CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY COMMUNICATION AND MEDIA STUDIES

Doctoral Dissertation

IMAGE RETRIEVAL: MODELLING KEYWORDS VIA LOW-LEVEL FEATURES

Zenonas Theodosiou

Limassol 2014

CYPRUS UNIVERSITY OF TECHNOLOGY
FACULTY OF COMMUNICATION AND MEDIA STUDIES
DEPARTMENT OF COMMUNICATION AND INTERNET STUDIES

IMAGE RETRIEVAL: MODELLING KEYWORDS VIA LOW-LEVEL KEYWORDS

Zenonas Theodosiou

Limassol 2014

APPROVAL CERTIFICATE

PhD thesis

Image Retrieval: Modelling Keywords via Low-level Features

Presented by

Zenonas Theodosiou

Supervisor: ______________________________

Nicolas Tsapatsoulis, Associate Professor
Cyprus University of Technology

Committee Member (President): ______________________________

Constantinos S. Pattichis, Professor
University of Cyprus

Committee Member: ______________________________

Andreas Lanitis, Associate Professor
Cyprus University of Technology

Cyprus University of Technology

May 2014

Acknowledgments

This work falls under the Cyprus Research Promotion Foundation's Framework Programme for Research, Technological Development and Innovation 2009-2010 (DESMI 2009-2010), co-funded by the Republic of Cyprus and the European Regional Development Fund, and specifically under Grant ΠΕΝΕΚ/0609/95.

Copyright Notice

The approval of the PhD dissertation by the Department of Communication and Internet Studies of the Cyprus University of Technology does not indicate necessarily that the Department accepts the positions of the author.

Abstract

Although Content Based Image Retrieval (CBIR) has attracted large amount of research interest, the difficulties in querying by an example propel ultimate users towards text queries. Searching by text queries yields more effective and accurate results that meet the needs of the users while at the same time preserves their familiarity with the way traditional search engines operate. However, text-based image retrieval requires images to be annotated i.e. they are related to text information. In recent years, much effort has been invested on automatic image annotation methods, since the manual assignment of keywords (which is necessary for text-based image retrieval) is a time consuming and labour intensive procedure. This thesis focuses on image retrieval under the perspective of machine learning and covers different aspects in this area. It discusses and presents several studies referring to: (a) low-level feature extraction and selection for the task of automatic annotation of images, (b) training algorithms that can be utilized for keyword modeling based on visual content, and (c) the creation of appropriate and reliable training data, to be used with the training scheme, using the least manual effort. The main contribution is a new framework that can be used to address the key issues in automatic keyword extraction by creating separate visual models for all available keywords using the one-against-all paradigm to account for the scalability and multiple keyword assignment problems. The prospective reader of this thesis would be equipped with the ability to identify the key issues in automatic image annotation and would be triggered to think ahead to propose alternative solutions. Furthermore, this thesis can serve as a guide for researchers who want to experiment with automatic keyword assignment to digital images.

‘When words become unclear, I shall focus with photographs.
When images become inadequate, I shall be content with silence.’

Ansel Adams

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the advent of cheap digital recording and storage devices and the rapidly increasing popularity of online social networks that make extended use of visual information, like Facebook and Instagram, image retrieval regained great attention among the researchers in the areas of image indexing and information retrieval. Thus, a large amount of research has been carried out on image retrieval the last decades. Image retrieval methods are mainly falling into content-based and text-based frameworks.

In the content-based approaches image low level features such as color, shape or texture are used for indexing and retrieving images. The user provides a target image and the system retrieves the best ranked images based on their similarity from the users query. Although it has been a long time since the scientists working on this area, content-based image retrieval still lacks semantic meaning.

Text-based methods are similar to document retrieval and retrieve images using keywords. In this direction, images must be somehow related with specific keywords or textual description. Commercial search engines utilize the textual information existing in web pages, such as image file names, anchor text, web-page keywords and, of course, surrounding text to retrieve web images. The huge amount of images that do not appear in web pages or they do not have clearly related context, either as surrounding text or keywords, creates the need of image annotation (i.e. assigning textual information to images).

Image annotation can be achieved using various approaches like free text descriptions, keywords chosen from controlled vocabularies etc. Manual image annotation has been proven to be costly and time consuming task. In any case, the annotation process remains a significant difficulty in image retrieval since the manual annotation seems to be the only way guarantying success. This is partially a reason explaining why the content-based image retrieval is still considered an option for accessing the enormous amount of digital images. Furthermore, manual annotations cannot always be considered as correct due to the visual information that always lets the possibility for contradicting interpretation and ambiguity [1].

In recent years, much effort has been invested on automatic image annotation in order to exploit the advantages of both the text-based and content-based image retrieval methods and compromise their drawbacks mentioned above. The ultimate goal is to allow keyword searching based on the image content [2]. The principal idea of automatic image annotation methods is to train semantic concept models from large number of image samples and use the models to label new ones. Thus, automatic image annotation efforts try to mimic humans aiming to associate the visual features that describe the image content with semantic labels.

1.1 Aims and Objectives

This thesis focuses on image retrieval using keywords under the perspective of machine learning. It covers different aspects of the current research in this area including low-level feature extraction, creation of training sets and development of machine learning methodologies. It also proposes the idea of addressing automatic image annotation by creating visual models, one for each available keyword, and presents several examples of the proposed idea by comparing different features and machine learning algorithms in creating visual models for keywords referring to the athletics domain.

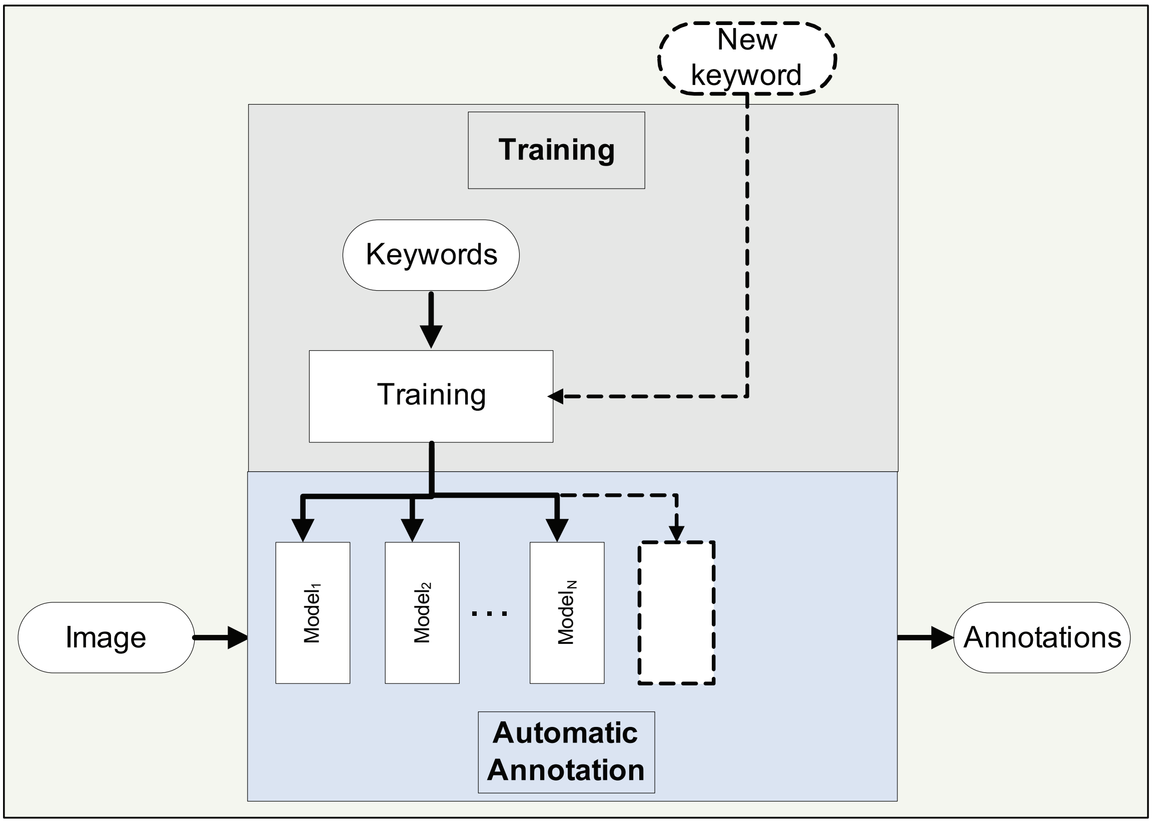


Figure 1.1: Automatic image annotation using visual models of keywords.

The idea of automatic image annotation through independent keyword visual models is illustrated in Fig. 1.1. The whole procedure is divided into two main parts: the training and automatic image annotation. In the first part, visual models for all available keywords are created, using the one-against-all training paradigm, while in the second part, annotations are produced for a given image based on the output of these models, once they are fed with a feature vector extracted from the input image.

An accurate manually annotated dataset containing pairs of images and annotations is prerequisite for a successful automatic image annotation. Since the manual annotations are likely to contain human judgment errors and subjectivity in interpreting the image, the current thesis investigates the factors that influence the creation of manually annotated image datasets through crowdsourcing. Furthermore, it proposes the idea of modeling the knowledge of several people by creating visual models using such training data, aiming to significantly improve the ultimate efficiency of image retrieval systems.

1.2 Major Contributions

The main contributions of this thesis are the following:

1. An in-depth investigation of the automatic image annotation methods, the identification of the key issues in automatic image annotation and the idea of addressing automatic image annotation by creating visual models, one for each available keyword. Creating independent keyword models, separately, appears to be a realistic solution to the drawbacks of the existing automatic image annotation methods. A given image could be associated with more than one keyword and a new keyword model can be trained irrespectively of the existing ones. This approach provides the required scalability for large scale text-based image retrieval. On the other hand, whenever a training data for new keyword are available, a new visual model is created for this keyword, and added into the unified framework.

2. A study on creating training examples for machine learning schemes. Since manual image annotation is a time-consuming and expensive task, creation of manual image annotations through crowdsourcing as well as the extraction of keywords from the surrounding text for web images are extensively explored. In this perspective:

   (a) We examined the influence of age and gender in manual image annotation. The image annotation is a social cognitive process and may vary among people based on their socio-demographic characteristics. This is reasonable to investigate the age and gender differences in manual image annotation using a controlled vocabulary and free keywords. The findings of the proposed study provide interesting insights into the relationship between annotator demographics and image annotation. The experiments reveal that there are significant age differences in the way that people annotate images for both vocabulary and free keywords. The gender, on the other hand, appears to not play a significant role in image annotation.

(b) We proposed a novel approach for investigating the manual image annotation quality aiming at: (a) identifying to which extend the use of structured lexicon and unstructured vocabularies improves annotation quality and at what cost (missing useful and valid annotations), (b) exploring to which extend and under what prerequisites free annotation can lead to valid and useful image annotation, and (c) inquiring the effect of image content itself on valid image annotation.

(c) We presented a web image indexing scheme that utilizes the surrounding textual information to extract keywords for images. The proposed method uses visual web-page parsing and specific distance metrics to assign textual segments to images. Key terms are located within the assigned segments using language models and used to index the corresponding web images.

3. Among a variety of feature extraction approaches, special attention has been given to the SIFT algorithm which delivers good results for many applications. However, the non-fixed and huge dimensionality of the extracted SIFT feature vector cause certain limitations when it is used in machine learning frameworks. We introduce the Spatial Histogram of Keypoints (SHiK), which keeps the spatial information of localized keypoints, on an effort to overcome this limitation. The proposed technique partitions the image into a fixed number of ordered sub-regions based on the Hilbert space-filling curve and counts the localized keypoints found inside each sub-region. The resulting spatial histogram is a compact and discriminative low-level feature vector that shows significantly improved performance on classification tasks. The proposed method achieves high accuracy on different datasets and performs significantly better on scene datasets compared to a similar method.

4. Inspired by the fact that the majority of tomorrow users of search engines are non-experts, we propose the idea of modeling the knowledge of several people rather than an expert. Different low-level feature extraction algorithms were

applied on training data collected through crowdsourcing, and visual models were created using several machine techniques. The experimental results indicate that nearly all models are able to assign the right keywords to unseen images. Finally, a co-training algorithm was utilized to overcome the limitations caused by the few training examples when used in classifications schemes. In this case, the developed visual models can obtain very high classification scores regardless of the image content.

A list of publications derived from this thesis is shown in Appendix A.

1.3 Structure of the Thesis

- Chapter 2 presents the background and related work on image retrieval. It covers different aspects of the current research and identifies the key issues in this area.
- Chapter 3 outlines the idea of modelling keywords via low level features
- Chapter 4 details the process of manual image annotation. It focuses on manual image annotation through crowdsourcing, and keyword extraction from surrounding html text in case of web images.
- Chapter 5 presents a study conducted to investigate the age and gender differences in manual image annotation.
- Chapter 6 details a study conducted to investigate the factors that influence the quality of manual image annotation.
- Chapter 7 presents a study conducted to extract keywords from web images.
- Chapter 8 details several available low-level features, focusing on the most successful ones when used in machine learning frameworks.
- Chapter 9 presents a study conducted to evaluate the performance of MPEG-7 descriptors in keyword extraction.

- Chapter 10 introduces the new low-level feature extraction algorithm Spatial Histogram of Keypoints (SHiK). The algorithm was developed aiming to overcome the limitations of the SIFT algorithm when used in machine learning schemes.

- Chapter 11 explores several available machine learning techniques starting from simple supervised learning algorithms and ending to more advanced learning schemes.

- Chapter 12 presents some examples of the idea of addressing automatic image annotation by creating visual models, one for each available keyword. Different features and machine learning algorithms are compared in creating visual models for keywords referring to the athletics domain.

- Chapter 13 summarizes the major findings of the thesis and proposes future research directions.

Chapter 2

Basic Theory and Problem Formulation

This chapter covers different aspects of the current research in this area, including the semantic gap, image segmentation, low-level image feature extraction, machine learning techniques and concludes with the major issues of the domain.

2.1 Image Interpretation Through Visual Content

The interpretative system proposed by Panofsky [3] for subject matter in visual art, was initially applied to identify and reconstruct the meaning of images [4, 5, 6]. The proposed system consists of three levels, where each level corresponds to content description, analysis and interpretation. The first or "pre-iconographical" level refers to the primary or natural subject matter and can be subdivided into "factual" and "expressional" sections. It refers to identifiable objects, people, or events that constitute the class of primary or natural meanings. "Iconography" was Panofsky's term for the secondary or conventional subject matter, and "iconology" the term for intrinsic meaning or content. At the second level, meaning is discerned by the interpretation of actions or gestures. The third level of meaning in pictures is related to interpretation of the image, predicated on knowledge and erudition, as well as substantial cultural background.

An interesting approach for classifying the content of an image for indexing purposes is presented in [7]. The pictorial information is classified into hard and soft indexing. Hard indexing deals with the relatively objective description of what can be observed in a picture. Soft indexing, on the other hand, is related to the subjective meaning and the personal response, which it evokes. It extends the indexing to reveal the message behind the image. Shatford [8] presented the idea of interpreting images using four attribute categories. The first attribute category refers to the biography of image and concerns the creator, the time and place of a picture's creation, any name or title given to it by its creator, etc. Although these attributes are relatively objective, in some cases they are trivial or unknown. The second category contains more subjective attributes related to the meaning of the image, while the third refers to the physical format of the images (ex. poster, photograph, etc.). Finally the last category covers the relationship attributes of the image with other images or textual work.

The relation between a word and an image is not a stable but a dynamic process linked to wider social and cultural issues [9]. This relation resists the stabilization as a binary opposition, shifting and transforming itself from the conceptual level to another, and shuttles between relations of contrariety and identity, difference and sameness. Holland [10] states that the main social and cultural changes in the world are the interpretive meanings the people bring to the pictures. As an example, he indicated the growth of the advertisement the last years that resulted to intimate images that invite the viewer to almost imagine a story rather than just see the presenting objects [11].

The way an image is encoded by the creator and how is it decoded by the viewer is explained in [12]. The transfer of a meaning works only if there are compatible systems of signs and symbols that both the creator and the viewer use within their life. Some demographic factors like the gender, age, education, etc., affect the interpretation of these signs and symbols and define the understanding of images. Messages are not always read as they were intended, since the various cultural, social and political background lead to different interpretations. Hall [12] suggests three possible readings of an image. The dominant reading, which complies with the meaning intended by the creator, the negotiated reading, which partly complies with the intended meaning, and the oppositional reading, which is in total conflict with the meaning. The type of meaning is primarily determined through associated verbal description and the context in which the image is used. There is no single meaning for an image, but rather an emergent meaning, within which the subject-matter of the image is but one element [13].

Concerning the visual information retrieval problem, each image can be related to widely varying expressions of interest. Its potential relevance, appropriateness, or usefulness is inherently unpredictable [14]. In the case of personal repositories, a system that sorts automatically the images in chronological order, and displays a large number of thumbnails at once is enough to allow the owners to find what they are looking. According to Rodden and Wood [15], the availability of text-based retrieval did not provide the creators with enough extra motivation to invest the effort in annotating their images. However, the findings in that study are based on personal image collections and cannot be generalized to the generic image retrieval problem.

Despite that querying using text-based terms remains the most practical way for browsing and retrieving images, only a few studies have addressed the user needs and queries for visual information. An analysis of user's queries in terms of image retrieval is also presented in [15]. The authors conclude that text-based queries contain a high incidence of terms representing specific or general persons and things, as well as geographical and chronological terms. Stvilia et al. [16] presented the relationships between user demographics and users' perceptions of the value of socially-created terms as well as the relationships between user demographics and indexing quality as measured by the number of the given tags.

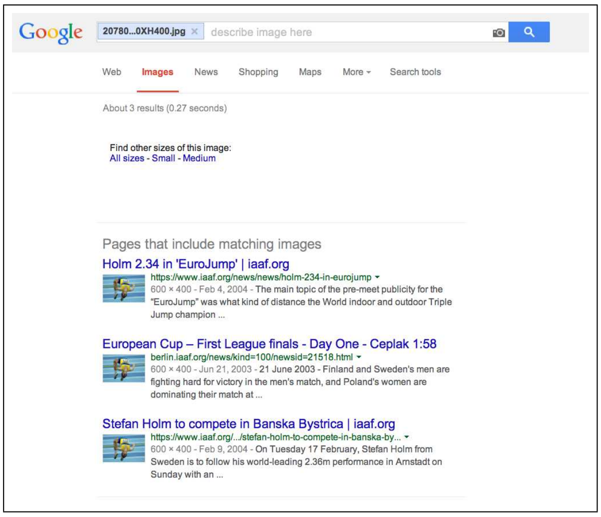
The socioeconomic and demographic differences in the use of internet have become increasingly critical to economic success [17]. Thus, some studies have already examined the differences in Internet access and use across a wide range of socioeconomic and demographic groups, including gender [18, 19, 20] and age [21]. The preceding studies provide valuable insight on how people manage their personal images, as well as on how they search for images. They also provide ample evidence that the interpretation of an image is directly connected with the background of the viewer. Thus, additional research is needed to investigate if there are differences in the way people annotate (shortly describe) images. In the current study we investigate further this issue and we try to uncover the strong relevance between image interpretation and image annotation.

2.2 Image Retrieval

Since the beginning of the World Wide Web (WWW) and the development of cheap digital recording and storage devices the amount of available on-line digital images, continuously increases. The increasing popularity of online social networks, like Instagram, that are based on visual information push further this tendency. As a result, effective and efficient web image indexing and retrieval schemes are of high importance and a lot of research has been devoted towards this end.

Image retrieval refers to the problem of selecting, from a repository of images, those images fulfilling to the maximum extent some criterion specified by an end user. Information retrieval and computer vision communities study image retrieval from different angles and their efforts are, mainly, falling into text-based and content-based frameworks. Content-based methods retrieve images by analyzing and comparing the content of a given image example as a starting point. Text-based methods are similar to document retrieval and retrieve images using keywords. Fig. 2.1 and Fig. 2.2 show examples of a search for images, using image and text query respectively.

The first text-based image retrieval efforts started in the late 1970s. Manually annotated images were retrieved through text-based database management systems [22], [23].



About 3 results (0.27 seconds)

Find other sizes of this image:
All sizes - Small - Medium

Pages that include matching images

Holm 2.34 in 'EuroJump' | iaaf.org
https://www.iaaf.org/news/news/holm-234-in-eurojump
600 × 400 - Feb 4, 2004 - The main topic of the pre-meet publicity for the "EuroJump" was what kind of distance the World indoor and outdoor Triple Jump champion ...

European Cup – First League finals - Day One - Ceplak 1:58
berlin.iaaf.org/news/kind=100/newsid=21518.html
600 × 400 - Jun 21, 2003 - 21 June 2003 - Finland and Sweden's men are fighting hard for victory in the men's match, and Poland's women are dominating their match at ...

Stefan Holm to compete in Banska Bystrica | iaaf.org
https://www.iaaf.org/.../stefan-holm-to-compete-in-banska-by...
600 × 400 - Feb 9, 2004 - On Tuesday 17 February, Stefan Holm from Sweden is to follow his world-leading 2.36m performance in Arnstadt on Sunday with an ...

Figure 2.1: Content-based image retrieval.

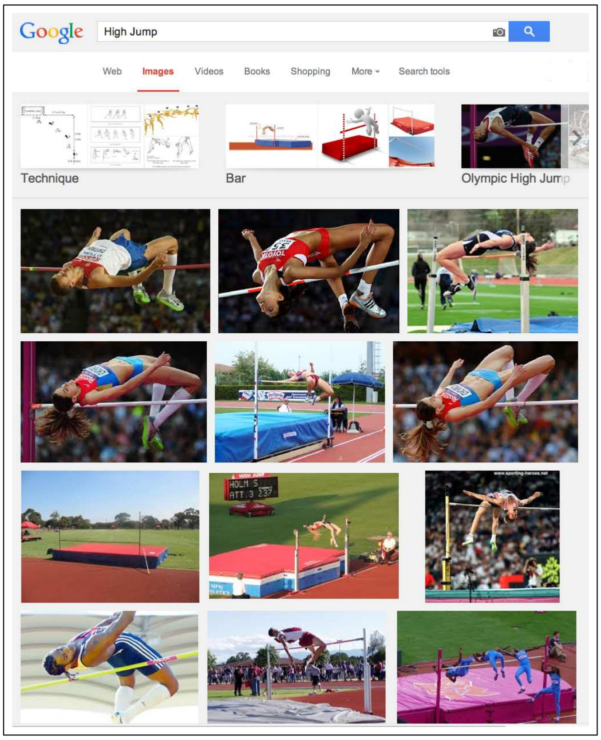


Figure 2.2: Text-based image retrieval.

The vast amount of effort required for image annotation as well as the rich content in the images and the subjectivity of human perception were the main reasons led the researchers to content-based image retrieval in the early 1990s. Instead of being manually annotated, images are indexed and retrieved by their own visual content, such as color, texture and shape [24].

Nevertheless, text-based is the approach of preference both for ordinary users and search engine engineers. Besides the fact that the majority of users are familiar with text-based queries, content-based image retrieval lacks semantic meaning. Furthermore, image examples that have to be given, as a query, are rarely available. From the search engine perspective, text-based image retrieval methods get advantage of the well-established techniques for document indexing and are integrated into a unified document retrieval framework. However, for text-based image retrieval to be feasible, images must be somehow related with specific keywords or textual description.

In contemporary search engines textual description of images is, usually, obtained from the web page, or the document, containing the corresponding images and includes HTML alternative text, the file names of the images, captions, surrounding text, metadata tags or keywords extracted from web page as a whole [25]. Despite the fact that this type of information is not directly related to images content it can be utilized only in web-page image retrieval. As a result, image retrieval from dedicated image collections can be done either by content-based methods or by explicitly annotating images by assigning tags to them to allow text-based search. The latter process is collectively known as ‘image annotation’ or ‘image tagging’.

2.3 Content-based Image Retrieval

Initially, the problem of searching the enormous number of digital image collections that are available through the Web or in personal repositories was tackled by efficient and intelligent schemes for content-based image retrieval (CBIR) [23]. An example of

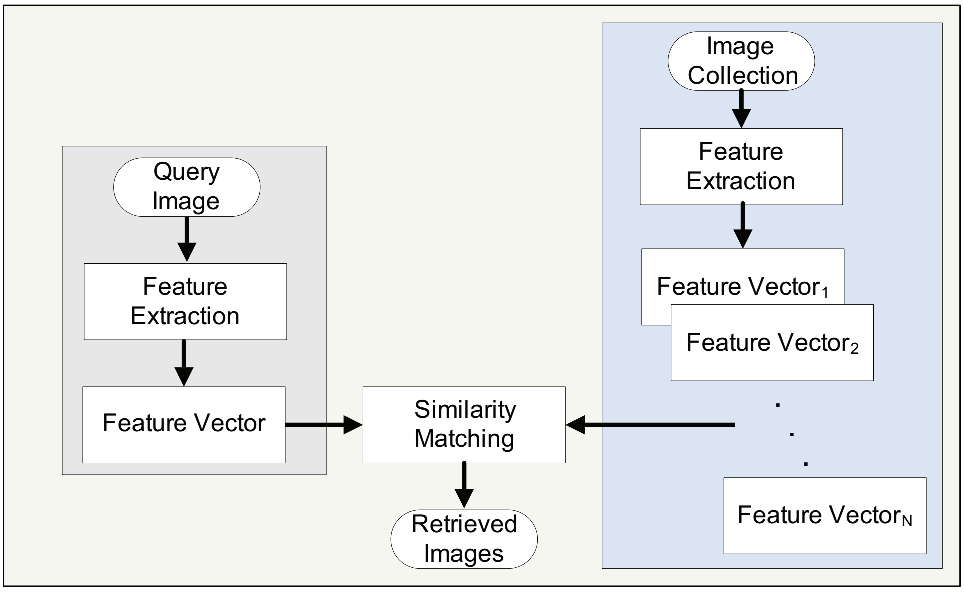


Figure 2.3: Content-based image retrieval.

CBIR is presented in Fig. 2.3. CBIR computes relevance based on the visual similarity of image low-level features such as color, texture and shape [26]. Early CBIR systems were based on the query by-example paradigm [27], which defines image retrieval as the search for the best database match to a user-provided query image. Under this framework, and in order to maximize the effectiveness of CBIR systems, it soon became necessary to mark specific regions in the images so as to model particular objects.

Image segmentation in general is defined as the process of "partitioning an image into homogenous groups such that each region is homogenous but the union of two adjacent regions is inhomogeneous" [28]. Segmentation is usually used to detect objects in an image and thus varies depending on the application or the content of an image. Moreover the detection of the presented objects, the spatial relationships between them may be used to represent the content of the image [29].

Several algorithms have been proposed for segmentation in image retrieval approaches. One of the simplest and fastest algorithms for image segmentation is based on the grid and divides the image into equal rectangular parts [30], [31], [32], [33]. The feature extraction is applied on each segmented part that consists of different objects. The effectiveness of grid segmentation depends on the size of the grid and on the nature of the images. The determination of the size of the grid is difficult, thus a novel approach is presented in [33], where several sizes of the grid were used to improve the nal segmentation.

Other approaches utilize clustering algorithms to segment the images into regions [34, 35, 36]. Usually, the image is divided into small parts that are grouped into clusters using a clustering algorithm. Visual features as color and texture are extracted from each part and then k-means algorithm clusters the part into a cluster. Parts of each cluster are then joined to create each region in an image. The prior assumption of the number of clusters remains the basic drawback of clustering based segmentation. In order to overcome this limitation, other methods use active contours to detect the objects' boundaries in an image [37], [38]. However, the limitations come along with the prior edge detection renders the content-based segmentation to be applicable in a limited range of applications.

More computational expensive methods use statistical models [39], [40], [41] or graph partitioning for segmentation [42], [43]. In [39], the joint distribution of color, texture and position features is modeled with a mixture of Gaussians. The pixels are grouped into regions by modeling the features with a mixture of Gaussians. The Expectation-Maximization (EM) algorithm estimates the number of regions and models parameters. The final segmentation is based on the resulting pixel-region memberships. A segmentation algorithm presented in [42] treats the input image as graph whose vertices are the image pixels and its edges weights are the feature similarities between pixels. The normalized cut (NCut) algorithm partitions the vertices into disjoint sets so that the similarity among the vertices in a set is high and, across different sets is low. The large number of pixels makes the computation of the optimal partition difficult and expensive. A significant improvement of the NCut is presented in [43] where a prior segmentation based on mean shift algorithm segments the input images into regions. These regions are then used as the vertices of the graph.

Region-based algorithms have been also used for image segmentation in image retrieval [44], [45], [46]. Deng and Manjunath [44] presented the JSEG algorithm for unsupervised segmentation of color-texture regions. It consists of two independent steps: First, colors in the image are quantized into classes and pixels are replaced by their corresponding color class labels. Then a region growing method is used to segment the image by setting as seed points the pixels with more homogenous neighbors representing the seed points.

2.3.1 Key Issues in Content-based Image Retrieval

The content-based image retrieval methods face some issues related to the image query: (a) The users can not be always provided with the appropriate input image, and (b) it is difficult to search for high level terms using visual content. Humans usually use high-level semantics including keywords and text descriptors to interpret and describe the content of an image. On the other hand, the visual content of an image in content-based approaches is represented by low-level features. In general, there is no direct link between the high-level semantics and the low-level features [47]. Although, there are many algorithms for sophisticated extraction of low-level features, no one can model the high-level semantics adequately when dealing with a broad content image datasets [48]. As presented in [49], the description of the high-level semantics in user's mind by low-level feature is not an easy task.

The interpretation inconsistency between image low-level descriptors and high-level semantics is known as 'semantic gap' [23] or 'perceptual gap' [50]. The 'semantic gap' was initially introduced by Smeulders et al. [23] as the "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". Recent research focuses on new low-level feature extraction algorithms to bridge the gap between the simplicity of available visual features and the richness of the user semantics.

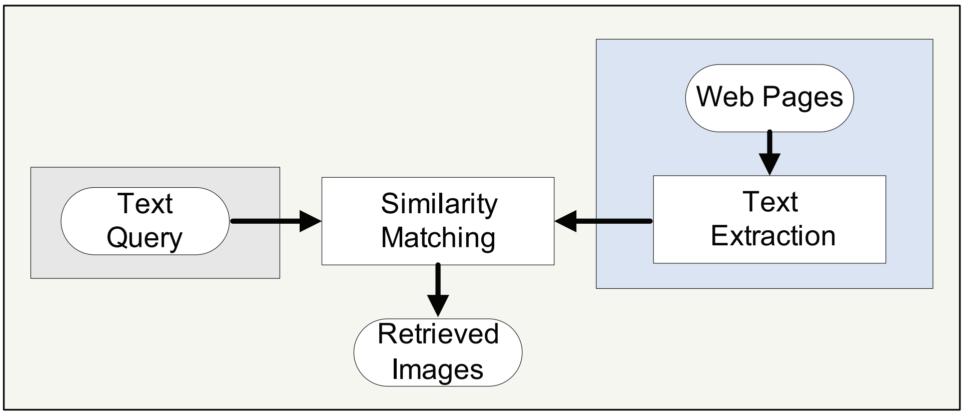


Figure 2.4: Web-based image retrieval.

2.4 Web-based Image Retrieval

Web-based image retrieval methods utilize the textual information, included in web pages, in order to retrieve those web images that are related somehow with the given text-query (Fig. 2.4). Moreover the HTML alternative text, file names, metadata tags etc., the surrounding text plays also significant role in these methods. Surrounding text, is the text that surrounds a Web image inside an HTML document. This text is, indeed, a very important source of semantic information for the image. However, automatic localization of surrounding text is by no means easy mainly due to the modern web-page layout formatting techniques which are based on external files (stylesheets). As a result, visual segmentation (parsing) of the rendered web-page is required in order to identify the surrounding text of an image.

The need to automatically extract the semantically related, to an image, textual blocks and assign them to this image led to what we call Web Image Context Extraction (WICE). In that terms, WICE is the process of automatically assigning the textual blocks of a web document to the images of the same document they refer to [51].

A variety of methods have been proposed for annotating automatically web images using the plethora of textual information existing in web pages. A number of methods have been combined this information with visual features for more accurate image annotation and retrieval. The SIEVE algorithm was proposed by Liu et al. [52] to improve the text-based Web image search. The algorithm utilizes visual features to discard irrelevant images from the image list returned by a text-based image search engine. Wang et al. [53] presented a clustering-based method for grouping homogeneously the retrieved images using visual features. The nearest image to each centroid is set then as the as the entry for the corresponding cluster to guide users browsing.

In [54], the AnnoSearch system is proposed which leverages the search technology to annotate WWW images. By giving an image and a keyword that describes its content, the system utilizes both text-based and content-based methods to annotate the given image. First, the semantically similar images to the input keyword are retrieved accompanied by their textual descriptions. Then, the top ranked images that are visually similar to the input image are identified by mapping the visual features to hash codes. Finally, the textual descriptions of the selected images are clustered and the words from the top scored clusters are given as annotations to the input image.

In [55], a hierarchical clustering method is proposed for image annotation using visual, textual and link analysis. The webpage is partitioned into semantic blocks using the Vision-based Page Segmentation (VIPS) algorithm [56], and textual and hyperlink information of an image is extracted from the semantic block containing that image. Initially, a clustering method based on textual and link information classifies images into different semantic clusters. Since the images clustered in each semantic cluster may have visual differences, a second clustering is applied in each semantic cluster using low level visual features. Although the proposed method seems promising, the unreliable textual and link features according to the existing search engines have a negative effect on the final results.

An alternative way of indexing web images is through surrounding text usually leads to annotated images with some irrelevant keywords. In [57], an attempt to increase the annotation accuracy by pruning the irrelevant keywords given during annotation is presented. The idea is based on the calculation of the semantic similarity between keywords given for an image using the WordNet lexicon. The keyword having semantic similarity to the other keywords below a threshold is discarded. Wang et al. [58] proposed an algorithm to re-rank the candidate annotations and prune the irrelevant ones. The algorithm is based on Random Walk with Restarts (RWR) and leverages co-occurrence-based similarity confidence scores of the original annotations. In [59], the authors proposed the content-based image annotation refinement (CIAR) algorithm to re-rank the candidate annotations, leveraging both the corpus information and the content feature of the query image. In both methods [58], [59], only the top ranked annotations are reserved as the final annotations.

2.4.1 Key Issues in Web-based Image Retrieval

Although many WICE methods have been proposed, the research is far from the ideal system. The high diversity of designing patterns in web pages, the noisy environment (advertisements, graphics, navigational objects etc.), and the too much textual and visual information in single documents, are some of the issues that would be addressed for efficient and effective web-based image retrieval. Furthermore, the huge amount of images which do no appear in web-pages or they do not have a clearly related context, either as surrounding text or as specific keywords, puts another challenge.

2.5 Image Annotation

Image annotation is the assignment of textual descriptions to images and assists the text-based image retrieval. Annotation can be done using various approaches like free text descriptions, keywords chosen from controlled vocabularies or taxonomies based on ontologies [1]. Despite the plethora of available tools, manual annotation is an extremely difficult and elaborate task since the keyword assignment is performed on image basis. Manual annotation of large collections is often prohibitive and manual annotations are known to be imprecise, ambiguous, inconsistent and subject to many variations [60]. A possible way to alleviate these problems and improve the annotation quality is to obtain multiple annotations per image by assigning several annotators into the task.

Joachims et al. [61] have discovered that differences between implicit and explicit relevance judgments are not so far as they were thought to be. This innovative finding opened a new way, where implicit relevance judgments were implicitly incorporated for annotating images [62] or considered as training data for various machine learning-based improvements to information retrieval [63], [64]. Problems can be addressed very quickly, at little cost, and the task provider might exploit a wider range of people [65]. Implicit crowdsourcing data can be easily collected and used as training data [66] in various tasks and their collection introduces no additional cognitive burden on users performing the queries. However, because participants carry out experiments without supervision, they may give erroneous feedback perfunctorily, carelessly, or dishonestly, even if they receive a reward for each experiment [67].

Consistency is a large problem for each annotation project and inter-annotator and intra-annotator agreements are very important [68]. The inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators while the intra-annotator agreement describes how consistent is a single annotator. The annotators and vocabulary used during annotation assessment have to be chosen with care while the resources should be used effectively [69].

2.6 Automatic Image Annotation

Automatic image annotation is the process of assigning annotations to images automatically and has been a topic of on-going research for more than a decade. Several interesting techniques have been proposed during this period [70]. Although appears to be a particularly complex problem for researchers and despite the fact that annotation obtained automatically is not expected to reach the same level of detail as the one obtained by humans, it remains a research hot topic. The reason is obvious: Manual annotation of the enormous amount of images created and uploaded to the web every day is not only impractical; it is simply impossible. Therefore, automatic assignment of keywords to images for retrieval purposes is highly desirable. The proposed methods towards this direction attempted to address first, the difficulty of relating high-level human interpretations with low-level visual features and second, the lack of correspondence between the keywords and image regions in the (training) data. Systems that automatically assign one or multiple keywords to an image have been developed [71], [51]. Nevertheless, the automated image annotation cannot be expected to perform at the same level of detail as a human annotator.

In automatic image annotation, a manually annotated set of data is used to train a system for the identification of joint or conditional probability of an annotation occurring together with a certain distribution of feature vectors corresponding to image content [72]. Different models and machine learning techniques were developed to learn the correlation between image features and textual words based on examples of annotated images. Learned models of this correlation are then applied to predict keywords for unseen images [73]. Although the low-level features extracted from an image cannot be automatically translated reliably into high-level semantics [74], the selection of visual features that better describe the content of an image is an essential step for the automatic image annotation.

2.6.1 Low-level Features used in Automatic Image Retrieval

The visual content of images in automatic image annotation and retrieval is represented by low-level features. The extraction of low-level features can be applied either on the entire image or on image's regions as created after image segmentation. Many sophisticated feature extraction algorithms have been proposed that are mainly based on color, texture and shape features of the image.

Color features are defined on a specified color space such as RGB, LUV, HSV and HMMD [75], [76], [77]. Different color features have been proposed for image retrieval such as the color moments [32], [78], color histogram [78], [79], color coherence vector [80], color correlogram [58], [81], etc. Moreover, there are several color features standardized by the MPEG-7 [82] such as the dominant color descriptor (DCD), color layout descriptor (CLD), color structure descriptor (CSD), and scalable color descriptor (SCD) [83].

Since texture provides important information about image's content and has discriminative capability, texture features have been widely used in image retrieval. Texture features can be extracted either on spectral or spatial domain. In the first case, image is transformed into the frequency domain and features are extracted using Gabor filtering [83], Wavelet transform [34], etc. On the other hand, spatial features are extracted on original image's domain by utilizing structural, statistical and model based techniques [2].

Features describing objects' shapes have also been used to represent images in image retrieval. Since image segmentation may cause small changes in the shapes presented in an image, the majority of shape based techniques extract features from the entire image (i.e. they consider image as an object). The shape features include area, moments, circularity [84], [79], and eccentricity [36]. Moreover, more complex shape features have been also used for specific applications involving Fourier descriptor [85] and contour shape descriptor standardized by MPEG-7 [86].

The low-level feature extraction is a basic step for accurate image annotation so an extended presentation of the sophisticated algorithms for feature extraction is given in chapter 8.

2.6.2 Single Labeling

In binary classification approaches, image classifiers are constructed with the aid of training data (pairs of images and keywords), and are applied to classify a given image into one of several classes. Each class usually corresponds to a particular keyword.

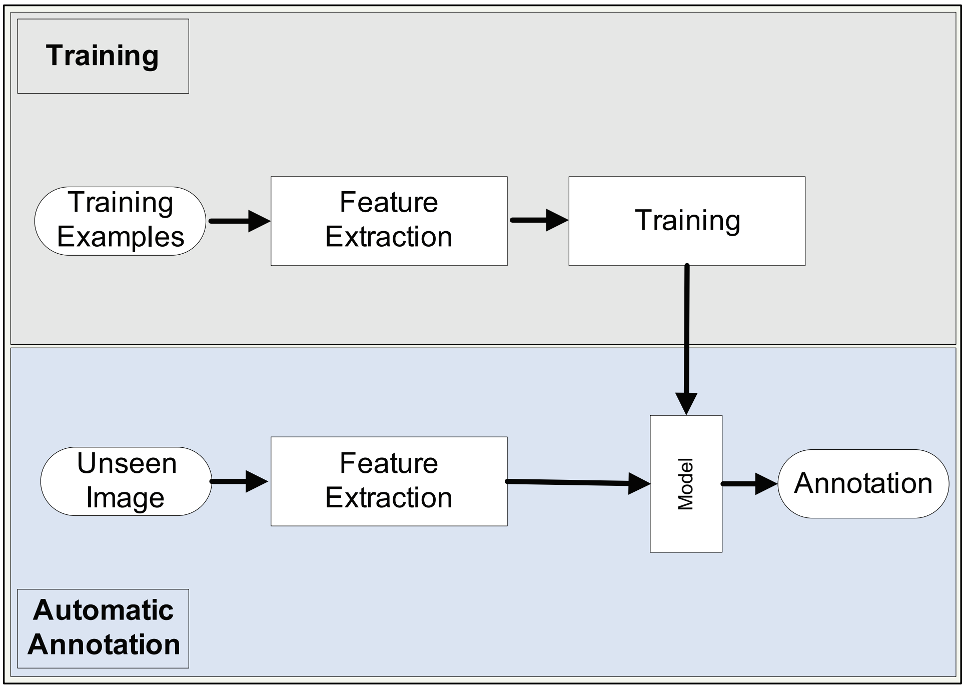


Figure 2.5: Automatic image annotation based on classification.

An example of the whole procedure is given in Fig. 2.5. In the first part, the model of each keyword is created using the features extracted from the training data, while in the second part annotation is produced for a given image based on the output of the model. Several machine learning algorithms have been used for image classification into keyword classes. Support Vector Machine (SVM) [31], Hidden Markov Models [87], Decision Trees [88], are some of them. An extensive review on machine learning classifiers in image annotation is given in chapter 11 while the general principal of machine learning utilization is revisited in chapter 12.

Although binary classification based annotation methods give promising results, they are designed for small-scale datasets and they use a small number of keyword classes. As a result the trained classifiers do not generalize smoothly to allow accurate classification, of the large amount of images that are missing annotations, to the available classes. Furthermore, the limited number of manually annotated data (few positive and negative examples) that are used during training lead to ineffective keywords models without generalization ability. The limited number of classes, on the other the hand, restricts the number of text queries that could derive results and search is applied thought the specific keywords (classes). Users, in general, are reluctant to adopt search interfaces that are based on predefined sets of keywords because they are familiar with the free text searching paradigm used in web-search engines. Finally, in binary classification based annotation, classifiers relate images with a single keyword while it is obvious that image content can be associated with many keywords.

2.6.3 Multiple Labeling

The multiple labeling assigns an image to multiple classes using probabilistic approaches such as Bayesian methods [30], [89]. The Bayesian methods are based on the posterior probability that an image is related to any particular concept, given the observation of certain features from the image or an image's region. This makes it possible to assign an image to multiple concepts and many images with the same concept according to the probabilities.

The co-occurrence model proposed by Mori et al. [30] can be considered as the first automatic image annotation approach. This model tries to capture the correlations between images and keywords (assigned to them) and consists of two main stages. In the first stage, every image is divided into sub-regions and a global descriptor for each sub-region is calculated. In the second stage, feature vectors extracted from sub-regions are clustered using vector quantization. The probability of a label related to a cluster is estimated by the co-occurrence of the label and the sub-regions within the cluster. Duygulu et al. [84] proposed a model of object recognition as a machine translator in order to annotate images automatically. Every image is segmented into object shape regions, called 'blobs', and a visual vocabulary is created by feature quantization of the extracted feature vectors of the regions. Finally, the correspondence between blobs and words is found by utilizing the Expectation-Maximization algorithm. The Cross Media Relevance Model (CMRM) was introduced by Jeon et al. [89] in order to improve the machine translator model. They followed the same procedure for calculating the blob representation of images as Duygulu et al. and then utilized the

CMRM to learn the joint distribution of blobs and words in a given image. The loss of useful information during the quantization from continuous features into discrete blobs that occurred on the translation model and CMRM, was treated by Lavrenko et al. [90]. The proposed Continuous Relevance Model (CRM) does not require an intermediate clustering stage and associates directly continuous features with words. Further improvement on annotation results was obtained by the Multiple Bernoulli Relevance Model (MBRM) [32], where the word probabilities are estimated using a multiple Bernoulli model and the image feature probabilities using a non-parametric kernel density estimate.

The computational cost of parameter estimation is probably one of the drawbacks of using statistical models in automatic image annotation approaches since the learning of parameters lasts several hours. Nevertheless, object recognition based methods for image annotation are of limited scope because object recognition itself is a very hard problem and is solved only under strict constraints.

The multi-instance multi-label learning (MIML) proposed in [91], [92] where each training example is described by multiple instances and associated with multiple class labels tries to eliminate this problem. Although this method gives a fair solution to the problem of assigning more than one keyword to a given image, it has also several limitations. In order to model a valid probability of labels it is necessary to compute a summation over all possible label assignments leading to high computational cost. Furthermore, there is no provision to add new labels (keywords). The initial set of keywords remains unchanged while erroneous tagging is accumulated since the labels that are assigned to a particular image depend on its content similarity with other images that already have this label. Therefore, in case where the initial label of an image is erroneous the error is propagated to all images having similar content.

2.7 Major Issues in Automatic Image Annotation

There are some issues commonly encountered in automatic image annotation systems either image retrieval is approached using the content-based or the text-based paradigm.

During automatic annotation which is based on image segmentation, labels corresponding to object classes are assigned every time a particular object instance is encounter in the input image. This object instance almost always corresponds to an image region (part of the image). Therefore, region-based features must be computed and used for object modeling. Under this perspective, semantic labeling using object class labels is actually an object detection task. Unfortunately, object detection is a very hard task itself and is solvable only for limited cases and under strict constraints.

The low performance of CBIR systems along with their limited scope led researchers to investigate alternative schemes for image retrieval. It was quickly realized that the ultimate users were willing to search for images using their familiar text-based search interface. Therefore, the design of fully functional image retrieval systems would require support for semantic queries [93]. In such systems, images are annotated with semantic labels, enabling the user to specify the query through a natural language description of the visual concepts of interest. This realization, combined with the cost of manual image labeling, generated significant interest in the problem of automatically extracting semantic descriptors from images.

The automatic annotation of digital images with semantic labels is traditionally coped by utilizing machine learning emphasizing on classification. In that case, semantic labels may refer to an abstract term, such as indoor, outdoor, athletics, or to an object class such as human, car, tree, foam mats, etc. The latter case reinforces the object detection problem encountered in CBIR systems. In order to learn a particular object class and create a classifier to recognize it automatically you need accurate and region specific features of the training examples. Indeed, there exist some approaches [94] that use some kind of object detection in order to assign semantic labels to images [95]. In contrary to object classes, abstract terms cannot be related to specific image regions. In the literature of automatic semantic image annotation, proposed approaches tend to classify images using only abstract terms or using holistic image features for both abstract terms and object classes.

The extraction and selection of low-level features, either holistic or from particular image areas is of primary importance for automatic image annotation. This is true either for the content-based or for the text-based retrieval paradigm. In the former case the use of appropriate low-level features leads to accurate and effective object class models used in object detection while in the latter case, the better the low-level features are, the easier the learning of keyword models is. The fusion of several types of features is applied in many cases in order to represent as many images as possible. The low-level feature algorithms as well as the feature fusion may lead to high dimensional features vectors. A huge dimensional feature vector causes certain limitations in the performance of the machine learning techniques. Therefore, the dimensionality reduction is an essential step for selecting the appropriate features and their dimensionality for annotation. Low-level feature extraction is one of the key issues in automatic image annotation and is examined extensively in chapter 8.

The intent of the image classification is to categorize the content of the input image to one of several keyword classes. A proper image annotation may contain more than one keyword that is relevant to the image content, so a reclassification process is required in this case, as well as whenever a new keyword class is added to the classification scheme. The creation of separate visual models for all keyword classes adds a significant value in automatic image annotation since several keywords can be assigned to the input image. As the number of keyword classes increases the number of keywords assigned to the images also increases too and there is no need for reclassification. However, the keyword modeling incurred various issues such as the large amount of manual effort required in developing the training data, the differences in interpretation of image contents, and the inconsistency of the keyword assignments among different annotators. These key issues are also examined in detail in subsequent chapters.

Chapter 3

Keyword Modelling via Low Level Features

This chapter presents the idea of modelling keywords via low-level features as an attempt to solve the key issues of the existing automatic image annotation methods.

3.1 Proposed Framework

The main contribution of this thesis is the idea of indexing non-annotated images using visual models of keywords. The proposed framework imposes the creation of a separate visual model for each available keyword. Whenever a training data for new keyword are available, a new visual model is created irrespectively of the existing ones, and added into the unified framework. The proposed framework appears to be a realistic solution to the drawbacks of the automatic image annotation methods presented in chapter 2. It assigns more than one keyword to an unseen image and provides the required scalability for large scale text-based image retrieval. The whole procedure consists of two main parts. The first part refers to the training while the second part refers to the automatic image annotation. The unified framework of creating visual models for keywords is illustrated in Fig. 3.1.

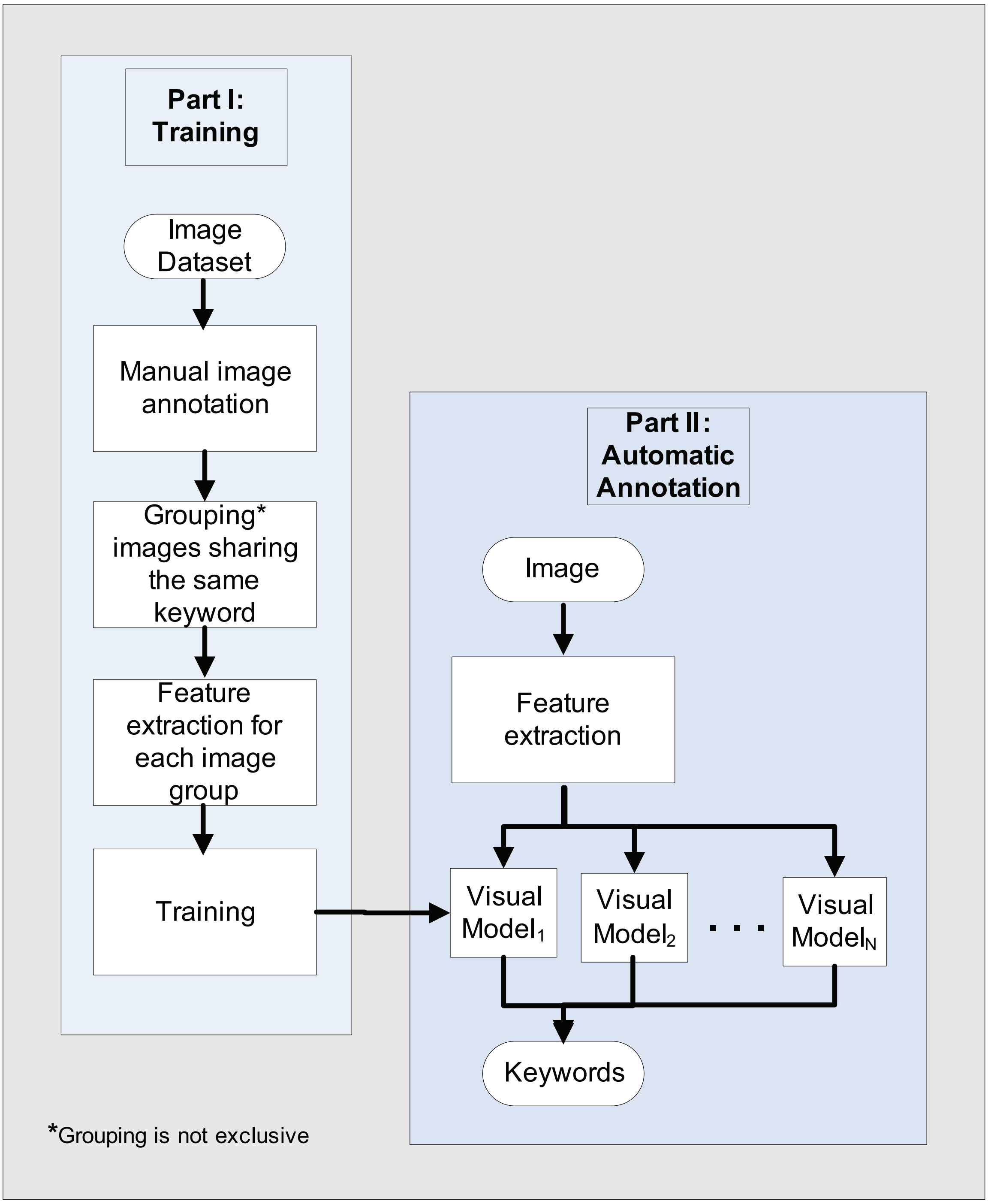


Figure 3.1: Automatic image annotation by modeling keywords.

In the first part, visual models for all available keywords are created. Initially, training data are collected and the most representative keywords are selected to create the visual models. Let assume that the set of the selected keywords is represented by $\boldsymbol{K} = \{K_1, ..., K_N\}$. The K_i indicates the *i-th* keyword while the total number of keywords is denoted by N. Images sharing the same keyword are grouped together and create the set of groups $\boldsymbol{G} = \{G_1, ..., G_N\}$, where G_i denotes the *i-th* group of a total number of N groups of images. Low level features are extracted from images and used to create the set of visual models $\boldsymbol{V} = \{V_1, ..., V_N\}$, where V_i indicates the visual model for the keyword K_i. Visual models are created using machine learning techniques following the one-against-all training paradigm. In the second part, annotations are assigned to an unseen image based on the output of these models, once they fed by a feature vector extracted from the input image. A detailed example of the automatic image annotation using visual models is given in Fig. 3.2 and Fig. 3.3.

3.2 Open Issues

The development of automatic image annotation using visual models can be divided into three main steps: (i) the dataset creation, (ii) the low level feature extraction, and (iii) the creation of visual models. Creating accurate visual models for keywords depends not only on the training data, but also on the low level feature set and machine learning technique that is used. Availability of training data and the use of especially designed learning algorithms for feature extraction and visual modelling, are three important factors that were carefully investigated in the framework of this thesis.

Concerning the training data, manual image annotation is costly and time consuming task. Furthermore, it is usually assigned to experts of the domain that images' content refers to. On the other hand, end-users of the search engines are non-experts. The current thesis investigates and presents alternative ways of creating training data aiming to address these issues in chapter 4. First, it proposes the manual image annotation

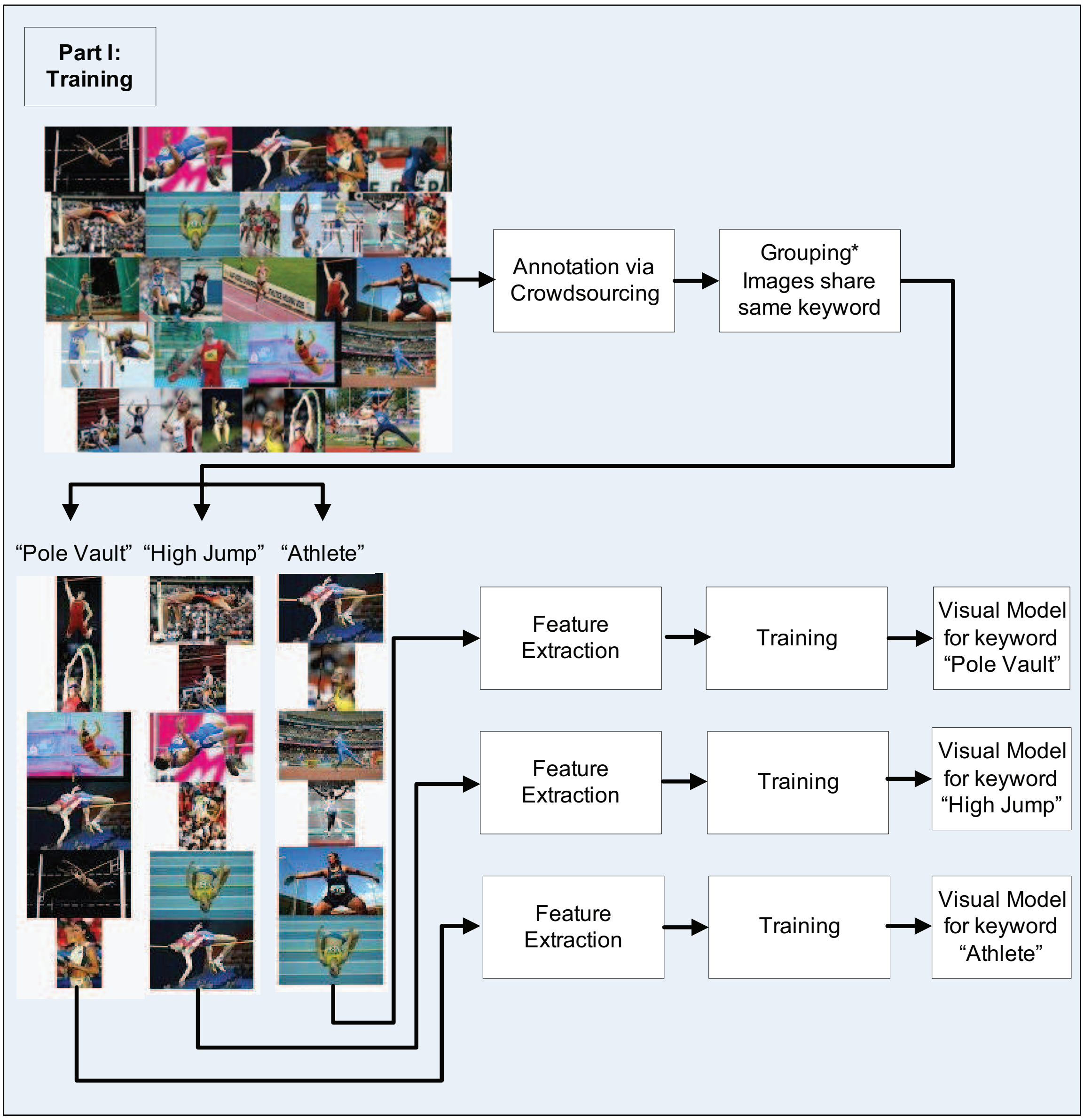

Part I: Training
Annotation via Crowdsourcing
Grouping* Images share same keyword
"Pole Vault"
"High Jump"
"Athlete"
Feature Extraction
Training
Visual Model for keyword "Pole Vault"
Feature Extraction
Training
Visual Model for keyword "High Jump"
Feature Extraction
Training
Visual Model for keyword "Athlete"

Figure 3.2: Example of creating visual models.

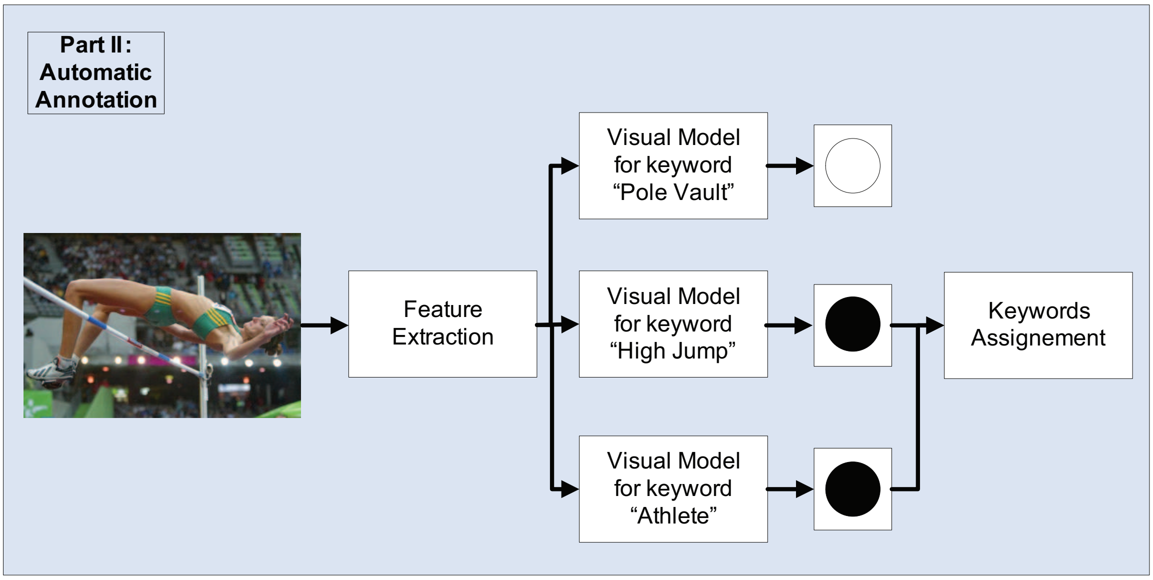

Part II: Automatic Annotation
Feature Extraction
Visual Model for keyword “Pole Vault”
Visual Model for keyword “High Jump”
Visual Model for keyword “Athlete”
Keywords Assignement

Figure 3.3: Example of automatic image annotation using visual models.

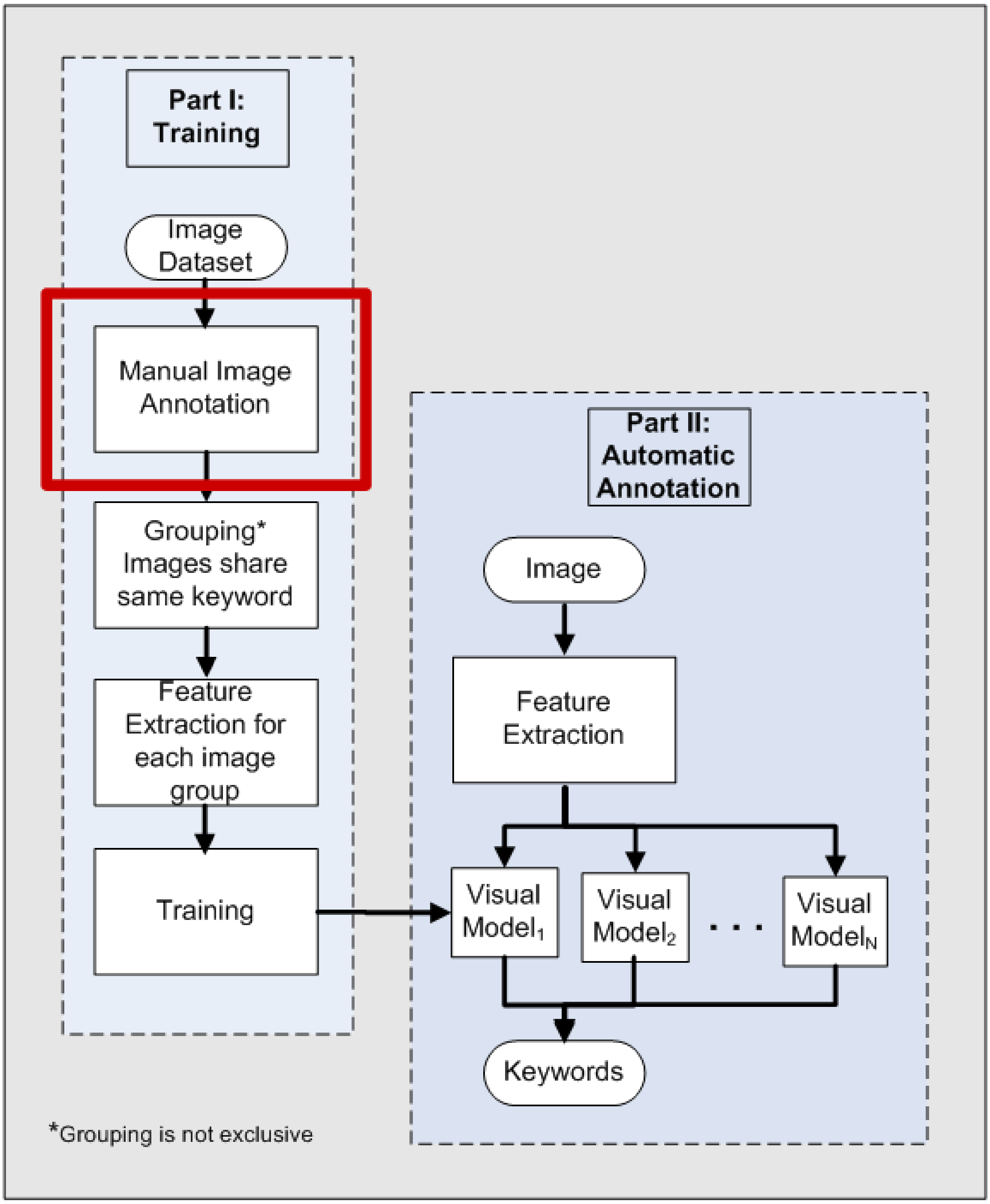
through crowdsourcing and investigates several factors that may affect the created annotations. The gender and age influence is examined in chapter 5 while the quality of the annotation is investigated in chapter 6. Second, it proposes the automated extraction of keywords for web pages using the surrounding text included in HTML web pages. A study for assigning the relevant textual segments to web images is presented in chapter 7.

The selection of the appropriate low level feature extraction algorithm is a crucial step for the creation of the visual models. The selected algorithm should be able to identify relevant features in the given images, and extract a feature vector with fixed and limited dimension (i.e. non-fixed and huge dimensionality causes certain problems when used in machine learning schemes). Chapter 8 presents several available low level feature extraction algorithms. Chapter 9 investigates the evaluation performance of MPEG-7 descriptors on keyword extraction. Finally, a new low-level feature extraction algorithm is presented in chapter 10.

Regarding the modelling process, the efficiency, the robustness to the variation of learning parameters and the effectiveness of the selected learning algorithm must be taken into consideration. Furthermore, the problem of the limited training data plays significant role in keyword modelling. Chapter 11 details several available learning techniques and proposes the co-training scheme as an option to overcome this issue.

Part I

Creating Training Examples


Part I: Training
Image Dataset
Manual Image Annotation
Grouping* Images share same keyword
Feature Extraction for each image group
Training
Part II: Automatic Annotation
Image
Feature Extraction
Visual Model$_1$
Visual Model$_2$
. . .
Visual Model$_N$
Keywords
*Grouping is not exclusive

Chapter 4

Dataset Creation

The collection of training data is a crucial step for the creation of accurate visual models. Training examples that are used for creating visual models for keywords are pairs of images and keywords. The low-level feature vector extracted from the image is considered as an example of the visual representation of keywords assigned to this image. Aggregating feature vectors across many images eliminates the case of having several keywords sharing exactly the same training examples. However, collection of manually annotated images to be used for creating the keyword visual models is a costly and tedious procedure. Manual annotations are likely to contain human judgment errors and subjectivity in interpreting the image due to differences in visual perception and prior knowledge. As result is a common practice nowadays to use multiple annotations per image obtained from different people to alleviate this subjectivity as well as for detecting outliers or erroneous annotations. In the past, manually annotated datasets were obtained by experts. Since the majority of tomorrow users of search engines are non-experts, the idea of modeling the knowledge of several people rather than an expert can significantly improve the ultimate efficiency of image retrieval systems. This chapter presents the collection of training data through crowdsourcing following this direction. Furthermore, it proposes the idea of extracting keywords from the surrounding text for web images, as an alternative solution for the problem of dataset creation.

4.1 Collecting Image Annotations Through Crowdsourcing

Crowdsourcing, and the act of outsourcing work to a large crowd of workers, is a specific form of harvesting wisdom of the crowd and contributions of users [96]. The distributed model of crowdsourcing assigns tasks traditionally undertaken by employees or contractors to an undefined crowd [97], [98]. Although crowdsourcing annotation is a fairly recent development, it is recognized as a growing and burgeoning research area, as evidenced by several works that have produced an overview of these methods from different perspectives [96]. Crowdsourcing has attracted the interest of several researchers and companies since, among others, it is a very attractive solution to the problem of cheaply and quickly acquiring annotations. The Amazon Mechanical Turk (MTurk) [99], extends the interactivity of crowdsourcing tasks by more comprehensive user interfaces and micro-payment mechanisms. MTurk is an online labor market where workers are paid small amounts of money to complete small tasks. It is possible to assign annotation jobs to hundreds, even thousands, of computer-literate workers and get results back in a matter of hours [100]. The quality of annotations provided by MTurk workers has been explored for a wide range of annotation types [101]. Sorokin et al. [102] presented a data annotation framework to obtain project-specific annotations very quickly on a large scale via MTurk. During the creation of the ImageNet dataset [103], MTurk workers were utilized to verify that each collected image contains objects from a given set of multiple words or phrases.

The Crowdcrafting platform [104], unlike the MTurk, supports volunteer-driven projects without handling payment or money. Crowdcrafting is a new, free, open-source platform that enables people to create and run crowdsourcing and micro-tasking applications. Although several image annotation applications are currently running on Crowdcrafting platform, there no relevant publications due to its recent establishment.

Annotation quality obtained through crowdsourcing varies. Sometimes annotators provide random or bad quality labels in the hope that they will go unnoticed and still be paid [105]. Others may have good intentions but completely misunderstand the task at hand [100]. Several studies have been presented that investigate the annotation quality obtained with crowdsourcing approaches. A study from Snow et al. [106] showed that crowdsourced annotators are not as effective individually as experts but, when non-expert opinions are aggregated together, it is possible to produce high-quality annotations. So their work establishes the merit of annotations' aggregation, suggesting that using a large number of untrained annotators can yield annotations of quality comparable to those produced by a smaller number of trained annotators on multiple-choice labeling tasks. Non-experts have been proven as poor annotators for video annotation. The experimental results in [107] indicate that traditional crowdsourced micro-tasks are not suitable for such a case since video annotation requires specialized skill. Thereby, a small group of experts is necessary for a high quality video labeling.

In [108] the problem of training a supervised learning system in the absence of ground truth data is addressed, when all that are available are noisy label information from non-expert annotators. The authors suggest that having effective annotators is more important than data coverage, and emphasize the use of multiple annotations for each item, in conjunction with weights for annotators based on their agreement with the induced ground truth. The paper aims at estimating the sensitivity and specificity of each of the annotators, and also annotates unlabeled examples. In [109] the difficulty of evaluating tasks where annotations are available from multiple annotators, but no ground truth is available as a reference, is also analyzed. The authors try to integrate the opinions of many experts to determine a gold standard, proposing also that annotator consensus can be used as a proxy in order to measure annotation quality. Annotator quality is also modeled in [110], where it is also showed how repeated and selective labeling increased the overall labeling quality on synthetic data. Furthermore in [111] a method for combining prioritized lists obtained from different annotators is proposed. Annotator consistency to obtain ground truth has also been used in the context of paired games and CAPTCHAs [112], [113]. In [114] two issues are considered:

the difficulty of non-expert annotation and the ability of annotators, while in [115] a system is proposed which actively asks for image labels that are the most informative and cost effective.

The assignment of several raters into annotation task introduces the problem to decide whether an annotation is a positive or as negative example if the ratings disagree. The inter-agreement is a good indicator for annotation quality and a high agreement rate can lead to more accurate and reliable annotation. Kilgarriff [69] presents a detailed methodology for creating gold standard datasets and states the necessity of more than one person to create the dataset, that one should calculate the inter-annotator agreement and determine whether it is high enough. He also underlines basic reasons for ambiguous annotations like the poor definition of annotation scheme, mistakes of annotators due to lack of motivation or knowledge. Snow et al. [106] examined the accuracy of labels created using Mechanical Turk for a variety of natural language processing tasks and measured the quality of non-expert annotations by comparing them against labels that had been previously created by expert annotators. They report inter-annotator agreement between expert and non-expert annotators, and show that the average of many non-experts converges on performance of a single expert for many of their tasks. Callison-Burch in [116] evaluated the translation quality using MTurk and found out that a combination of non-expert judgments has a high-level of agreement with the existing gold-standard judgments of machine translation quality, and correlates more strongly with expert judgments than Bleu does. Brants in [68] investigated the inter-annotator agreement for part-of-speech and structural syntactic annotations in the NEGRA (a syntactically annotated corpus of German newspaper texts) by determining the accuracy and F-score between annotated corpus of two annotators. Veronis presented a systematic study of polysemy judgements and inter-annotator agreement for word sense disambiguation [117], while Chklovski and Mihalcea studied the agreement of web users who contribute the word sense annotation [118]. In [119] an interesting approach is presented on how much several sets of expert annotations differ from each other and if the non-expert annotations are reliable enough to provide ground truth annotations. Four experiments on inter-annotator agreement are conducted applied to the annotation of images from MIR Flickr that were annotated first from 11 different experts and then the set was distributed over MTurk to nine non-expert annotators. The inter-annotator agreement is computed at an image-based and concept-based level using majority vote, accuracy and k statistics.

Having the above in mind, crowdsourcing annotation opened a new way on human computation. Given that crowdsourcing-based annotation is the method of preference for manual image annotation, we have explored this issue further emphasizing on: (i) the effects of the selected annotation method, (ii) image content itself, and (iii) the used lexicon on the quality of annotation. The findings of this study are reported on chapter 6.

4.2 Getting Image Annotations Through Web Documents

An alternative to manual annotation to create training data is to explore the successful mechanisms of automatic keyword extraction in text-based documents adopted by contemporary search engines. The large amount of web images located in text documents and web-pages can be used for that purpose. The text that surrounds these images inside the web documents provides important semantic information that can be used for keyword extraction. Web Image Context Extraction (WICE) denotes the process of determining the textual contents of web document that are semantically related to an image and associates them with that image. WICE uses the associate text as a source for deriving the content of images. In text-based image retrieval, the user provides keywords or key phrases and text retrieval techniques are used for retrieval of the best ranked image. Successful web image search engines like the Google images[1] and Yahoo!Image Search[2] are well known WICE examples.

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

[2] http://images.search.yahoo.com/.

HTML document content structure such as image file names, anchor texts, surrounding paragraphs or even the whole text of the hosting web page are usually used as a textual content in WICE applications. The initial research efforts focus on exploiting the range of HTML document content structure [120], [121]. Although, they give effective results for a subset of images they cannot be used for general purpose since such textual information often do not give sufficient information for the visual content. More sophisticated methods divide the textual content into text blocks and relevant blocks are extracted for each image. These text blocks are then used to extract the keywords. Moreover, as presented in the previous section, other methods leverage the visual content together with the textual for more accurate annotation.

The diversity of the designing concepts, the noisy environment (advertisements, navigation bars, etc.) as well as the too much textual and visual information, affect the accurate assignment of the text blocks to web images. Thereby, the assignment of textual content to an image remains one of the key issues in WICE applications. Fortunately, there is regularity to the appearance of relevant surrounding text with respect to the position of an image in an HTML text. Several approaches have been proposed that exploit this regularity to extract the text blocks as concept sources for images.

A bootstrapping approach to automatically annotate web images based on predefined list of concepts by fusing evidences from images content and their associated HTML text in terms of a fixed size of sequence is presented in [122]. The authors assume that the relevant surrounding text is contained in the nearby text tagged by HTML structural tags and appears to the left or right of the image. If the extracted text is more than 32 words then only the first 32 words are used as surrounding text. On the other hand, the visual classifier is trained utilizing traditional visual content features such as color histogram, DCT texture and statistical shape features. Both textual and visual classifiers are used in a co-training framework for annotating web images. The main drawback of the proposed method is the fixed size of sequence that lead to low annotation performance since extracted text may be irrelevant to the corresponding image, or on the other hand, important parts of the relevant text may be discarded.

In [123], [124], the Document Object Model (DOM) tree structure of the web page is utilized to extract the surrounding text of the images. Fauzi et al. [123] proposed a DOM tree-based segmentation algorithm that utilizes the image characteristics as it appears on HTML document and extracts image segments. Alcic and Conrad [124] used the hierarchical structure of the DOM tree to calculate the distance of the web contents. The distance measure is used to extract image content based on 1D clustering where each image is associated with the textual contents of the cluster it belongs to. In general the DOM tree structure based methods are not adaptive and they are designed for specific design patterns.

The Vision based Page Segmentation (VIPS) [56], a heuristic DOM-based segmentation algorithm, can also be used to segment the Web page into a number of semantic blocks where each block contains semantically related information. The degree of correlation between the contents within a block is determined by the Degree of Coherence (DoC). The DoC is calculated using heuristic rules on the DOM tree structure and visual cues and ranges from 1 to 10. The higher the DoC value, the higher the correlation between the contents within the block. For each image, the block containing this image is identified and the surrounding text is extracted as its textual information.

In [125], the VIPS was utilized to accurately model the semantic relationships among images on the Web. An image graph was created using the page-to-block, block-to-image, block-to-page relationships. Techniques based on spectral graph and Markov Chain theory are then used for image ranking clustering and embedding. Web page segmentation is indeed a more adequate solution to the problem of text extraction since it is adaptable to different web page styles and depends on the visual cues that form each web page. Most of the proposed algorithms with this approach though, are not designed specifically for the problem of image indexing and therefore often deliver poor results [126].

4.3 Refinement of keyword extraction using semantics

An interesting approach was presented in [51] where the context extraction is achieved by utilizing VIPS algorithm and semantic representation of text blocks. In the proposed method, the whole text found in the web page is used as a source to extract content information for the web images. More than the structural text blocks extracted using VIPS, text fragments are assigned to images after their semantic representation. This representation is achieved using the WordNet project [127].

WordNet is a popular research product that attempts to model the lexical knowledge that a native English speaker possesses. WordNet models each unique meaning of a word as a synset (*i.e* synonym set), which is always accompanied by an annotation which defines the exact concept of this word or set. An example of such a synonym set is the group {car, auto, automobile, motorcar} which contains words that native English speakers would consider synonyms. WordNet not only provides these groups and the corresponding definitions, but also important knowledge on the semantic relations that appear between different sets. These relations, namely the “IS-A”, “Has-Part”, “Part-of” and “Member-Of” form a hierarchy of words. A part of this taxonomy, which contains the synset {car, auto, automobile, motorcar}, is presented in Fig. 4.1.

WordNet contains several Part Of Speech (POS) items, such as nouns, verbs, adverbs and adjectives. However, only nouns and verbs are organized into hierarchies based on “Is-A” relations as those which are presented in Fig. 4.1. WordNet does not cross POS boundaries and therefore no judgements between concepts of different POS can be made. This property facilitates the creation of flexible semantic similarity measures between synsets of the same POS.

In the literature, several methods attempt to quantize the similarity between WordNet synsets. The Path length (S_{pal}) metric is a simple node-counting scheme as the similarity score it describes is inversely proportional to the number of nodes along the

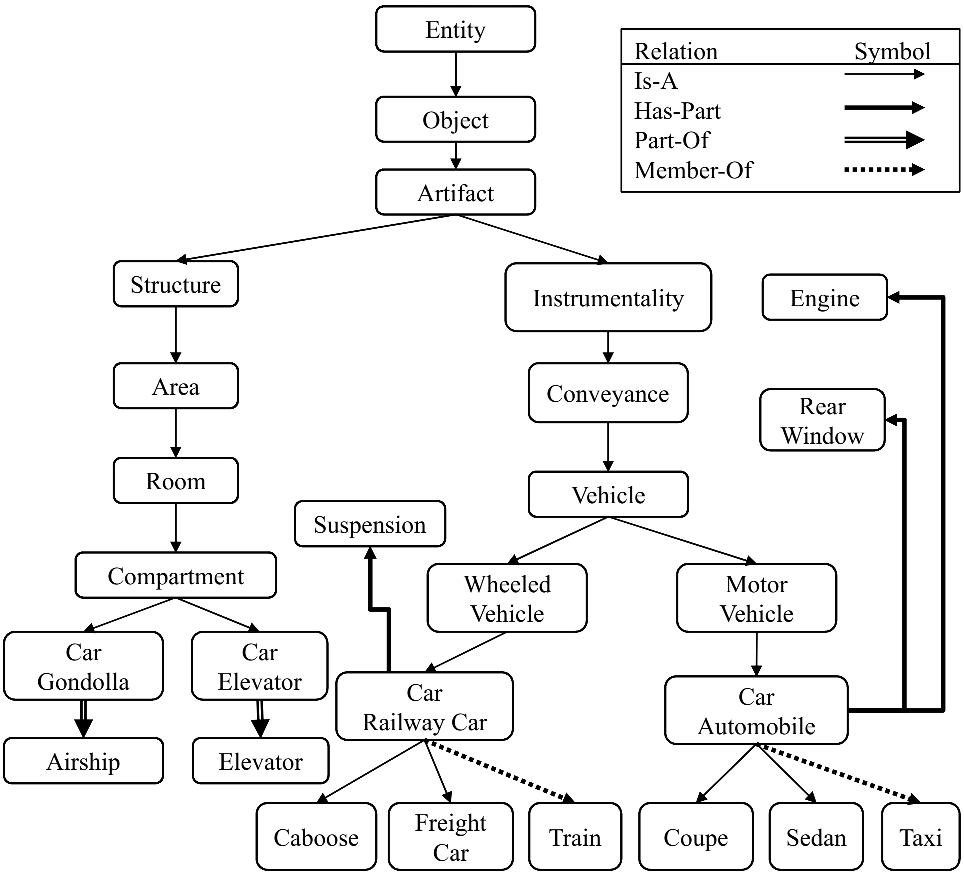


Figure 4.1: A part of the WordNet hierarchy.

shortest path, $L(s_1, s_2)$ between two synsets s_1 and s_2:

$$S_{pal} = \frac{1}{L(s_1, s_2)} \quad . \tag{4.1}$$

On the other hand, the similarity measure proposed by Leacock and Chodorow in [128] is given by:

$$S_{lch} = -log\frac{L(s_1, s_2)}{2D_{max}} \quad , \tag{4.2}$$

where D_{max} is the maximum depth in the WordNet taxonomy.

In [129], Wu and Palmer proposed an alternative measure that calculates the similarity between two synsets s_1 and s_2 by considering their depths in the WordNet hierarchy, along with the depth of their least common subsumer (LCS) s_{lcs}:

$$S_{wup} = \frac{2D(s_{lcs})}{D(s_1) + D(s_2)} \quad , \tag{4.3}$$

where $D(s_1)$ and $D(s_2)$ the depth of the synsets s_1 and s_2 in the WordNet taxonomy. In Tab. 4.1 the similarity values for the two synset pairs P_1 ={structure, instrumentality} and P_2 ={wheeled vehicle, taxi} are presented. The given values for the depths of the nodes and their LCS may not correspond to the actual depth when the whole WordNet hierarchy is used, however the comparison among the three different measures is still valid since the given part of the taxonomy may be considered autonomous. The S_{pal} assigned the lowest similarities to the pairs compared to the other two measures. For the synsets of pair P_1, which are siblings in the taxonomy, the metric resulted to a rather low score, while for the pair P_2, which is located quite deep into the taxonomy, and therefore it is formed by less abstract terms, with similar meanings, resulted to an even lower score.

On the other hand, the S_{lch} measure takes into account the depth of the taxonomy

Table 4.1: The similarity values between two synset pairs found in Fig. 4.1.The maximum depth of the taxonomy is 9.

Synsets	$L(s_1, s_2)$	$D(s_1)$	$D(s_2)$	$D(s_{lcs})$	S_{pal}	S_{lch}	S_{wup}
P_1 ={structure, instrumentality}	3	4	4	3	0.33	0.78	0.75
P_2 ={wheeled vehicle, taxi}	5	7	9	6	0.2	0.56	0.8

in which the synsets are found and therefore it is affected by the presence or absence of a unique root node. This form does not ensure a value range; it is not possible to conduct a meaningful comparison between the numbers 0.78 and 0.56, if we are not aware of the maximum values these similarities could have. Moreover, this measure still rates P_2 lower in terms of similarity while one would expect the opposite. It seems that measure S_{wup}, which ensures a range $0 < S_{wup} \leq 1$ and also takes into consideration how abstract the two terms are, gives the best results in comparing the similarity between two synsets. Among other similarity measures, the above mentioned ones are implemented in the Perl module WordNet::Similarity [130], which is used in the developed system as described below.

In this thesis we have also examined on alternative way of extracting image annotation from web pages. The method is presented in chapter 7.

Chapter 5

Age and Gender Differences in Manual Image Annotation

Manual image annotation is an essential step for the development of methods for automatic image annotation which are based on the learning by example paradigm. However, manual image annotation contains human judgment errors and subjectivity in interpreting the image due to differences in visual perception and prior knowledge. Multiple annotations per image obtained from different people can definitely alleviate this subjectivity and facilitate the whole procedure. Image annotation can be approached as a social cognitive process and, thus, it varies among people based on their socio-demographic characteristics. In the current study we investigate the effect of age and gender of annotators in manual image tagging. We consider the case when a controlled vocabulary is used versus free keywords annotation. The findings of this study provide interesting insights into the relationship between annotator demographics and image annotation. The experiments reveal that age affects seriously the way people annotate images either with the aid of a vocabulary or via free keywords. On the other hand gender appears not to play a significant role in image annotation in contrary to our initial assumption.

Since judgments of meaning are more or less subjective, image tagging is also expected to vary across people. Thus, there is no single correct label for an image. Labels assigned to an image by different annotators may vary and may be related to annotators' knowledge, age, gender, intuitions, background, etc [131]. Crowdsourcing [96], the act of outsourcing work to a large crowd of workers, has attracted the interest of several researchers and companies. Among others, it is a very attractive solution to the problem of cheaply and quickly acquiring annotations. In such a case, the assignment of several raters into annotation task introduces the problem to decide whether an annotation is a positive or a negative example if the ratings disagree. The research community has focused on bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics [132], [52] while the semantic gap between users has not been examined.

5.1 Research Questions

In this study we aim to examine the existence of the semantic gap between people and in such a case, what is the influence of some demographic factors like age and gender. In other words, this study examines the differences in the way people with diverse gender and age annotate images.

Our study focuses on the following research questions:

- Do age and gender affect the way people annotate images using vocabulary keywords?
- Do age and gender affect reliability agreement?
- Do age and gender affect the way people annotate images using free keywords?
- Can we assume a strong relevance between image interpretation and image annotation?

The underlying assumption is that the way humans annotate images is strongly related with the way they evaluate image retrieval results. Thus, differences in image annotation indicate differences in the way they interpret the results.

5.2 Data Collection

A survey of 40 annotators was conducted between March 1^{st} to 31^{st}, 2013, in Cyprus. The participants were Cypriot citizens aged between 15 and 55. We aimed for an even distribution of participants concerning age and gender groups, therefore the set was divided into two age groups consisting of people aged 15 - 35 and 36 - 55 years old. Age seperation was based on the idea that the participants of the first age group are more familiar with the Internet which it became popular in Cyprus in the mid 1990s. We tried to have an equal number of females and males in each group. The first age group (15-35) comprised 11 females and 9 males, while the second group (36-55) contains equal number of males and females. Survey population characteristics are summarized in Tab. 5.1.

Table 5.1: Description of participants.

Age	***Gender***				***Total***
	Female		***Male***		
	#	%	#	%	
15-35	11	55	9	45	20
36-55	10	50	10	50	20

Participants were asked to fill in a questionnaire consisting of two parts. In the first part, we asked the participants to fill in their demographic information including age and gender. The second part included the image annotation content. The average time required to complete the questionnaire was about 60 minutes.

5.3 Image Annotation

Survey's participants annotated a set of images using vocabulary and free keywords. The image dataset was formed by 50 images downloaded from the web covering different subjects. The preselected set of vocabulary keywords consisted of 52 keywords from 5 different main topic categories (Fig. 5.1). The participants were asked to annotate the image dataset by selecting 1 to 5 keywords from the controlled vocabulary and adding 1 to 5 free keywords.

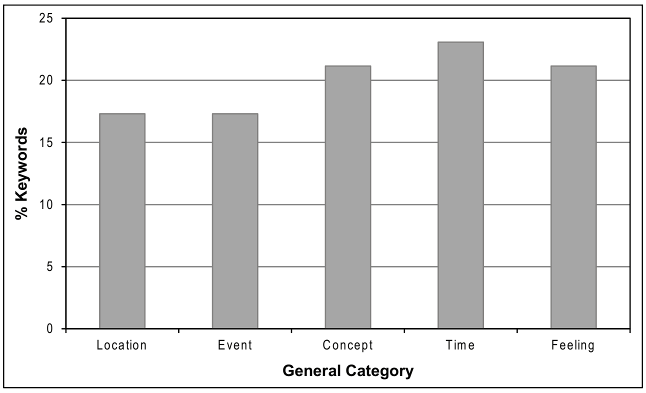


Figure 5.1: The categorization of the vocabulary keywords into general categories.

After collecting and saving the annotations we applied the following measurements in order to investigate the differences in the way that people annotate images using vocabulary keywords and to calculate the reliability of the agreement among the participants of different age and gender groups. First, for each category we calculated and compared the total number of keywords that were selected from participants for all age and gender groups and then we computed the reliability of the agreement among the participants on category-keyword basis using the kappa statistics.

This statistical measure was first proposed by Cohen [133] and reports the inter-rater agreement as a range of values between 0, which stands for the level of agreement that is expected from random assignment, and 1, when there is perfect agreement. Negative values of kappa indicate agreement below the expected from random assignment. A study from Landis et Koch [134] states that a kappa value above 0.6 represents an adequate annotator agreement while a value above 0.8 is considered as almost perfect. The free marginal kappa statistic [135], which can be utilized for any number of participants that are not forced to assign a certain number of images to each keyword, was exploited separately in a binary scenario for each one of the 5 keyword categories.

In order to obtain more nuanced description of the intrinsic quality of the terms given as free keywords, we categorized them into the cognitive categories of nouns using the Prototype Theory [136]. Prototype Theory is a product of cognitive psychology that provides different levels of inclusiveness for categorizing concepts. It has also applied into cognitive linguistics and suggests three category levels of concepts: (a) “Basic”, (b) “Superordinate”, and (c) “Subordinate”. The “Basic” is the most inclusive level among the others where its members have common attributes and a single mental image can be formed. “Basic” level concepts are usually labeled with the most occurring, contextually neutral, shortest terms that are countable nouns [137]. Concepts falling into “Superordinate” level share fewer attributes among each other and are named with mass nouns. Finally, the “Subordinate” level contains concepts with many overlapped attributes that are morphologically complex nouns. The “Superordinate” corresponds to the highest abstract level while the “Subordinate” corresponds to the lowest one.

The nouns given as free keywords were categorized into “Basic”, “Superordinate”, and “Subordinate” categories. The total number of keywords falling into the three cognitive categories was calculated for each annotator separately and then the age and gender differences were investigated.

5.4 Results

5.4.1 Age and Gender Influences in the Way People Use Vocabulary Keywords in Image Annotation

A comparison among the used keywords was made for both age and gender groups. Tab. 5.2 outlines the total number of selected keywords for each category when comparing female and male participants as described above. It is evident that both groups show quite similar behavior, selecting more keywords from “Feeling”, “Concept” and “Time” categories. Despite the number of keyword selection being slightly differing and women are using more keywords for annotation, the only category with a significant difference is “Feeling” (p-value = 0.03). In the case of comparing the total number of selected keywords for all participants in both groups, the p-value is too high to make their values significantly different (p-value = 0.29).

Concerning the two age groups, it is clear from Tab. 5.3 that the participants used to select keywords mainly from “Feeling”, “Concept” and “Time” categories with a significant differences in the way of using keywords from 3 categories: “Concept” (p-value = 0.02), “Time” (p-value = 0.03) and “Event” (p-value = 0.02). There is also a significant difference in the total number of selected keywords from all categories (p-value = 0.04), with the first age group having the highest score.

Table 5.2: Gender differences in the use of vocabulary keywords.

Keyword Category	Location	Event	Concept	Time	Feeling	Total
Female (n=21)	516	301	622	614	716	2769
Male (n=20)	279	270	457	439	407	1852
p-value	0.55	0.32	0.24	0.70	0.003	0.29
Interpretation	n.s.(not significant)	n.s.	n.s.	n.s.	s. (significant)	n.s.

Table 5.3: Age differences in the use of vocabulary keywords.

Keyword Category	Location	Event	Concept	Time	Feeling	Total
15-35 (n=20)	468	301	584	570	628	2551
36-55 (n=20)	327	280	495	484	514	2100
p-value	0.39	0.02	0.02	0.03	0.65	0.04
Interpretation	n.s.(not significant)	s. (significant)	s.	s.	n.s.	s.

5.4.2 Age and Gender Effects in Reliability Agreement

Inter-participant agreement was calculated using an online kappa calculator tool [138]. Fig. 5.2 presents the kappa statistics for the 5 keyword categories. It is evident, that there are no basic gender or age influences in the inter-annotator agreement of participants. On average the 21 females agree with the value of 0.65, with adequate agreement in all categories except the “Time” category which presents the lowest agreement value (0.56) among female participants. On the other hand, the 19 males agree with a quite higher average value (0.67) and show an adequate agreement in all categories. Regarding the age, both groups agree with the same kappa value equal to 0.66. Younger participants adequately agree in 4 of the 5 categories (except "Time"), while older ones show values greater than 0.6 in all categories. The "Time" category presents the lowest agreement among the participants with values lower than 0.6 in female and 15-35 groups. The "Event" category presents the highest values, reaching almost the perfect agreement for all age and gender groups.

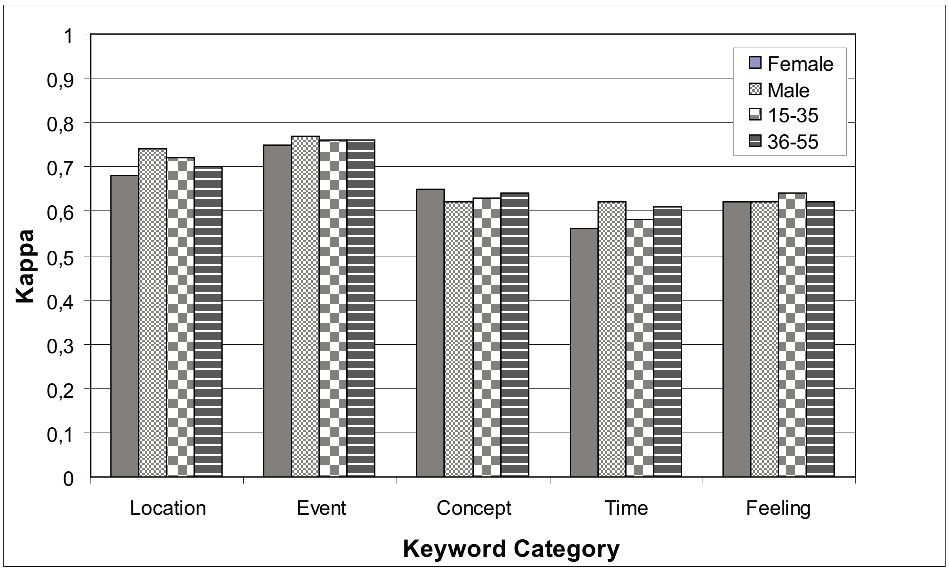


Figure 5.2: The Kappa values for the the keyword categories.

5.4.3 Age and Gender Influences in the Use of Free Keywords

After a manual identification and correction (spelling or typing errors, synonymy, etc) of the terms given as free keywords, a total of 529 different keywords were collected and used in our experiments. Categorization of the keywords into the cognitive categories mentioned earlier, is presented in Fig. 5.3. Keywords related to other parts of speech such as verbs, adverbs, etc, were categorized as "Other" that could be considered as the fourth category in our experiments. The 56.52% of the keywords were categorized as "Basic", the 26.84% as "Subordinate", the 9.45% as "Superordinate", while the 7.18% was assigned to the "Other" category.

The total number of times that each keyword was given during the annotation process was calculated separately for both age and gender groups. As shown in Tab. 5.4, there is no significant age influence in the use of keywords for all cognitive categories. However, women added more free keywords than men. Both women and men added mostly terms from the "Basic" category. The 65.26% of terms given by women and the 68.20% of terms given by men were categorized as "Basic" while the rest were categorized into the remaining categories.

Table 5.4: Gender differences in the use of free keywords.

Cognitive Category	Basic	Superordinate	Subordinate	Total
Female (n=21)	883	90	380	1353
Male (n=20)	755	76	276	1107
p-value	0.52	0.71	0.09	0.20
Interpretation	n.s.(not significant)	n.s.	n.s.	n.s.

Concerning the age influence, the younger group used more free keywords to annotate the dataset (Tab. 5.5). There are considerable differences in the way that people use free keywords in terms of cognitive categories. There are significant differences for "Basic" (p-value=0.03) and "Subordinate" (p-value=0.02) categories while there is no significant difference for "Superordinate". There is also a significant difference in the total number of given free keywords (p-value=0.001), regardless the category, with the young people having the highest score.

Table 5.5: Age differences in the use of free keywords.

Cognitive Category	Basic	Superordinate	Subordinate	Total
15-35 (n=20)	893	81	400	1374
36-55 (n=20)	745	85	256	1086
p-value	0.03	0.79	0.02	0.001
Interpretation	s.(significant)	n.s. (not significant)	s.	s.

5.5 Discussion

The way that people understand and describe the content of an image was investigated in this study. The participants were asked to annotate an image dataset by assigning

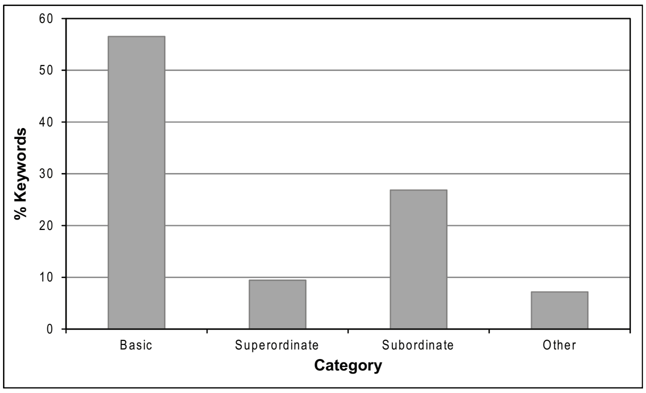


Figure 5.3: The categorization of free keywords into cognitive categories.

1 to 5 vocabulary keywords and supplement the description by adding 1 to 5 free keywords. The age and gender gaps in image annotation were examined through different measurements.

Concerning the vocabulary keywords, the way how women and men annotate images is not significantly different. Regarding the gender, a significant difference was occurred only in the use of keywords related to the “Feeling” category, in which women tend to select more keywords than men. This finding is perfectly justified since women are considered more emotional than men [139]. When examining the impact of age on the way of people annotating images, the results show that there are differences between younger and older participants. Significantly, younger people used in general more keywords to annotate images and there are also significant differences in the way of annotating images using keywords from the “Concept”, “Time” and “Event” categories. Overall, females and younger participants used more vocabulary keywords to annotate the image dataset. Women used in total 2769 keywords (2.64 per image), while people in 15-35 group used 2551 keywords (2.43 per image).

The fact that there are no age or gender differences in inter-annotator agreement is also a significant finding of this study. On average, participants of both gender and age groups adequately agree with a value kappa of 0.65-0.67. The majority of keyword categories presents an adequate agreement in all but the “Time” category which presents the lowest agreement value among female (0.56) and young (0.58) participants, but this may be due to the abstract nature of the keywords in this category.

The participants gave in general 529 valid free keywords where the 56.52% of them were classified into the “Basic” category. The results are in full agreement with the idea drawn by Laokff [140] which indicates that images are identified most accurately and categorized fastest at the “Basic” level. The experimental results indicate that there are no any significant gender differences in the way people give free keywords in terms of the cognitive categories. At the same time, the younger participants have significant differences compared to the older ones. Considerable differences were revealed in the total number of the given free keywords, as well as, in “Basic” and “Superordinate” categories. Overall, the annotators assigned more vocabulary than free keywords to the image dataset. The average number of given free keywords per image is in the range of 1-1.4 for all gender and age groups.

The results of this study reveal that there is significant age influence in image interpretation and description through vocabulary and free keywords. However, there is no considerable gender influence (except on the use of vocabulary keywords related to the “Feelings” category). The results are justified by related studies conducted on gender and age differences in performing several tasks on internet, including the search. Singer et al. [141] confirm that users age impacts behavior and search performance significantly, while gender influences were smaller than expected.

The findings of the proposed study provide interesting insights into the relationship between annotator demographics and image annotation, as well as, the relationship between annotator and cognitive categories of nouns. They are lead to better understanding how the people annotate images since the annotators and vocabulary used during annotation assessment have to be chosen with care while the resources should be used effectively [69]. The identification of the demographic factors that may influenced the manual image annotation is essential for choosing and training future annotators, creating controlled vocabularies and golden standard image datasets. The presenting findings definitely will play a key role not only in future research of the domain, but in the development of more efficient and accurate text-based image retrieval systems.

5.6 Conclusions and Remarks

This chapter presents an attempt to investigate the semantic gap between people in the way of annotating an image dataset using vocabulary and free keywords. The study also aimed at looking at the gender and age influences in inter-annotator agreement using the vocabulary keywords. The experiments were conducted on a dataset that consists of 50 images, downloaded from the web covering different subjects. Forty participants were asked to annotate the image dataset by selecting 1 to 5 keywords from a controlled vocabulary and adding 1 to 5 free keywords. The results revealed that age affects the way people annotate images using both vocabulary and free keywords. Concerning the inter-annotator agreement, there is an adequate agreement among the gender and age groups. The used image dataset as well as the annotations assigned during this study are stored in a database[1] and they are both available for research experiments and cooperative evaluations.

[1] http://cis.cut.ac.cy/∼z.theodosiou/Database1.zip.

Chapter 6

Quantifying the Quality of Manual Image Annotation

In the framework of this thesis a study was conducted for investigating the manual image annotation quality aiming at: (a) identifying to which extend the use of structured lexicon and unstructured vocabularies improves annotation quality and at what cost (missing useful and valid annotations), (b) exploring to which extend and under what prerequisites free annotation can lead to valid and useful image annotation, and (c) inquiring the effect of image content itself on valid image annotation. For the experimental setup, 500 images were manually annotated using three annotation methods. The annotations were evaluated for each method independently and the results lead to important conclusions and revealed very interesting issues for further study.

6.1 Commandaria Platform

The Commandaria platform [1] has been developed in the framework of Commandaria project [2] and offers along with the user profiling, the feasibility of uploading, anno-

[1] http://cis.cut.ac.cy/CommandariaPortal.

[2] "The History of Commandaria: Digital Journeys Back to Time", project funded by the Cyprus Research Promotion Foundation (CRPF) under the contract ANTHRO/0308(BE)/04.

tating and searching data. In this project users from Cyprus and around the world were invited to: (a) provide multimedia content which relates the famous Commandaria wine with the history and culture of Cyprus, and, (b) annotate existing content. The Commandaria data collection consists approximately of 7500 files related to the Commandaria Cypriot wine. The 3500 files are digitized manuscripts and scanned papers from books, journals and official legislating documents, while the remaining 4000 files are images and videos. Proper annotation of collected data items was an absolute requirement in order to allow effective information retrieval related to Commandaria for various categories of users. The value of the collected information is priceless for Cyprus heritage, therefore the collection, proper preservation and easy access to this information is a task of tremendous importance [142]. Effective indexing and retrieval of this information requires accurate and rich annotation of data; according to our initial assumptions the annotation quality may be affected by various factors.

6.2 Annotation Methods

There are many different types of information that can be associated with images. In our method we focused mostly on data which directly or indirectly referred to the visual content of the image and were divided into the following categories: a) Content-descriptive metadata, which refer to the actual semantic contents presented on the image like emotions or meanings of the visual signs, or the actual scenes presented on the image [1], and b) Content-independent metadata, which refer to the content of the image without describing it directly, like date, location, etc. [1]. The two categories were specified using one or more of vocabulary keywords using taxonomy, hierarchical vocabulary keywords and free keywords.

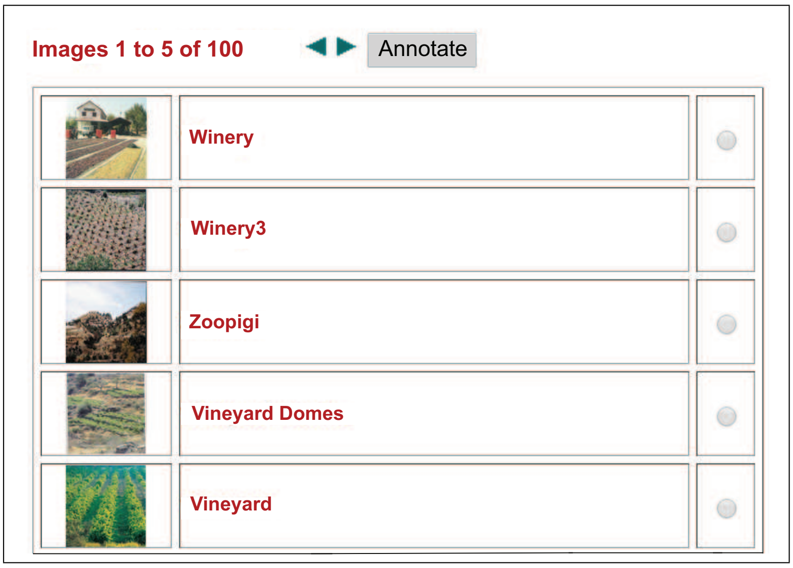
The first annotation approach restricted the annotator to use pre-defined keywords from a lexicon created using the Commandaria taxonomy [142]. Furthermore, the pre-defined set of keywords was also offered in an hierarchical structure by the second approach. The fact that there are many ways of classifying the concepts of an image, depending on culture, age, knowledge, etc, remains the main problem of these approaches [1]. Finally, the third approach tried to overcome the limitations of the first and second approach, by allowing the use of free keywords. This approach does not provide any restrictions, nevertheless suffer from a series of challenges. Spelling and typing mistakes are the most common problems of the specific approach, which can be addressed by an intelligent spell checker and/or by using an ontology. Each of the annotation approaches provides some benefits along with some limitations, the combination of three approaches leads to a complete annotation proposal. These methods cover more or less the whole spectrum of manually annotation methods.

6.2.1 Dataset

The experiments were conducted on a dataset that consists of 500 images, randomly selected from an large amount of data, collected in the framework of Commandaria project. 28 pre-selected keywords were used to create the vocabulary and hierarchical vocabulary keywords provided by the Commandaria platform. The 500 images were distributed over the Commandaria platform and annotated by non experts in form of mini-jobs. Non-expert annotators were students (21-23 years old) enrolled in Digitisation of Cultural Heritage courses at Cyprus University of Technology. Each annotator received partial credit toward completion of the course.

6.2.2 Task Design

The annotation process was divided into mini-tasks, where each mini-task consisted of the annotation of one image using one or more of the 3 proposed methods. The mini-tasks are presented as a list of thumbnails (Fig. 6.1) and the user can choose which image wants to annotate. The annotators are presented with a list of instructions and are asked to classify the image into categories “abstract” and “specific” based on



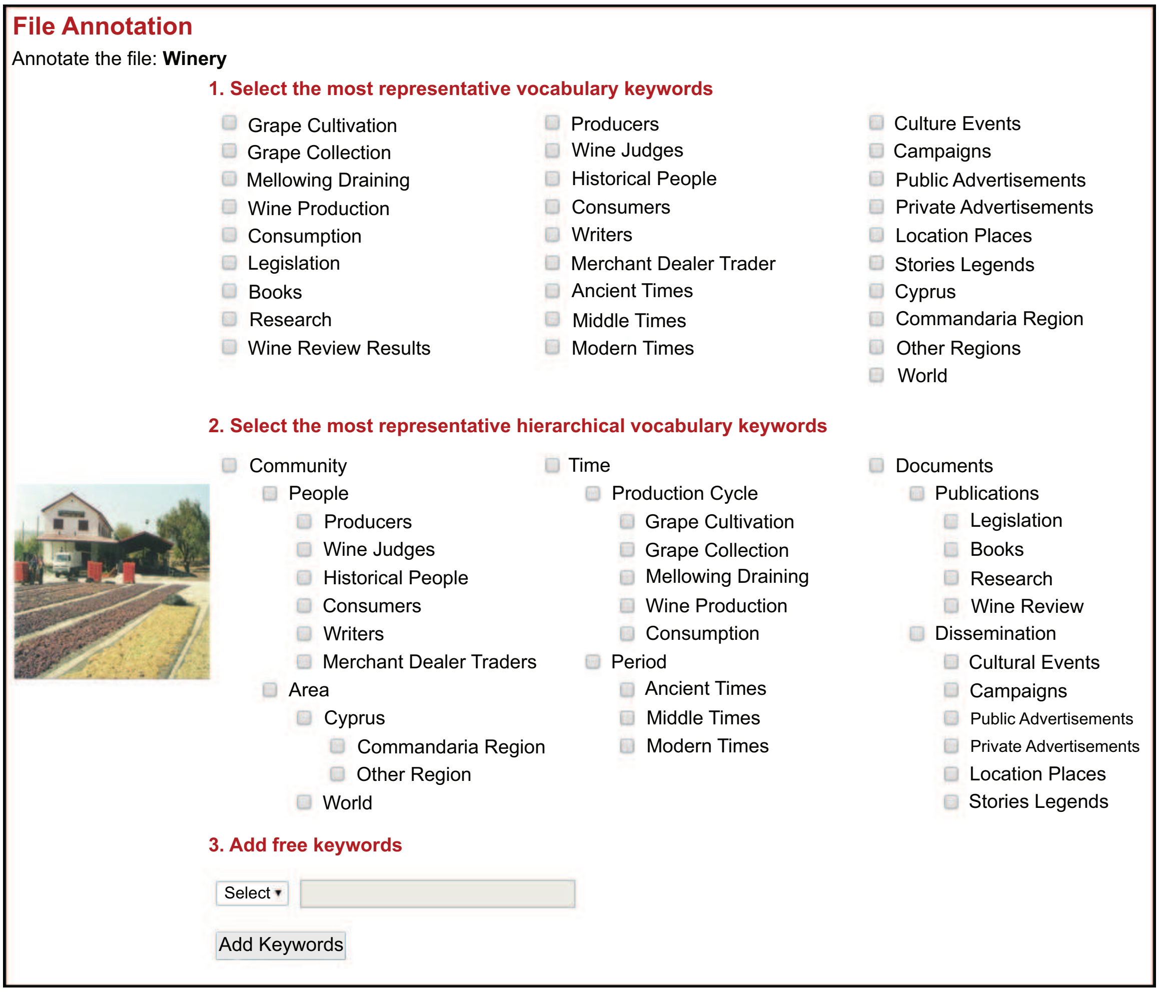
Images 1 to 5 of 100

	Winery	○
	Winery3	○
	Zoopigi	○
	Vineyard Domes	○
	Vineyard	○

Figure 6.1: The list of mini-tasks.

their content before starting the annotation. The instructions described the annotation process; how to use each annotation method and the minimum time required to fulfill a mini-task (it was set to 60 sec based on expert's estimation).

Fig. 6.2 illustrates an example of a mini-task. For the first method, the 28 pre-selected keywords of the annotation taxonomy compiled by Commandaria team were presented and the user could annotate the image file by clicking the most appropriate keywords from the list of check boxes. Furthermore, the users could choose keywords from the hierarchical vocabulary that offered in the second method. The 37 hierarchical vocabulary keywords were classified in three main categories. Every main category was further divided to a number of subcategories and any subcategory to a number of nodes and so on, providing an hierarchical annotation tree. The users could also add free keywords using the second annotation method. First, they chose their preferred language between English and Greek and then typed the keyword in the corresponding text box. By pressing the "Add Keywords" button all the suggested keywords were stored in the platform. Fig. 6.3 shows the structure of the relational database used for storing the collected annotations.



File Annotation

Annotate the file: **Winery**

1. Select the most representative vocabulary keywords

□ Grape Cultivation
□ Grape Collection
□ Mellowing Draining
□ Wine Production
□ Consumption
□ Legislation
□ Books
□ Research
□ Wine Review Results

□ Producers
□ Wine Judges
□ Historical People
□ Consumers
□ Writers
□ Merchant Dealer Trader
□ Ancient Times
□ Middle Times
□ Modern Times

□ Culture Events
□ Campaigns
□ Public Advertisements
□ Private Advertisements
□ Location Places
□ Stories Legends
□ Cyprus
□ Commandaria Region
□ Other Regions
□ World

2. Select the most representative hierarchical vocabulary keywords

- □ Community
  - □ People
    - □ Producers
    - □ Wine Judges
    - □ Historical People
    - □ Consumers
    - □ Writers
    - □ Merchant Dealer Traders
  - □ Area
    - □ Cyprus
      - □ Commandaria Region
      - □ Other Region
    - □ World

- □ Time
  - □ Production Cycle
    - □ Grape Cultivation
    - □ Grape Collection
    - □ Mellowing Draining
    - □ Wine Production
    - □ Consumption
  - □ Period
    - □ Ancient Times
    - □ Middle Times
    - □ Modern Times

- □ Documents
  - □ Publications
    - □ Legislation
    - □ Books
    - □ Research
    - □ Wine Review
  - □ Dissemination
    - □ Cultural Events
    - □ Campaigns
    - □ Public Advertisements
    - □ Private Advertisements
    - □ Location Places
    - □ Stories Legends

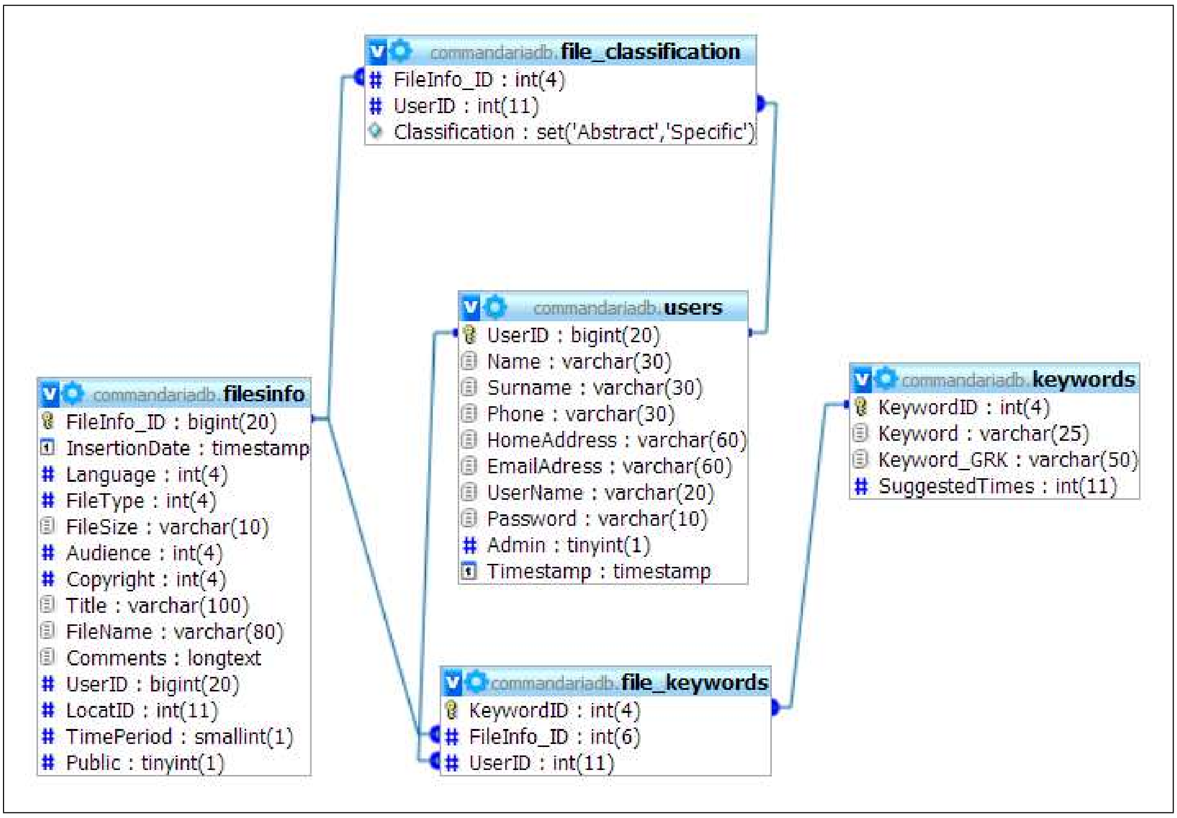
3. Add free keywords

Figure 6.2: An example of an image annotation using: (1)Vocabulary Keywords, (2) Hierarchical Vocabulary Keywords, and (3) Free Keywords.

6.3 Evaluation Process

6.3.1 Mathematical Background

The mathematical background of the evaluation process is as follows. We denote by A^i the *i-th* annotator (i=*1,...,*N_A). I^j indicates the *j-th* image (j=*1,...,*N_I) in the image dataset, while N_I denotes the total number of images in this dataset. t^{ij} indicates the set of keywords suggested by annotator A^i for image I^j. The total number of keywords suggested by the *i-th* annotator, T_A^i, and the total number of keywords submitted for the *j-th* image, T_I^j, are computed by equations (6.1) and (6.2), respectively:



commandariadb.file_classification
FileInfo_ID : int(4)
UserID : int(11)
Classification : set('Abstract','Specific')

commandariadb.users
UserID : bigint(20)
Name : varchar(30)
Surname : varchar(30)
Phone : varchar(30)
HomeAddress : varchar(60)
EmailAdress : varchar(60)
UserName : varchar(20)
Password : varchar(10)
Admin : tinyint(1)
Timestamp : timestamp

commandariadb.filesinfo
FileInfo_ID : bigint(20)
InsertionDate : timestamp
Language : int(4)
FileType : int(4)
FileSize : varchar(10)
Audience : int(4)
Copyright : int(4)
Title : varchar(100)
FileName : varchar(80)
Comments : longtext
UserID : bigint(20)
LocatID : int(11)
TimePeriod : smallint(1)
Public : tinyint(1)

commandariadb.keywords
KeywordID : int(4)
Keyword : varchar(25)
Keyword_GRK : varchar(50)
SuggestedTimes : int(11)

commandariadb.file_keywords
KeywordID : int(4)
FileInfo_ID : int(6)
UserID : int(11)

Figure 6.3: The structure of the relational database.

$$T_A^i = \bigcup_{j=1}^{N_I} t^{ij} \tag{6.1}$$

$$T_I^j = \bigcup_{i=1}^{N_A} t^{ij} \tag{6.2}$$

A valid keyword for an image is every keyword that was being suggested either by an expert or by the majority of the annotators or by more than one annotator (the aim is to exclude keywords suggested by mistake). The set of valid keywords for the *j-th* image is denoted by K^j={K_1^j,..., K_n^j}. Finally, the intersection between the valid keywords K^j for image I^j and the keywords that the annotator A^i suggested for the same image, denoted as v^{ij}, indicates the set of valid keywords that suggested by the *i-th* annotator for the *j-th* image. Therefore $v^{ij} \subseteq t^{ij}$ and $v^{ij} \subseteq K^j$. The Venn diagram explaining the relations between T_A^i, T_I^j, K^j and v^{ij} sets is shown in Fig. 6.4. Finally, for clarity purpose, the notations used so far for the evaluation process are listed in

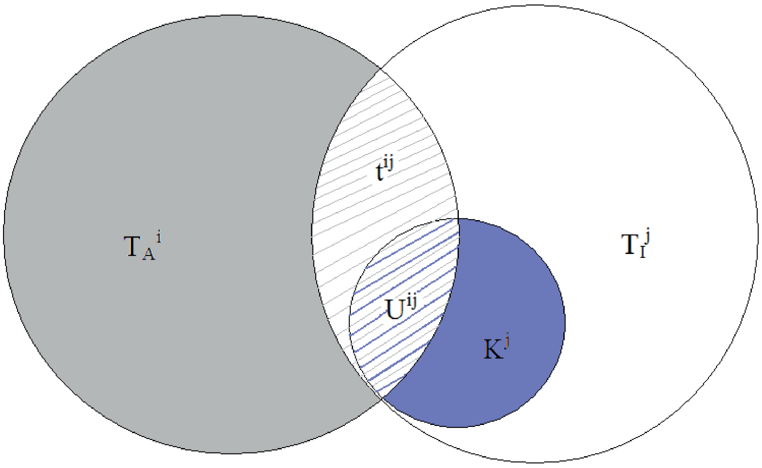


Figure 6.4: The relations between T_A^i, T_I^j, K^j and v^{ij} sets.

Table 6.1: List of notations

Notation	*Description*
A^i	*i-th* annotator
N_A	Total number of annotators
I^j	*j-th* image
N_I	Total number of images
t^{ij}	Set of keywords suggested by *i-th* annotator for the *j-th* image
T_A^i	Total number of keywords suggested by the *i-th* annotator
T_I^j	Total number of keywords submitted for the *j-th* image
K^j	Set of valid keywords for the *j-th* image
V_j	Total number of valid keywords given for the *j-th* image
v^{ij}	Set of valid keywords suggested by the *i-th* annotator for the *j-th* image
C_A^i	Overall consistency of the annotator A^i

Tab. 6.1.

6.3.2 Evaluation Metrics

Following manual identification and correction (spelling or typing errors, synonymy, etc) of the keywords which were submitted as free text, we tried to answer the questions set above by utilizing the following measurements:

1. The annotators consistency

The first measurement estimates the annotators consistency by comparing the valid keywords submitted for each image with those submitted by each annotator for the same image. The overall consistency C^i_A, of the annotator A^i is given by summing its consistency across all images he/she annotated:

$$C^i_A = \sum_{j,t^{ij} \neq \emptyset} \frac{|(v^{ij})|}{|(K^j)|} \tag{6.3}$$

2. Agreement analysis between expert and non-experts

   The second measurement determines the accuracy agreement between the expert and non-experts on image basis. Following the formula used by Brants [68], the accuracy between the annotations given by non-experts and the annotations given by the expert for the dataset I can calculated as follows:

$$Accuracy(I) = \frac{1}{N_I} \sum_{j=1}^{N_I} \frac{|(K^j)|}{|T^j_I|} \tag{6.4}$$

   Where, the T^j_I denotes the total number of keywords given by the non-experts for the *j-th* image and K^j denotes the valid keywords given by the expert for the same image.

6.4 Results and Discussion

The 500 mini-tasks were assigned to 50 annotators resulting in 25000 annotation sets. Annotations completed in less than the 60 seconds were rejected. An overall of 36 annotation sets related to 20 different images were rejected. From the 50 annotators involved in the experiment, 31 annotators chose to describe the set of 500 images with vocabulary keywords, while the remaining 19 used the hierarchical vocabulary keywords. Moreover, the majority of the annotators tried to enhance the image description by adding also free keywords. A total number of 818 different free keywords were proposed by 45 annotators and 263 keywords suggested twice or more. Some of

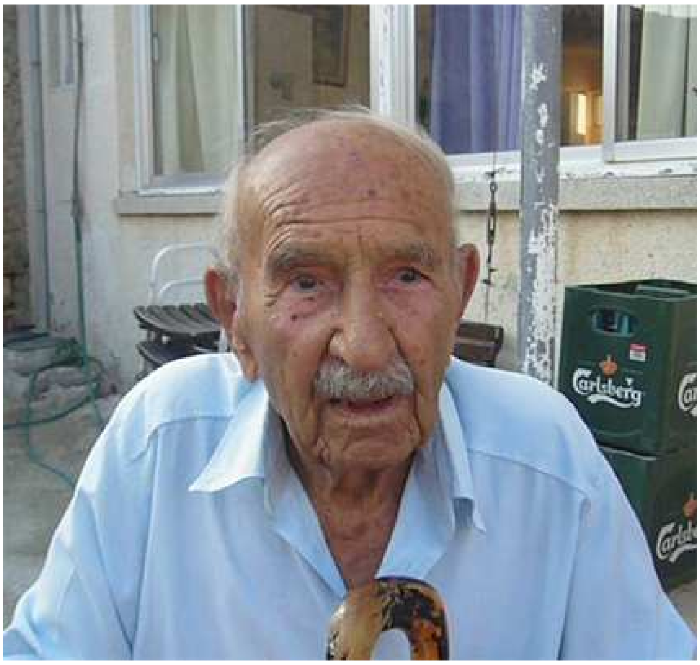
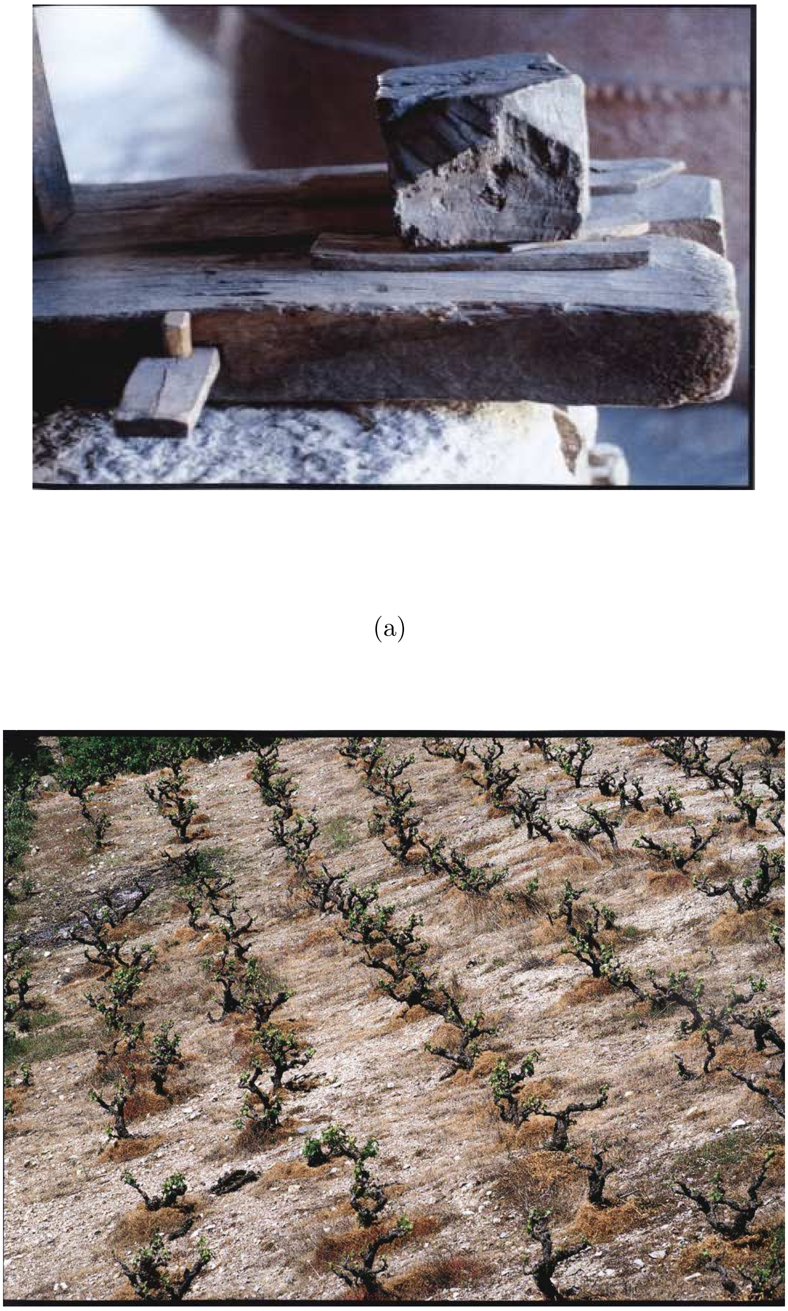


Figure 6.5: An image sample used for the experimental setup.

them received a high number of suggestions indicating their importance for annotating the set of images used for the experimental setup. The example image presented in Fig. 6.5 was annotated by some annotators using abstract terms such as “producer”, “wine judge”, or “historical people” while other annotators preferred to annotate it based on their visual interpretation and submitted free keywords like “old man” or “grandfather”.

During the annotation process the annotators were asked to classify the 500 images into two categories “abstract” and “specific”, based on images’ content. 330 images were classified as “abstract” while the remaining 170 were classified as “specific”. Examples of the images classified in these two categories are shown in Fig. 6.6. From the bulk of the collected data we first evaluate the consistency of annotators and compared the efficiency of the three proposed annotation methods. The consistency was calculated separately for abstract and specific images and the results are presented in Tab. 6.2.

Concerning the effectiveness of the various approaches for defining the valid keywords,



(a)

(b)

Figure 6.6: Two images used for experimental setup which were classified based on their content as: (a) “abstract”, (b) “specific”.

the experimental results indicate that the use of majority vote perform better than the other two. The average consistency obtained using valid keywords based on majority vote is very good for both abstract and specific images. The lowest consistency was obtained when the keywords suggested by more than one annotator were set as valid. The results are perfectly justified since this approach set as valid the majority of the given keywords for each image making it even more difficult for annotators to be consistent. The performance of the valid keywords regarding the expert annotations is also noteworthy, where the average consistency is quite good for the various combinations.

The average annotators' consistency, in the case of the annotation using the vocabulary keywords, is low while the gap between consistency for "abstract" and "specific" images is bigger than the other two annotation methods indicating the weakness of the annotators to understand the meaning of the proposed keywords especially for "abstract" images. The use of free keywords improves the annotation score but a significant improvement occurs when the annotators use the hierarchical vocabulary keywords. The annotators consistency obtained using this method is quite good in comparison with the other two and has average consistency values in the range of 0.51-0.76 for "abstract" and 0.54-0.76 for "specific" images. The results indicate that hierarchical structure of the vocabulary keeps the annotation score high even if some annotators did not suggest the same keywords for a specific image. The hierarchical lexicon improves the annotators consistency while normalizes the differences between the annotation of "abstract" and "specific" images, and it appears to be a good method to use for crowdsourcing-wise manual image annotation.

Finally, the t-test was applied to assess whether there is enough empirical evidence to claim a difference between the way that users annotate "abstract" and "specific" images. This statistical test compares the mean values of the two distributions to verify whether the hypothesis that there is no difference between the two methods is rejected. Using the average annotators consistency for the two types of images, the t-test at the 5% confidence level [143] was used to test the significance of the difference. The results in Tab. 6.2 show that the users performed significantly different between the

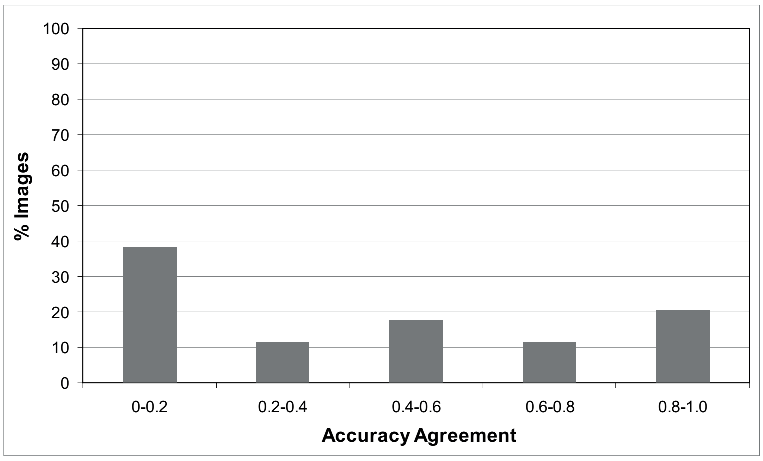
Table 6.2: Average annotation consistency

Annotation Method	*Type of Images*	*Annotators Consistency*		
		≥2	*Expert*	*Majority Vote*
Vocabulary Keywords	Abstract	0.37	0.57	0.76
	Specific	0.42	0.71	0.89
		n.s.	s.	s.
Hierarchical Vocabulary Keywords	Abstract	0.51	0.61	0.76
	Specific	0.54	0.68	0.78
		n.s.	n.s.	n.s.
Free Keywords	Abstract	0.42	0.59	0.69
	Specific	0.44	0.62	0.64
		n.s.	n.s.	n.s.
n.s. (not significant), s. (significant)				

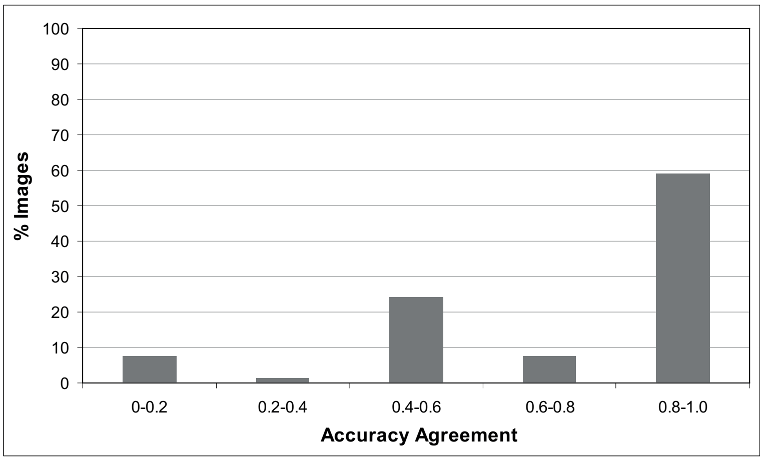
“abstract” and “specific” images only in case of vocabulary keywords. The consistency using vocabulary keywords is significantly different when is calculated based on the expert and majority vote.

The accuracy agreement between expert and non-experts is examined on image basis separately for “abstract” and “specific” images for vocabulary and hierarchical vocabulary keywords. As the previous measure, the accuracy agreement for free keywords has not been examined due to the low probability of using the same free keyword both the expert and non-experts for the same image. As shown in Fig. 6.7, the accuracy agreement using the vocabulary keywords is very low for the “abstract”, while is much better for the “specific” images. The average accuracy agreement is 0.42 for “abstract”, and 0.76 for the “specific”, respectively. Comparing the accuracy agreement between “specific” and “abstract” images using vocabulary keywords, the results are in full agreement with the conclusion drawn by Fujisawa [144] who indicates that the terminology and description of cultural heritage are often too technical and difficult for nonprofessional users of the domain.

As shown in Fig. 6.8, the corresponding values in the case of the hierarchical vocabulary keywords are higher, and the difference between the agreement of “abstract” and “specific” is normalized. The accuracy values are in the range of 0.8-1 for the 68% of the “abstract” and for the 78% of the “specific” images, respectively. The average ac-



(a)



(b)

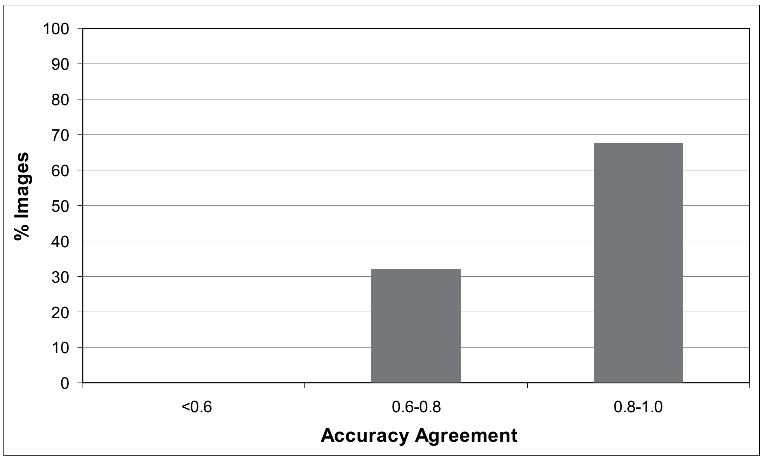
Figure 6.7: The accuracy agreement between expert and non-experts using vocabulary keywords for: (a) “abstract”, (b) “specific”, images.

curacy agreement between expert and non- experts is 0.77 for “abstract”, and 0.82 for “specific” images. The use of hierarchical vocabulary keywords improves the quality of the manual annotations regardless the content of the image and helps the non-experts to assign keywords to the image dataset almost like the expert.

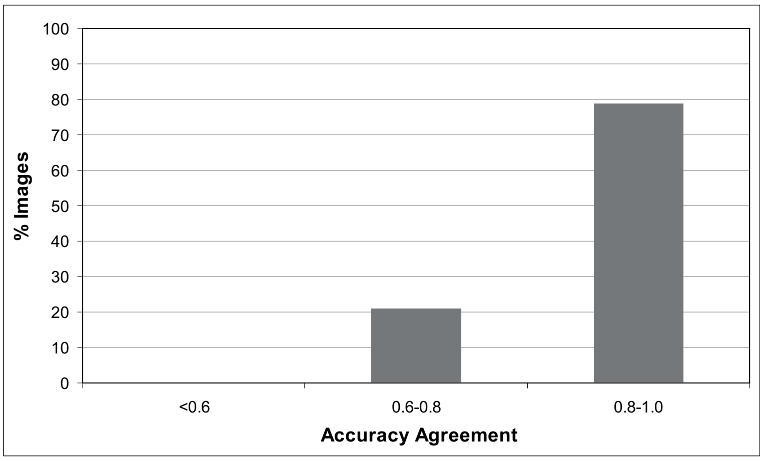
6.5 Conclusions and Remarks

In this chapter we investigate the factors that influence the quality of annotation created through crowdsourcing-wise methods. A set of 500 images were manually annotated by 50 annotators using: (i) a pre-selected set of keywords in a form of lexicon, (ii) an hierarchical lexicon, and (iii) free keywords. The results indicate that the image content itself and the used lexicon affect the annotation quality. The frequent use of free keywords implies the inability of non-expert annotators to fully understand the meaning of given keywords, and the inability of the given lexicon to describe the content of the image dataset. Definitely, free keywords could be used for the creation or enhancement of an existing vocabulary/hierarchical lexicon. The data (images and annotations) collected in the framework of this study are available to the research community for further experiments[3].

[3] http://cis.cut.ac.cy/∼z.theodosiou/Database2.zip.



(a)



(b)

Figure 6.8: The accuracy agreement between expert and non-experts using hierarchical vocabulary keywords for: (a) “abstract”, (b) “specific”, images.

Chapter 7

Extracting Keywords for Web Images

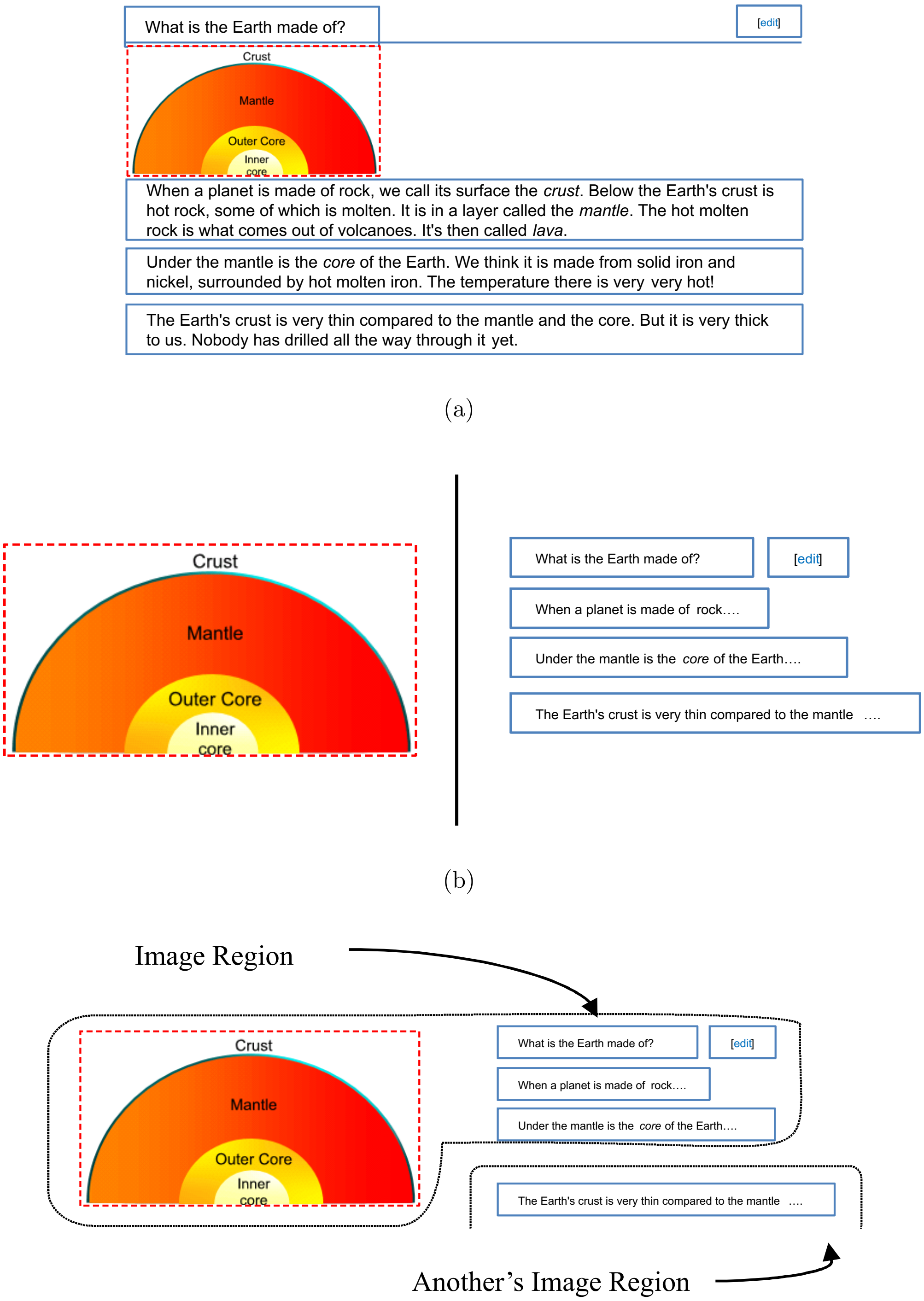
Among the most challenging scientific interests of the past years, special attention has been given to the task of web image information mining. Web images exist in huge amounts on the web and several methods for their efficient description and representation have been proposed so far. In many of the exploited algorithms, web image information is extracted from textual sources such as image file names, anchor texts, existing keywords and, of course, surrounding text. However, the systems that attempt to mine information for images using surrounding text suffer from several problems, such as the inability to correctly assign all relevant text to an image and discard the irrelevant text at the same time. A novel method for indexing web images is discussed in this section. The proposed system uses visual cues in order to obtain a web page segmentation. The segments are represented with semantic metrics and a k-means clustering assigns these segments to the web image they refer to. The evaluation procedure indicates that the semantic representation method of the visual segments delivers a good description for the web images.

7.1 Proposed Method

In the proposed method, the whole text that is found in the web page is used as a source to extract content information for the web images that exist in the same web document. The structural text blocks of the web page are extracted using Vision based page segmentation (VIPS) [56], but several text fragments are discarded at an early stage of the processing, where we attempt to eliminate the noise of the web page. The text blocks are finally assigned to images after their semantic representation which is achieved using the WordNet project [127]. This semantic representation ensures that the text blocks assigned to a single are semantically uniform; in other words they share similar content.

The proposed system performs the extraction of web image content information following the four steps that are presented in Fig. 7.1. The input web page is initially processed using the VIPS algorithm in order to obtain the constructing blocks of the web page. Therefore, this step results to a set of components that consist of visually indivisible contents. Making use of the HTML source code of each component, it is possible to separate them into two groups: one comprising the web images of the web page and one the textual blocks. In the following step, the task of assigning text blocks to the images they refer to, and discard those that do not refer to any image, is addressed. In [145], two different methods for estimating image regions were proposed and evaluated. In the first approach, the Euclidean distances between each text block and image are calculated and the text blocks are assigned to the closest image. In order to discard text blocks that may refer to no images an adaptive statistical threshold is used.

In the second approach, the Vector Space Model (VSM) [146] is used, representing text blocks with feature vectors where each dimension corresponds to a different vocabulary term found in the web page. If a term occurs in a certain text block, its value (*i.e*, weight) in the vector is 1, while if it does not appear the weight is 0 (*i.e*, binary weighting). Based on this formulation the weight vectors are clustered using the k-



What is the Earth made of?

When a planet is made of rock, we call its surface the *crust*. Below the Earth's crust is hot rock, some of which is molten. It is in a layer called the *mantle*. The hot molten rock is what comes out of volcanoes. It's then called *lava*.

Under the mantle is the *core* of the Earth. We think it is made from solid iron and nickel, surrounded by hot molten iron. The temperature there is very very hot!

The Earth's crust is very thin compared to the mantle and the core. But it is very thick to us. Nobody has drilled all the way through it yet.

[edit]
Crust
Mantle
Outer Core
Inner core
When a planet is made of rock....
Under the mantle is the *core* of the Earth....
The Earth's crust is very thin compared to the mantle ....
Image Region
Another's Image Region

(a)

(b)

(c)

Figure 7.1: An example of estimating the image's regions: (a) the input web page is divided into constructing components, (b) These components are seperated into images and text blocks, (c) The distances between every image/text block pair is calculated and regions are estimated.

means algorithm. The number of clusters is equal to the number of images found in the web page, plus one, so that any text that refers to no image is discarded.

In [145], it was shown that when assigning textual blocks to web images, the k-means clustering of the above described VSM representation of blocks leads to poor results, compared to the naive approach of assigning text to the closest image according to the spatial Euclidean distance, and discarding those that are not "close enough", according to some empirical threshold. This may result from the fact that when two neighboring text blocks refer to the same image and contain similar semantic content, the web page creator would select synonyms in order to express the same meaning. This action reduces the occurrences of each vocabulary term in the weight vectors and results to sparse weight vectors since the length of the vocabulary is not proportional to the size of the each text block. It is important to reduce this sparsity of the vectors in order to obtain good clustering of the blocks based on their semantic representation. This may be achieved by merging words that share similar semantic concepts into new vocabulary terms. The extracted vocabulary was therefore processed using WordNet and the WordNet::Similarity module.

The processing that took place is illustrated in Fig. 7.2. As it is shown there, a recursive method was employed in order to transform the vocabulary using semantic information. The vocabulary was initially reduced to the terms that correspond only to noun (or verb) synsets. The S_{wup} [129] measure between every noun-to-noun (or verb-to-verb) synset pair was then computed. The pairs that shared the highest semantic similarity were considered candidates for the merging step. In order to decide which candidate pair would be merged in each step of the recursion, the depth of each synset in the WordNet taxonomy and their document frequency in the corpus was taken into account as follows. Among the candidates those that shared the highest average depth in WordNet hierarchy were kept. If more than one pairs shared the same average depth, the pair that appeared more often in the corpus was merged. The process of merging candidate pairs continued while the rank of the matrix comprising the weight vectors (*i.e* the maximum number of linearly independent weight vectors) did

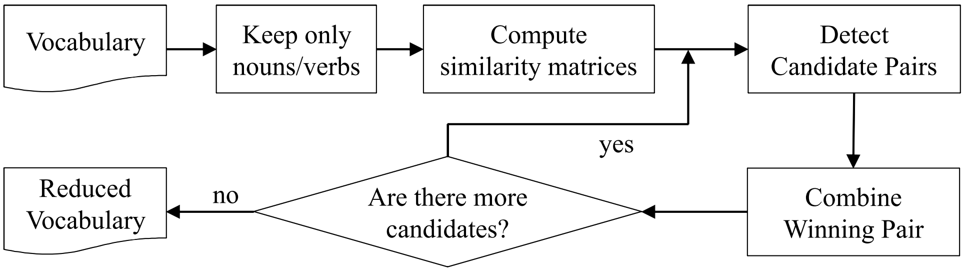


Figure 7.2: A part of the WordNet hierarchy.

not decrease. Although the reduction of the vocabulary size was critical in order to eliminate the sparsity of the weight vectors, it was also important not to introduce linear dependencies among them.

An example of the above described recursive method is shown in Tab. 7.1. The vocabulary reduction is illustrated in a small fragment of the vocabulary as obtained using the text blocks shown in Fig. 7.2. From the 15 nouns that appear in these four text blocks, five representative samples were selected and form the weight vectors that are shown in Tab. 7.1(a). In the same table, the terms' depth in WordNet hierarchy and their document frequency (*df*) are presented. The document frequency of each term is calculated as the number of text blocks that contain this term, normalized to the total number of text blocks in the corpus. The application of the S_{wup} similarity measure on these terms results to the matrix that is presented in Tab. 7.1(b). Among the three term pairs that share the maximum similarity (*i. e.* candidate pairs: $cp_1 = \{t_2, t_4\}$, $cp_2 = \{t_3, t_4\}$ and $cp_3 = \{t_4, t_5\}$) and also the same average depth in the WordNet hierarchy ($\tilde{d} = 7.5$), the pair cp_3 is the one that has the maximum average document frequency ($\tilde{df} = 0.5$) so it is merged, into the new vocabulary term t'_4, as appears in Tab. 7.1(c). The similarity value between this new term and the rest of the vocabulary is given by:

$$S_{wup}(t'_4, t_i) = \min\{S_{wup}(t_4, t_i), S_{wup}(t_5, t_i)\} \tag{7.1}$$

The k-means algorithm, applied on the weight vectors of the new dictionary, is used for the clustering of the text blocks into M regions. Since each text block may refer to any image but it may also be irrelevant to every image, the number of clusters M is equal to the total number of images found in the web page plus one for text blocks irrelevant to every image. The clusters are initialized with text to image assignment based on Euclidean distances.

	Weight Vectors					
Terms	*b1*	*b2*	*b3*	*b4*	Depth	Df
t_1: earth	1	0	1	1	7	0.75
t_2: surface	0	1	0	0	8	0.25
t_3: crust	0	1	0	1	8	0.5
t_4: layer	0	1	0	0	7	0.25
t_5: mantle	0	1	1	1	8	0.75

(a) The weight vectors for the 4 text blocks.

Term	t_1	t_2	t_3	t_4	t_5
t_1	1	0.77	0.71	0.76	0.77
t_2	0.77	1	0.87	0.93	0.87
t_3	0.71	0.87	1	0.93	0.87
t_4	0.76	0.93	0.93	1	0.93
t_5	0.77	0.87	0.87	0.93	1

(b) The similarity matrix.

Term	t_1	t_2	t_3	t'_4
t_1	1	0.77	0.71	0.76
t_2	0.77	1	0.87	0.87
t_3	0.71	0.87	1	0.87
t'_4	0.76	0.87	0.87	1

(c) The similarity matrix after the first iteration of dictionary reduction. The term t'_4 corresponds to the winning pair $\{t_4, t_5\}$.

Table 7.1: An example of the vocabulary reduction.

7.2 Evaluation

For the evaluation of the proposed method, a corpus that consists of 40 real world web pages was created. The dataset which contains a total of 131 images was manually

Table 7.2: The results of text block to image assignment using the original VSM dictionary and its reduced version.

Method	Precision	Recall	F-score
Original Vocabulary	0.3576	0.4528	0.4533
Reduced Vocabulary	0.7103	0.8705	0.7632

labeled by an annotator who was asked to assign to each image found in the web pages, the text fragments that refer to it. The evaluation measures Precision, Recall and F-score as described for a similar task in [126], were applied on the output of the proposed system, giving the results that are presented in Tab. 7.2. As it is shown there, the implemented vocabulary reduction offers an important improvement to the clustering of the web page segments. The execution of the proposed algorithm yields an average content F-score equal to 0.7632 for the total of the 131 annotated images, when the vocabulary is reduced with the use of WordNet similarity measures. The same metric is 0.4533 when the original vocabulary, as obtained using the VSM is used.

The combination of web page segmentation, k-means clustering and natural language processing is a solution for the problem of web image context extraction, which opposed to many state-of-the-art methods on the field, does not depend on the layout of the web page (since the DOM structure is not used). Furthermore, the use of WordNet improves the k-means clustering results as expected, since semantic information should not be omitted when it comes to web image context extraction. It is critical however, for our future experimentation, to test more similarity measures, which are available through the WordNet::Similarity project.

7.3 Conclusions and Remarks

This chapter presents a new framework for automatically extracting keywords for web images using the surrounding html text. The Vision based page segmentation (VIPS) algorithm was initially used for segmenting web page into images and text fragments.

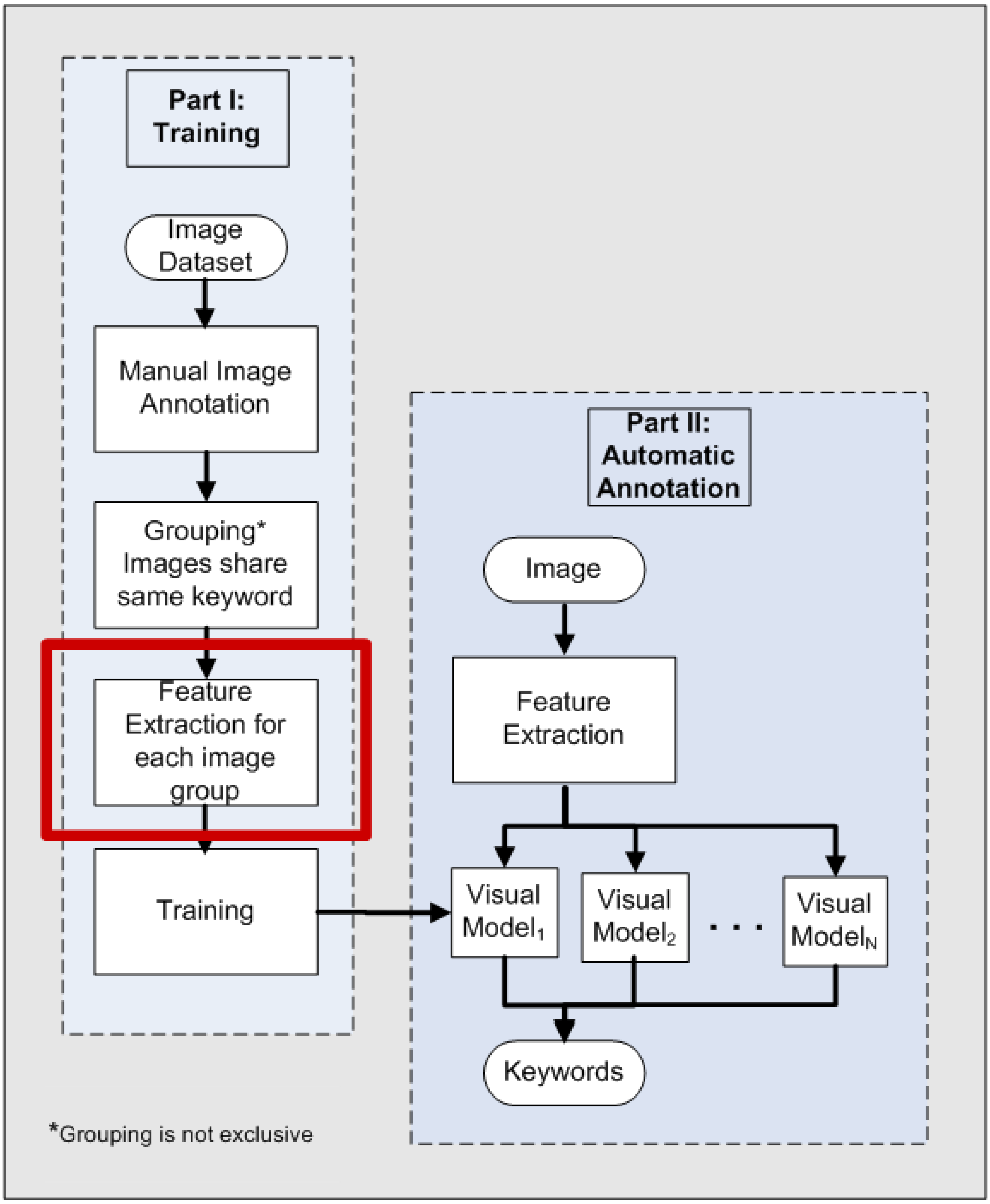
The text fragments are represented with semantic metrics and a k-means clustering assigns them to the web image they refer to. The semantic representation of text fragments was achieved using the WordNet project. Both VIPS algorithm[1] and WordNet project[2] are free available software which can be used for research purposes.

[1] http://www.cad.zju.edu.cn/home/dengcai/VIPS/VIPS.html.

[2] http://wordnet.princeton.edu/.

Part II

Low Level Feature Extraction


Part I: Training
Image Dataset
Manual Image Annotation
Grouping* Images share same keyword
Feature Extraction for each image group
Training
Part II: Automatic Annotation
Image
Feature Extraction
Visual Model₁
Visual Model₂
. . .
Visual Model_N
Keywords
*Grouping is not exclusive

Chapter 8

Low-Level Feature Extraction for Automatic Image Annotation

Low-level feature extraction is the first crucial step either in content-based retrieval or in the automatic image annotation. It aims at capturing the important characteristics of the visual content of images. The low-level features are defined to be those basic features that can be extracted automatically from an image without any information about spatial relationships [147]. They can be broadly divided into two main types: (a) Local or domain-specific features, and (b) Global or holistic features. Selection of the most appropriate subset of features plays a significant role in effective classification schemes as well as in visual modeling of keywords, which is a necessary step in learning-based automatic image annotation methods. Feature extraction and selection can be evaluated from three different perspectives: First, in terms of their ability to identify relevant (appropriate) features, second in terms of the performance of the created classifiers and third, in terms of the reduction of the number of features. The research in feature extraction is rich and dozens of low-level feature types have been proposed.

8.1 Local Features

Local features are image patterns that differ from their immediate neighborhood. They are usually associated with a change of an image property or several properties simultaneously, although they are not necessarily localized exactly on this change. Image properties commonly considered for local feature derivation are intensity, colour, and texture. Local invariant features not only allow finding correspondences in spite of large changes in viewing conditions, occlusions, and image clutter, but also yield an interesting description of the image content. Ideal local features should be characterized by repeatability, distinctiveness, locality, quantity, accuracy and efficiency [148]. Local features were first introduced by Schiele and Crowely [149], and Schmid and Mohr [150] and soon became very popular especially in machine learning frameworks.

The Scale Invariant Features Transform (SIFT) [151] and Histogram of Gradients (HOG) [152] are two of the most successful local features categories considered for a variety of tasks including object detection and object recognition. They are based on histograms of gradient orientations weighted by gradient magnitudes. The two methods differ slightly in the type of spatial bins that they use. The SIFT, proposed by Lowe [151], transforms image data into scale-invariant coordinates relative to local features and computes a set of features that are not affected by object scaling and rotation. Key points are detected as the maxima of an image pyramid built using difference-of-Gaussians. The multi-scale approach results in features that are detected across different scales of images. For each detected keypoint, a 128 dimensional feature vector is computed describing the gradient orientations around the keypoint. The strongest gradient orientation is selected as reference, thus giving rotation invariance to SIFT features. On the other hand, HOG uses a more sophisticated way for binning. The image is divided into small connected regions and a histogram of gradient directions or edge orientations within each region is compiled. For the implementation of HOG, each pixel within the region casts a weighted vote for an orientation-based histogram channel.

Due to the large number of SIFT keypoints contained in an image, various approaches have been used to reduce the dimensionality or prune the number of detected keypoints before using them in learning based environments. Another difficulty in using the original SIFT features in machine learning frameworks is that the number of keypoints and consequently the dimensionality of input vector is image dependent. As a result they cannot directly employed for creating and feeding models using a learning by a example paradigm. The PCA-SIFT was proposed in [153] as a solution this problem. It utilizes Principal Component Analysis (PCA) to normalized gradient patches to achieve fast matching and invariance to image deformations. Mikolajczyk and Schmid [154] presented an extension of the SIFT descriptor, the significance of the Gradient Location and Orientation Histogram (GLOH) which applies also the PCA on SIFT features for dimensionality reduction. Instead of PCA, the Linear Discriminate Analysis (LDA) has also been applied to create a low-dimensional representation of the SIFT descriptors [155].

The effectiveness of SIFT and GLOH features led to several modifications that try to combine their advantages. Recently, the Speeded Up Robust Features (SURF) descriptor that approximates the SIFT and GLOH by using integral images to compute the histograms bins has been proposed [156]. This method is computationally efficient with respect to computing the descriptor values at every pixel and differs from SIFT's spatial weighting scheme. In particular, all gradients contribute equally to their respective bins, which results in damaging artifacts when used for dense keypoints computation. The Daisy descriptor [157], on the other hand, retains the robustness of SIFT and GLOH and can be computed quickly at every single image pixel.

Other approaches, use clustering techniques to manage the thousands of local descriptors produced by SIFT. Bag-Of-Features (BOF) methods represent an image as orderless collection of local features [158], [159], [160]. Usually, the k-means clustering algorithm groups visual patches into clusters and creates a visual vocabulary. For each image the number of occurrences of each word is counted to form a histogram representation. Besides the advantages of the BOF representation, these methods have important descriptive limitation because they disregard the spatial information of the local features. Lazebnik et al. [161] extended the BOF approach and proposed the Spatial Pyramid Matching method which partitions the image into increasingly fine sub-regions and computes histograms of local features found inside each sub-region to retain the spatial information.

SIFT features were originally proposed for object detection and recognition tasks. In these tasks a dedicated matching scheme is used to compare images or image regions. In machine learning environments this is not the case. The SIFT feature vector feeds the keyword visual models to produce an output indicating whether or not the corresponding keyword can be assigned to the image corresponding to this input vector. This difference, along with the dimensionality reduction, which is applied to produce SIFT based vectors of fixed dimensionality, lead to deteriorate performance in image retrieval compared to other types of features, like the MPEG-7 descriptors [162].

8.2 Global or Holistic Features

Global features provide different information than local ones since they are extracted from the entire image. Statistical properties such as histograms, moments, contour representations, texture features and features derived from image transforms like Fourier, Cosine and Wavelets can be considered as global features. Global features cannot separate foreground from background information; they combine information from both parts together [148]. Therefore, they are considered as more appropriate than local features in image indexing and retrieval. These features can be used when there is interest for the overall composition of the image, rather than a foreground object. However, in some cases, global features have been also applied for object recognition [163], [164]. The feature set in these approaches are obtained from the projections to the eigenspace created by computing the prominent eigenvectors based on the Principal Component Analysis of the image training sets.

Recently, the Compact Composite Descriptors (CCDs) [165] which capture more than one types of information at the same time in a very compact representation have been used for image retrieval applications [166], [167]. The Fuzzy Color and Texture Histogram (FCTH) [168] and the Color and Edge Directivity Descriptor (CEDD) [169] are determined for natural color images and combine color and texture information in a single histogram. The Brightness and Texture Directionality Histogram (BTDH) descriptor [170] describes grayscale images and captures both brightness and texture characteristics in a 1D histogram. Finally, the Spatial Color Distribution Descriptor (SpCD) [171] combines color and spatial color distribution information and can be used for artificial images. The performance of CCDs has been evaluated using several databases and experimental results indicated high accuracy in image retrieval task achieving, in some cases, better performance than other commonly used features for image retrieval such as the MPEG-7 descriptors.

The MPEG-7 visual descriptors [82] use standardized description of image content and they were especially designed for image retrieval in the content-based retrieval

Table 8.1: MPEG-7 visual descriptors.

Descriptor	Type	#Features
Color	DC coefficient of DCT (Y channel)	1
	DC coefficient of DCT (Cb channel)	1
	DC coefficient of DCT (Cr channel)	1
	AC coefficients of DCT (Y channel)	5
	AC coefficients of DCT (Cb channel)	2
	AC coefficients of DCT (Cr channel)	2
	Dominant colors	Varies
	Scalable color	16
	Structure	32
Texture	Intensity average	1
	Intensity standard deviation	1
	Energy distribution	30
	Deviation of energy's distribution	30
	Regularity	1
	Direction	1 or 2
	Scale	1 or 2
	Edge histogram	80
Shape	Region shape	35
	Global curvature	2
	Prototype curvature	2
	Highest peak	1
	Curvature peaks	Varies

paradigm. Their main property is the description of global image characteristics based on color, texture or shape distribution, among others. A total of 22 different kinds of features (known as descriptors) are included: nine for color, eight for texture and five for shape. These feature types are shown in Tab. 8.1. The number of features, shown in the third column of this table, in most cases is not fixed and depends on user choice. The dominant color descriptor includes color value, percentage and variance and requires especially designed metrics for similarity matching. Furthermore, the number of features included in this descriptor is not known a priori since they are image dependent (for example an image may be composed from a single color whereas others vary in color distribution). The previously mentioned difficulties cannot be easily handled in machine learning schemes and as a result the dominant color descriptor is rarely used in keyword modeling and classification schemes. The region shape descriptor features are computed only on specific image regions (and therefore they are not used in holistic image description). The number of peaks values of the contour shape descriptor varies depending on the form of an input object. Furthermore, they require a specifically designed metric for similarity matching because they are computed based on the HighestPeak value. The remaining of the MPEG-7 descriptors shown in Table 8.1 can be easily employed in machine learning schemes and since they are especially designed for image retrieval they are an obvious choice for keyword modeling.

Global features are a natural choice for image retrieval that is based on machine learning. Since they are extracted from the image as whole they are also appropriate for creating visual models for keywords. This is because training data can be created by defining the keywords that are related with the images used for training and there is no need to define specific regions in these images (which is by far more tedious). However, the choices of global features from which one can select is unlimited and in some cases depend on the type of images. Despite the fact that the MPEG-7 descriptors were initially proposed for CBIR systems they perform excellent within the machine learning paradigm used either in classification based keyword extraction or in keyword modeling. As a result they provide a good starting point in experimentation dealing with automatic image annotation and should be used as a benchmark test before adopting different feature types.

8.3 Feature Fusion

Feature fusion is of primary importance in case where multiple features types are used in training keyword models. Fusion can derive and gain the most effective and least dimensional feature vectors that benefit final classification [172]. Usually for each keyword group, various feature vectors are normalized and combined together into a feature union-vector whose dimension is equal to the sum of the dimensions of the individual low-level feature vectors. Dimensionality reduction methods are then applied to extract the linear features from the integrated union vector and reduce the dimensionality. Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two widely used approaches in this framework.

The PCA is a well-established technique for dimensionality reduction which converts a number of correlated variables into several uncorrelated variables called principal components. For a set of observed d-dimensional data vectors X_i, $i \in \{1, ..., \mathrm{N}\}$, the M principal components p_j, $j \in \{1, ..., \mathrm{M}\}$ are given by the M eigenvectors with the largest associated eigenvalues λ_j of the covariance matrix:

$$S = \frac{1}{N} \sum_i (X_i - \overline{X})(X_i - \overline{X})^T \tag{8.1}$$

where $\overline{X}$ is the data sample mean and Sp_j=$\lambda_j p_j$. The M principal components of the observed vector X_i are given by the vector:

$$c_i = P^T(X_i - \overline{X}) \tag{8.2}$$

where P=$\{p_1, p_2, ..., p_M\}$. The variables c_j are uncorrelated because the covariance matrix S is diagonal with elements λ_j. Usually cross-validation is performed to estimate the minimum number of features required to yield the highest classification accuracy. However, the computational cost of cross-validating is prohibitive so other approaches such as the maximum likelihood estimator (MLE) [173] are employed to estimate the intrinsic dimensionality of the fused feature vector by PCA.

LDA follows a supervised method to map a set of observed d-dimensional data vectors X_i, $i \in \{1, ..., \mathrm{N}\}$ to a transformed space using a function $Y = wX$. The w is given by the maximum eigenvector of the $S_w^{-1}S_b$ where S_w is the average within-class scatter matrix and S_b is the between-class covariance matrix of X_i.

The matrix w is determined such that the Fisher criterion of between-class scatter over average within-class scatter is maximized [174]. The original Fisher criterion function applied in the LDA is,

$$J = \frac{wS_bw^T}{wS_ww^T} \tag{8.3}$$

Obviously there are several fusion techniques that can be used to select the best feature set for training visual models for keywords. However, both PCA and LDA are based on a strong mathematical background and should investigate before examining alternatives. Nonlinear fusion methods, on the other hand, might be proved more efficient in some cases.

Chapter 9

Evaluation of MPEG-7 Visual Descriptors on Keyword Extraction

Much of the attention paid to automatic image annotation and CBIR systems is due to the MPEG-7 visual content description interface, which provides a unified framework for experimentation. Furthermore, the MPEG-7 experimentation model [175] provides practical ways for the computation of the MPEG-7 descriptors.

The performance of the MPEG-7 visual descriptors in terms of image retrieval, however, was not examined in detail. Although inclusion of these particular descriptors in the MPEG-7 protocol stack was based on experimental evaluation, the results were not published and the experiments cannot be recreated. Investigation of the performance of color and texture descriptors was reported in [77] but the main discussion there was devoted to the introduction of these descriptors to the research community rather than to experimental evaluation. The same holds for the work of Bober [176], which deals with the shape descriptors. A very interesting study on the MPEG-7 visual descriptors was conducted by Eidenberger in [177]. The descriptors are evaluated using statistics obtained by three different datasets including the one used during the MPEG-7 tests. One of the aims of the study reported in this chapter, was to investigate experimentally whether the conclusions made by Eidenberger are valid in a different dataset and by using a variety of classifiers. Spyrou in [178] investigates a variety of methods for fusing the MPEG-7 visual descriptors for image classification. The idea is interesting but the dataset used is small and the experiments cannot be recreated based on the description given in the corresponding paper.

In this chapter we deal with the experimental evaluation of the performance of the MPEG-7 descriptors [82] in terms of object classification. None of the works reported in the previous paragraph deals with object classification. This is quite logical since the MPEG-7 visual descriptors were defined primarily for image classification and not for object detection and classification. Furthermore, manual annotation of image objects through definition of the blob area is much harder than image annotation. In our study we get advantage of the availability of a large dataset of manually annotated objects created during the FP6 BOEMIE project[1] to perform extended experiments. We have used publicly available tools for the computation the MPEG-7 descriptors [175] and the object model creation (the Weka tool [179] and the libSVM [180] library integrated with Weka).

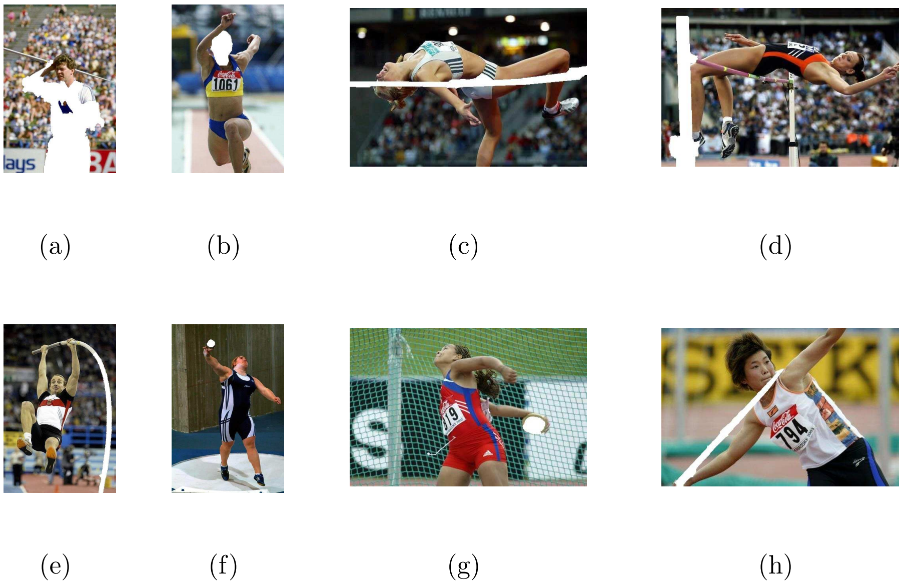
Object classification, on the other hand, can be related to keyword extraction in cases were the identified objects are used as keywords. This consideration allows multiple keywords to be extracted from a single image and, therefore, is a better choice than an image classification scheme.

9.1 Dataset Creation and Object Modelling

For dataset creation 1952 images from the athletics domain were used. These images were collected in the framework of the FP6 BOEMIE project and objects, corresponding to humans and athletic instruments, were manually marked by humans creating blobs. Example of such blobs overlayed on the original images are shown in Fig. 9.1. A total of 7686 manually annotated object instances corresponding to eight different class

[1] The images were randomly selected from a large dataset collected in the framework of FP6 BOEMIE project.

objects were used in our experiments. The eight object classes are: Person Body, Person Face, Horizontal Bar, Pole, Pillar, Discus, Hammer and Javelin. The training set contains 2597 instances while the remaining 5089 were used for test. The distribution of the various object instances in the training and test sets is presented in Tab. 9.1.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 9.1: Images from the athletic domain showing the detected objects (a)Person Body, (b) Person Face, (c)Horizontal Bar, (d)Pillar, (e)Pole, (f)Hammer, (g)Discus, (h)Javelin

Object models were created using Weka tool [179]. Among a variety of possible classifiers we decided to use (1) libSVM [180], (2) Sequential Minimal Optimization (SMO) [181], [182] and (3) Radial Basis Function networks [183]. The latter is a reasonable choice when dealing with multidimensional and multiclustered data while libSVM and SMO are state of the art implementations of Support Vector Machines. These algorithms have been reported in several publications as the best performing machine learning algorithms for a variety of classification tasks.

During training some parameters were optimized via experimentation in order to obtain the best performing model for each descriptor. Cost, Gamma, and Epsilon were optimally selected for the libSVM models. For SMO models we have experimented on the complexity constant C and then based on the chosen kernel type, we try to get

Table 9.1: Dataset

Objects	*Number of instances*	*Training Set*	*Test Set*
Person Body	3180	1062	2118
Person Face	3209	1044	2165
Horizontal Bar	493	164	329
Pole	229	94	135
Pillar	138	51	87
Discus	132	49	83
Hammer	142	56	86
Javelin	163	77	86
Total	**7686**	**2597**	**5089**

the optimum values the exponent of the polynomial kernel or the Gamma for the RBF kernel respectively. Finally, for the RBF models, the number of clusters and ridge were tuned for each one of the MPEG-7 descriptors.

In addition to the construction of individual models for each MPEG-7 descriptor we also trained models for several descriptor combinations using feature fusion. The parameter optimization followed was the same as the one described earlier.

9.2 Experimental Results

We used the dataset and object modeling process described in the previous section to examine the classification performance, of the eight object classes, in terms of precision and recall values. Tab. 9.2 summarizes the results for the models of the individual MPEG-7 descriptors while Tab. 9.3 shows the corresponding figures obtained using descriptor combinations. The results shown in these tables can be examined under two perspectives: First, in terms of the efficiency of the various descriptors as far as the object classification task is concerned. Second, in terms of the ability of the machine learning algorithms to create effective object class models for classification.

Concerning the classification efficiency of the individual MPEG-7 descriptors it is evident from Tab. 9.2 that the most reliable descriptor is Edge Histogram. Not only has the ability to discriminate the whole range of the eight classes used but the precision and recall values obtained using this descriptor are quite good irrespectively of the training algorithm used. This result is in full agreement with the conclusion drawn by Eidenberger [177] who examines the efficiency of the MPEG-7 descriptors using statistical analysis on different datasets. The second most reliable descriptor for object classification is Color Structure. Although the precision and recall values obtained for the classes with few training examples (that is, all classes but Person Body and Person Face) are rather low this descriptor has the potential to discriminate multiple classes irrespectively of the training algorithm used. The Contour Shape descriptor is effective for classification of objects having a well defined shape such as Horizontal Bar, Pole and Pillar. In contrary, it cannot be used for the classification of Discus and Hammer. These two classes although in principle they must have a circular shape their inaccurate segmentation, as created by the human annotators, make them appearing extremely variable in shape. Furthermore, they have been easily confused with Person Face as far as the shape is concerned. The most disappointing classification performance is achieved by the Region Shape descriptor. Although it contains much more features than the Contour Shape descriptor, it is only able to discriminate Person Body and Person Face. These two classes have a high population of training samples and are easily discriminated by all descriptors (with some variance mainly in the precision values).

Combinations of MPEG-7 descriptors are shown in Tab. 9.3. There, it can be seen that classification performance is increased through the use of feature based fusion for the majority of descriptor combinations. However, improvement in recall and precision values is not as significant as one might expect. This can be attributed to the variance of the feature values among different descriptors.

The performance of the training algorithms is examined through the effectiveness of the created models, the time required to train the models and the robustness to the variation of learning parameters. The libSVM algorithm requires by far the lower time and effort to create an effective model. This is true, however, if an RBF or a polynomial kernel is used. In such a case learning takes no more than a few seconds for

Table 9.2: Object classification results using the MPEG-7 visual descriptors and various data classifiers

Classifier	*Descriptor*	*Measure*	*Object Class*							
			Person Body	*Person Face*	*Horizontal Bar*	*Pole*	*Pillar*	*Discus*	*Hammer*	*Javelin*
libSVM	Color Layout	Recall	0.818	0.876	0.152	0.252	0.241	0.060	0.151	0.163
	(CL)	Precision	0.772	0.796	0.370	0.301	0.236	0.208	0.342	0.219
	Color Structure	Recall	0.990	0.819	0.176	0.185	0.287	0.241	0.233	0.279
	(CSt)	Precision	0.750	0.921	0.527	0.439	0.556	0.465	0.588	0.308
	Scalable Color	Recall	0.817	0.847	0.313	0.400	0.333	0.145	0.314	0.070
	(SC)	Precision	0.813	0.895	0.256	0.214	0.240	0.333	0.375	0.222
	Contour Shape	Recall	0.901	0.899	0.565	0.311	0.379	0.000	0.081	0.349
	(CS)	Precision	0.875	0.848	0.699	0.609	0.317	0.000	0.636	0.201
	Region Shape	Recall	0.516	0.541	0.334	0.000	0.264	0.000	0.000	0.000
	(RS)	Precision	0.475	0.513	0.298	0.000	0.200	0.000	0.000	0.000
	Edge Histogram	Recall	0.986	0.870	0.818	0.615	0.529	0.349	0.209	0.651
	(EH)	Precision	0.864	0.931	0.906	0.669	0.767	0.744	0.720	0.549
	Homogenous Texture	Recall	0.968	0.762	0.252	0.104	0.460	0.325	0.291	0.093
	(HT)	Precision	0.783	0.824	0.653	0.304	0.444	0.297	0.379	0.170
SMO	Color Layout	Recall	0.906	0.866	0.195	0.200	0.184	0.012	0.093	0.198
	(CL)	Precision	0.758	0.830	0.547	0.391	0.333	0.200	0.500	0.370
	Color Structure	Recall	0.992	0.763	0.179	0.111	0.184	0.133	0.349	0.221
	(CSt)	Precision	0.720	0.931	0.476	0.283	0.410	0.500	0.612	0.171
	Scalable Color	Recall	0.996	0.306	0.000	0.000	0.000	0.000	0.012	0.000
	(SC)	Precision	0.482	0.934	0.000	0.000	0.000	0.000	1.000	0.000
	Contour Shape	Recall	0.903	0.892	0.714	0.289	0.506	0.000	0.000	0.081
	(CS)	Precision	0.868	0.840	0.685	0.639	0.270	0.000	0.000	0.467
	Region Shape	Recall	0.973	0.274	0.000	0.000	0.000	0.000	0.000	0.000
	(RS)	Precision	0.482	0.730	0.000	0.000	0.000	0.000	0.000	0.000
	Edge Histogram	Recall	0.981	0.876	0.828	0.556	0.586	0.325	0.244	0.698
	(EH)	Precision	0.874	0.928	0.906	0.688	0.680	0.692	0.636	0.546
	Homogenous Texture	Recall	0.943	0.635	0.256	0.091	0.325	0.102	0.232	0.000
	(HT)	Precision	0.738	0.741	0.606	0.320	0.331	0.215	0.220	0.000
RBF Network	Color Layout	Recall	0.830	0.869	0.228	0.296	0.230	0.121	0.140	0.256
	(CL)	Precision	0.797	0.825	0.346	0.342	0.198	0.227	0.200	0.339
	Color Structure	Recall	0.949	0.818	0.374	0.311	0.322	0.133	0.326	0.326
	(CSt)	Precision	0.842	0.915	0.547	0.269	0.235	0.220	0.467	0.181
	Scalable Color	Recall	0.282	0.881	0.167	0.052	0.172	0.000	0.000	0.000
	(SC)	Precision	0.615	0.527	0.200	0.119	0.119	0.000	0.000	0.000
	Contour Shape	Recall	0.914	0.883	0.559	0.311	0.379	0.000	0.058	0.302
	(CS)	Precision	0.870	0.854	0.669	0.618	0.260	0.000	0.556	0.193
	Region Shape	Recall	0.515	0.407	0.207	0.007	0.149	0.000	0.000	0.000
	(RS)	Precision	0.458	0.464	0.192	0.004	0.086	0.000	0.000	0.000
	Edge Histogram	Recall	0.985	0.785	0.520	0.637	0.482	0.361	0.670	0.581
	(EH)	Precision	0.798	0.909	0.945	0.601	0.646	0.411	0.697	0.471
	Homogenous Texture	Recall	0.953	0.695	0.204	0.074	0.402	0.157	0.244	0.000
	(HT)	Precision	0.743	0.765	0.632	0.312	0.321	0.245	0.212	0.000

Table 9.3: Object classification results using selected combinations of the MPEG-7 visual descriptors and various data classifiers

Classifier	*Descriptors Combination*	*Measure*	*Object Class*							
			Person Body	*Person Face*	*Horizontal Bar*	*Pole*	*Pillar*	*Discus*	*Hammer*	*Javelin*
libSVM	SC and CS	Recall	0.910	0.901	0.580	0.421	0.382	0.150	0.336	0.352
		Precision	0.881	0.899	0.706	0.622	0.325	0.342	0.640	0.301
	SC and EH	Recall	0.991	0.882	0.825	0.631	0.542	0.361	0.329	0.662
		Precision	0.872	0.945	0.916	0.681	0.786	0.766	0.736	0.561
	CS and EH	Recall	0.995	0.990	0.831	0.634	0.536	0.359	0.230	0.669
		Precision	0.892	0.940	0.912	0.683	0.771	0.740	0.731	0.553
	SC and CS and EH	Recall	0.997	0.994	0.841	0.642	0.550	0.401	0.346	0.672
		Precision	0.895	0.951	0.922	0.689	0.801	0.770	0.742	0.571
SMO	SC and CS	Recall	0.998	0.895	0.725	0.291	0.520	0.000	0.020	0.092
		Precision	0.872	0.941	0.691	0.649	0.281	0.000	1.000	0.475
	SC and EH	Recall	0.999	0.881	0.835	0.568	0.589	0.331	0.251	0.703
		Precision	0.875	0.942	0.910	0.691	0.689	0.699	1.000	0.559
	CS and EH	Recall	0.982	0.899	0.832	0.560	0.591	0.335	0.251	0.702
		Precision	0.880	0.939	0.912	0.695	0.692	0.701	0.642	0.560
	SC and CS and EH	Recall	0.999	0.901	0.840	0.571	0.601	0.341	0.259	0.712
		Precision	0.882	0.945	0.915	0.680	0.682	0.701	1.000	0.565
RBF Network	SC and CS	Recall	0.915	0.888	0.669	0.325	0.388	0.000	0.062	0.306
		Precision	0.872	0.862	0.660	0.617	0.271	0.000	0.550	0.192
	SC and EH	Recall	0.988	0.895	0.529	0.652	0.488	0.370	0.679	0.592
		Precision	0.802	0.909	0.952	0.601	0.652	0.412	0.709	0.469
	CS and EH	Recall	0.985	0.890	0.572	0.642	0.492	0.360	0.682	0.592
		Precision	0.872	0.892	0.950	0.629	0.654	0.421	0.701	0.479
	SC and CS and EH	Recall	0.901	0.899	0.662	0.661	0.495	0.371	0.685	0.598
		Precision	0.879	0.912	0.960	0.631	0.659	0.431	0.712	0.481

the majority of the descriptor models. Furthermore, the fluctuation in classification performance during parameters' tuning is significantly lower than that of the other two training algorithms. The models created using libSVM are the ones that are able to discriminate between multiple classes for all individual descriptors used. A characteristic example is the model created for the Scalable Color descriptor. The libSVM model for this descriptor can be used for the discrimination between the seven of the eight object classes (Javelin class is an exception) while the corresponding SMO and RBF network models are only able to discriminate between three classes at most.

Chapter 10

Spatial Histogram of Keypoints

Among a variety of feature extraction approaches, special attention has been given to the SIFT algorithm which delivers good results for many applications. However, the non fixed and huge dimensionality of the extracted SIFT feature vector cause certain limitations when it is used in machine learning frameworks. In this chapter we tried to overcome the problems of the dimensionality reduction as well the lack of the spatial information of the Bag-Of-Features method by introducing the Spatial Histograms of Keypoints (SHiK) algorithm. The Spatial Histogram of Keypoints (SHiK) keeps the spatial information of localized keypoints, on an effort to overcome this limitation. The proposed technique partitions the image into a fixed number of ordered sub-regions based on the Hilbert space-filling curve and counts the localized keypoints found inside each sub-region. The resulting spatial histogram is a compact and discriminative low-level feature vector that shows significantly improved performance on classification tasks. The proposed method achieves high accuracy on different datasets and performs significantly better on scene datasets compared to the Spatial Pyramid Matching method. Specifically, we utilized the Pyramid Histogram Of visual Words (PHOW) scheme as presented in [184]. PHOW is based on spatial pyramid matching and obtains better categorization performance than the original BoW.

10.1 Scale Invariant Feature Transform

The Scale Invariant Feature Transform (SIFT) extract a keypoint descriptor that is robustly invariant to general image transformations (rotation, translation and scale) and partially invariant to affine distortion, illumination change and noise [151]. The algorithm consists of 4 main steps: (a) Detection of scale-space extrema, (b) keypoint localization, (c) orientation assignment, and (d) computation of keypoint descriptor.

10.1.1 Detection of Scale-Space Extrema

The first step of the algorithm focuses on the identification of the potential interest points that are invariant to scale and orientation. The identification is accomplished by detecting local extrema of Difference-of-Gaussian function at different scales. The convolution of an input image $I(x, y)$ with the variable-scale Gaussian, $G(x, y, \sigma)$, gives the scale space function, $L(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{10.1}$$

Where,

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{10.2}$$

The Difference-of-Gaussians, $D(x, y,)$, is calculated from the difference of two adjacent scales separated by a constant multiplicative factor k:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{10.3}$$

For the detection of local extrema (maxima and minima) each sample point is compared to its eight neighbors in the current image and the nine neighbors in the scale above and below. In case that the sample point is the larger or the smaller than all of them, is identified as potential interest point.

10.1.2 Keypoint Localization

In the second step, the accurate localization of all potential interest points identified in the previous step is occurred. The points comprising low contrast or corresponding to responses along edges are rejected for selecting the more stable ones. In order to detect the points with low contrast, the interpolated location is determined by utilizing the Taylor expansion of the scale-space function, $D(x, y, \sigma)$, shifted so that the origin is at the sample point:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathbf{T}}\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{10.4}$$

Where D and its derivatives are computed at the candidate point and $\mathbf{x} = (x, y, \sigma)^T$ is the offset from this point. The location of the $\widehat{\mathbf{x}}$ is calculated by equating the derivative of the previous equation with respect to $\mathbf{x}$ to zero:

$$\widehat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2}\frac{\partial D}{\partial \mathbf{x}} \tag{10.5}$$

If the $\widehat{\mathbf{x}}$ is larger than 0.5 in any direction then the current extremum lies closer to another candidate sample point. In such a case, the sample point is changed and the interpolation perfomed instead about that point. The final offset $\widehat{\mathbf{x}}$ is added to the location of its sample point to get the interpolated estimate for the location of the extremum.

The detection of unstable extremas with low contrast is achieved by replacing the location of the extremun, $\widehat{\mathbf{x}}$, in the Taylor expansion:

$$D(\widehat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \mathbf{x}}\widehat{\mathbf{x}} \tag{10.6}$$

Extremas having $D(\widehat{\mathbf{x}})$ less than 0.03 are rejected.

The rejection of candidate points corresponding to responses along edges is based on the assumption that a poorly defined peak in the DoG function has a large principle curvature across the edge but a small one in the perpendicular direction. These points are detected using the following formula:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} < \frac{(r+1)^2}{r} \tag{10.7}$$

Where the $Tr(\mathbf{H})$ represents the sum and $Det(\mathbf{H})$ the product of the eigenvalues from the Hessian matrix $\mathbf{H}$. $\mathbf{H}$ is computed at the location and scale of the keypoint:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{10.8}$$

Where derivatives are estimated by taking differences of neighboring sample points. According to Lowe [151], the r value in eq. 10.7 was set equal to 10 and therefore keypoints having ratio between the principal curvatures greater than 10 were eliminated.

10.1.3 Orientation Assignment

In the third step, each keypoint is assigned an orientation based on local image properties. This step makes the final keypoint descriptor to be invariant to image rotation since it can be represented relative to the assigned orientation. The gradient magnitude and orientation are calculated using the following formulas:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{10.9}$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1))/(L(x+1, y) - L(x-1, y))) \qquad (10.10)$$

Where $L(x, y)$ is the Gaussian smoothed image sample at the scale of the keypoint.

For each keypoint, a 36-bin histogram that covers the whole range of orientations is created using the gradient orientations of sample points that are around its region. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. The highest peak in the histogram with any other local peak that is within 80% of the highest peak is used to create a keypoint with that orientation.

10.1.4 Keypoint Descriptor

After assigning the image location, scale and orientation, the keypoint descriptor is created. The gradient magnitude and orientation at each image sample point in a region ($16x16$ sample array) around the keypoint is calculated. Each sample point is weighted by a Gaussian weighting function with equal to 0.5 of the descriptor window. The samples are accumulated into $4x4$ array of orientation histograms with 8 bins in each. The feature vector dimensionality for each keypoint corresponds to $4x4x8 = 128$.

10.2 Spatial Pyramid Matching Method

Spatial Pyramid Matching is a kernel-based method used by Lazebnik et al. [161] for scene recognition. The method utilizes a modification of the pyramid match kernels proposed by Grauman and Darrell [185]. The image is partitioned into increasingly fine sub-regions and local features found inside each sub-region are accumulated into the corresponding histogram. The pyramid matching is performed in two-dimensional image space and features vectors are quantized into M discrete types. Each m channel represents the coordinates of the m-type features in the corresponding images and consists of two vectors (X_m, Y_m). Using the pyramid match kernel:

$$k^L(X,Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{10.11}$$

where $l = 0, ..., L$ is the level of resolution for the grid and I is the histogram intersection function. The final kernel is equal to the sum of the separate channel kernels:

$$K^L(X,Y) = \sum_{m=1}^{M} k^L(X_m, Y_m) \tag{10.12}$$

Bosch et al. [184] generalized the spatial pyramid matching method from an image to a region of interest (ROI) and classified images into object categories. They applied spatial pyramid matching to both appearance and local shape. In case of appearance, the SIFT descriptors were computed at points on a regular grid with spacing M=10 pixels. At each grid point the descriptors were computed over four 4 support patches with different radii (*i.e.* radii=4, 8, 12, 16) and each point is represented by four SIFT descriptors. Finally, the k-means algorihm quantized the feature vector into 300 visual words. The local shape was represented using the HOG features within an image subregion quantized into K bins. Each bin indicated the accumulation of edges having orientations within a certain angular range. Two shape descriptors were created for orientations in the range [0 180] and [0 360] respectively. The histogram of the first descriptor was discretized into $K = 20$ bins, while the histogram of the second one into $K = 40$. The appearance and shape descriptors were combined with spatial layout of the image and gave two representations: (a) the Pyramid Histogram Of Visual Words (PHOW), and (b) the Pyramid HOG (PHOG) descriptors.

10.3 Hilbert's Space-Filling Curve

In the proposed method we utilized the natural idea of locality, which support that points which are close together on the Euclidean distance space are grouped together in the resulting ordering and applied the Hilbert curve to create the spatial histogram of the keypoints. The use of space-filling curves is a widely used method for ordering data. Such functions are continuous and self similar and tend to be good at preserving locality points that are close together. Points that are close using the Euclidean distance in an n-dimensional set tend to be close together in the linear ordering defined by the curve.

Jordan [186] defined a curve in one-dimension as a set of points $(\phi(t), \psi(t))$, where ϕ and ψ are continous functions on some closed interval (*e.g.* [0, 1]). The t can be considered as the time and the curve as the path of a particle starting at $(\phi(0), \psi(0))$ (*i.e.* t=0) and ending at $(\phi(1), \psi(1))$ (*i.e.* t=1). Peano [187] presented a curve whose range contains the entire 2-dimensional unit square, known as space-filling curve. Hilbert [188] presented a variation of the space-filling curve based on a geometric approach.

The Hilbert fractal curves have strong locality properties [189] and have extensively used for spatial groupings of data in a variety of applications, including visual feature extraction [190]. The way in which the Hilbert curve is drawn is shown in Fig. 10.1 where the first two steps of the process is illustrated for the 2-dimensional case. Let $\boldsymbol{I}$=$\{z \mid 0 \leqslant z \leqslant 1\}$ denote the unit interval and $\boldsymbol{S}$=$\{(x,y) \mid 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1\}$ the unit square. For each n integer, $\boldsymbol{I}$ is subdivided into 4^n closed subintervals of length 4^{-n} and $\boldsymbol{S}$ into 4^n closed quadrants of side 2^{-n}, respectively. The correspondence between the subintervals and quadrants amounts to numbering the quadrants so that the adjacency and nesting conditions are satisfied [191]:

Adjacency Condition Adjacent subintervals correspond to adjacent quadrants (with an edge in common).

Nesting Condition If at the n-th partition, the subinterval I_{n_k} corresponds to a quadrant S_{n_k} then at $(n+1)$-st partition the 4 subintervals of I_{n_k} must correspond to the 4 quadrants of S_{n_k}.

For each n, the 4^n subintervals are labeled in their natural order from left to right. The centers of the quadrants are connected by consecutive straight lines in the manner indicated in Fig. 10.1. The first quadrant is always in the left corner and the last in the lower right one. The Hilbert space-filling curve starts at (0,0) at t=0 and ends at (1,0) at $t = 1$. Since the first and last quadrants of each partition are determined, only one enumeration of the quadrants can be achieved that satisfies the adjaceny and nesting conditions.

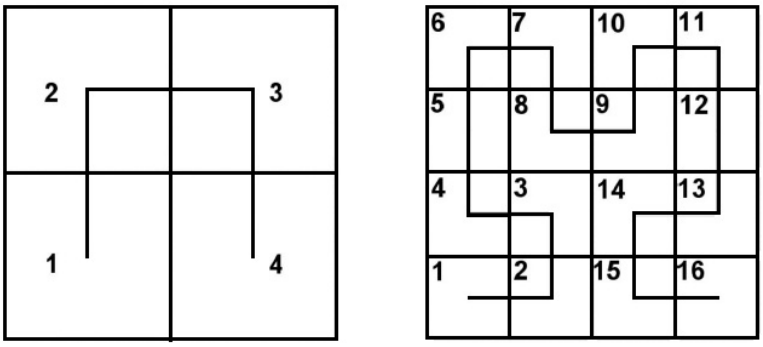


Figure 10.1: Hilbert curves of order 1 and 2.

The way the sub-regions are ordered are very important because it preserves the spatial relations between localized keypoints. This is an important improvement compared to similar approaches (e.g. Lazebnik et al. [161]). Furthermore, the use of Hilbert curves allows a multiresolution representation in case where space-filling curves of increasing order n are combined together.

10.4 Spatial Histogram of Keypoints (SHiK) Algorithm

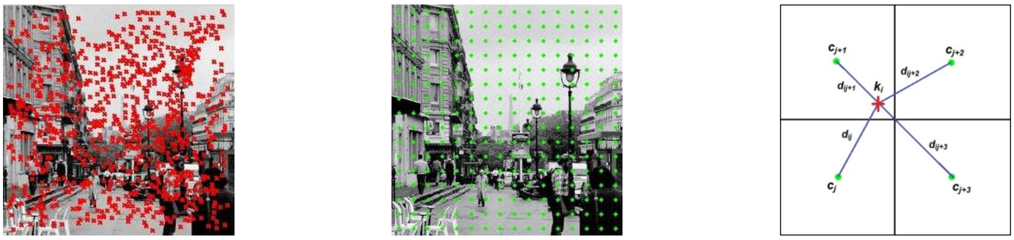
The proposed algorithm named Spatial Histogram of Keypoints (SHiK) utilizes the same procedure as the SIFT algorithm to localize the keypoints and then computes their spatial histogram based on the Hilbert fractal geometry.

The localization of the keypoints is achieved in the first two steps of the SIFT algorithm. The first step searches for scale-space peaks over all scales and image locations in a series of Difference of Gaussian (DoG). The local minima and maxima are detected by comparing each sample point to its 26 neighbors in 3x3 regions at the current and adjacent scales. In the second step, a detailed model is fit to determine location, scale and ratio of principal curvatures at each candidate location. The keypoints are selected based on measures of their stability by rejecting those having low contrast or being poorly localized along an edge.

The spatial histogram of localized keypoints for a given image is calculated based on their location in relation to Hilbert space-filling curve. A space-filling curve of order n passes through 4^n sub-regions and their center points comprises a set of points $\boldsymbol{C}$=$\{c_1$, ..., $c_{4^n}\}$. The set of keypoints extracted by the SIFT approach is represented by $\boldsymbol{K}$=$\{k_1$, ..., $k_N\}$. The k_i indicates the *i-th* keypoint while the N denotes the total number of keypoints. A keypoint k_i is assigned to the center point c_j if the distance between them, d_{ij} satisfies the following criterion:

$$d_{ij} = \underset{1\leqslant z\leqslant 4^n}{\arg\min}\{d_{iz}\} \qquad (10.13)$$

The distance (Euclidean distance) between the keypoint k_i and center point c_z is computed using the formula:



(a) (b) (c)

Figure 10.2: (a): Keypoints Localization for a given image. (b): Center points of the sub-regions of a Hilbert space-filling curve of order n=4. (c) An example of the distance calculation between the keypoint K_i and its four neighbor center points.

$$d_{iz} = [(k_{i_x} - c_{z_x})^2) + (k_{i_y} - c_{z_y})^2]^{1/2} \tag{10.14}$$

where $k_i : (k_{i_x}, k_{i_y})$ and $c_z : (c_{z_x}, c_{z_y})$.

The spatial histogram $\boldsymbol{H}$ of the image contains 4^n bins. The $\boldsymbol{H}_j$ bin is equal to the number of keypoints that are closest to the c_j center point. For our experiments we utilized a Hilbert space-filling curve of n=4 which results in a histogram of 256 bins. Comprising a multiresolution representation by utilizing various orders of Hilbert space-filling curves is also possible and it has been examined in other studies. An example of the algorithm is shown in Fig. 10.2.

10.5 Experiments

A series of experiments was conducted to evaluate the performance of the proposed algorithm on three different datasets. The first dataset consists of 400 images from the athletics domain [192] separated in 8 category classes, 50 images in each class. The second dataset contains 13 scene categories [193] (8 categories were created from Oliva and Torralba [194]) and each category has 210 to 410 images. Finally, the third dataset was created using 20 object categories from Caltech-101 [195] where each category contains 59 to 800 images.

To overcome the multiclass classification problem and facilitate effective and efficient learning, each category is treated as a separate binary classification case with the support vector machine (SVM) training scheme. We have followed the one-against-rest approach [196] and we have built a total number of N models, one for each category. The feature vectors of each category class were split into two groups, called the training (80%) and testing (20%) set. Each mode is trained and tested between one class and the *N-1* other classes. The training and testing set for each model contain the feature vectors of the corresponding category class and the same number of randomly selected feature vectors of the rest *N-1* classes.

Tab. 10.1 summarizes the classification results for the three diverse datasets while Tab. 10.2 compares the performance of SHiK and PHOW algorithms. Concerning the first dataset, the experimental results indicate that SHiK features perform better than PHOW. The classification accuracy obtained using these features is quite good achieving an average classification accuracy values in the range of 75%- 90%, with a total average classification accuracy equal to 81.88%. The best classification performance (90%) was occurred for “Discus” and “Triple Jump” categories. This may happens because the content of the images belonging to these categories have a unique object like the purpose of “Discus” or unique human posture as in “Triple Jump” which presents also the best performance when using PHOW (80%). On the other hand, “Hurdles” category presents the lowest performance for both SHiK (75%) and PHOW (60%) features.

The behavior of the SHiK features on the second dataset is also noteworthy: With the exception of the “Suburb” class, where a high performance is obtained, the classification accuracy for the outdoor classes varies from 67.36% to 73.78% while the results for indoor classes (“Bedroom”, “Kitchen”, “Living Room”, “Office”) are quite perfect and higher than 80%. The performance of PHOW was disappointing on this dataset using

Table 10.1: Classification results using SHiK and PHOW features.

Datasets		***Algorithms***			
		SHiK		***PHOW***	
		Rate (%)	***F-Measure (%)***	***Rate (%)***	***F-Measure (%)***
Athletics	*Discus*	90.0	0.91	70.0	0.70
	Hammer	85.0	0.82	70.0	0.67
	High Jump	80.0	0.82	75.0	0.74
	Hurdles	75.0	0.67	60.0	0.67
	Javelin	75.0	0.78	70.0	0.67
	Long Jump	80.0	0.82	80.0	0.80
	Running	80.0	0.80	65.0	0.70
	Triple Jump	90.0	0.91	80.0	0.80
Scenes	*Bedroom*	99.63	0.92	53.13	0.52
	Suburb	88.54	0.89	43.02	0.40
	Kitchen	81.82	0.82	69.57	0.72
	Living Room	89.13	0.89	63.04	0.64
	Coast	67.36	0.73	44.44	0.47
	Forest	68.73	0.74	38.64	0.36
	Highway	72.12	0.77	53.85	0.60
	Inside City	70.73	0.78	35.48	0.39
	Mountain	72.0	0.76	49.33	0.48
	Open Country	73.78	0.80	36.59	0.28
	Street	70.09	0.71	37.61	0.39
	Tall Building	71.83	0.76	50.35	0.44
	Office	87.88	0.89	34.33	0.33
Caltech-101	*Airplanes*	98.25	0.98	95.65	0.96
	Background	72.06	0.71	97.79	0.98
	Bonsai	63.42	0.72	95.24	0.95
	Brain	68.75	0.69	93.94	0.94
	Buddha	72.73	0.73	86.97	0.80
	Butterfly	84.62	0.86	96.30	0.97
	Chandelier	66.67	0.71	91.67	0.91
	Faces	99.43	0.99	96.55	0.97
	Faces Easy	90.23	0.92	95.4	0.96
	Ketch	77.42	0.84	93.75	0.95
	Laptop	70.97	0.73	90.32	0.89
	Leopards	96.25	0.96	93.75	0.96
	Llama	73.91	0.79	95.65	0.95
	Menorah	69.23	0.64	88.46	0.89
	Scorpion	70.0	0.73	90.0	0.90
	Soccer Ball	69.57	0.63	95.65	0.95
	Umbrella	70.83	0.74	100.0	1.0
	Watch	66.04	0.61	96.23	0.95
	Wheelchair	76.19	0.80	95.45	0.67
	Yin Yang	78.26	0.74	95.65	0.95

Table 10.2: Performance comparison of SHiK and PHOW features.

Datasets	***Algorithms***								***Interpretation***
	SHiK				***PHOW***				
	Rate(%)	F-measure	TPR (%)	FPR (%)	Rate (%)	F-measure	TPR (%)	FPR (%)	
Athletics	81.88	0.82	0.84	0.2	71.25	0.72	0.73	0.27	s. (significant)
Scenes	77.28	0.80	0.70	0.07	46.87	0.46	0.45	0.49	s.
Caltech-101	76.74	0.78	0.82	0.28	94.22	0.92	0.87	0.002	s.

the proposed classification protocol. The overall classification accuracy is lower than 50% with the indoor classes having an important role in experimental results. The three of the four indoor classes have the highest accuracy while the fourth one presents the lowest (34.33%) among the others.

PHOW performs better than SHiK features on the Caltech-101 dataset where the average classification varies from 86.97% to 100% on the twenty object classes. The SHiK features can achieve quite good levels of accuracy for many categories while the highest levels occur in the purpose of "Faces" (99.43%), "Faces Easy" (90.23%) and "Airplanes" (98.25%) categories. The results are perfectly justified since the specific categories contain solid objects covering the largest part of the image and this gives adequate values to the spatial histogram bins, especially those represent areas around image boundaries.

The t-test applied to assess whether there is enough empirical evidence to claim a difference between the SHiK and PHOW algorithm. The specific statistical test compares the mean values of the two distributions to verify whether the hypothesis that there is no difference between the two algorithms is rejected. Using the F-measure values of each algorithms performance, the t-test at the 5% confidence level [143] was used to test the significance of the differences in performance for each dataset separately. The results showed that the proposed algorithm (SHiK) performed significantly better than PHOW for two of the three datasets (Athletics, Scenes).

10.6 Conclusions and Remarks

In this chapter we presented the SHiK algorithm which maintains the spatial information of the SIFT keypoints and results in a feature vector with fixed and low dimensionality. In this framework we assign the localized keypoints of SIFT algorithm to the closest center point of the ordered sub-regions of the image obtained by applying the Hilbert space-filling curve. The keypoints assignment results in a spatial histogram with a number of bins equal to the number of sub-regions. The new feature vector was tested in classifying images in three diverse datasets and showed very promising results, especially in scene categorization. The first dataset consisted of 400 images from the athletics domain separated in 8 categories. The second dataset contained 13 scene categories[1]. The third dataset contained images from 20 object categories from Caltech101[2]. The SHiK algorithm is available to the research community for further experiments[3].

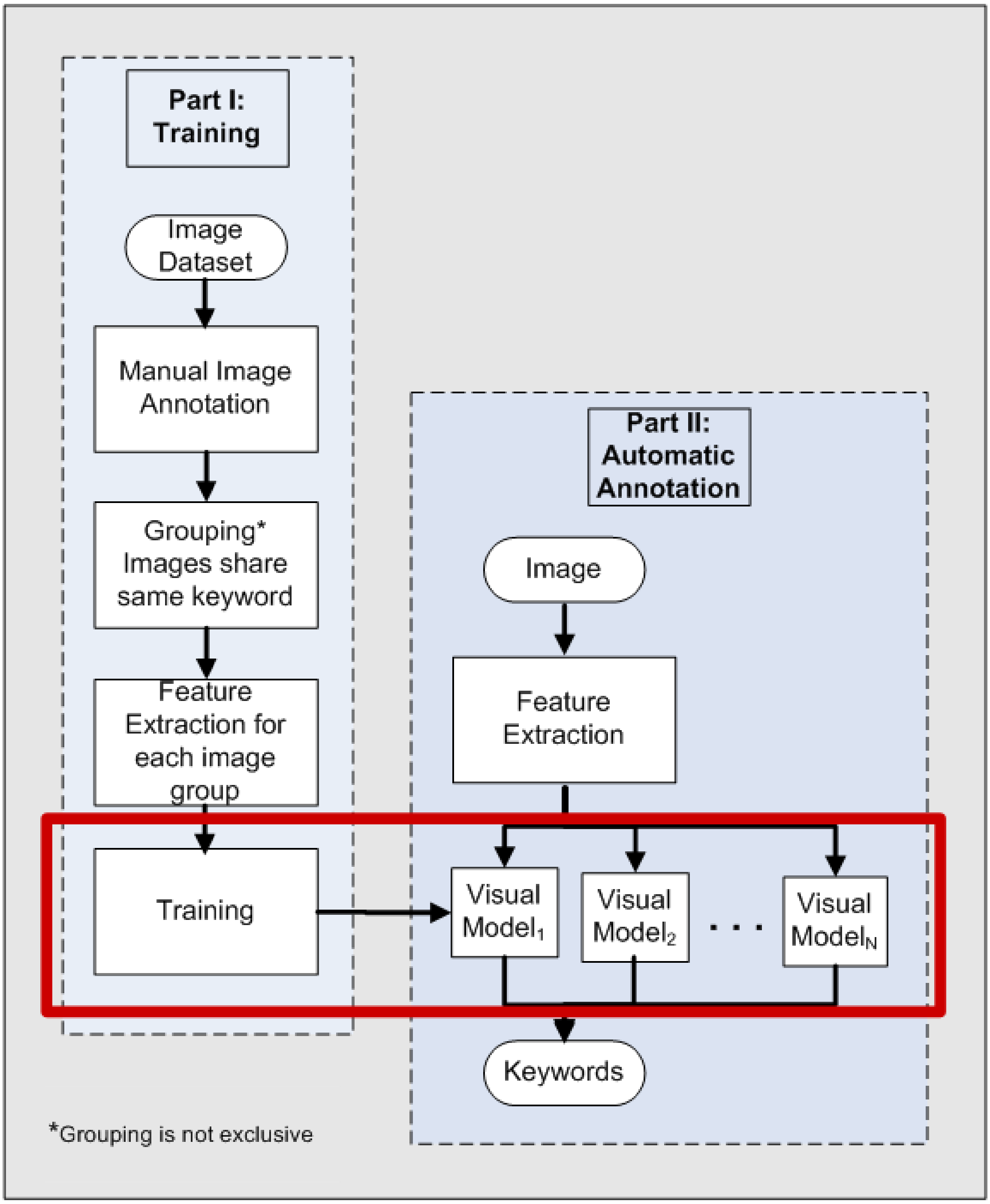
[1] http://vision.stanford.edu/resources_links.html.

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/.

[3] http://cis.cut.ac.cy/∼z.theodosiou/shik.zip.

Part III

Creating Visual Models using Machine Learning Techniques


Part I: Training
Image Dataset
Manual Image Annotation
Grouping* Images share same keyword
Feature Extraction for each image group
Training
Part II: Automatic Annotation
Image
Feature Extraction
Visual Model1
Visual Model2
. . .
Visual ModelN
Keywords
*Grouping is not exclusive

Chapter 11

Machine Learning Techniques

Machine learning methods play an important role in automatic image annotation schemes. Machine learning involves algorithms that build general hypotheses based on supplied instances and then use them to make predictions about future instances (known also as the "learning by example paradigm"). Classification algorithms are based on the assumption that input data belong to one of several classes that may be specified either by an analyst or automatically clustered. Many analysts combine supervised and unsupervised classification processes to develop final output analysis and classified maps.

Supervised image classification organizes instances into classes by analyzing the properties of the supplied image visual features where each instance is represented by the same number of features. Training instances are split into training and test sets. Initially, the characteristic properties of the visual features of the training instances are isolated and class learning finds the description that is shared by all positives instances. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. Several supervised methods based on probabilistic classifiers, rules, neural networks, and statistical learning have been utilized for classifying images into class labels as well as for keyword model creation.

11.1 Supervised Learning

Probabilistic classifiers are derived from generative models and are product distributions over the original attribute space or more involved spaces. Naive Bayes and Bayesian networks are examples of probabilistic classifiers. These classifiers are based on the posterior probability that an image is related to any particular concept, given the observation of certain features from the image or a region. Given a set of images and a set of classes, a variety of statistical models try to determine the posterior probability from the conditional probabilities and the priors. The most common model used in classification schemes is the Naive Bayes which ignores possible dependencies, correlations, among the inputs and reduces a multivariate problem to a group of univariate problems [197].

Decision trees are logic-based learning algorithms that sort instances according to feature values based on the divide-and-conquer approach. They are developed by algorithms that split the input set of visual features into branch-like segments (nodes). A decision tree consists of internal decision nodes where a test function is applied and the proper branch is taken based on the outcome. The process starts at the root node and it is repeated until a leaf node is achieved. There is a unique path for data to enter class that is defined by each leaf node and this path is used to classify unseen data. A variety of decision trees methods have been used in classification tasks such as CART [198], ID3 decision tree [199], its extension C4.5 [200] that has shown a good balance between speed and error rate [201], and the newest Random Forest [202].

Although, decision trees offer a very fast processing and training phase comparing to other machine learning approaches, suffer from the problem of overfitting to the training data, resulting in some cases in excessively detailed trees and low predictive power for previously unseen data (lack of generalization). Furthermore, decision trees were designed for classification tasks: Every input entering the tree's root is classified to only one of its leafs. Assuming that the leafs correspond to keywords and the input is a low-level vector extracted from an input image then this image is assigned at most one keyword during prediction. The keyword models, on the other hand, are based on the one-against-all paradigm. For each keyword there is a dedicated predictor which decides, based on the low-level feature vector it fed with, whether the corresponding image must be assigned the particular keyword or not.

The rules created for each path from the root to a leaf in a decision tree can also be used directly for classification. Rule based algorithms aim to construct the smallest rule-set that is consistent with the training data. In comparison with decision trees, rule-based learning evaluates the quality of the set of instances that is covered by the candidate rule while the former evaluates the average quality of a number of disjoint sets [203]. However, rule-based learning faces problems with noisy data. More efficient learners have been proposed such as the IPER (Incremental Reduced Error Pruning) [204] and Ripper (Repeated Incremental Pruning to Produce Error Reduction) [205] to overcome these drawbacks.

Assuming that every keyword is modeled with a rule, then rule-based learning is appropriate for creating visual models for keywords since it provides the required scalability. That is, every time a new keyword must be modeled a new rule is constructed, based on available training data, without affecting the existing keyword models (rules). Unfortunately, the case is not so simple. Rule based systems perform well in cases where the dimensionality of input is limited. However, the low-level features that are used to capture the visual content of an image or image region are inherently of high dimensionality. Thus, despite their scalability rule-based systems lack in performance compared to other keyword modeling schemes.

Neural Networks (NNs) have incredible generalization and good learning ability for classification tasks. NNs use the input data and train a network to learn a complex mapping for classification. A NN is supplied by the input instances and actual outputs and then compares the predicted class with the actual class and estimates the error to modify the weights. There are numerous NNs based algorithms with significant research interest in Radial Basis Function (RBF) networks [206] since in comparison with the multilayer perceptrons, the RBF trains faster and their hidden layer has easier interpretation. Furthermore, in a comparative study on object classification [207] for keyword extraction purposes, RBFs proved to be the more robust and with the highest predicting performance among several state of the art NN classifiers.

An RBF network consists on an input layer, a hidden layer with a RBF activation function and a linear output layer. Tuning the activation function to achieve the best performance is a little bit tricky, quite arbitrary and dependent on the training data. Thus, despite their significant abilities in prediction and generalization RBF networks are not as popular as the statistical learning approaches discussed next.

Support Vector Machines (SVMs), a machine learning scheme which is based on statistical learning theory, is one of the most popular approaches to data modeling and classification [208]. SVMs, with the aid of kernel mapping, transform input data into a high dimensional feature space and try to find the hyperplane that separates positive from negative samples. The kernel can be linear, polynomial, Gaussian, Radial Basis Function (RBF), etc. The hyperplane is chosen such as it keeps the distance between the nearest positive and negative examples as high as possible. The number of features encountered in the training data does not affect the model complexity of an SVM, so SVMs deal perfectly with learning tasks where the dimensionality of feature space is large with respect to the number of training instances. Training a classifier using SVM has less probability of losing important information because SVM constructs the optimal hyperplane using dot products of the training feature vectors with the hyperplane. Sequential Minimal Optimization (SMO) [181], [182] and LibSVM [180] are two of the state of the art implementations of the SVMs with high classification performance.

As far as the creation of visual models from keywords is concerned, the SVMs have many desirable properties. First, they are designed to deal with binary problems (they make decisions on whether an input instance belongs or not to a particular class) which provides the required scalability to train new keyword models independently of the existing ones. Second, they deal effectively with the large dimensionality of the input space created by low-level feature vectors extracted from images. Finally, they do not require so many training examples as other machine learning methods. As we have already mentioned, in automatic image annotation the availability of training data is a key issue. Therefore, methods that are conservative in this requirement are highly preferable.

More sophisticated training frameworks utilized more than one classifier in order to improve the classification performance. Usually several classifiers are created and their outputs are combined under a combination rule to produce an overall output. The ensemble obtains higher classication accuracy when there is a significant diversity among the classifiers [209]. Several studies have been proposed for analyzing the diversity measures for classifier ensembles [209], [210], [211]. The classifiers can be trained using different examples or different features of the same training set or even different learning models trained by the same training examples. The combination rule is an interesting issue in the research community. Among a variety of methods such as LSE-based weighting and double-layer hierarchical combining [212], the majority vote can be considered as the simplest one. In such a case, the votes received from the individual classifiers are counted for each class. The class which receives the largest number of votes is selected as the consensus decision [213].

Kim et al. presented the SVM ensemble in [212] in order to overcome the limitations of the SVM on multi-class classification and large scale data. Each SVM classifier is trained via the bagging (bootstrap aggregating) or boosting method. Bootstrapping method builds a number of replicate training data sets by randomly re-sampling with replacement the given training data. Therefore, each training example may appear in any particular but not in all training datasets. Different SVM classifiers are created independently using the training data sets and aggregated together under a combination rule. On the other hand, the creation of the training data sets using a boosting algorithm (e.g. AdaBoost algorithm [214]) is quite different from bagging. In boosting, the SVM networks are trained sequentially. Initially, all training example are assigned to have the same value of weight. A number of training examples is selected to build a classifier which is then tested on the whole training set. The weight value of each example is updated based on the classification results: the values of the incorrectly training examples are increased while the values of the correctly classified are decreased. The main idea is to select more frequently those samples that are hard for classification that the others. The updated weights are then used to build the training data set for the next classifier. The sampling procedure is repeated until the number of desired SVM classifiers is built.

11.2 Co-Training

The creation of visual models requires an enormous number of manually annotated data. Since the manual image annotation is a time-consuming and expensive task, the co-training seems to be a good solution for the creation of visual models using a limited number of training data. Co-training was first introduced by Blum and Mitchell [215] to boost the performance of a learning algorithm using unlabeled data when only a small set of labeled data is available. The algorithm can be applied to datasets that have natural separation of their features into two disjoint sets. Two separate classifiers are built, one for each feature set, and their predictions are combined to reduce the classification error.

The method was initially tested on classifying web pages [215], where the two different views consist of a bag of words, appearing on the web page, and on a bag of words, underlined in all links pointing into the web page from other pages in the database. The classifiers were trained using the naive Bayes algorithm. In [216], an object tracking algorithm is presented using a co-training SVM framework. The algorithm was based on color histograms and HOG features. Each feature set was used to train a support vector machine and their outputs combined for the final classification result. Zhang and Lee [217] applied co-training for Web image classification using both text and image data. More recently, a new method based on co-training, named co-graph, was presented also for web image classification [218]. Three different views were used for image representation including global, local and textual features and a separate graph was constructed for each view.

Using the Probably Approximately Correct (PAC) learning theory [219], the algorithm supports effective learning using both labeled and unlabeled data under two assumptions: (1) each set of features is sufficient for classification, and (2) the feature sets of each example are conditionally independent given the class. Therefore, the co-training algorithm requires two independent views X^1, X^2 (e.g. different feature sets) of the data X. Each view should provide different complementary information and simultaneously should be by itself sufficient for correct classification. Although the use of two independent views yields to fewer generalization errors [220], it might not be possible in some cases.

The process of the co-training algorithm is illustrated in Fig. 11.1. Co-training is an iterative process where the classifiers are incrementally built using the two feature sets. Initially, each classifier is built using just few labeled data. At each round, the highest confidence estimates on unlabeled data of each classifier are used to enlarge the training set of the other. Each classifier is then rebuilt from the augmented labeled set.

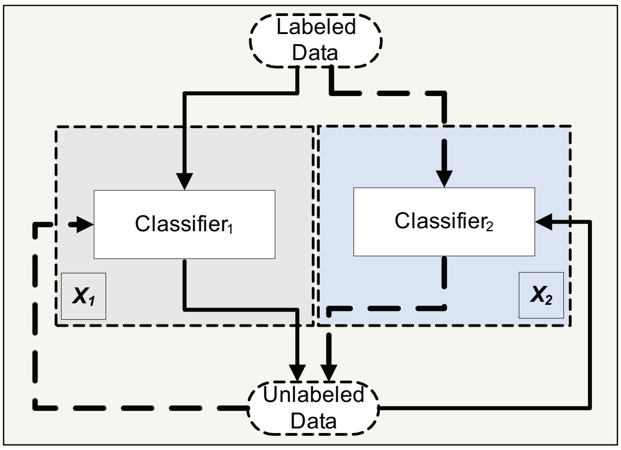


Figure 11.1: The co-training procedure.

11.3 Confident Co-Training with Data Editing

Standard co-training and its variations measure the labeling confidence on unlabeled examples implicitly. In [215], the confidence measurement is based on the posterior probability of the classifiers outputs, while in [221] and [222] the cross-validation is performed on the original labeled examples to compute the correct predictions. Zhou and Li [223], proposed an algorithm which generates three classifiers from the original labeled example set. These classifiers are then refined using unlabeled examples in the tri-training process, where each unlabeled example is labeled for a classier if the other two classiers agree on the labeling, under certain conditions.

Zhang and Zhou [224], proposed the Confident Co-Training with Data Editing (CoTrade) algorithm which estimates explicitly the confidence on unlabeled examples using specific data editing techniques. A number of examples that predicted with high confidence from one classifier are then passed to the other. In each co-training round, the algorithm implements two steps consecutively. In the first step, a graph is constructed over the labeled and unlabeled examples and the cut edge weight statistic [225] [226] is utilized to evaluate the labeling confidence of each example. In the second step, the training set is updated by optimizing the expected error rate with the aid of the classification noise rate.

11.3.1 Data Editing

For a given labeled examples set, $Z = \{(z_i, y_i)|i = 1, 2, ..., Z\}$, where z_i and y_i indicating the example and label respectively, an undirected neighborhood graph G_Z is constructed based on the k-nearest criterion. Each labeled example, (z_i, y_i), corresponds to a vertex in G_Z and is connected with a vertex (z_j, y_j) with an edge $\overline{ij}$, if the z_j is among the k-nearest neighbors of z_i or the z_i is among the k-nearest neighbors of z_j. The edge $\overline{ij}$ is associated with a weight $w_{ij} \in$ [0,1]:

$$w_{ij} = \frac{1}{1 + d(z_i, z_j)} \tag{11.1}$$

Where $d(z_i, z_j)$ indicates the Euclidean distance between the example z_i and z_j.

The labeling confidence is calculated using the assumption that a correctly labeled example should have the same label with its neighbors. Therefore, a *cut edge* is every edge on the G_z whose vertices have different labels. The labeling confidence for the example (z_i, y_i) is calculated based on the *cut edge statistic*, using the following formula:

$$J_i = \sum_{z_j \in C_i} w_{ij} I_{ij} \tag{11.2}$$

Where C_i denotes the set of examples that are connected with z_i in the G_Z, and I_{ij} denotes the independent and identically distributed Bernoulli random variable which is equal to 1 when label y_i is not equal to label y_j (edge cut). When the size of C_i is sufficiently large, then according to the central limit theorem, J_i can be modeled by a normal distribution $N(0,1)$. Based on the left unilateral p-value of J_i^s with the respect to $N(0,1)$, the labeling confidence of (z_i, z_j) is calculated:

$$CF_Z(z_i, z_j) = 1 - \Phi(J_i^s) \tag{11.3}$$

Where $\Phi(J_i^s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{J_i^s} e^{-\frac{t^2}{2}dt}$ indicates the p-value of J_i^s under standard normal distribution.

11.3.2 Labeling Information Exchange

The CoTrade algorithm utilizes the learning from noisy examples [227] to handle the labeling information exchange. According to learning from noisy examples, the number of noisy examples m that are needed to create a classifier with low rate error satisfies the following inequality:

$$m \geq \frac{2}{\epsilon^2(1-2n^b)^n} \ln(\frac{2N}{\delta}) \tag{11.4}$$

Where ϵ indicates the tolerance parameter, δ indicates the confidence parameter, and n^b denotes the upper bound on the noise rate. N denotes the size of finite hypothesis space, $H = \{H_i | i = 1, 2, ..., N\}$, where each hypothesis H_i maps the input space to the out space. The CoTrade algorithm identifies the optimal choice of unlabeled examples for labeling information exchange that would give the smallest classification error.

Chapter 12

Visual keyword Modelling: An overall evaluation

This chapter presents an overall evaluation of the visual keyword modelling using several low-level features and machine learning techniques.

12.1 Dataset Creation Used in Experimental Setup

Initially 500 images[1] related to the athletics domain was used for our experiments. The images were manually annotated by 15 users using a predefined vocabulary of 33 keywords [162]. Manual annotation was performed with the aid of the MuLVAT annotation tool [228].

For the experiments eight representative keywords have been selected and for each keyword, 50 images that were annotated with this keyword were chosen. Twelve different visual models were created for each keyword class by combining three different feature types and four different machine learning algorithms. The keywords modeled are: “Discus”, “Hammer”, “High Jump”, “Hurdles”, “Javelin”, “Long Jump”, “Run-

[1] The images were randomly selected from a large dataset collected in the framework of FP6 BOEMIE project.

Table 12.1: Low-level feature extraction.

Vector	***# Features***
HOG	225
SIFT	100
MPEG-7	186
SURF	64
SHiK	256

ning", and "Triple Jump". The performance and effectiveness of the created models are evaluated utilizing the accuracy of correctly classified instances.

12.1.1 Feature Extraction

Five different low-level feature types, HOG, SIFT, MPEG-7, SURF and SHiK were extracted from each image group and used to create the visual models. In the case of HOG, the implementation proposed in [229] was used with the aid of 25 rectangular cells and 9 bins histogram per cell. The 16 histograms with 9 bins were then concatenated to make a 225-dimensional feature vector. In the case of SIFT, the large number of extracted keypoints was quantized into a 100-dimensional feature vector using k-means clustering. Regarding the MPEG-7, after an extensive experimentation on MPEG-7 descriptors (for details see also [207]) the Color Layout(CL), Color Structure (CS), Edge Histogram (EH) and Homogenous Texture (HT) descriptors were chosen. The combination of the selected descriptors creates a 186-dimensional feature vector. Regarding the SURF features, a 64-dimensional feature vector was created [156]. Finally, a Hilbert space-filling curve of n=4 was utilized for SHiK feature extraction which results in a 256-dimensional vector. The dimensionalities of the extracted feature vectors are presented in Tab. 12.1.

12.1.2 Keywords Modelling

As already mentioned in the previous chapters, in order to ensure scalability and to fulfill the multiple keyword assignment per image, keyword models should be developed using the one-against-all training paradigm [196]. Thus, the creation of a visual model for each keyword was treated as a binary classification problem. The feature vectors of each keyword class were split into two groups: 80% were used for training models and the remaining 20% for testing the performance of these models. Positive examples were chosen from the corresponding keyword class while the negative ones were randomly taken from the seven remaining classes.

For the learning process five different algorithms were used: decision trees (in particular the Random Forest variation), induction rules (Ripper), Neural Networks (RBFNetwork), Support Vector Machines (SMO) and statistical classifiers (Naive Bayes). During training some parameters were optimized via experimentation in order to obtain the best performing model for each feature vector. The number of trees was optimally selected for the Random Forest models. The minimal weights of instances within a split and the number of rounds of optimization were examined for Ripper models. The number of clusters and ridge were tuned for each one of the feature vectors for the RBFNetwork. Finally, we experimented with the complexity constant and type of kernel for the SMO. Since most statistical classifiers do not require model selection and are estimated directly from the training data, Naive Bayes classifier was utilized without any experimentation.

12.1.3 Experimental Results

Fig. 12.1, 12.2, 12.3, 12.4, 12.5 show the accuracy of correctly classified instances per keyword class using the five different machine learning algorithms mentioned earlier. The results shown in these figures can be examined under three perspectives: First, in terms of the efficiency and effectiveness of the various learning algorithms in modelling

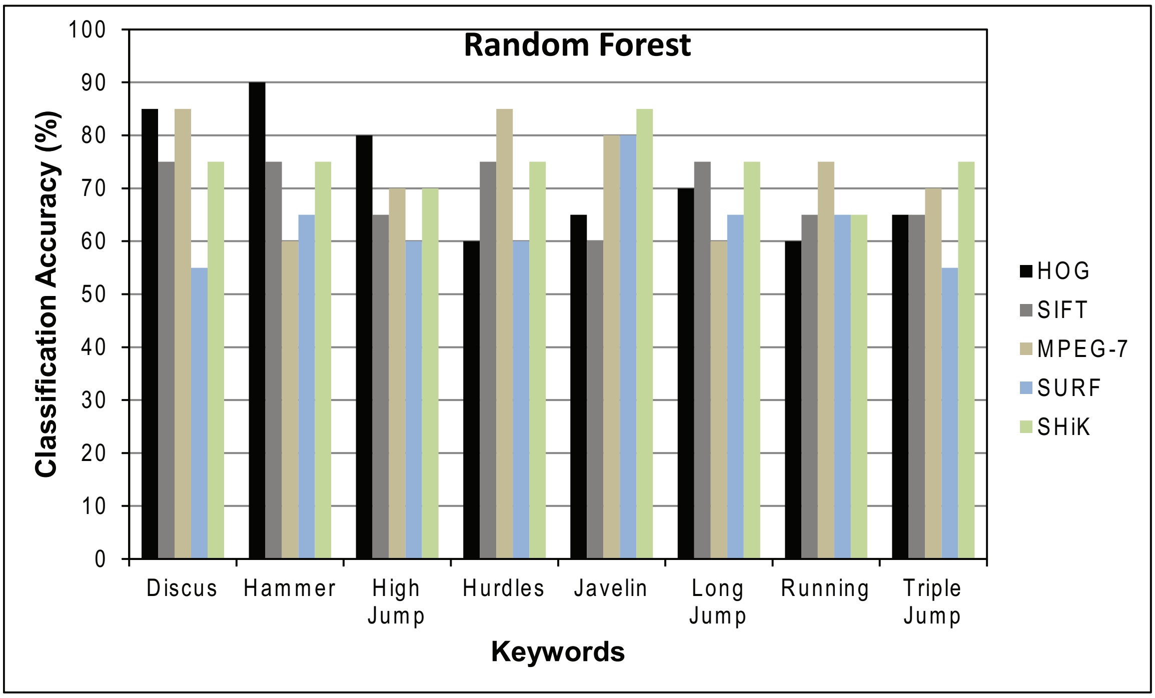


Figure 12.1: Evaluation performance of visual models using Random Forest decision tree.

Table 12.2: Average classification accuracy(%) values.

Classifier	HOG	SIFT	MPEG-7	SURF	SHiK	Overall
Random Forest	71,875	69,375	73,125	63,125	74,375	70,375
Ripper	65,625	58,125	72,5	56,875	65	63,625
RBFNetwork	75	64,375	75,625	56,875	72,5	68,875
SMO	72,5	69,375	81,25	61,25	81,875	73,25
Naive Bayes	71,25	69,375	76,25	61,25	70	69,625

crowdsourced keywords, second, in terms of the appropriateness of the low-level features to accurately describe the visual content of images in distinctive manner, and third, in terms of the ability of the created models to classify the images into the corresponding classes and assign to them the right keywords.

The performance of the learning algorithms is examined through the time required to train the models (efficiency), the robustness to the variation of learning parameters and the effectiveness of the created models to identify the correct keywords for unseen input images.

The learning takes no more than a few seconds for the majority of the keyword models for all the machine learning algorithms examined. Apart from Naive Bayes, the fluc-

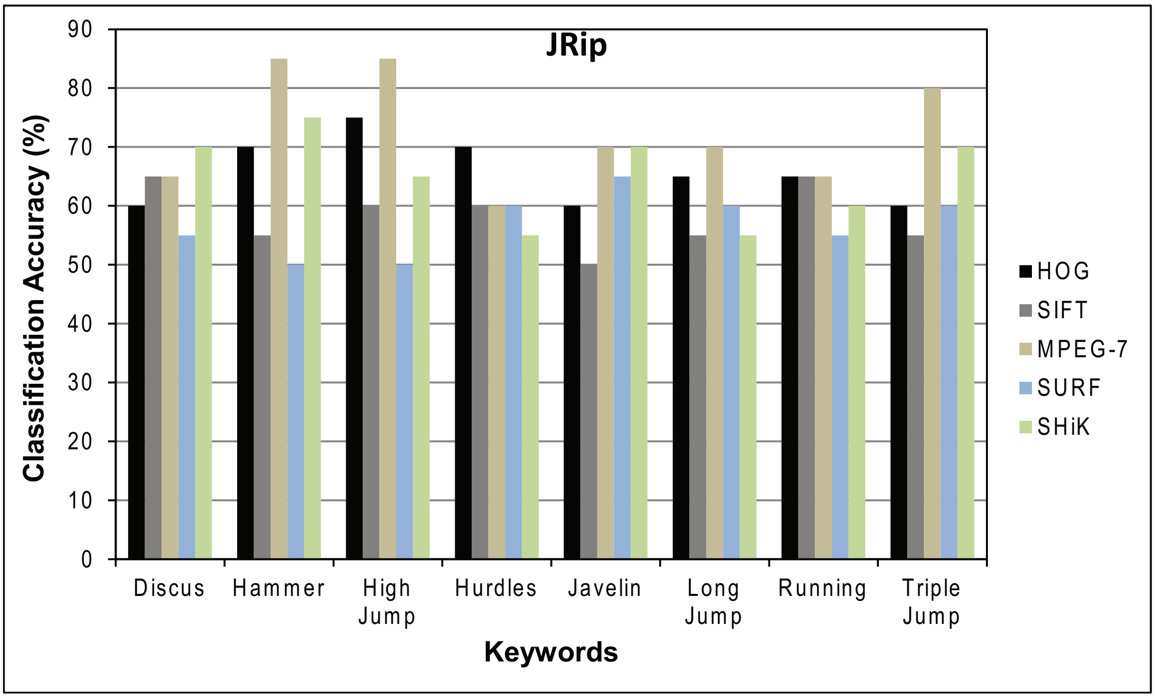


Figure 12.2: Evaluation performance of visual models using Ripper induction rule.

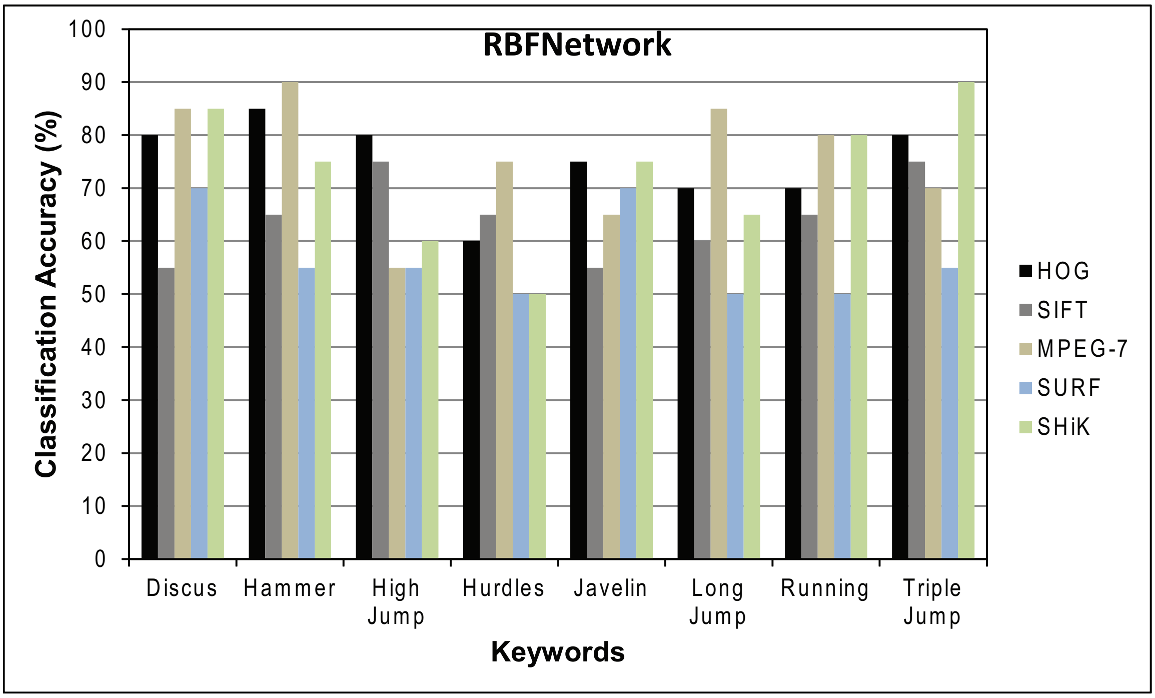


Figure 12.3: Evaluation performance of visual models using RBFNetwork.

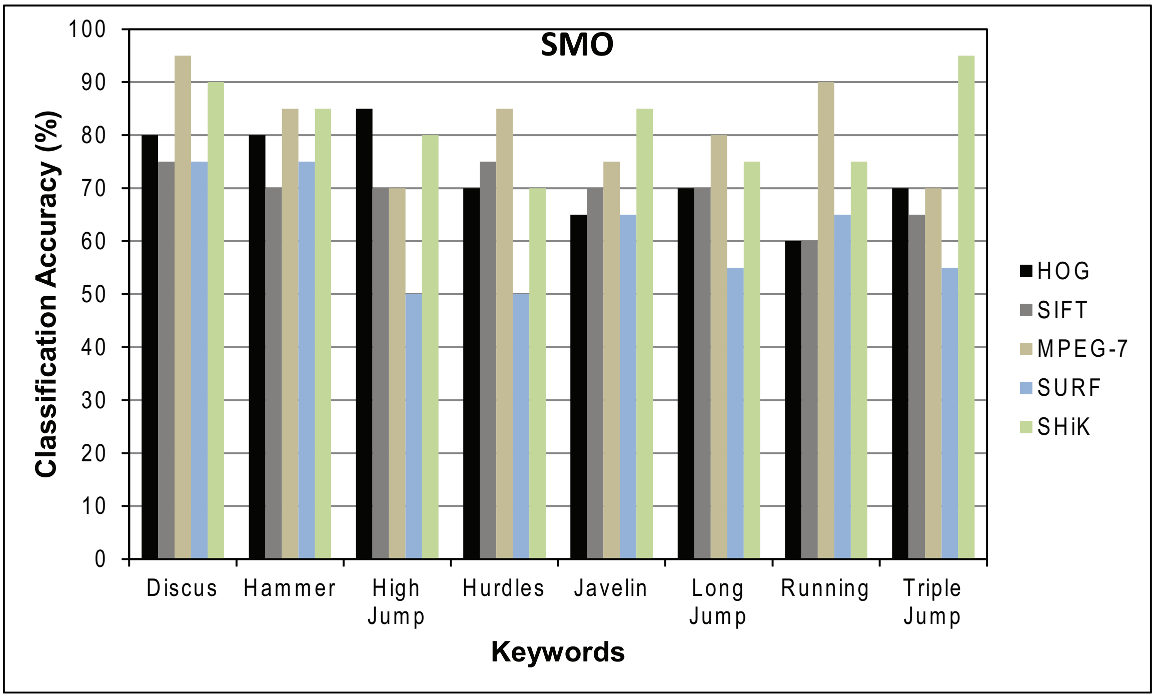


Figure 12.4: Evaluation performance of visual models using SMO support vector machine.

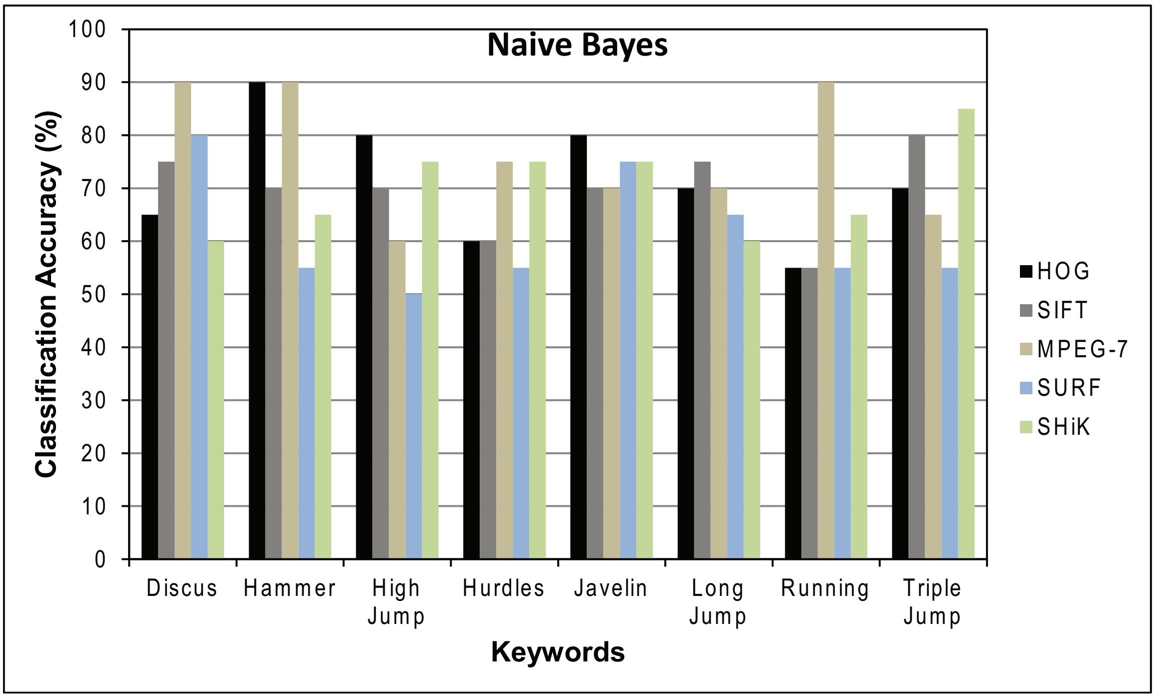


Figure 12.5: Evaluation performance of visual models using Naive Bayes.

tuation in classification performance during parameters tuning is significantly lower in Random Forest, Ripper and SMO than that of the RBFNetwork. This is something expected. As already discussed, tuning the activation function layer in RBF networks is a bit tricky and depends on the training data.

As far as the effectiveness is concerned, there is a significant difference on the performance of the models created using the individual learners. It is evident from Tab. 12.2 that SMO is the most reliable learner with total average classification accuracy equal to 73,25%. The Random Forest, Naive Bayes and RBFNetwork, obtain nearly the same average classification accuracy with the first performing better when fed with SHiK features and second and third when fed with MPEG-7 features. Furthermore, the Random Forest performs well for HOG, MPEG-7 and SIFT features and quite moderately for SURF features. The Naive Bayes obtains good average classification scores for the rest feature types (HOG, SIFT, SURF, SHiK). The RBFNetwork performs fairly well when using HOG and SHiK features, quite moderately for SIFT type features and poor when fed with SURF features. The Ripper inductive rule algorithm obtains the worst average classification accuracy score. The overall best classification accuracy occurs when combining the SMO classifier with SHiK features while the worst performance occurs when combining the Ripper and RBFNetwork with the SURF features. The majority, if not all, of the above results are in agreement with previous studies. SVM based algorithms perform well in learning tasks where the dimensionality of input space is high with respect to the number of training examples. Rule-based classifiers, on the other hand, face difficulties whenever the dimensionality of input space is high.

Concerning the effectiveness of the various low-level feature types, the experimental results indicate that the MPEG-7 features perform better than the others. The classification accuracy obtained using MPEG-7 is better than the others independently of the training algorithm used. The second more reliable low-level feature type for modeling keywords is the SHiK. SHiK features when combined with the SMO classifier achieve an average classification accuracy of 81,875% which is the highest among the others. HOG features achieve quite high average classification accuracies with the highest occurring when used with RBFNetwork classifier (75%). The average scores for SIFT features are very low but the worst performance is obtained by the SURF type features. It only achieves maximum classification accuracy over 70% for specific keyword classes. In particular, when combined with Random Forest decision tree it reaches an accuracy score 80% for the “Javelin” keyword class (see also Fig. 12.1) while when combined with SMO it reaches the score of 75% for “Discus” and “Hammer” keyword classes (see also Fig. 12.4). Additionally, obtained high scores for “Discus” and “Javelin” keyword classes equal to 80% and 75% respectively, when combined with Naive Bayes classifier. It is interesting to note that the best performance of the SURF features is obtained for keyword classes corresponding to objects with well-defined shape characteristics.

Once again, the results referring to the feature types are quite predictable. MPEG-7 descriptors are especially designed features to accommodate content based image retrieval. They were selected based on extended experimentation and comparative studies with other feature types. SIFT features on the other hand, were primarily defined for object detection and object modeling tasks. Furthermore, in an attempt to fix the dimensionality of input space so as to be used in machine learning frameworks, SIFT keypoints are grouped together using either histograms or clustering methods. This grouping discards the information about the spatial distribution of keypoints and deteriorates significantly their visual content description power. The high average classification scores obtained by SHiK features indicates the effective contribution of spatial information in image classification.

Nearly all models are able to assign the right keywords to unseen images. The overall accuracy scores are in the range 50%-95%. The best scores are obtained for keyword classes corresponding to objects with a well-defined shape such as “Discus” and “Hammer”. In contrary, keyword classes corresponding to more abstract terms, such as “Running” and “Triple Jump” achieve relatively poor scores in some cases. Thus, keywords that are related with the actual content of the images can be more easily modeled, and as a result, automatic annotation of input images with such keywords is both feasible and realistic. On the other hand, modeling keyword classes which are not clearly related with the image content is by far more difficult. This is because the content of images corresponding to these keywords has many similarities with the content of images corresponding to other keywords.

12.2 Creating Visual Models Using Fusion of Low-level Features

The fusion plays an important role when multiple features are used in classification process and can derive and gain the most effective and least dimensional feature vectors that benefit the final classification [172]. By fusing HOG, SIFT and MPEG-7 features, different visual models were created for the 8 crowdsourced keywords described above. For each keyword group, the several feature vectors are normalized and combined together into the feature union-vector whose dimension is 511 equal to the sum of the dimensions of the individual low-level feature vectors. PCA is applied to extract the linear features from the integrated union vector and reduce the dimensionality.

The maximum likelihood estimator (MLE) [173] is employed to estimate the intrinsic dimensionality of the fused feature vector by PCA. The fusion process is also applied on the single vectors separately and a second fused feature vector is created by summing the 3 resulted intrinsic dimensionalities. Tab. 12.3 summarizes the estimated intrinsic dimensionalities for the feature vectors used in the experiments: (1) PCA 65, which is created applying the PCA on the union-vector, and (2) PCA 133, which created by summing the results extracted after applying the PCA separately on the 3 single low-level feature vectors.

12.2.1 Experimental Results

Following the same keyword modelling process described above, different visual keyword models were created using the two fused feature vectors. The average classifi-

Table 12.3: Intrinsic dimensionality after applying PCA and MLE.

Vector	*Initial Dimensionality*	*Intrinsic Dimensionality*
HOG	225	60
SIFT	100	41
MPEG-7	186	32
PCA 65 *(HOG + SIFT + MPEG-7)*	511	65
PCA 133 *(60 HOG + 41 SIFT + 32 MPEG-7)*	-	133

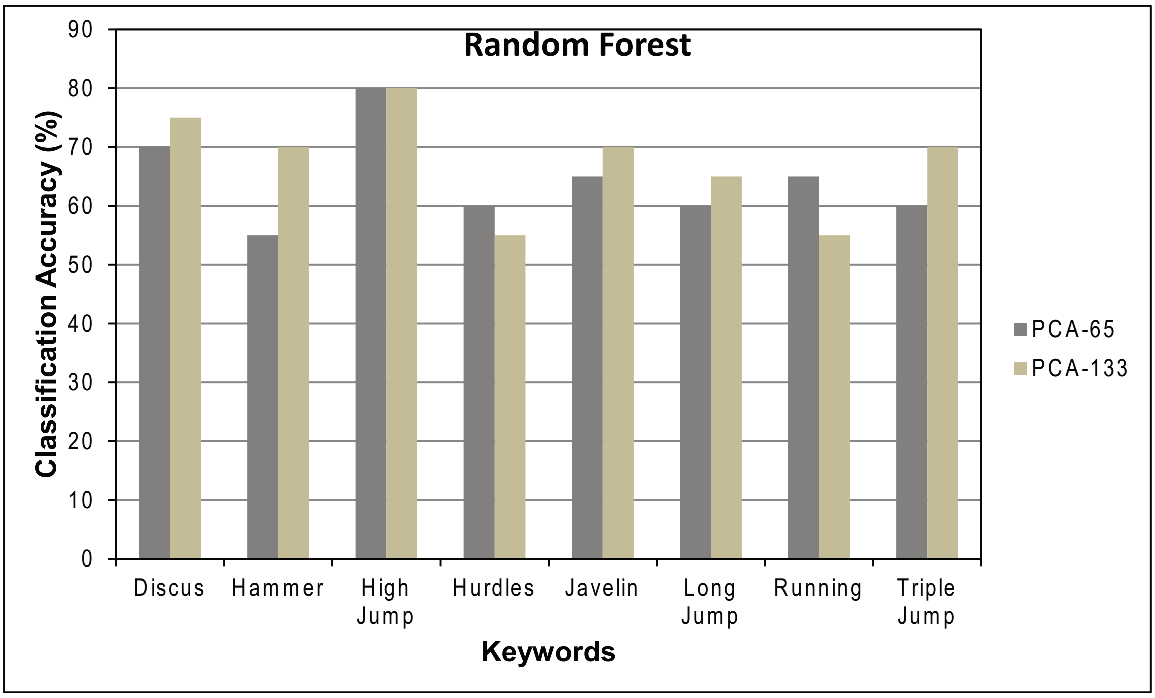


Figure 12.6: Evaluation performance of visual models using Random Forest decision tree.

cation accuracy scores are concentrated in Fig. 12.6, 12.7, 12.8, 12.9, 12.10 while the total average accuracy scores per classifier are presented in Tab. 12.4.

Concerning the fused feature vectors, Naive Bayes is the most reliable learner with total average classification accuracy equal to 73,125% when combined with PCA-65 and 79,375% when combined with PCA-133. Its worth to mention that the average accuracy score obtained for PCA-133 is the highest one among all visual models created either using single or fused feature vectors. The SMO classifier also performs very well for both feature vectors, obtaining average scores equal to 75% and 75,625% respectively. The average scores obtained by RBFNetwork are quite good (70,625% - 71,25%) while

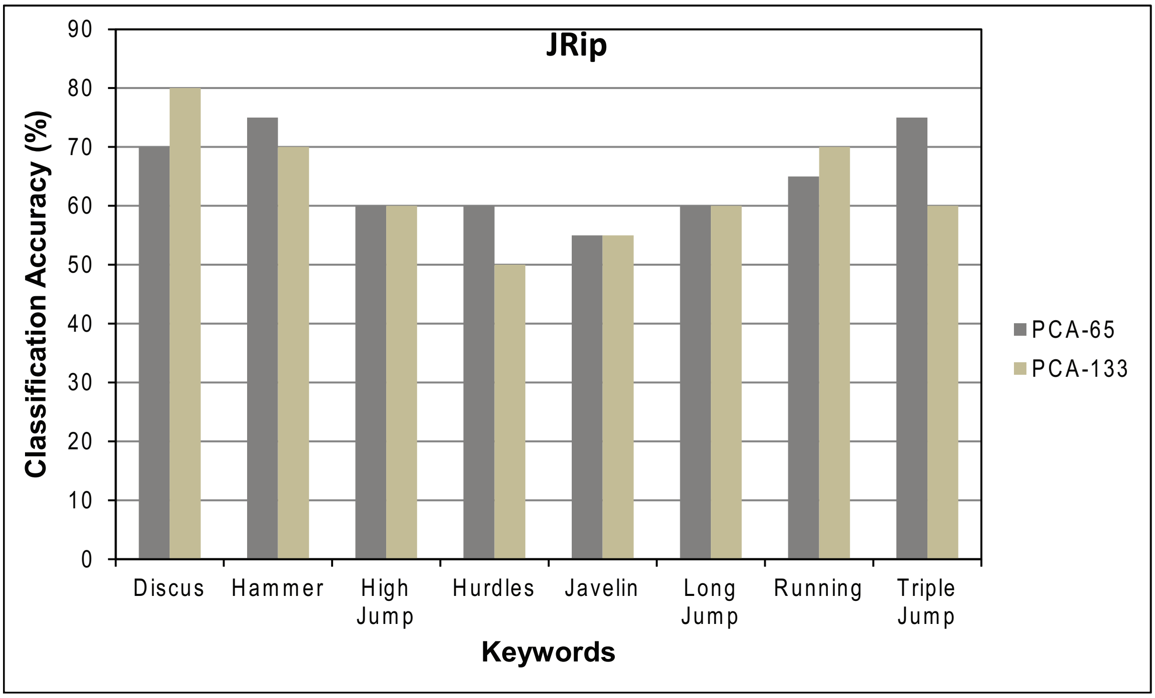


Figure 12.7: Evaluation performance of visual models using Ripper induction rule.

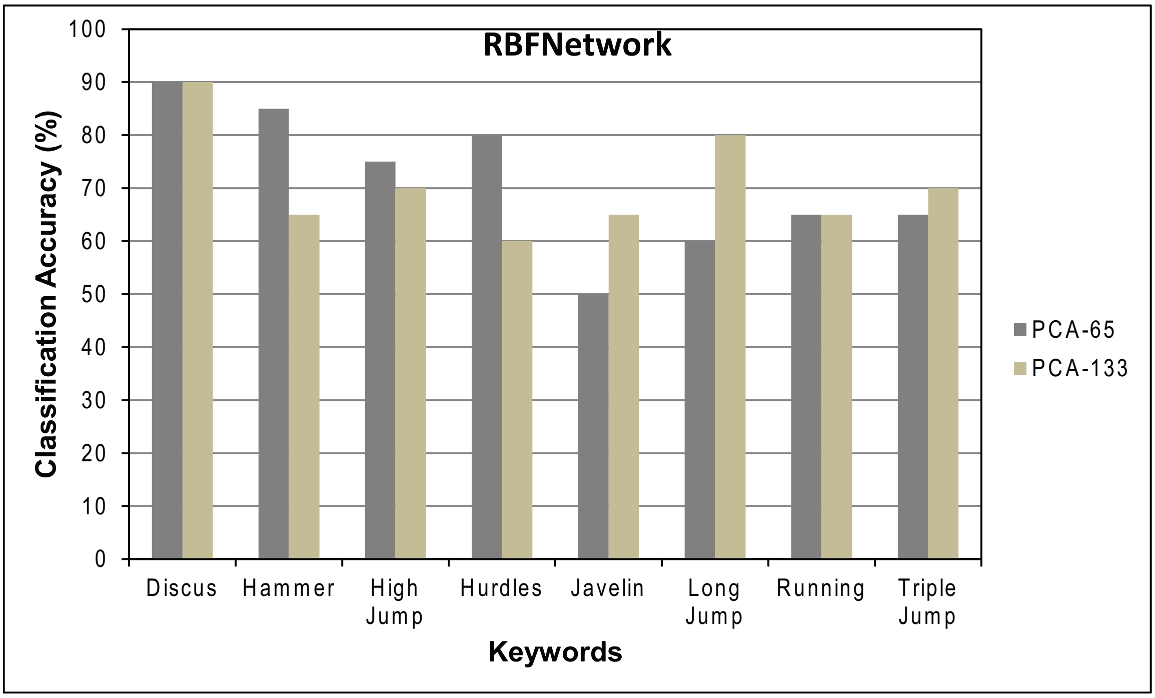


Figure 12.8: Evaluation performance of visual models using RBFNetwork.

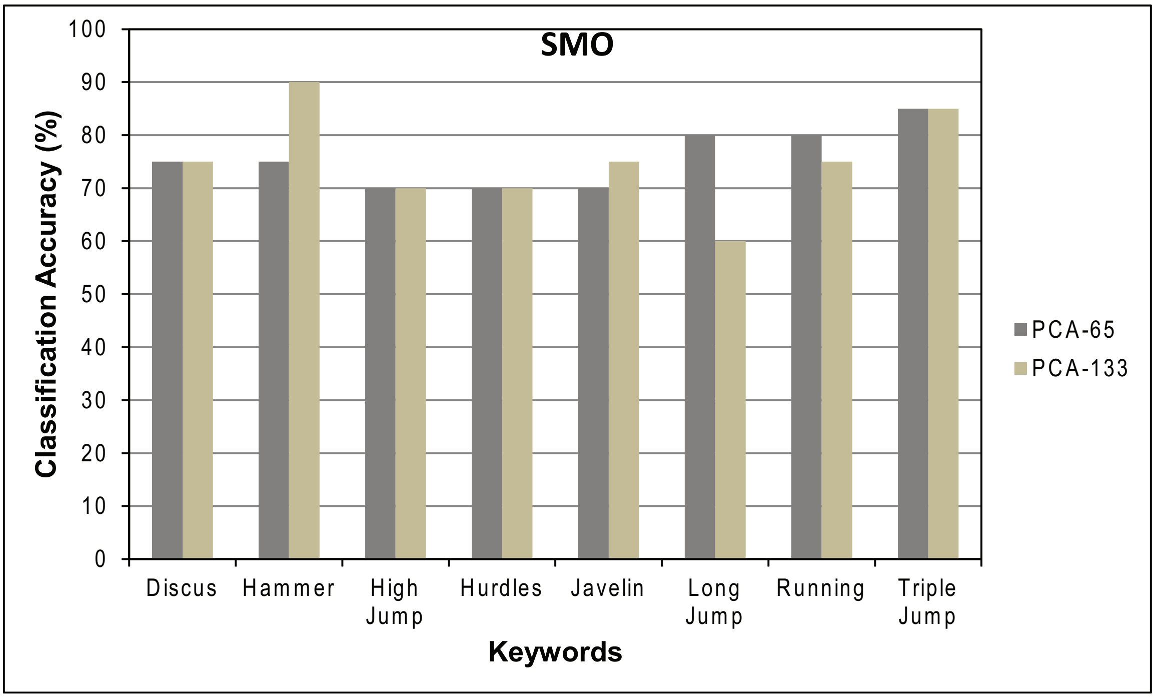


Figure 12.9: Evaluation performance of visual models using SMO support vector machine.

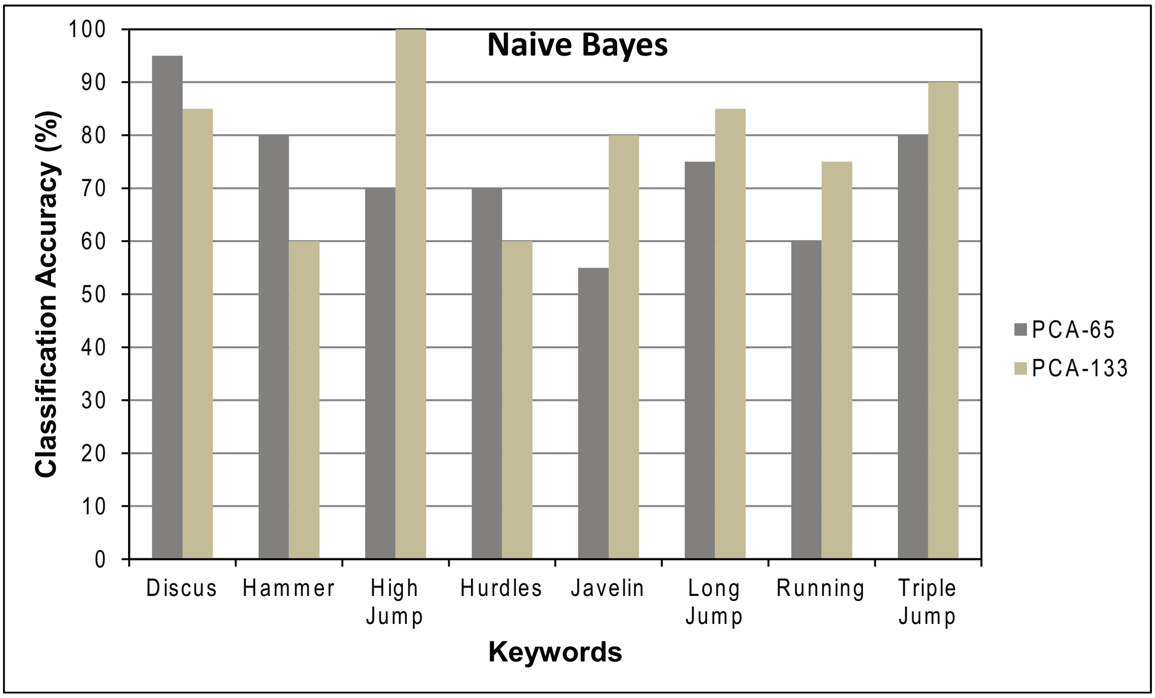


Figure 12.10: Evaluation performance of visual models using Naive Bayes.

Table 12.4: Average classification accuracy(%) values.

Classifier	PCA-65	PCA-133	Overall
Random Forest	64,375	67,5	65,9375
Ripper	65	63,125	64,0625
RBFNetwork	71,25	70,625	70,9375
SMO	75,625	75	75,3125
Naive Bayes	73,125	79,375	76,25

the ones obtained by Random Forest and Ripper are quite moderate and lower than 70%. The worst performance occurs when combining the Ripper with the PCA-133 feature vector.

The experimental results indicate that the PCA-133 feature vector performs slightly better than the PCA-65. The highest performance occurred when is combined with Naive Bayes classifier and the lowest when combined with Ripper with average accuracy scores in the range 63,125%-79,375%. In particular, when combined with Naive Bayes it reaches an accuracy score of 100% for the "High Jump" and 90% for "Triple Jump" keyword class (see also Fig. 12.6). Additionally, it reaches the score of 90% for "Discus" keyword class when combined with RBFNetwork (see also Fig. 12.8), and for "Hammer" keyword class when combined with SMO (see also Fig. 12.9). On the other hand, the PCA-65 obtained very high accuracy for "Discus" keyword class when combined with Naive Bayes and RBFNetwork classifiers with score values equal to 95% and 90% respectively.

The obtained high classification accuracy indicates the ability of the created models to successfully assign keywords to unseen images. The overall accuracy scores are in the range 50%-100%. In contrast to the models created using the single feature vectors, the models created using the fused vectors obtained high scores for keyword classes corresponding to either objects with a well-defined shape or not. In particular, the eight keyword classes can achieve very high classification scores in the range 80-100%. The results are in full agreement with previous studies that prove the benefit feature fusion when used in classification schemes.

12.3 Creating Visual Models Using Co-Training

In order to overcome the limitations introduced by the small amount of labeled data and boost the performance of the learners, a co-training algorithm was also used to create visual models. In particular, the CoTrade [224] algorithm was utilized to create visual models for the following crowdsourced keywords: “Discus”, “Hammer”, “High Jump”, “Hurdles”, “Javelin”, “Pole Vault”, “Long Jump”, “Running”, and “Triple Jump”.

12.3.1 Dataset Creation

A total number of 1040 images were used to create visual keywords models through co-training. The dataset consists of the 500 images used for the creation of visual models in the previous subsection, and of new 540 images that were collected and annotated in the same manner. In particular the new images were annotated through the TISAMUC portal[2], where 10 users annotate the dataset using 28 different keywords. Images annotated with the keywords, “Discus”, “Hammer”, “High Jump”, “Hurdles”, “Javelin”, “Pole Vault”, “Long Jump”, “Running”, and “Triple Jump” were selected to used for the experiments. A detailed description of the final dataset is presented in Tab. 12.5.

12.3.2 Experiments

Two different views of the image dataset were created and used for the experimental setup. HOG and SHiK features were extracted following the same process described in the previous subsection. Thereby, each image was represented by two feature vectors of 225 and 256- dimensionality respectively. The visual model for each one of the nine keywords was treated as a binary classification problem following the one-against-all

[2] http://cis.cut.ac.cy/tisamuc/.

Table 12.5: The image dataset used for the creation of visual models through co-training.

A/A	***Keyword Class***	***# of images***
1	Discus	149
2	Hammer	153
3	High Jump	158
4	Hurdles	50
5	Javelin	153
6	Long Jump	50
7	Running	50
8	Triple Jump	163
9	Pole Vault	114
Total		***1040***

training paradigm [196]. For each keyword, two classifiers were trained separately for each view using the Naive Bayes algorithm. For each view, 260 (25%) of the 1040 images were randomly selected as a test data. From the remaining 780 images, 1 positive and 3 negative examples were randomly selected to generate the labeled data. The remaining 776 images were used as unlabeled data for the co-training scheme. Using the labeled and unlabeled sets, the visual models were created through the CoTrade [224] algorithm. The process of model creation stopped when no more examples from the unlabelled data were available or when 50 training rounds were reached. Finally, the test set was used to evaluate the performance of the created models and the results are presented in Tab. 12.6.

Table 12.6: Evaluation performance of visual models using co-training.

Keyword	***Accuracy (%)***
Discus	0,873
Hammer	0,869
High Jump	0,865
Hurdles	0,958
Javelin	0,869
Long Jump	0,958
Running	0,958
Triple Jump	0,862
Pole Vault	0,904
Average	***0,902***

The models created through co-training are able to assign the right keywords to unseen images with high accuracy. The overall accuracy scores are in the range 0,862%-0,958%. The highest classification accuracy was obtained for keyword classes "Hurdles", "Long Jump" and "Running" and the lowest for "Triple Jump". The classification accuracy scores achieved by the current models are higher in comparison to the models created using single or fused feature vectors. The results verify the successful performance of co-training on classification tasks when only few labeled data are available [215] [218], and lead to the conclusion that a co-training algorithm can lead to accurate visual models.

12.4 Conclusions and Remarks

This chapter presents an overall evaluation of the idea of addressing the problem of automatic image annotation by creating visual models via low level features. In this framework, keyword models were created using several low level features and machine learning techniques. In particular, MPEG-7, SIFT, SURF, HOG and SHiK features were utilized to create visual models. HOG features were extracted using the algorithm developed by the Oswaldo Ludwig[3]. The extraction of SIFT features was achieved using the free software developed by Andrea Vedaldi[4]. The extraction of MPEG-7 features was based on a developed software using the MPEG-7 experimentation model[5]. SURF features were extracted using the free software developed by Chris Evans[6]. The SHiK features were extracted using the algorithm presented in Chapter 10[7]. For the comparison of the various machine learning techniques that have been utilized for the models creation, we have utilized the Weka open source software[8]. Finally, for the creation of visual models using the CoTrade algorithm, the free software provided by

[3] http://www.mathworks.com/matlabcentral/fileexchange/28689-hog-descriptor-for-matlab.

[4] http://www.robots.ox.ac.uk/ vedaldi/code/sift.html

[5] http://cis.cut.ac.cy/∼z.theodosiou/mpeg7.zip

[6] http://www.chrisevansdev.com/

[7] http://cis.cut.ac.cy/ z.theodosiou/shik.zip.

[8] http://www.cs.waikato.ac.nz/ml/weka/.

the authors[9] was utilized.

[9] http://cse.seu.edu.cn/people/zhangml/Resources.htm.

Chapter 13

Conclusions and Future Work

This chapter summarizes the main conclusions of the current thesis and gives some future research directions.

13.1 Summary and Conclusions

The enormous increase of the available digital images generated the need of automatic image annotation to help search engines to better retrieve desired images in response to text queries. Since manual annotation is a difficult and costly task, significant interest has been generated in relating high-level human interpretations with low-level visual features. This thesis concentrates on image retrieval under the perspective of modelling keywords using low-level features. In this framework, several studies related to dataset creation, low-level feature extraction and modelling techniques have been conducted along with the basic conceptualization of modelling keywords with low level features, which is the contribution of this thesis to image retrieval field. Visual modelling of keywords allows for automatic and is scalable since a separate model is created for each available keyword. The proposed scheme settles the limitations of previous studies on automatic image annotation such as the restricted scalability and the inability of assigning more than one keyword to each image. A more detailed description of the conducted work is follows.

In chapter 2 we have presented and discussed the basic theory and formulized the problem of automatic image annotation. First, the framework of image retrieval is outlined and then the basic steps are studied. In particular, the content-based image retrieval, the text-based image retrieval focusing mainly on web-based approaches, the task of image annotation and automatic image annotation methods are examined. The drawbacks of the existing approaches are extracted and led to the conclusion that a proper image annotation may contain more than one keyword that is relevant to the image content, so a reclassification process is required in this case, as well as whenever a new keyword class is added to the classification scheme.

The idea of creating separate visual models for all keyword classes adds a significant value in automatic image annotation. More than one keyword can be assigned to the input image and new keyword classes can be added into annotation scheme without reclassification. The idea of modelling keywords via low level features as an attempt to overcome the limitations of the existing automatic image annotation methods is detailed in chapter 3. The key issues of the proposed idea, as well as the directions which the current thesis followed to overcome them are also outlined in the same chapter.

Availability of training data is a basic prerequisite for creating accurate visual models for keywords. Training examples that are used for creating visual models for keywords are pairs of images and keywords and their collection incurs various issues like the large amount of manual effort and time required in developing the training data, the differences in interpretation of image contents, and the inconsistency of the keyword assignments among different annotators. An exhausting review on alternative methods for creating training data using the least manual effort is presented in chapter 4. Acquiring annotations through crowdsourcing or extracting keywords from the surrounding text in the purpose of web images seem to be attractive solutions to the problem of collecting manual annotations.

The image annotation is a social cognitive process and may vary among people based on their socio-demographic characteristics. Since, the raw image data can not readily transferred to high-level semantics, image annotation is dependent on both humans and image content. Chapter 5 presents a study which focuses on interpretation differences appeared in image annotation and examines whether these differences are related to the demographic factors such as age and gender. In other words, this study examines how the people of different age and gender interpret the meaning of an image and how they describe its content by assigning vocabulary and free keywords. The vocabulary keywords were categorized into five main topic categories, while the given free keywords were categorized into the cognitive categories of nouns using the Prototype Theory. The study also aimed at looking at the gender and age differences in inter-annotator agreement using the vocabulary keywords. The results of this study reveal that there are age differences in the way that people annotate images using both vocabulary and free keywords. The gender gap is smaller than the initial assumptions with significant difference only in the way that women use vocabulary keywords related to the “Feeling” category. Concerning the inter-annotator agreement, there is an adequate agreement among the gender and age groups with the highest agreement occurred in the use of “Location” category.

Chapter 6 presents an experimental framework to investigate how various factors like the content, lexicon and annotation method affect the crowdsourcing-wise annotations. The experimental setup was based on image dataset which was annotated by several annotators using vocabulary, hierarchical vocabulary and free keywords. The experimental results show that the hierarchical vocabulary keywords lead to more consistent annotation with normalized difference between abstract and specific images. Although the hierarchical structure reduces the inter-annotator agreement on keyword basis, it gives high accuracy agreement between expert and non-experts. The frequent use of free keywords implies first the inability of non-expert annotators to fully understand the meaning of vocabulary keywords and, second the inability of the selected vocabulary keywords to cover the content of the image dataset. However, the mixture of the three different annotations approaches can achieve competitive results. The results obtained in these experiments are quite promising and show that researchers can outsource image annotation to an Internet crowd or use tags created through social platforms without compromising the quality of the results and at the same time achieve wider participant diversity. The valuable set of annotated images can lead to effective information retrieval related to several research purposes in the field of crowdsourcing annotation and image retrieval. Furthermore, the findings can lead to better understanding of the factors that may affect the quality of the annotations coming from the social media platforms.

Chapter 7 presents a new web image indexing framework. In the proposed method, the whole text that is found in the web page is used as a source to extract content information for the web images that exist in the same web document. The structural text blocks of the web page are extracted and assigned to images after their semantic representation which is achieved using language models. This semantic representation ensures that the text blocks assigned to a single are semantically uniform; in other words they share similar content. The experimental results are encouraging and indicate that automatic web image indexing can be used for collecting training data.

Low-level feature extraction is the first crucial step in the keyword modeling process aiming at capturing the important characteristics of the visual content of images. Chapter 8 presents and discusses several algorithms for low-level feature extraction. The research in this domain is rich and several methods have been proposed. Although both local and global feature sets are used for image retrieval purposes, global features are a natural choice for image retrieval that is based on machine learning. Since they are extracted from the image as whole they are also appropriate for creating visual models for keywords. MPEG-7 descriptors perform excellent within the machine learning paradigm used either in classification based keyword extraction or in keyword modeling. Chapter 9 evaluates the performance of the MPEG-7 descriptors in keyword extraction. Experimental results show that there is a significant variation on the performance of various descriptors and combination of descriptors increase the classification performance.

SIFT features appear to be a state-of-art reliable feature representation sufficiently used in object recognition and image retrieval. However, the non-fixed and huge dimensionality of the extracted SIFT feature vector cause certain limitations when it is used in machine learning frameworks. An attempt was made to overcome the limitations of the SIFT feature vector when it is used in image classification tasks. Chapter 10 presents a new feature extraction algorithm, the Spatial Histogram of Keypoints, which maintains the spatial information of the SIFT keypoints and results in a feature vector with fixed and low dimensionality. The algorithm localizes the keypoints by utilizing the first two steps of the SIFT algorithm and then partitions the image into ordered sub-regions based on the Hilbert geometry. The proposed method, it has shown very promising results on three diverse datasets. The algorithm was also compared with one of the most reliable feature extraction methods which also maintains the spatial information, the Spatial Pyramid Matching method, and shows significantly better performance in scenes datasets. The algorithm performs better in scene rather than object recognition but it has to be further examined such as the majority of images coming from object datasets, contain accumulated spatial information in the center and scattered information on the borders.

The final step for the creation of visual models is the use of appropriate learning method for organizing instances into classes by analyzing the properties of the supplied image visual features. Chapter 11 explores several supervised methods that have been utilized for classifying images into class labels as well as for keyword model creation, and chapter 12 presents an overall evaluation of the idea of creating visual models. In this manner, different features and machine learning algorithms were compared in creating visual models for crowdsourcing originated keywords.

Different keywords within the athletic domain were modeled using various low-level features and data classifiers. Visual models were created using single and fused feature vectors. Concerning the single feature vectors, nearly all created models can classify the images into the keyword classes with medium to high classification accuracy. The best scores are obtained for keyword classes corresponding to objects with a well-defined shape. Although there is a significant variation on the efficiency of the various classifiers with the SMO having the highest performance, a great improvement can achieved when the SHiK features are used. The high average classification scores obtained by SHiK features indicate the effective contribution of spatial information in image classification. In contrast to the models created using the single feature vectors, the models created using the fused vectors obtained high scores for keyword classes corresponding to either objects with a well-defined or not.

A co-training algorithm was also used to create visual models in an effort to boost the performance of the classifiers when a small amount of training data is available. Visual models for several keywords within the athletics domain were created using two different feature sets as the two different views of the dataset. The models created through co-training are able to assign the right keywords to unseen images with high accuracy. The classification accuracy scores achieved by the current models are higher in comparison to the models created using single or fused feature vectors. The findings of this chapter can serve as a guide for researchers who want to experiment with automatic keyword assignment to digital images.

13.2 Future Work

This thesis explores the image retrieval under the perspective of modeling keywords using low-level features. Different aspects of this area are explored and new ideas and techniques are proposed. This chapter highlights some avenues for possible future research based on the findings of this study.

13.2.1 Dataset Creation

The crowdsourcing seems to be an attractive solution for collecting annotations in a cheaply and quickly manner. In the framework of this work, a new web platform was created for collecting crowdsourcing-wise annotations using vocabulary, hierarchical lexicon and free keywords. Although, the experimental results indicate that the quality of such annotations is quite satisfactory some future work could definitely explore further this framework. Questions about the interface of the annotation platform as well as the annotation task can be given to the users in order to evaluate and improve the method. The frequent use of free keywords indicates the inability of non-experts users to describe image content with vocabulary or hierarchical keywords. The enhancement of the predefined keywords with the free keywords that suggested frequently by annotators, improves the overall annotation quality. Additionally, the annotation platform can be expanded to store more information during the annotation process like the date and time when a keyword is suggested for a specific image. The new data can lead to further measures for annotators consistency since the keywords suggested first may have more liabilities to be relevant to the image content unlike the keywords suggested later.

The annotation quality can further be improved if some problems derived from crowdsourcing are addressed. Since the annotations are collected without supervision, the quality should be ensured through diverse methods by filtering those annotations given randomly or by a mistake. Furthermore, the motivation plays significant role in crowdsourcing schemes and an investigation of how the people can be motivated to work for the annotation task (money or other incentives), could definitely add further improvements into the whole procedure.

13.2.2 Image Interpretation

The subjectivity on image interpretation by humans deserves also more study. Starting by the interpretative system proposed by Panofsky [3] for subject matter in visual art, several studies have been conducted to investigate the diversity in extracting the meaning of visual data. More recent studies state that the main social and cultural changes in the world influence the meaning that people give to images [10], and emphasize the dynamic relation between words and images which is linked to wider social and cultural issues [9]. The experiments conducted in the framework of the current thesis indicate the significant age differences and negligible gender differences in manual image annotation. The study in gender and age gaps in image annotation along with the image search will also may lead to more important insights. The effects of more demographics variables like education, job, sexuality, religion and ethnicity, should also be examined on the way that people annotate images. Furthermore, a study on how people annotate images related to different domains can generalize the whole process.

Research interest also exists on what exactly set an image to have a specific or abstract meaning by humans. In this manner, a new algorithm can be developed using manually created training examples aiming to classify automatically an unseen image into one of these two categories. Such a method will definitely have a wide range of applications in different domains related to visual content and make the process of selecting the suitable visual data simpler.

13.2.3 Low-level Feature extraction

The extraction of low-level feature is an essential step in several applications in the domain of image processing and computer vision. The SHiK algorithm is proposed to overcome the limitations derived from the high and non-fixed dimensionality of SIFT features when used in machine learning frameworks. The experimental results indicate the importance of the spatial information preserved by utilizing the Hilbert Fractal which improves significantly the visual content description power. In this manner, different ways for concentrating the localized keypoints with the related spatial information can also be examined in order to ensure the repeatability, distinctiveness, locality, quantity, accuracy and efficiency of the new local features. Furthermore, the orientation information extracted by SIFT features can also be kept by the new features without extending or mislaying the fixed and low-dimensionality of the SHiK feature vector.

In addition, the algorithm could also to be examined in larger and different datasets as well as with different classification schemes and matching approaches. The very promising results of SHiK algorithm when used on face datasets encourages for further study and experimentation on face recognition and verification. Finally, the use of Hilbert curves allows a multiresolution representation in case where space-filling curves of increasing order are combined together. Thereby, the development of multiresolution SHiK algorithm can also be a future task.

13.2.4 Visual models creation using machine learning techniques

The accurate creation of visual models requires further experimentation of additional training algorithms and other classifications schemes. In addition, the efficiency of more low-level features in creation of visual models could be investigated. The performance of created visual models using various feature vectors and machine learning techniques varies based on: (a) the visual content of the image, (b) the visual content description power of the low-level features, and (c) on the discriminative ability of the machine learning technique. Keeping this in mind, an ensemble of several classifiers appears to be a good solution for obtaining high accuracy for all available keywords. Furthermore, the high accuracy values occurred when the co-training algorithm was applied generates a special need to investigate in-depth the performance as of co-training in creating visual models for keywords created through crowdsourcing. Special attention should be given on the maintenance of a large disagreement between the base learners and on the accurate measurement of the labeling confidence on the unlabelled examples.

13.2.5 Modelling keywords via low level features

The enormous amount of available images which do not appear in web-pages or they are not manually annotated, creates the need for efficient and effective automatic image annotation. Although idea of automatic image annotation by keywords modelling is very promising, there is a long way to go before this method can be utilized to solve the problem of automatic image annotation. Undoubtedly, a further evaluation on larger and different datasets is essential to the generalization of the proposed idea in the field. Experimental results showed that the accuracy of the proposed idea depends largely on image content. Therefore, the widest range of different datasets used, the better conclusions can be achieved.

Future perspectives also involve the experimentation of the proposed framework in annotating different types of multimedia, such as video and 3D images. The number of available video clips has been steadily growing with the advent of social media and self-broadcasting online services like YouTube. As a result, intelligent automated annotation tools are essential for accurate searching, indexing and retrieval purposes. By segmenting video into shots and determining the keyframes for each shot, the proposed framework for automatic image annotation can be applied without any modification.

On the other hand, 3D images are increasingly used for a wide range of tasks in computer vision (ex. 3D cinema, television, gaming, mobile video, modeling tools, etc.). The 3D image retrieval has become a hot topic of interest and special attention given to text-based retrieval methods. A variety of automatic and semi-automatic 3D image annotation methods has been proposed. Thus, the experimentation of the proposed framework on 3D image datasets is definitely of utmost importance.

Appendix A

List of Publications

The following papers have been accepted for publication or under review as a direct or indirect result of the research discussed in this thesis.

Journal Papers

Z. Theodosiou, N. Tsapatsoulis, “The Importance of Content and Lexicon in Manual Image Annotation”, (submitted to the IEEE Transactions on Cybernetics).

Z. Theodosiou, N. Tsapatsoulis, “Age and Gender Differences in Manual Image Annotation”, (submitted to the Cognitive Computation).

Conference Papers

Z. Theodosiou, Nicolas Tsapatsoulis, “Spatial Histogram of Keypoints (SHiK)”, in Proceedings of the IEEE International Conference on Image Processing, pp. 2924-2928, Melbourne Australia, September 2013.

Z. Theodosiou, Nicolas Tsapatsoulis, “Semantic Gap Between People: An Experimental Investigation based on Image Annotation”, in Proceedings of the 7th International Workshop on Semantic Media Adaptation and Personalization, pp. 73-77, Luxembourg, December 2012.

G. Tryfou, **Z. Theodosiou**, N. Tsapatsoulis, “Web image context extraction based on semantic representation of web page visual segments", in Proceedings of the 7th International Workshop on Semantic Media Adaptation and Personalization, pp. 63-67, Luxembourg, December 2012.

Z. Theodosiou, N. Tsapatsoulis, "Modelling Crowdsourcing Originated Keywords within the Athletics Domain", Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology, Vol. 381, pp. 404-413, 2012.

Z. Theodosiou, N. Tsapatsoulis, "On the Creation of Visual Models for Keywords through Crowdsourcing", Recent Researches in Applications of Electrical and Computer Engineering, pp.96-100, Athens Greece, March 2012.

Z. Theodosiou, O. Georgiou and N. Tsapatsoulis, "Evaluating annotators consistency with the aid of an innovative database schema", in Proceedings of the 6th International Workshop on Semantic Media Adaptation and Personalization, pp. 74-78, Vigo, Spain, December 2011.

Z. Theodosiou, N. Tsapatsoulis, "Crowdsourcing Annotation: Modelling keywords using low level features", in Proceedings of the 5th International Conference on Internet Multimedia Systems Architecture and Application, Bangalore, India, December 2011.

Z. Theodosiou, N. Tsapatsoulis, S. Bujak-Pietrek, I. Szadkowska-Stanczyk, "Airborne asbestos fibers detection in microscope images using re?initialization free active contours", in Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.4785-4788, Buenos Aires, Argentina, September 2010.

N. Tsapatsoulis, **Z. Theodosiou**, "Object Classification using the MPEG-7 visual descriptors: An experimental evaluation using state of the art data classifiers", in Proceedings of the 19th International Conference on Artificial Neural Networks, Limassol, Cyprus, September 2009.

Z. Theodosiou, A. Kounoudes, N. Tsapatsoulis, M. Milis, "MuLVAT: A Video Annotation Tool based on XML-dictionaries and Shot Clustering”, in Proceedings of the 19th International Conference on Artificial Neural Networks, Limassol, Cyprus, September 2009.

Book Chapters

Z. Theodosiou, N. Tsapatsoulis, “Image Retrieval Using Keywords: The Machine Learning Perspective”, in Semantic Multimedia Analysis and Processing, Ed. By E. Spyrou, D. Iakovides, P. Mylonas, CRC Press / Taylor & Francis (May 2014).

Bibliography

[1] A. Hanbury. **A survey of methods for image annotation**. *Journal of Visual Languages & Computing*, **19**(5):617–627, 2008.

[2] D. Zhang, M. M. Islam, and G. Lu. **A review on automatic image annotation techniques**. *Pattern Recognition*, **45**:346–362, 2012.

[3] E. Panofsky. *Meaning in the visual arts*. Garden City, NY: Doubleday, New York, NY, USA, 1955.

[4] K. Markey. **Computer-assisted construction of a thematic catalog of primary and secondary subject matter**. *Visual Resources*, **3**(1):16–49, 1983.

[5] K. Markey. **Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials**. *An International Journal on Library and Information Science Research*, **6**(2):155–177, 1984.

[6] S. Shatford. **Analyzing the subject of a picture: A theoretical approach**. *Cataloging & Classification Quarterly*, **6**(3):39–62, 1986.

[7] M. G. Krause. **Intellectual problems of indexing picture collections**. *Audiovisual Librarian*, **14**(2):73–81, 1988.

[8] L. S. Shatford. **Some issues in the indexing of images**. *Journal of the American Society for Information Science*, **45**(8):583–588, 1994.

[9] W. J. T Mitchell. *What do pictures want? The lives and loves of images.* University Of Chicago Press, Chicago, IL, USA, 2005.

[10] P. Holland. **'Sweet it is to scan': personal photographs and popular photography. 2nd ed.** In L. Wells, editor, *Photography: a critical introduction*, pages 117–164. Routledge, London, UK, 2000.

[11] D. Ward. *Photography of advertising.* Little, Brown and Company, New York, NY, USA, 1990.

[12] S. Hall. **Encoding, Decoding**. In S. During, editor, *The cultural studies reader*, pages 90–103. Routledge, London, UK, 1993.

[13] M. Price. *The photograph: A strange, confined space.* Stanford University Press, Stanford, CA, USA, 1997.

[14] P. G. B. Enser. **Pictorial information retrieval**. *Journal of Documentation*, **51**(2):126–170, 1995.

[15] K. Rodden and K. R. Wood. **How do people manage their digital photographs?** In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 409–416, 2003.

[16] B. Stvilia, C. Jörgensen, and S. Wu. **Establishing the value of socially-created metadata to image indexing**. *Library & Information Science Research*, **34**(2):99–109, 2012.

[17] J. McConnaughey, D. Everett, T. Reynolds, and W. Ladner. **Falling through the Net: Defining the digital divide**. Technical report, National Telecommunications and Information Administration, Washington DC, USA, 1999.

[18] H. Ono and M. Zavodny. **Gender and the Internet***. *Social Science Quarterly*, **84**(1):111–121, 2003.

[19] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G Gay. **The influence of task and gender on search and evaluation behavior using Google**. *Information Processing & Management*, **42**(4):1123–1131, 2006.

[20] S. Park. **Concentration of internet usage and its relation to exposure to negative content: Does the gender gap differ among adults and adolescents?** *Women's Studies International Forum*, **32**(2):98–107, 2009.

[21] A. J. Stronge, W. A. Rogers, and A. D. Fisk. **Web-based information search and retrieval: Effects of strategy use and age on search success**. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **48**(3):434–446, 2006.

[22] N.-S. Chang and K. sun Fu. **Query-by-pictorial-example**. *IEEE Transactions on Software Engineering*, **6**(6):519–524, 1980.

[23] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. **Content-based image retrieval at the end of the early years**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(2):1349–1380, 2000.

[24] Y. Rui and T. S. Huang. **Image retrieval: Current techniques, promising directions and open issues**. *Journal of Visual Communication and Image Representation*, **10**:39–62, 1999.

[25] J. Li and J.Z. Wang. **Real-time computerized annotation of pictures**. In *Proceedings of the ACM Multimedia Conference*, pages 911–920, 2006.

[26] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. **QBIC project: querying images by content, using color, texture, and shape**. In *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, pages 173–187, 1993.

[27] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. **Efficient color histogram indexing for quadratic form distance functions**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(7):729–736, 1995.

[28] N. R. Pal and S. K. Pal. **A review on image segmentation techniques**. *Pattern Recognition*, **26**(9):1277–1294, 1993.

[29] Y. A. Aslandogan and C. T. Yu. **Techniques and systems for image and video retrieval**. *IEEE Transactions on Knowledge and Data Engineering*, **11**(1):56–63, 1999.

[30] Y. Mori, H Takahashi, and R. Oka. **Image-to-word transformation based on dividing and vector quantizing images with words**. In *Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*, pages 405–409, 1999.

[31] C. Cusano, G. Ciocca, and R. Schettini. **Image annotation using SVM**. In *Proceedings of Internet Imaging V*, **SPIE 5304**, pages 330–338, 2003.

[32] S. L. Feng, R. Manmatha, and V. Lavrenko. **Multiple Bernoulli relevance models for image and video annotation**. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, 2004.

[33] G. Arellano, L. E. Sucar, and E. Morales. **Automatic image annotation using multiple grid segmentation**. In G. Sidorov, A. Hernndez Aguirre, and C. A. Reyes Garca, editors, *Advances in artificial intelligence*, **6437** of *Lecture Notes in Computer Science*, pages 278–289. Springer Berlin Heidelberg, 2010.

[34] J. Z. Wang, J. Li, and G. Wiederhold. **SIMPLIcity: semantics-sensitive integrated matching for picture libraries**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(9):947–963, 2001.

[35] K. Kuroda and M. Hagiwara. **An image retrieval system by impression words and specific object names-IRIS**. *Neurocomputing*, **43**(14):259–276, 2002.

[36] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. **An ontology approach to object-based image retrieval**. In *IEEE International Conference on Image Processing*, **2**, pages 511–514, 2003.

[37] T.F. Chan and L.A. Vese. **Active contours without edges**. *IEEE Transactions on Image Processing*, **10**(2):266–277, 2001.

[38] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. **Image segmentation using active contours: Calculus of variations or shape gradients?** *SIAM Journal on Applied Mathematics*, **63**(6):2128–2154, 2003.

[39] C. Carson, S. Belongie, H. Greenspan, and J. Malik. **Blobworld: image segmentation using expectation-maximization and its application to image querying**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(8):1026–1038, 2002.

[40] H. Feng and T.-S. Chua. **A bootstrapping approach to annotating large image collection**. In *5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 55–62, 2003.

[41] R. Shi, H. Feng, T.-S. Chua, and C.-H. Lee. **An adaptive image content representation and segmentation approach to automatic image annotation**. In P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, editors, *Image and video retrieval*, **3115** of *Lecture Notes in Computer Science*, pages 545–554. Springer Berlin Heidelberg, 2004.

[42] J. Shi and J. Malik. **Normalized cuts and image segmentation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8):888–905, 2000.

[43] W. Tao, H. Jin, and Y. Zhang. **Color image segmentation based on mean shift and normalized cuts**. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **37**(5):1382–1389, 2007.

[44] Y. Deng and B.S. Manjunath. **Unsupervised segmentation of color-texture regions in images and video**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **23**(8):800–810, 2001.

[45] Y. Liu, D. Zhang, and G. Lu. **Region-based image retrieval with high-level semantics using decision tree learning**. *Pattern Recognition*, **41**(8):2554–2570, 2008.

[46] M.M. Islam, Dengsheng Zhang, and Guojun Lu. **Automatic categorization of image regions using dominant color based vector quantization**. In *Digital image computing: Techniques and applications (DICTA)*, pages 191–198, 2008.

[47] I. K. Sethi, I. L. Coman, and D. Stan. **Mining association rules between low-level image features and high-level concepts**, 2001.

[48] A. Mojsilovic and B. Rogowitz. **Capturing image semantics with low-level descriptors**. In *IEEE International Conference in Image Processing*, pages 18–21, Thessaloniki, Greece, 2001.

[49] X. S. Zhou and T. S. Huang. **CBIR: From low-level features to high-level semantics**. In *Proc. SPIE Image and Video Communication and Processing*, pages 24–28, 2000.

[50] R. Jaimes, M. Christel, S. Gilles, R. Sarukkai, and W.-Y. Ma. **Multimedia information retrieval: What is it, and why isn't anyone using it?** In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 3–8, 2005.

[51] G. Tryfou, Z. Theodosiou, and N. Tsapatsoulis. **Web image context extraction based on semantic representation of web page visual**

segments. In *Proc. of International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 63–67, 2012.

[52] Y. Liu, D. Zhang, and G. Lu. **SIEVE-Search images effectively through visual elimination**. In N. Sebe, Y. Liu, Y. Zhuang, and T. S. Huang, editors, *Multimedia content analysis and mining*, **4577** of *Lecture Notes in Computer Science*, pages 381–390. Springer Berlin Heidelberg, 2007.

[53] X.-J. Wang, W.-Y. Ma, Q.-C. He, and X. Li. **Grouping web image search result**. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 436–439, 2004.

[54] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. **AnnoSearch: Image auto-annotation by search**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**, pages 1483–1490, 2006.

[55] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. **Hierarchical clustering of WWW image search results using visual, textual and link information**. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 952–959, 2004.

[56] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. **VIPS: a Vision based page segmentation algorithm**. Technical report, Microsoft Research, 2003.

[57] Y. Jin, L. Khan, L. Wang, and M. Awad. **Image annotations by combining multiple evidence & wordNet**. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 706–715, 2005.

[58] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. **Image annotation refinement using random walk with restarts**. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pages 647–650, 2006.

[59] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. **Content-based image annotation refinement**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[60] K. K. Matusiak. **Towards user-centered indexing in digital image collections**. *OCLC Systems & Services*, **22**(4):283–298, 2006.

[61] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and Gay G. **Accurately interpreting clickthrough data as implicit feedback**. In *Proc. of the 28th Annual International ACM SIGIR Conference*, pages 154–161, 2005.

[62] K. Ntalianis, N. Tsapatsoulis, A. Doulamis, and N. Matsatsinis. **Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution**. *Multimedia Tools and Applications*, 2012.

[63] C. Macdonald and I. Ounis. **Usefulness of quality clickthrough data for training**. In *Proc. of the 2009 Workshop on Web Search Click Data*, pages 75–79, 2009.

[64] T. Tsikrika, C. Diou, A. P De Vries, and A. Delopoulos. **Image annotation using clickthrough data**. In *Proc. of the 8th International Conference on Image and Video Retrieval*, pages 1–8, 2009.

[65] A. Kittur and R. E Kraut. **Harnessing the wisdom of crowds in wikipedia: quality through coordination**. In *Proc. of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46, 2008.

[66] Z. Theodosiou and N. Tsapatsoulis. **Crowdsourcing annotation: Modelling keywords using low level features**. In *Proc. of the 5th International Conference on Internet Multimedia Systems Architecture and Application*, 2011.

[67] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. **A crowdsourceable QoE evaluation framework for multimedia content**. In *Proc. of the 17th ACM international conference on Multimedia*, pages 491–500, 2009.

[68] T. Brants. **Inter-annotator agreement for a German newspaper corpus**. In *Proc. of the 2nd International Conference on Language Resources and Evaluation*, pages 1–5, 2000.

[69] A. Kilgarriff. **Gold standard datasets for evaluating word sense disambiguation programs**. *Computer Speech and Language*, **12**(3):453–472, 1998.

[70] A. Makadia, V. Pavlovic, and S. Kumar. **A new baseline for image annotation**. In *Proceedings of European Conference on Computer Vision*, pages 316–329, 2008.

[71] C-F. Tsai, K. McGarry, and J. Tait. **Qualitative evaluation of automatic assignment of keywords to images**. *Information Processing and Management*, **42**(1):136–154, 2006.

[72] K. Athanasakos, V. Stathopoulos, and J. Jose. **A framework for evaluating automatic image annotation algorithms**. *Lecture Notes in Computer Science*, **5993**:217–228, 2010.

[73] R. Zhang, Z. Zhang, M. Li, and H. J. Zhang. **A probabilistic semantic model for image annotation and multi-modal image retrieval**. *Multimedia Systems*, pages 27–33, 2006.

[74] R. Datta, D. Joshi, J. Li, and J. Z. Wang. **Image retrieval: Ideas, influences, and trends of the new age**. *ACM Computing Surveys*, **40**(2):5:1–5:60, 2008.

[75] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Computer graphics: Principles and practice (2Nd Ed.)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.

[76] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[77] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. **Color and texture descriptors**. *IEEE Transactions on Circuits and Systems for Video Technology*, **11**(6):703–715, 2001.

[78] K.-S. Goh, E.Y. Chang, and Beitao Li. **Using one-class and two-class SVMs for multiclass image annotation**. *IEEE Transactions on Knowledge and Data Engineering*, **17**(10):1333–1346, 2005.

[79] C. Yang, M. Dong, and F. Fotouhi. **Image content annotation using Bayesian framework and complement components analysis**. In *IEEE International Conference on Image Processing*, **1**, pages I–1193–6, 2005.

[80] G. Pass and R. Zabih. **Histogram refinement for content-based image retrieval**. In *IEEE Workshop on Applications of Computer Vision*, pages 96–102, 1996.

[81] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. **Image indexing using color correlograms**. In *Conference on Computer Vision and Pattern Recognition*, page 762, 1997.

[82] ISO/IEC 15938-3:2001 Information Technology - Multimedia Content Description Interface - Part 3: Visual, Ver. 1.

[83] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia content description interface.* John Wiley & Sons, Inc., New York, NY, USA, 2002.

[84] P. Duygulu, K. Barnard, J. F. G. de Freitas, and A. D. Forsyth. **Object recognition as machine translation: learning a lexicon for a fixed image vocabulary**. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, May 2002.

[85] Y. Liu, J. Zhang, D. W. Tjondronegoro, and S. Geva. **A shape ontology framework for bird classification**. In *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications*, 2007.

[86] K.-W. Park, J.-W. Jeong, and D.-H. Lee. **OLYBIA: Ontology-based automatic image annotation system using semantic inference rules**. In

R. Kotagiri, P. R. Krishna, M. Mohania, and E. Nantajeewarawat, editors, *Advances in databases: Concepts, systems and applications*, **4443** of *Lecture Notes in Computer Science*, pages 485–496. Springer Berlin Heidelberg, 2007.

[87] A. Ghoshal, P. Ircing, and S. Khudanpur. **Hidden Markov models for automatic annotation and content-based retrieval of images and video**. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–551, 2005.

[88] L. Jiang, J. Hou, Z. Chen, and D. Zhang. **Automatic image annotation based on decision tree machine learning**. In *Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 170–175, 2009.

[89] J. Jeon, V. Lavrenko, and R. Manmatha. **Automatic image annotation and retrieval using cross-media relevance models**. In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 119–126, 2003.

[90] V. Lavrenko, R. Manmatha, and J. Jeon. **A model for learning the semantics of pictures**. In *Proceedings of Advances in Neural Information Processing Systems*, 2003.

[91] Z. H. Zhou and M. L. Zhang. **Multi-instance multi-label learning with application to scene classification**. In *Advances in Neural Information Processing Systems 19*, pages 1609–1616, 2006.

[92] S. Zhu and X. Tan. **A novel automatic image annotation method based on multi-instance learning**. *Procedia Engineering*, **15**:3439–3444, 2011.

[93] R. W. Picard. **Light-years from Lena: video and image libraries of the future**. In *Proceedings of International Conference on Image Processing*, pages 310–313, 1995.

[94] J. Yang, S, S. K. Kim, K. S. Seo, Y. M. Ro, J.-Y. Kim, and Y. S. Seo. **Semantic categorization of digital home photo using photographic region templates**. *Information Processing & Management*, **43**(2):503–514, 2007.

[95] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. **Supervised learning of semantic classes for image annotation and retrieval**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**(3):394–410, 2007.

[96] J. Howe. *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Business, 2008.

[97] J. Howe. **The rise of crowdsourcing**. *Wired Magazine*, **14**(6):176–183, 2006.

[98] D. Brabham. **Crowdsourcing as a model for problem solving: An Introduction and cases**. *Convergence*, **14**(1):75–90, 2008.

[99] A Kittur, E. H. Chi, and B. Suh. **Crowdsourcing user studies with mechanical turk**. In *Proc. Of CHI conference on Human Factors in Computing Systems*, pages 453–456, 2008.

[100] P. Welinder and P. Perona. **Online crowdsourcing: rating annotators and obtaining cost effective labels**. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 25–32, 2010.

[101] B. C. Russel, A. Torralba, K. P. Murphy, and W. T. Freeman. **Labelme: A database and web-based tool for image annotation**. *International Journal of Computer Vision*, **77**(1-3):157–173, 2008.

[102] A. Sorokin and D. Forsyth. **Utility data annotation with Amazon Mechanical Turk**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[103] J. Deng, W. Dong, R Socher, L. Li, K. Li, and L. Fei-Fei. **ImageNet: a large-scale hierarchical image database**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 710–719, 2009.

[104] http://crowdcrafting.org/.

[105] C. Eickhoff and A. P. Vries. **Increasing cheat robustness of crowdsourcing tasks**. *Information Retrieval*, **16**(2):121–137, 2013.

[106] R. Snow, B. O Connor, D. Jurafsky, and A. Ng. **Cheap and fast - But is it Good? Evaluating non-expert annotations for natural language tasks**. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[107] C. Vondrick, D. Patterson, and D. Ramanan. **Efficiently scaling up crowdsourced video annotation**. *International Journal of Computer Vision*, **101**(1):184–204, 2013.

[108] V. Raykar, S. Zhao, L. Yu, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy. **Supervised learning from multiple experts: Whom to trust when everyone lies a bit**. In *Proc. 26th Annual International Conference on Machine Learning*, pages 889–896, 2009.

[109] P. Smyth, M. Fayyad, U. amd Burl, P. Perona, and P. Baldi. **Inferring ground truth from subjective labeling of Venus images**. *Advances in Neural Information Processing Systems*, **7**:1085–1092, 1995.

[110] V. S. Sheng, F. Provost, and P. G. Ipeirotis. **Get another label? Improving data quality and data mining using multiple, noisy labelers**. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 614–622, 2008.

[111] M. Spain and P. Perona. **Some objects are more equal than others: Measuring and predicting importance**. In *Proc. 10th European Conference on Computer Vision*, pages 523–536, 2008.

[112] L. V. Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. **reCAPTCHA: Human-based character recognition via web security measures**. *Science*, **321**(5895):1465–1468, 2008.

[113] V. L Ahn and L. Dabbish. **Labeling images with a computer game**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004.

[114] J. Whitehill, P. Ruvolo, J. Bergsma, T. Wu, and J. Movellan. **Whose vote should count more: Optimal integration of labels from labelers of unknown expertise**. In *Proc. 23rd Annual Conference on Neural Information Processing Systems*, pages 2035–2043, 2009.

[115] S. Vijayanarasimhan and K. Grauman. **What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations**. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2262–2269, 2009.

[116] C. Callison-Burch. **Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk**. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 286–295, 2009.

[117] J. Veronis. **A study of polysemy judgements and interannotator agreement**. In *Programme and advanced papers of the Senseval Workshop*, Herstmonceux Castle, England, 1998.

[118] T. Chklovski and R. Mihalcea. **Exploiting agreement and disagreement of human annotators for word sense disambiguation**. In *Proc. of International Conference on Recent Advances in Natural Language Processing*, 2003.

[119] S. Nowak and S. Ruger. **How reliable are annotations via crowdsourcing a study about inter-annotator agreement for multi-label image an-**

notation. In *Proc. of the International Conference on Multimedia Information Retrieval*, pages 557–566, 2010.

[120] H. T. Shen, B. C. Ooi, and K.-L. Tan. **Giving meanings to WWW images**. In *Proc. of the Eighth ACM International Conference on Multimedia*, pages 39–47, 2000.

[121] K. Yanai. **Generic image classification using visual knowledge on the web**. In *Proc. of the 11th ACM International Conference on Multimedia*, pages 167–176, 2003.

[122] H. Feng, R. Shi, and T.-S. Chua. **A bootstrapping framework for annotating and retrieving WWW images**. In *Proc. of the 12th Annual ACM International Conference on Multimedia*, pages 960–967, 2004.

[123] F. Fauzi, J.-L. Hong, and M. Belkhatir. **Webpage segmentation for extracting images and their surrounding contextual information**. In *Proc. of the ACM International Conference on Multimedia*, pages 649–652, 2009.

[124] S. Alcic and S. Conrad. **A clustering-based approach to web image context extraction**. In *Proc. of 3rd International Conferences on Advances in Multimedia*, pages 74–79, 2011.

[125] X. He, D. Cai, J.-R. Wen, W.-Y. Ma, and H.-J. Zhang. **Clustering and searching WWW images using link and page layout analysis**. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **3**(2), 2007.

[126] S. Alcic and S. Conrad. **Measuring performance of web image context extraction**. In *Proc. of the 10th International Workshop on Multimedia Data Mining*, pages 8:1–8:8, 2010.

[127] C. Fellbaum, editor. *WordNet: an electronic lexical database.* MIT Press, Cambridge, USA, 1998.

[128] C. Leacock and M. Chodorow. **Combining local context and WordNet similarity for word sense identification**. In C. Fellfaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts, 1998.

[129] Z. Wu and M. Palmer. **Verb semantics and lexical selection**. In *Proc. of the 32nd Annual Meeting of the Associations for Computational Linguistics*, pages 133–138, 1994.

[130] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. **WordNet::Similarity: Measuring the relatedness of concepts**. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, 2004.

[131] V. Bhardwaj, R. J. Passonneau, A. Salleb-Aouissi, and N. Ide. **Anveshan: a framework for analysis of multiple annotators' labeling behavior**. In *Proc. of the 4th Linguistic Annotation Workshop*, pages 47–55, 2010.

[132] C. Wang, L. Zhang, and H.-J. Zhang. **Learning to reduce the semantic gap in web image retrieval and annotation**. In *Proc. of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 355–362, 2008.

[133] J. Cohen. **A coefficient of agreement for nomimal scales**. *Educational and Phsychological Measurement*, **20**(1):37–46, 1960.

[134] J. R. Landis and G. K Koch. **The measurement of observer agreement for categorical data**. *Biometrics*, **33**:159–174, 1977.

[135] J. J. Randolph. **Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa**. In *In Joensuu University Learning and Instruction Symposium*, 2005.

[136] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. **Basic objects in natural categories**. *Cognitive Psychology*, **8**(3):382–439, 1976.

[137] B. Stvilia and C. Jörgensen. **Member activities and quality of tags in a collection of historical photographs in Flickr**. *Journal of the American Society for Information Science and Technology*, **61**(12):2477–2489, 2010.

[138] Randolph, J. J.(2008). Online Kappa Calculator. Retrieved October 5, 2012, from http://justus.randolph.name/kappa/.

[139] L. R. Brody and J. A. Hall. **Gender, emotion, and socialization**. In J. C. Chrisler and D. R. McCreary, editors, *Handbook of Gender Research in Psychology*, pages 429–454. Springer New York, 2010.

[140] G. Lakoff. **Cognitive semantics**. In P. Violi U. Eco, M. Santambrogio, editor, *Meaning and mental representations*, pages 119–154. Indiana University Press, Bloomington, IN, USA, 1988.

[141] G. Singer, U. Norbisrath, and D Lewandowski. **Impact of gender and age on performing search tasks online**. *CoRR*, 2012.

[142] K. Papadopoulos, N. Tsapatsoulis, A. Lanitis, and A. Kounoudes. **The history of Commandaria: Digital journeys back to time**. In *Proc. of the 14th International Conference on Virtual Systems and Multimedia*, 2008.

[143] M. Cowles and C. Davis. **On the origins of the .05 level of statistical significance**. *American Psychologist*, **37**(5):553–558, 1982.

[144] S. Fujisawa. **Automatic creation and enhancement of metadata for cultural heritage**. In *Bull. IEEE Tech. Committee on Digital Libraries (TCDL)*, 2007.

[145] G. Tryfou and N. Tsapatsoulis. **Extraction of web image information: Semantic or visual cues?** In L. Iliadis, I. Maglogiannis, and H. Papadopoulos, editors, *Artificial Intelligence Applications and Innovations*, **381** of *IFIP Advances in Information and Communication Technology*, pages 368–373. Springer Berlin Heidelberg, 2012.

[146] G. Salton, A. Wong, and C. S. Yang. **A Vector space model for automatic indexing**. *Communications of the ACM*, **18**(11):613–620, 1975.

[147] M. Nixon and A. S. Aguado. *Feature extraction & image processing.* Academic Press, San Diego, CA, USA, second edition, January 2008.

[148] T. Tuytelaars and K. Mikolajczyk. **Local invariant feature detectors: A survey**. *Computer Graphics and Vision*, **3**(3):177–280, 2008.

[149] B. Schiele and J.L. Crowley. **Probabilistic object recognition using multidimensional receptive field histograms**. In *Proc. of the 13th International Conference on Pattern Recognition*, **2**, pages 50–54, 1996.

[150] C. Schmid and R. Mohr. **Local grayvalue invariants for image retrieval**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(5):530–535, 1997.

[151] D. G. Lowe. **Distinctive image features from scale invariant keypoints**. *International Journal of Computer Vision*, **60**(2):91–110, 2004.

[152] N. Dalal and B. Triggs. **Histograms of oriented gradients for human detection**. In *Proc. of International Conference on Computer Vision & Pattern Recognition*, pages 886–893, 2005.

[153] Y. Ke and R. Sutkthankar. **PCA-SIFT: A more distinctive representation for local image descriptors**. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, **2**, pages II–506–II–513, 2004.

[154] K. Mikolajczyk and C. Schmid. **A performance evaluation of local descriptors**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(10):1615–1630, 2005.

[155] B. Ayers and M. Boutell. **Home interior classification using SIFT keypoint histograms**. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[156] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. **SURF: Speeded up robust features**. *Computer Vision and Image Understanding*, **110**(3):346–359, 2008.

[157] E. Tola, V. Lepetit, and P. Fua. **DAISY: An efficient dense descriptor applied to wide-baseline stereo**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(5):815–830, 2010.

[158] J. Sivic and A. Zisserman. **Video google: A text retrieval approach to object matching in videos**. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[159] M. Zhang and A. A. Sawchuk. **Motion primitive-based human activity recognition using a bag-of-features approach**. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 631–640, 2012.

[160] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua. **Contextual bag-of-words for visual categorization**. *IEEE Transactions on Circuits and Systems for Video Technology*, **21**(4):381–392, 2011.

[161] S. Lazebnik, C. Schmid, and J. Ponce. **Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories**. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**, pages 2169–2178, 2006.

[162] Z. Theodosiou and N. Tsapatsoulis. **Modelling crowdsourcing originated keywords within the athletics domain**. In L. Iliadis, I. Maglogiannis, and H. Papadopoulos, editors, *Artificial Intelligence Applications and Innovations*, **381** of *IFIP Advances in Information and Communication Technology*, pages 404–413. Springer Berlin Heidelberg, 2012.

[163] M. A. Turk and A. P. Pentland. **Face recognition using eigenfaces**. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[164] H. Murase and S. Nayar. **Visual learning and recognition of 3-d objects from appearance**. *International Journal of Computer Vision*, **14**(1):5–24, 1995.

[165] S. A. Chatzichristofis and Y. S. Boutalis. *Compact composite descriptors for content based image retrieval: Basics, concepts, tools.* VDM Verlag, Saarbrucken, Germany, 2011.

[166] A. Arampatzis, K. Zagoris, and S. A. Chatzichristofis. **Dynamic two-stage image retrieval from large multimedia databases**. *Information Processing & Management*, **49**(1):274–285, 2013.

[167] S. A. Chatzichristofis, A. Arampatzis, and Y. S. Boutalis. **Investigating the behavior of compact composite descriptors in early fusion, late fusion and distributed image retrieval**. *Radioengineering*, **19**(4):725–733, 2010.

[168] S. A. Chatzichristofis and Y. S. Boutalis. **FCTH: fuzzy color and texture histogram - A low level feature for accurate image retrieval**. In *Proceedings of 9th International Workshop on the Image Analysis for Multimedia Interactive Services*, pages 191–196, 2008.

[169] S. A. Chatzichristofis and Y. S. Boutalis. **CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval**. In A. Gasteratos, M. Vincze, and J. K. Tsotsos, editors, *Computer Vision Systems*, **5008** of *Lecture Notes in Computer Science*, pages 312–322. Springer Berlin Heidelberg, 2008.

[170] S. A. Chatzichristofis and Y. S. Boutalis. **Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor**. *Multimedia Tools and Applications*, **46**:493–519, 2010.

[171] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. **SpCD - Spatial color distributiondDescriptor - A fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval**. In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence*, pages 58–63, 2010.

[172] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu. **Feature fusion: Parallel strategy vs. serial strategy**. *Pattern Recognition*, **33**:1369–1381, 2003.

[173] E. Levina and P. J. Bickel. **Maximum likelihood estimation of intrinsic dimension**. *Advances in Neural Information Processing Systems*, **17**:777–784, 2004.

[174] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[175] MPEG-7 Visual Experimentation Model (XM), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4063, 2001.

[176] M. Bober. **MPEG-7 visual shape descriptors**. *IEEE Transactions on Circuits and Systems for Video Technology*, **11**(6):716–719, 2001.

[177] H. Eidenberger. **How good are the visual MPEG-7 features?** In *Proceedings SPIE Visual Communications and Image Processing Conference*, **5150**, pages 476–488, 2003.

[178] E. Spyrou, H. L. Borgne, T. P. Mailis, E. Cooke, Y. S. A., and N. E. O'Connor. **Fusing MPEG-7 visual descriptors for image classification**. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, editors, *ICANN (2)*, **3697** of *Lecture Notes in Computer Science*, pages 847–852. Springer, 2005.

[179] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques, second edition (Morgan Kaufmann series in data management systems).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[180] R.-E. Fan, P.-H. Chen, and C.-J. Lin. **Working set selection using second order information for training support vector machines**. *Journal of Machine Learning Research*, **6**:1889–1918, 2005.

[181] J. C. Platt. **Advances in kernel methods**. chapter Fast training of support vector machines using sequential minimal optimization, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[182] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. **Improvements to Platt's smo algorithm for svm classifier design**. *Neural Computation*, **13**(3):637–649, 2001.

[183] C. M. Bishop. *Pattern recognition and machine learning (Information science and statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[184] A. Bosch, A. Zisserman, and X. Munoz. **Image Classification using Random Forests and Ferns**. In *Proc. of IEEE 11th Intenational Conference on Computer Vision*, pages 1–8, 2007.

[185] K. Grauman and T. Darrel. **The pyramid match kernel: Discriminative classification with sets of image features**. In *Proc. of the International Conference on Computer Vision*, pages 1458–1465, 2005.

[186] C. Jordan. *Cours D'analyse De L'Ecole Polytechnique*. Gauthier-Villars Et Fils, Paris, 1887.

[187] G. Peano. **Sur une courbe, qui remplit toute une aire plane**. *Mathematische Annalen*, **36**:157–160, 1890.

[188] D. Hilbert. **Üeber die stetige Abbildung einer linie auf ein Flächenstück**. *Mathematische Annalen*, **38**:459–460, 1891.

[189] B. Moon, H. Jagadish, C. Faloutsos, and J. H. Salz. **Analysis of the clustering properties of the Hilbert space-filling curve**. *Knowledge and Data Engineering*, **13**(1):124–141, 2001.

[190] D. Guliato, W. A. A de Oliveira, and C. Traina. **A new feature descriptor derived from Hilbert space-filling curve to assist breast cancer classification**. In *Proc. IEEE 23rd International Symposium on Computer-Based Medical Systems*, pages 303–308, 2010.

[191] N. J. Rose. **Hilbert-type space-filling curves**, 2001. http://www4.ncsu.edu/ njrose/pdfFiles/HilbertCurve.pdf.

[192] **BOEMIE - Bootstrapping Ontology Evolution with Multimedia Information Extraction**. http://www.boemie.org.

[193] L. Fei-Fei and P. Perona. **A bayesian hierarchical model for learning natural scene categories**. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2**, pages 524–531, 2005.

[194] A. Oliva and A. Torralba. **Modelling the shape of the scene: A holistic representation of the spatial envelope**. *International Journal of Computer Vision*, **42**:145–175, 2001.

[195] L. Fei-Fei, R. Fergus, and P. Perona. **Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories**. In *Proc. of CVPR Workshop on Generative-Model Based Vision*, 2004.

[196] D. M. J. Tax and R .P .W. Duin. **Using two class classifiers for multiclass classification**. In *Proc. of 16th International conference of Pattern Recognition*, **2**, pages 124–127, 2002.

[197] E. Alpaydin. *Introduction to machine learning.* The MIT Press, 2nd edition, 2010.

[198] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees.* Statistics/Probability Series. Chapman & Hall, New York, NY, USA, 1984.

[199] J. R. Quinlan. **Discovering rules by induction from large collections of examples**. In D. Michie, editor, *Expert systems in the micro-electronic age*, pages 168–201. Edinburgh University Press, Edinburgh, UK, 1979.

[200] J. R. Quinlan. *C4.5: Programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, October 1992.

[201] T. S. Lim, W.-Y. LOH, and W. Cohen. **A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms**. *Machine Learning*, **40**:203–228, 2000.

[202] L. Breiman. **Random forests**. *Machine Learning*, **45**:5–32, 2001.

[203] S. B. Kotsiantis. **Supervised machine learning: A review of classification techniques**. *Informatica*, **31**:249–268, 2007.

[204] J. Fürnkranz and G. Widmer. **Incremental reduced error pruning**. In *Proc. of International Conference on Machine Learning*, pages 70–77, 1994.

[205] W. W. Cohen. **Fast effective rule induction**. In *Proc. of 12th International Conference on Machine Learning*, pages 115–123, 1995.

[206] R. J. Howlett and Jain L. C. *Radial basis function networks 2: New advances in design.* Physica-Verlag, Heidelberg, Germany, March 2001.

[207] N. Tsapatsoulis and Z. Theodosiou. **Object classification using the MPEG-7 visual descriptors: An experimental evaluation using state of the art data classifiers**. In C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, *Artificial Neural Networks II ICANN 2009*, **5769** of *Lecture Notes in Computer Science*, pages 905–912. Springer Berlin Heidelberg, 2009.

[208] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, NY, USA, 1995.

[209] L. I. Kuncheva and C. J. Whitaker. **Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy**. *Machine Learning*, **51**(2):181–207, 2003.

[210] A. Narasimhamurthy. **Evaluation of diversity measures for binary classifier ensembles**. In N. C. Oza, R. Polikar, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems*, **3541** of *Lecture Notes in Computer Science*, pages 267–277. Springer Berlin Heidelberg, 2005.

[211] S. Wang and X. Yao. **Relationships between diversity of classification ensembles and single-class performance measures**. *IEEE Transactions on Knowledge and Data Engineering*, **25**(1):206–219, 2013.

[212] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Y. Bang. **Constructing support vector machine ensemble**. *Pattern Recognition*, **36**(12):2757–2767, 2003.

[213] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. **On combining classifiers**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(3):226–239, 1998.

[214] Y. Freund and R. E. Schapire. **A Decision-theoretic generalization of on-line learning and an application to boosting**. *Journal of Computer and System Sciences*, **55**(1):119–139, 1997.

[215] A. Blum and T. Mitchell. **Combining labeled and unlabeled data with co-training**. In *Proc. of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[216] T. Feng, S. Brennan, Z. Qi, and T. Hai. **Co-tracking using semi-supervised support vector machines**. In *Proc. of the IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[217] D. Zhang and W. S. Lee. **Validating co-training models for web image classification**. In *SMA Annual Symposium, National University of Singapore*, 2005.

[218] Y. Du, Q. Li, Z. Cai, and X. Guan. **Multi-view semi-supervised web image classification via co-graph**. *Neurocomputing*, **122**:430–440, 2013.

[219] L. G. Valiant. **A theory of the learnable**. *Communications of the ACM*, **27**(11):1134–1142, 1984.

[220] S. Dasgupta, M. L. Littman, and D. A. Mcallester. **PAC Generalization bounds for co-training.** In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, pages 375–382, 2001.

[221] S. Goldman and Y. Zhou. **Enhancing supervised learning with unlabeled data**. In *Proc. of the 17th International Conference on Machine Learning*, pages 327–334, 2000.

[222] Y. Zhou and S. Goldman. **Democratic co-learning**. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pages 594–602, 2004.

[223] Z.-H. Zhou and M. Li. **Tri-training: Exploiting unlabeled data using three classifiers**. *IEEE Transactions on Knowledge and Data Engineering*, **17**(11):1529–1541, 2005.

[224] M.-L. Zhang and Z.-H. Zhou. **CoTrade: Confident co-training with data editing**. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **41**(6):1612–1626, 2011.

[225] F. Muhlenbach, S. Lallich, and D. A. Zighed. **Identifying and handling mislabelled instances**. *Journal of Intelligence Information Systems*, **22**(1):89–109, 2004.

[226] D. A. Zighed, S. Lallich, and F. Muhlenbach. **Separability index in supervised learning**. **2431** of *Lecture Notes in Computer Science*, pages 475–487. Springer, 2002.

[227] D. Angluin and P. Laird. **Learning from noisy examples**. *Machine Learning*, **2**(4):343–370, 1988.

[228] Z. Theodosiou, A. Kounoudes, N. Tsapatsoulis, and M. Milis. **MuLVAT: A video annotation tool based on XML-dictionaries and shot clustering**. In C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, *Artificial Neural Networks II ICANN 2009*, **5769** of *Lecture Notes in Computer Science*, pages 913–922. Springer Berlin Heidelberg, 2009.

[229] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes. **Trainable classifier-fusion schemes: An application to pedestrian detection**. In *Proc. of 12th International IEEE Conference on Intelligent Transportation Systems*, pages 432–437, 2009.