. Blanken. **Multimodal content-based video retrieval**. In H.M. Blanken, A.P. de Vries, H.E. Blok, and L. Feng, editors, *Multimedia Retrieval*, Data-Centric Systems and Applications, pages 271–294. Springer Verlag, Berlin, 2007. 66

[117] G.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders. **The Semantic Pathfinder: Using an Authoring Metaphor for Generic Multimedia Indexing**. *IEEE Transactions on Pattern. Anal. and Mach. Intel.,*, **28**(10):1678–1689, Oct. 2006. 66

[118] W. H. Adams, Giridharan Iyengar, Ching-Yung Lin, Milind Ramesh Naphade, Chalapathy Neti, Harriet J. Nock, and John R. Smith. **Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues**. *EURASIP Journal on Applied Signal Processing*, **2003**(2):170–185, 2003. 66

[119] Martin Možina, Claudio Giuliano, and Ivan Bratko. **Arguments Extracted from Text in Argument Based Machine Learning**. In *Proceedings of 1st Asia Conference on Intelligent Information and Database Systems*, 2009. 66

[120] Joao Magalhaes and Stefan Rüger. **Information-theoretic semantic multimedia indexing**. In *CIVR '07*, pages 619–626, New York, USA, 2007. ACM. 66, 134

[121] Z. Wu and M. Palmer. **Verm semantics and lexical selection**. In *Proceedings of the 32nd annual meeting of the association for computational linguistics*, pages 133–138, New Mexiko, USA, 1994. 66

[122] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. **Multimedia content processing through cross-modal association**. In *MULTIMEDIA '03*, pages 604–611, New York, USA, 2003. ACM. 66, 134

[123] Spiros Nikolopoulos, Christina Lakka, Ioannis Kompatsiaris, Christos Varytimidis, Konstantinos Rapantzikos, and Yannis Avrithis. **Compound Document Analysis by Fusing Evidence Across Media**. *International Workshop on Content-Based Multimedia Indexing*, **0**:175–180, 2009. 66

[124] Christina Lakka, Spiros Nikolopoulos, Christos Varytimidis, and Ioannis Kompatsiaris. **A Bayesian network modeling approach for cross media analysis**. *Signal Processing: Image Communication*, **26**(3):175 – 193, 2011. 66

[125] Jose Iria, Spiros Nikolopoulos, and Martin Mozina. **Cross-Media Knowledge Extraction in the Car Manufacturing Industry**. *IEEE International Conference on Tools with Artificial Intelligence*, **0**:219–223, 2009. 67

[126] A. Laender, B. Ribeiro-Neto, A. Silva, and J. Teixeira. **A Brief Survey of Web Data Extraction Tools**. In *SIGMOD Record*, **31**, June 2002. 68

[127] A. Arasu and A. H. Garcia-Molina. **Extracting structured data from Web pages**. In *ACM SIGMOD International Conference on Management of Data*, San Diego, California, USA, 2003. 68

[128] B. Rosenfeld, R. Feldman, and J. Aumann. **Structural extraction from visual layout of documents**. In *ACM Conference on Information and Knowledge Management (CIKM)*, 2002. 68

[129] Paul Viola and Michael Jones. **Rapid Object Detection using a Boosted Cascade of Simple Features**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **1**:511, 2001. 69

[130] Yoav Freund and Robert E. Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. In *Proceedings of the Second European Conference on Computational Learning Theory*, pages 23–37, London, UK, 1995. Springer-Verlag. 69

[131] Constantine P. Papageorgiou, Michael Oren, and Tomaso Poggio. **A General Framework for Object Detection**. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 555, Washington, DC, USA, 1998. IEEE Computer Society. 69

[132] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. 71, 137

[133] Claudia Leacock and Martin Chodorow. **Combining local context with WordNet similarity for word sense identification**. In Christiane Fellbaum, editor, *WordNet: A Lexical Reference System and its Application*. The MIT Press, 1998. 72

[134] J. J. Jiang and D. W. Conrath. **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**. In *International Conference Research on Computational Linguistics*, September 1997. 72

[135] Dekang Lin. **An Information-Theoretic Definition of Similarity**. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. 72

[136] G. Hirst and D. St. Onge. **Lexical chains as representations of context for the detection and correction of malapropisms**. In Christiane Fellbaum, editor, *WordNet: A Lexical Reference System and its Application*. The MIT Press, 1998. 72

[137] Philip Resnik. **Using Information Content to Evaluate Semantic Similarity in a Taxonomy**. In *IJCAI*, pages 448–453, 1995. 72

[138] Satanjeev Banerjee. **Extended gloss overlaps as a measure of semantic relatedness**. In *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003. 72

[139] Siddharth Patwardhan. *Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness*. Master's thesis, August 2003. 72

[140] Gregory F. Cooper and Edward Herskovits. **A Bayesian Method for the Induction of Probabilistic Networks from Data**. *Mach. Learn.*, **9**(4):309–347, 1992. 74, 84

[141] B. Scholkopf, A. Smola, R. Williamson, and P. Bartlett. **New support vector algorithms**. *Neural Networks*, **22**:1083–1121, 2000. 80, 88, 105

[142] Thorsten Joachims. **Transductive Learning via Spectral Graph Partitioning**. In *In ICML*, pages 290–297, 2003. 81

[143] David Heckerman, Dan Geiger, and David M. Chickering. **Learning Bayesian networks: The combination of knowledge and statistical data**. *Machine Learning*, **20**:197–243, 1995. 10.1007/BF00994016. 84

[144] Koen van de Sande, Theo Gevers, and Cees Snoek. **Evaluating Color Descriptors for Object and Scene Recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **99**(1), 2008. 88, 103

[145] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. **Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study**. *Int. J. Comput. Vision*, **73**(2):213–238, 2007. 88

[146] David G. Lowe. **Distinctive Image Features from Scale-Invariant Keypoints**. *Int. J. Comput. Vision*, **60**(2):91–110, 2004. 88

[147] Josef Sivic and Andrew Zisserman. **Video Google: A Text Retrieval Approach to Object Matching in Videos**. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society. 88, 136

[148] **TREC Video Retrieval Evaluation Notebook Papers and Slides**, November 2010. 90

[149] Emine Yilmaz and Javed A. Aslam. **Estimating average precision when judgments are incomplete**. *Knowl. Inf. Syst.*, **16**(2):173–211, 2008. 91

[150] Spiros Nikolopoulos, Elisavet Chatzilari, Eirini Giannakidou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Athena Vakali. **Leveraging Massive User Contributions for Knowledge Extraction**. In Nik Bessis and Fatos Xhafa, editors, *Next Generation Data Technologies for Collective Computational Intelligence*, **352** of *Studies in Computational Intelligence*, pages 415–443. Springer Berlin / Heidelberg, 2011. 95

[151] Elisavet Chatzilari, Spiros Nikolopoulos, Ioannis Patras, and Ioannis Kompatsiaris. **Enhancing Computer Vision Using the Collective Intelligence of Social Media**. In Athena Vakali and Lakhmi Jain, editors, *New Directions in Web Data Management 1*, **331** of *Studies in Computational Intelligence*, pages 235–271. Springer Berlin / Heidelberg, 2011. 95

[152] E. Chatzilari, S. Nikolopoulos, I. Patras, and I. Kompatsiaris. **Leveraging social media for scalable object detection**. *Pattern Recognition*, **45**(8):2962 – 2979, 2012. 95

[153] E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris, E. Giannakidou, and A. Vakali. **Leveraging social media for training object detectors**. In *Proceedings of the 16th international conference on Digital Signal Processing*, DSP'09, pages 232–239, Piscataway, NJ, USA, 2009. IEEE Press. 95

[154] S. Nikolopoulos, E. Chatzilari, E. Giannakidou, and I. Kompatsiaris. **Towards fully un-supervised methods for generating object detection classifiers using social data**. In *Image Analysis for Multimedia*

Interactive Services, 2009. WIAMIS '09. 10th Workshop on, pages 230 –233, may 2009. 95

[155] Paul A. Viola and Michael J. Jones. **Rapid Object Detection using a Boosted Cascade of Simple Features**. In *CVPR (1)*, pages 511–518, 2001. 95, 96

[156] Bastian Leibe, Ales Leonardis, and Bernt Schiele. **An Implicit Shape Model for Combined Object Categorization and Segmentation**. In *Toward Category-Level Object Recognition*, pages 508–524, 2006. 95, 96

[157] Kah Kay Sung and Tomaso Poggio. **Example-Based Learning for View-Based Human Face Detection**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**(1):39–51, 1998. 95, 96

[158] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. **Discovering Objects and their Localization in Images**. In *ICCV*, pages 370–377, 2005. 95, 96

[159] Robert Fergus, Fei-Fei Li, Pietro Perona, and Andrew Zisserman. **Learning Object Categories from Google's Image Search**. In *ICCV*, pages 1816–1823, 2005. 96

[160] Fei-Fei Li, Pietro Perona, and California Institute of Technology. **A Bayesian Hierarchical Model for Learning Natural Scene Categories**. In *CVPR (2)*, pages 524–531, 2005. 96

[161] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. **Using Multiple Segmentations to Discover Objects and their Extent in Image Collections**. In *CVPR (2)*, pages 1605–1614, 2006. 96

[162] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. **HT06, tagging paper, taxonomy, Flickr, academic article, to read**. In *Hypertext*, pages 31–40, 2006. 97, 101

[163] Eirini Giannakidou, Ioannis Kompatsiaris, and Athena Vakali. **SEMSOC: SEMantic, SOcial and Content-Based Clustering in Multimedia Collaborative Tagging Systems**. In *ICSC*, pages 128–135, 2008. 101

[164] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. **Still Image Segmentation Tools For Object-Based Multimedia Applications**. *IJPRAI*, **18**(4):701–725, 2004. 103

[165] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. **Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study**. *Int. J. Comput. Vision*, **73**(2):213–238, 2007. 103

[166] David G. Lowe. **Distinctive Image Features from Scale-Invariant Keypoints**. *Int. J. Comput. Vision*, **60**(2):91–110, 2004. 103, 136

[167] Brendan J. Frey and Delbert Dueck. **Clustering by Passing Messages Between Data Points**. *Science*, **315**:972–976, 2007. 104

[168] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. **The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results**. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html. 123

[169] Spiros Nikolopoulos, Giannakidoy Eirini, Ioannis Kompatsiaris, Ioannis Patras, and Athina Vakali. **Combining multi-modal features for social media analysis**. In Steven Hoi, Jiebo Luo, Susanne Boll, Dong Xu, Rong Jin, and Irwin King, editors, *Social Media Modeling and Computing*. Springer Berlin / Heidelberg, 1st edition. 133

[170] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. **Optimal multimodal fusion for multimedia data analysis**. In *MULTIMEDIA '04*, pages 572–579, New York, USA, 2004. ACM. 134

[171] Thomas Hofmann. **Probabilistic Latent Semantic Analysis**. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999. 134, 138

[172] R. Lienhart and M. Slaney. **PLSA on Large Scale Image Databases**. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, **4**, pages IV–1217 –IV–1220, 2007. 134

[173] Eva Hörster, Rainer Lienhart, and Malcolm Slaney. **Image retrieval on large-scale image databases**. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 17–24, New York, NY, USA, 2007. ACM. 134

[174] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. **NUS-WIDE: a real-world web image database from National University of Singapore**. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–9, New York, NY, USA, 2009. ACM. 136, 137, 147, 151

[175] Sanghamitra Bandyopadhyay, Chris Giannella, Ujjwal Maulik, Hillol Kargupta, Kun Liu, and Souptik Datta. **Clustering distributed data streams in peer-to-peer environments**. *Information Sciences*, **176**(14):1952 – 1985, 2006. 144

[176] Bhaskar Mehta. **Learning from What Others Know: Privacy Preserving Cross System Personalization**. In *Proceedings of the 11th international conference on User Modeling*, UM '07, pages 57–66, Berlin, Heidelberg, 2007. Springer-Verlag. 144

[177] Jon Louis Bentley. **Multidimensional binary search trees used for associative searching**. *Commun. ACM*, **18**(9):509–517, 1975. 148

[178] Nguyen Xuan Vinh, Julien Epps, and James Bailey. **Information theoretic measures for clusterings comparison: is a correction for chance necessary?** In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM. 149

[179] E. Chatzilari, S. Nikolopoulos, S. Papadopoulos, C. Zigkolis, and Y. Kompatsiaris. **Semi-supervised object recognition using flickr images**. In *9th International Workshop on Content-Based Multimedia Indexing (CBMI), 2011*, pages 229 –234, june 2011. 162

Author's publication list covering this thesis

Journal articles

- S. Nikolopoulos, S. Zafeiriou, I. Patras, I. Kompatsiaris, "High Order pLSA for Indexing Tagged Images", Signal Processing (Elsevier) Special Issue on: "Indexing of Large-Scale Multimedia Signals", under revision.

- E. Chatzilari, S. Nikolopoulos, I. Patras, I. Kompatsiaris, "Leveraging social media for scalable object detection", Pattern Recognition Journal, Volume 45, Issue 8, August 2012, Pages 2962-2979, DOI: 10.1016/j.patcog.2012.02.006

- S. Nikolopoulos, G. Th. Papadopoulos, I. Kompatsiaris and I. Patras, "Evidence driven image interpretation by combining implicit and explicit knowledge in a bayesian network", IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics, vol.41, no.5, pp.1366-1381, Oct. 2011. DOI: 10.1109/TSMCB.2011.2147781

- C. Lakka, S. Nikolopoulos, C. Varytimidis and I. Kompatsiaris, "A bayesian network modeling approach for cross media analysis", Signal Processing: Image Communication 26 (2011) 175-193. DOI: 10.1016/j.image.2011.01.004

Book chapters

- S. Nikolopoulos, E. Giannakidou, I. Kompatsiaris, I. Patras, and A. Vakali, "Combining multi-modal features for social media analysis", in book *Social Media Modeling and Computing*, Steven Hoi, Jiebo Luo, Susanne Boll, Dong Xu, Rong Jin, Irwin King (Eds.), 1st Edition, VIII, 276 p, Springer 2011, ISBN: 978-0-85729-435-7.

- S. Nikolopoulos, E. Chatzilari, E. Giannakidou, S. Papadopoulos, I. Kompatsiaris, A. Vakali. "Leveraging Massive User Contributions for Knowledge Extraction". In book *Next Generation Data Technologies for Collective Computational Intelligence*, Bessis, Nik; Xhafa, Fatos (Eds.), book series: Studies in Computational Intelligence, vol. 352, 1st Edition., XVIII, 638 p. 211, Springer 2011, ISBN 978-3-642-20343-5.

- E. Chatzilari, S. Nikolopoulos, I. Patras and I. Kompatsiaris, "Enhancing Computer Vision Using the Collective Intelligence of Social Media", in book *New Directions in Web Data Management 1*, Athena Vakali, Lakhmi C Jain (Eds.), book series: *Studies in Computational Intelligence*, vol. 331, Springer 2011, ISBN: 978-3-642-17550-3.

Conference papers

- S. Nikolopoulos, G. Th. Papadopoulos, I. Kompatsiaris, and I. Patras, "Image interpretation by combining ontologies and bayesian networks", 7th Hellenic Conference on Artificial Intelligence (SETN 2012), 28-31 May, Lamia, Greece.
- E. Chatzilari, S. Nikolopoulos, S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, "Semi-Supervised object recognition using flickr images", Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011), Madrid, Spain, June 2011.
- S. Nikolopoulos, G. Th. Papadopoulos, I. Kompatsiaris, and I. Patras, "An evidence-driven probabilistic inference framework for semantic image understanding", International Conference on Machine Learning and Data Mining (MLDM 2009), 23-25 July 2009, Leipzig, Germany.