

HANOLISTIC: A HIERARCHICAL AUTOMATIC IMAGE ANNOTATION
SYSTEM USING HOLISTIC APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGE ÖZTİMUR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2008

Approval of the thesis

**HANOLISTIC: A HIERARCHICAL AUTOMATIC IMAGE
ANNOTATION SYSTEM USING HOLISTIC APPROACH**

submitted by **ÖZGE ÖZTİMUR** in partial fulfillment of the requirements for
the degree of **Master of Science in Computer Engineering, Middle East
Technical University** by,

Prof. Dr. Canan Özgen

Dean, **Graduate School of Natural and Applied Sciences**

Prof. Dr. Volkan Atalay

Head of Department, **Computer Engineering**

Prof. Dr. Fatoş Tünay Yarman-Vural

Supervisor, **Computer Engineering, METU**

Examining Committee Members:

Prof. Dr. Volkan Atalay

Computer Engineering, METU

Prof. Dr. Fatoş Tünay Yarman-Vural

Computer Engineering, METU

Prof. Dr. Neşe Yalabık

Computer Engineering, METU

Dr. Onur Tolga Şehitoğlu

Computer Engineering, METU

Yüksek Mühendis Ahmet Sayar

Computer Engineer, TUBİTAK,UZAY

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : ÖZGE ÖZTİMUR

Signature :

ABSTRACT

HANOLISTIC: A HIERARCHICAL AUTOMATIC IMAGE ANNOTATION SYSTEM
USING HOLISTIC APPROACH

ÖZTİMUR, ÖZGE

M.S., Department of Computer Engineering

Supervisor: Prof. Dr. Fatoş Tünay Yarman-Vural

January 2008, 54 pages

Automatic image annotation is the process of assigning keywords to digital images depending on the content information. In one sense, it is a mapping from the visual content information to the semantic context information. In this thesis, we propose a novel approach for automatic image annotation problem, where the annotation is formulated as a multivariate mapping from a set of independent descriptor spaces, representing a whole image, to a set of words, representing class labels. For this purpose, a hierarchical annotation architecture, named as HANOLISTIC (Hierarchical Image Annotation System Using Holistic Approach), is defined with two layers. At the first layer, called level-0 annotator, each annotator is fed by a set of distinct descriptor, extracted from the whole image. This enables us to represent the image at each annotator by a different visual property of a descriptor. Since, we use the whole image, the problematic segmentation process is avoided. Training of each annotator is accomplished by a supervised learning paradigm, where each word is represented by a class label. Note that, this approach is slightly different than the classical training approaches, where each data has a unique label. In the proposed system, since each image has one or more annotating words, we assume that an image belongs to more than one class. The output of the level-0 annotators indicate the membership values of the words in the vocabulary, to belong an image. These membership values from each annotator is, then, aggregated at the second layer by using various rules, to obtain meta-layer annotator.

The rules, employed in this study, involves summation and/or weighted summation of the output of layer-0 annotators. Finally, a set of words from the vocabulary is selected based on the ranking of the output of meta-layer. The hierarchical annotation system proposed in this thesis outperforms state of the art annotation systems based on segmental and holistic approaches. The proposed system is examined in-depth and compared to the other systems in the literature by means of using several performance criteria.

Keywords: Automatic Image Annotation, Holistic Approach, Combination of Image Annotators, MPEG-7 Descriptors, Hierarchical Architecture

ÖZ

BÜTÜNSEL YAKLAŞIMLA HİYERARŞİK OTOMATİK GÖRÜNTÜ AÇIKLAMA

ÖZTİMUR, ÖZGE

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Fatoş Tünay Yarman-Vural

Ocak 2008, 54 sayfa

Otomatik görüntü açıklama dijital bir görüntüye, görüntü ile ilgili anahtar kelimelerin verilmesi işlemidir. Görüntü açıklama bir anlamda görsel içerik bilgisinin anlamsal bilgiye dönüştürülmesi işlemi olarak düşünülebilir. Bu çalışmada, yeni bir otomatik görüntü açıklama yöntemi önerilmektedir. Görüntü açıklama çok değişkenli bir dönüşüm şeklinde formüle edilmiş olup, bu dönüşüm, görüntüyü bir bütün olarak ifade eden birbirinden bağımsız çok sayıda görsel betimleyici uzayından, sınıf isimlerine denk düşen bir grup kelimeyi atama işlemi olarak gerçekleştirilmektedir. Bu amaçla iki katmanlı sıradüzenli bir yapı tanımlanmaktadır. Birinci katmanda, çok sayıda görüntü açıklayıcısının herbiri görüntünün tamamından çıkarılan, birbirinden farklı betimleyicilere göre bütün kelimelerin ait olma olasılıkları hesaplanır. Görüntünün öznitelikleri bütün görüntüden çıkarıldığı için bölütleme ile ilgili sorunlardan kaçınılabilmektedir. Görüntü açıklayıcıları kelimelerin sınıf ismi olarak ele alındığı denetlenebilir öğrenme algoritmaları ile eğitilmiştir. Bu eğitimde, bir görüntünün birden çok sınıfa ait olduğu varsayılmıştır. Bu öğrenme yönteminde, öznitelik uzayındaki vektörler birden fazla sınıfa ait olabilirler. Birinci katmanın çıktıları verilen bir görüntü için, sözlükteki bir kelimenin ait olma olasılık değerini ifade etmektedir. İkinci katmanda, birinci katmandaki çok sayıda açıklayıcıdan gelen aitlik verileri çeşitli yöntemlerle birleştirilerek yeni aitlik değerleri hesaplanır. En sonunda, ikinci katmadan elde edilen sonuçlardan yararlanılarak görüntüyü açıklayan bir grup kelime seçilir. Önerilen çok katmanlı yapı, literatürdeki bütünsel ve

bölgesel yaklaşımla resim açıklama yöntemlerinden daha başarılı sonuçlar elde etmektedir. Bu çalışmada öne sürülen yöntem deneysel olarak incelenmiş ve pek çok performans kriteri göz önünde bulundurularak literatürdeki yöntemlerle karşılaştırılmıştır.

Anahtar Kelimeler: Otomatik Görüntü Açıklama, Bütünsel Yaklaşım, Resim Açıklayıcılarının Birleştirilmesi, MPEG-7 Betimleyicileri, Hiyerarşik Yapı

ACKNOWLEDGMENTS

I would like to thank Prof. Dr. T. Fatoş Yarman-Vural who supervised me throughout this work. I am very grateful to her for her self sacrificing involvement in this job and for her motivating approach. I would also like to thank Emre Akbaş who shared his studies as well as his ideas with me and to Cüneyt Mertayak for his comments and help. I would also like to thank Onur Tolga Şehitoğlu for his support on latex. I am very thankful to all my friends who shared my hard times. Additionally, I would like to express my gratitude to my parents who encourage me all the time and to my sister Selin who has always been by me with her brilliant ideas. Finally, I am very thankful to Çağrı Can Karadağ for his patience and sensibility.

To my parents, my sister and Can

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ	vi
ACKNOWLEDGMENTS	viii
DEDICATON	ix
TABLE OF CONTENTS	x
LIST OF FIGURES	xii
LIST OF TABLES	xiii
CHAPTER	
1 INTRODUCTION	1
1.1 State of the Art Automatic Image Annotation Systems	2
1.2 Problems in Automatic Image Annotation	3
1.3 Performance Measures	4
1.4 Our Approach	5
1.5 Thesis Outline	6
2 RELATED WORK	7
2.1 Studies Based on the Segmental Approach	8
2.2 Studies Based on the Holistic Approach	11
2.3 Discussion	12
2.3.1 Representation	12
2.3.2 Segmentation	13
2.3.3 Quantization of Feature Vectors	13
2.3.4 Training	13
2.3.5 Data Set	13

3	HANOLISTIC: A HIERARCHICAL IMAGE ANNOTATION SYSTEM	15
3.1	Introduction	15
3.2	Image Representation	16
3.2.1	Low Level Visual Descriptors for Content Information	17
3.2.2	Semantic Words for Context Information	18
3.3	Mathematical Definition of Image Annotation Problem	19
3.4	Hierarchical Architecture for Annotation	21
3.4.1	Level-0 Annotators	21
3.4.2	Meta-Level Annotator	22
3.5	Annotation Training	22
3.6	Automatic Annotation	23
3.7	Performance Measures	23
3.8	Discussion on the Pros and Cons of HANOLISTIC	25
3.9	Realization of the System	26
3.9.1	Realization of the Level-0 Annotators	26
3.9.2	Realization of the Meta-Level Annotator	28
3.9.3	Automatic Image Annotation by HANOLISTIC	30
4	EMPIRICAL STUDY	32
4.1	Experimental Setup	32
4.1.1	Data Set	32
4.1.2	Selection of Descriptors	34
4.1.3	Estimating the Best k -Values for Fuzzy-knn	34
4.1.4	Exploring the Performances of Individual Descriptors	34
4.1.5	Exploring the Overall Performance of Descriptors in HANOLISTIC	38
4.2	Automatic Annotation Performance	39
4.2.1	Setting Parameters for Fuzzy-knn at Level-0	39
4.2.2	Performance of Level-0 Annotators	40
4.2.3	Exploring the Performance of Meta-Level	41
4.2.4	Exploring the Overall System Performance	43
4.2.5	Result	44
4.3	Automatic Annotation Examples	44
5	CONCLUSION AND FUTURE WORK	48
	REFERENCES	52

LIST OF FIGURES

FIGURES

Figure 1.1	An annotation example from Corel Stock Photo Library.	1
Figure 1.2	An annotation example from Corel Stock Photo Library.	3
Figure 3.1	Level-0 Annotator.	21
Figure 3.2	System Architecture.	22

LIST OF TABLES

TABLES

Table 3.1	Notation for the description of the hierarchical architecture.	20
Table 4.1	Sample images from the 5000 image Corel Dataset whose annotations are directly related to the image content.	33
Table 4.2	Sample images from the 5000 image Corel Dataset whose annotations involve abstract level of information.	33
Table 4.3	Sample images from the 5000 image Corel Dataset whose annotations are not directly related to the image content.	33
Table 4.4	The best k -values determined on the test set.	34
Table 4.5	Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by color layout features.	35
Table 4.6	Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by color structure features.	36
Table 4.7	Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by scalable color features.	37
Table 4.8	Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by homogenous texture features.	37
Table 4.9	Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by edge histogram features.	38

Table 4.10 Sample words with relatively high precision and recall values for level-0 annotators based on the five descriptors and for HANOLISTIC with summation annotator at the meta-level.	39
Table 4.11 The parameter k determined by cross-validation for each descriptor. . .	40
Table 4.12 Performance of HANOLISTIC for a function of increasing k -values. . . .	40
Table 4.13 Performance of HANOLISTIC with summation annotator method at the meta-level for several values of m	41
Table 4.14 Performance of Level-0 annotators.	41
Table 4.15 Performances for several descriptor combinations.	42
Table 4.16 Performance of HANOLISTIC over 263 words for several meta-level techniques.	43
Table 4.17 Effect of increasing the number of assigned words on the performance of per-word based weighted summation annotator.	44
Table 4.18 Comparison of HANOLISTIC with other systems in the literature . . .	44
Table 4.19 Sample annotations from the test set by HANOLISTIC with overall weighted summation annotator at the meta-level.	46
Table 4.20 Sample annotations from the test set by HANOLISTIC with overall weighted summation annotator at the meta-level.	47

CHAPTER 1

INTRODUCTION

In the simplest form, automatic image annotation is defined as the process of assigning keywords to digital images by Wikipedia. An automatic image annotation system may provide one or more keywords for a given image. Assignment of the words to images depends on several criterion. In this study, we assume that the annotation is based on the content of the image. Figure 1.1 is an example annotation from the Corel Stock Photo collection.

Digital images are widely used in today's life in a variety of applications with drastically increasing amount of storage space. As a result, the organization and retrieval of the digital media becomes a critical issue. It is a labor intensive job to manually annotate the images in a large set of images. As a consequence, automatic image annotation has become an important alternative to manual annotation. The studies, conducted on content based image retrieval (CBIR) is closely related to the automatic annotation problem. The main goal in the annotation problem is to provide images with semantically meaningful keywords, so that a semantically meaningful retrieval can be achieved. On the other hand, content based image retrieval systems aim directly to retrieve semantically meaningful images for a given query. In CBIR, retrieval is done either by 'query by sketch' or 'query by example' [1]. With the additional information provided by the automatic annotation system it is possible to make



beach, sand, sky, water

Figure 1.1: An annotation example from Corel Stock Photo Library.

queries with words. In that sense, CBIR and annotation is closely related. On the other hand, CBIR systems contain only visual information about the images in the data set, while an annotation system provides both visual information and semantic information, provided by the words assigned to the image. An automatic image annotation system provides a semantic indexing of a set of images. Information provided by the annotation systems is considered as an invaluable knowledge in today's life since it is a demanding job to annotate a large set of digital images with semantically meaningful keywords.

Annotation problem can be considered as a multi-class classification problem, where the number of classes is equal to the number of annotation words or number of concepts in the data set. In a typical classification problem, a sample is assigned to one of a priori known classes. In the annotation problem, a sample may be assigned to one or more classes. Since classification is a well-explored problem in pattern recognition, a wide range of studies are available on both supervised and unsupervised methods. Depending on its nature, annotation problem can be considered both as a supervised and an unsupervised classification problem. In the supervised classification, one takes advantage of the class labels provided by the keywords while classifying the given images and, then, finds a set of annotation words based on the final grouping. While in the unsupervised case, one clusters the images based on visual features and then assign annotation words based on the final grouping without making use of the class labels.

1.1 State of the Art Automatic Image Annotation Systems

State of the art automatic image annotation systems can be analyzed and grouped from various point of views: The available studies differ in terms of description methods, learning techniques and the application domain. Description of images is based on the low level features obtained either from the whole image [2] or from the regions of image [3], [4]. Some approaches make use of both the low level features and the high level semantic information [5] in training of the system. In terms of learning techniques, there are approaches using supervised learning algorithms to train the pre-defined classes of image database [6] while some approaches assume no classes and consider the problem as an unsupervised classification problem [2]. In most of the studies [3], [4], [7] the application domain consists of a set of images with annotation words and a set of images without annotation words to be annotated. On the other hand, [6], [8] define the problem as a supervised classification problem where the images have class labels and a set of words is assigned to each class. Considering all



castle, mountain, Scotland, water

Figure 1.2: An annotation example from Corel Stock Photo Library.

these view points, we can group the annotation studies as follows;

1. **Segmental Approaches:** This group of studies consider the image as consisting of semantically meaningful parts and tries to find a probabilistic relation between the parts of the image and the keywords. For this purpose, images are segmented or parts are taken from the image and features are extracted from these parts. [3], [4] and [9] are examples to this approach.
2. **Holistic Approaches:** This group considers the image as a whole. Features are extracted from the whole image. And a relation is explored directly between the image and the annotation words, [2].

Both approaches bear many pros and cons, which depends highly on the application domain. The first approach starts by segmentation, which is problematic by itself. For example Figure 1.2 is an example to indicate the difficulty of segmentation. Also, it is not possible to find the annotation words of the image when the image regions are considered. Instead, one needs to consider the whole image, as the concept information is gathered, once the whole image is perceived. Furthermore, it may not be the case that the annotation of the image needs segmentation. Even if it is so, segmentation is an extremely difficult and unsolved problem, which brings an extra error to the annotation problem. The second approach avoids segmentation. However, it may not always be possible to extract the meaningful words from the whole image represented by low-level visual features.

1.2 Problems in Automatic Image Annotation

The first problem is that systems lack an effective representation of scene which is one of the earliest problems of vision. Although several studies have been conducted on the representation of information available on images, the existing solutions are application dependent.

The problem is not just about the representation, but it is also about how we process the available data, so that it can be used for recognition of objects and further for annotation of images. Some techniques make use of segmentation and find probabilistic relations between the keywords and image regions. However, in many practical applications there is no clear evidence that there exists meaningful relations between the image regions and annotation words.

The second problem is the semantic gap between the low level visual features and high level concepts in images. When compared to a CBIR system, annotation systems provide extra information by the words, which can be used to reduce the semantic gap.

Besides these general problems of image processing, annotation systems suffer from the human subjectivity problem. Even in the manual annotation, there is not a standard way of annotating an image. Group of words annotating an image may vary to some extent depending who annotates the image. Therefore, the available performance measures of an annotation system is also plausible since it is not really possible to talk about ground truth. Another problem is that the performances of the systems proposed can not be measured precisely as the performance depends on several criteria such as precision, recall and coverage percentage, which will be discussed in the next section.

Finally, systems need to be tested over several standard data sets to measure the performance of the systems accurately. However, due to the difficulty of creating the databases, almost all the researchers use the few sets such as Corel Stock Photo library or Getty Image Archive [2], [3], [10], [4], [11] all of which have problems about statistical stability. In order generate test data, researchers provide annotation tools on the world wide web to obtain annotated data sets. For example the open annotation tool LabelMe provided on "<http://labelme.csail.mit.edu/>" provides a detailed data set with image parts labeled with the annotation words.

1.3 Performance Measures

Most frequently used performance criteria in the annotation systems are precision, recall. These criterion are widely used in measuring the performance of CBIR systems and they are modified for image annotation. Originally, recall is defined as the fraction of the images that are relevant to the query that are successfully retrieved. And precision is defined as the fraction of images retrieved that are relevant to the user's information need. In annotation systems, recall and precision values are evaluated for each word and the mean of all words are

considered as the performance of the system. In that sense; recall of a word is the number of correct annotations with that word divided by the number of annotations with that word in the ground truth. And precision of a word is the number of correct annotations with that word divided by the number of annotations with that word.

Another criterion proposed is the an mean average precision, which emphasizes returning relevant images earlier. Given a word, a ranking of the images is provided. If related images for the word take place earlier in the ranking then the mean average precision increases implying a "good" annotation. Coverage percentage is another criterion. It measures how many of the true words are provided with the annotation made by the system. As the number of annotation words that the system returns, increases, then the coverage percentage increases. It lacks the ability to penalize wrong annotations. In that sense, it is not a reliable criteria. Nevertheless, it can be used as a measure for an early processing of an annotation for the purpose of decreasing the number of words pertaining to be used for annotation of an image. For example, if a system automatically annotates the words of an image with a coverage percentage over ninety percent then that system could be used for decreasing the number of words available for the annotation of the image.

It is critical to emphasize that providing only these criteria (recall, precision, mean average precision, coverage percentage) are not adequate to compare and contrast different systems. Other metrics, like the number of words used in the annotation or the number of words that are never used in annotation, should also be provided. Otherwise, the results would be misleading.

1.4 Our Approach

We approach the annotation problem as a multi-class classification problem where each annotation word is considered as a class label. Unlike a typical classification problem, in this approach each image may belong to more than class. Images are described by a set of low level visual descriptors and a set of high level semantic information corresponding to words. Information gathered from different description spaces are combined by means of a Stacked Generalization [12] architecture. At each layer of stacked generalization, visual and semantic information is processed together, and moving from one layer to another involves integration of new information. The system takes its power from it simplicity. Any kind of processing inevitably causes some information loss. For this reason, the proposed system avoids complex computations and provide possible solutions for the problem at each layer.

The proposed system in this thesis, named as **HANOLISTIC** (**H**ierarchical image **A**nnotation system using **HOLISTIC** approach), is a holistic approach to the annotation problem. In this approach, global features are extracted from the whole image to represent the content information. On the other hand, the context information refers to a set of words containing the true words obtained from manual annotation of the image.

During the annotation training the context information is assumed as the class labels of each manually annotated image and the classifiers at the first layer of the hierarchical architecture are trained. Therefore, image annotation problem is formulated as a multi-class classification problem for each annotator at the first layer. The output of these annotators are the class membership values of each word for a given image. Then, the image annotation problem reduces to the selection of these words, given an unknown image at the second layer. This task is accomplished by statistically computing the highest probability words. This novel approach to automatic image annotation avoids the very difficult and problematic process of segmentation. Representing the whole image by many descriptors provide various aspects of visual information about the same image. Rather than representing the regions of image by the same visual descriptor, in this study, we represent the same image (without any regions) by many descriptors. Each representation is processed independently at the first layer of the hierarchical architecture, yielding many alternative solutions to image annotation. Second layer successfully combine the results of the first layer to estimate the final annotation.

1.5 Thesis Outline

The thesis is organized as follows; Chapter 2 includes the literature survey on the available automatic image annotation systems. Chapter 3 explains the detailed description of the proposed system, HANOLISTIC, and discusses the contributions of this study. Chapter 4 shows the power of the proposed system on the empirical studies. It describes the experimental setup and explains the results obtained. Chapter 5 summarizes and concludes the thesis with some discussion on the future work.

CHAPTER 2

RELATED WORK

Automatic image annotation has become a popular research area since late 1990s. It has aroused during the studies on CBIR. Automatic image annotation is proposed as a way of eliminating the semantic gap problem of CBIR systems. CBIR systems retrieve the images by querying using several techniques. However, this querying techniques are not sufficient for today's large image data sets. Most of the time, the user needs to query by text and wants to retrieve semantically related images. Without an automatic image annotation system, manually annotation of the whole data set is needed, to be able to query by text. Manual annotation demands too much time and labor. For this reason it is not surprising that the problem has become so popular.

One of the earliest studies in annotation is the interactive semi-automatic annotation proposed by Picard and Minka [13]. In their study, user labels the several parts of an image with a word. System is trained with the labeled regions, based on both the low level texture and the high level description. Then a relevance feedback is provided by the user, once the system annotates some image.

Annotation problem is considered from different perspectives and researchers associate it with other problems. For example, Barnard [7] refers to the association of words with whole image as annotation and association of words with particular image substructures as correspondence. They consider the correspondence problem as a peculiar feature of object recognition. On the other hand, Monay [14] classifies the annotation studies into two major groups;

1. **Annotation by propagation (supervised annotation):** Initially, image classes are defined as concepts. One or more words corresponding to a concept are assigned to classes. Each class is trained with the labeled data. For a given image, class or classes it belongs to are determined and annotation is done by propagating the corresponding

class words.

2. **Annotation by inference (unsupervised annotation):** This group of techniques tries to discover a relation between the words and the visual features. The joint probability of words and regional image features are estimated. In this view, annotation is considered as statistical inference in a graphical model.

Examining the studies in the literature we group the available approaches as follows:

1. **Segmental Approaches:** Image is considered as consisting of semantically meaningful regions which can be obtained by a low-level segmentation algorithm. It is assumed that the relations between the image regions and annotation words can be established by statistical methods.
2. **Holistic Approaches:** Image is considered having a semantic meaning as a whole. These approaches are more likely to handle situations such as the annotation of a war scene, in which case war is not an object in the image but it is the general information obtained from the whole image.

Most of the studies are concentrated on the first approach. The following sections describe shortly how state of the art annotation systems approach the problem in the two perspectives.

2.1 Studies Based on the Segmental Approach

In this approach, it is assumed that there is a relation between the image regions and the annotation words. Some studies [3] assume that there is a one-to-one correspondence, while the others [4] assume more relaxed relations. General steps followed in this perspective can be explained shortly in Algorithm 1.

Algorithm 1 General annotation algorithm for the segmental approaches

- 1: Segment or partition the image into regions
 - 2: Extract features from the regions
 - 3: Quantize features into blobs (for discrete space)
 - 4: Model the relation between image regions and annotation words
-

Mori [15] suggested finding a correlation between annotation words and image parts instead of a correlation between the annotation words and the whole image. Their model

is called Co-occurrence model as it investigates the co-occurrence of words with rectangular image regions. For this purpose, they initially divide each image into rectangular parts. It is assumed that all words of an image are inherited into each of the divided parts. Next, features are extracted from image parts and feature vectors are clustered by vector quantization. Then, likelihood of each word is calculated for each cluster. In the annotation of a given image, image is initially divided into rectangular regions then for each region the nearest centroid is found, and an average of the nearest centroids is evaluated. Finally, words with the highest average value of likelihoods are selected. The disadvantage of this method is that frequent words dominate the result, that is frequent words are mapped to more than one blob.

An improved version of the region based annotation, which uses machine learning methods and treats annotation as a translation from a vocabulary of blobs to a vocabulary of words is proposed by Duygulu [3]. They aimed to find out which image regions give rise to which annotation words using a machine translation approach. For this purpose, they initially segment images into regions by normalized cut and extract features from each region. Then, region descriptions are vector quantized by k-means and blobs are obtained. Next, the probability of each word for each blob is evaluated. In the annotation of a given image, image is first segmented into regions, then its blobs are found. After that, for each blob of the image words with the highest probability are determined. This method assumes one-to-one correspondence between the blobs and the words in the image.

Cross media relevance model proposed by Jeon [4] learns the joint distribution of blobs and words. This system takes advantage of the fact that an image can be described both with image features (blobs) and text (words). This model assign words to entire image not to blobs. Contrary to the translation model, CMRM does not assume a one-to-one correspondence between the blobs and words in an image. In stead, they assume that a set of keywords is related to a set of objects represented by blobs. The relevance model mentioned here refers to the probability distribution of all possible blobs appearing in an image and all possible words belonging to image. For annotation of an image, using the training image set, the probability of observing a set of words given a set of blobs is estimated and the distribution is marginalized with respect to words.

Continues-space relevance model (CRM) proposed by Lavrenko [9], emphasizes the fact that image regions should be used to obtain a context knowledge. For this purpose, they associate continues features directly with words without clustering. CRM is very similar to CMRM. But there is a major difference between the two. CMRM is discrete while CRM is

continues. CMRM needs to quantize continues features into discrete vocabulary and for this reason, applies clustering to the features and obtains blobs. CRM first, segments image and compute the features of each region. Then, by a generative model, predict the probability of generating a word given the features computed over the image regions.

Blei and Jordan, [16] consider the problem of modeling annotated data as "data with multiple types where the instance of one type (such as a caption) serves as a description of the other type (such as an image)." Their model is called the correspondence latent Dirichlet allocation (Corr-LDA) and it finds conditional relationships between latent variable descriptions of sets of image regions and sets of words. They also eliminate the clustering step but their model is parametric while the CRM is a non-parametric model.

Multiple Bernoulli Relevance Model [11] is based on the CRM. It overcomes the two shortcomings of CRM.

1. Segmented regions are replaced by rectangular grids. Computational time is reduced and the annotation performance is also increased. Context information is better incorporated in this model.
2. Image annotation is modeled by multiple-Bernoulli distribution instead of the multinomial distribution. Presence or absence of words is more critical than their prominence and the multiple bernoulli model is more suitable for modeling the words. It is experimentally shown that the annotation performance is increased with this model.

Monay and Perez in 2003 proposed using two latent space models; PLSA (Probabilistic Latent Semantic Analysis) and LSA (Latent Semantic Analysis) in the annotation problem. PLSA which is an inference based approach performed better than LSA which is a propagation based approach.

In the previous latent space models semantic information and the visual information are considered to have equal importance. Monay in [5] proposed using probabilistic latent space models for modeling the multi-model co-occurrences by firstly ensuring that the semantic information is kept consistent. The idea is based on the fact that semantic features give more information than the visual features. In this study, they proposed using two linked PLSA models to represent the semantic and the visual features. They reported that this method outperformed the other latent space models in annotation.

In the recent study of Monay [17], they presented three versions of the PLSA method. Versions differ from each other according to the amount of dominance assigned to visual and semantic features. They reported their best result on PLSA-Words model which learns the

latent aspects of the model from the text captions. Their results outperformed state of the art annotation systems.

2.2 Studies Based on the Holistic Approach

Most of the studies on the first approach are based on segmentation. This may be due to the fact that the initial studies on annotation [15], [3], [4] explored this approach and most researchers find this approach promising. It is also possible that, the researchers intuitively, believe that human visual system follows similar processing steps with the segmental approach. That is, in perception of an image, human visual system initially extracts the regions then the perception of the whole image follows. Nevertheless, it is not clear whether this approach really conforms with the information processing in the human visual system. Does our visual system really process image starting from the distinct regions or does it process the image as a whole and obtain a representation considering the whole image? Keeping these questions in mind, this section explains the studies belonging to the second category, the holistic approach in the literature.

Li and Wang [6] introduced a new problem domain for automatic annotation. In their problem, they have a set of categorized images and each category is labeled with the same words. Categorizes can be thought of corresponding to concepts. System is trained for each concept with the training images. For this purpose, features are extracted from the image then a 2-dimensional Multiresolution Hidden Markov Model is trained for each category. For the annotation of a new image, initially features are extracted from the image and image is fed to the trained models of each category. Then most probable top five classes are selected and a subset of words belonging to the selected classes are used for annotation of the image.

Akbaş and Yarman Vural in [2] proposed Supervised Ensemble of Visual Descriptors (SEVD) for the same problem domain as Wang. They proposed using an ensemble of visual descriptors instead of using one. SEVD is trained with the available annotated images by a classifier for each descriptor separately. And stacked-generalization, which was proposed by Wolpert [12], is applied to combine the results of the classifiers. A test image can be annotated by first, feeding the image to all classifiers and determining the most probable five classes for that image. In the final step, a subset of the words belonging to top five classes are selected for annotation.

They also propose an unsupervised annotation method in [2] which is called the Unsupervised Ensemble of Visual Descriptors (UEVD). In this system, training images are clustered

in each description space. For annotation of a new image, first the cluster of the image for each descriptor is determined then a subset of the words belonging to that cluster is selected. In selection of the annotation words, they use the rare words selection scheme proposed in [6].

A recent study in this approach is [10]. They propose the Color Structure Descriptor propagation (CSD-Prop) method which is a very simple method. They used the color structure descriptor and for a given image, rank all the training images according to their similarity. Annotation is done by propagating from the most similar image going until a certain number of words are selected. Despite its simplicity, it outperforms state of the art automatic annotation systems.

2.3 Discussion

There are several issues to be considered for a successful image annotation system. Let us briefly discuss these issues available in the literature which are described above.

2.3.1 Representation

The major problem in all image annotation problems is the representation of the visual information. Low level information lacks the ability to describe the semantic information in the image. This is the major reason, explaining why automatic image annotation has arose during the studies on CBIR. CBIR systems are only considered in terms of visual similarity while in automatic image annotation it is possible to talk about the semantic relation between images considering the annotation words. Nevertheless, vision studies still look for a better representation of the information available in an image. There are a number of low level descriptors based on texture, color, shape. Some systems [10], [14] make use of one descriptor which best fits the problem domain, while the other systems [3], [7], [2], [6] make use of a combination of them. Considering the current techniques, it is more advantageous to make use of several descriptors. But, the critical issue at this point is how to combine them. It should be examined thoroughly if the combination method really increases the performance. For example, most studies [3], [4] combine descriptors by concatenation, which is prone to errors due to the necessity of normalization and resulting deformation of the feature space.

2.3.2 Segmentation

Most of the studies in the first group of annotation literature [3], [4], [7] that are based on segmental approach, face with the problem of segmentation. Segmentation itself is an unsolved problem. In real life applications, most of the time, firstly context information is processed before segmenting an image. Besides this, vision researchers are not sure that the perception of an image really starts by perception of image regions, [18]. It is quite probable that the human visual system initially gets the context information then perceives the single objects in a scene. So, why not use this approach for the annotation problem?

2.3.3 Quantization of Feature Vectors

The problem domain of the initial studies in the segmental approaches are discrete. Features are extracted from the image regions or parts then these features are clustered to obtain blobs in [3], [4]. In the annotation, relations of words with blobs are modeled. However, during the quantization step, some information is lost. To overcome this problem, continuous models [9], [16] are proposed to, which outperformed the previous discrete techniques.

2.3.4 Training

Automatic annotation problem is appropriate to be considered both as a supervised and as an unsupervised classification problem. If it is considered as an unsupervised classification problem, as stated in Unsupervised Ensemble of Visual Descriptors(UEVD) [2], then the visual features are used to cluster the images and words of clusters is used for annotation of an unseen image. A variety of word selection scheme can be used in this approach. But it is not possible to talk about training in unsupervised approach as the class labels does not exist. On the other hand, if annotation is considered as a supervised classification problem then the annotation words can be considered as class labels and an image may belong to one or more classes. Another supervised classification approach is proposed in ALIP [6] and SEVD [2] which considers concepts as class labels where a concept may correspond to one or more words. In the supervised approach a variety of algorithms can be used for training of the system.

2.3.5 Data Set

Most studies in the literature, made experiments on the images of Corel Stock Photo Library. Similarly, we used a set of images from the Corel Stock Photo Library. In order to be able to

evaluate the performance and compare it with the other systems in the literature, we used the same training and testing sets with [2], [3], [11]. Although Corel Stock images are used widely, the data set has some disadvantages. For example, it contains several examples to annotation of city names like Hawaii, which are too high level for the current annotation systems. Furthermore, the data set have images with very high visual similarity (almost the same). So, most of the time nearest neighbor approaches performs well in the images of the Corel Stock Photo Library, which is also stated in [10]. In fact, it may not be considered as a direct disadvantage since some application domain may come up with a similar case as well.

CHAPTER 3

HANOLISTIC: A HIERARCHICAL IMAGE ANNOTATION SYSTEM

In this chapter, the proposed hierarchical image annotation system is described. The system formulates the annotation problem as a multivariate transformation from the low level visual information domain to high level semantic words. This formulation can also be considered as different version of multi-class classification problem, where each annotation word is considered as a class and an image belongs to one or more classes.

3.1 Introduction

It is well-known that the major goal of the image annotation problem is to relate the content information, representing the visual features, to the context information corresponding to the semantic words. In this study, we employ a two-level hierarchical stacked generalization architecture proposed by Wolpert [12], which consists of a set of annotators, each of which learns the probability mass function of the annotation words. For this purpose, we first extract the content information by variety of visual descriptors. Then, we train a set of annotators at the first level to estimate the probabilities of each word to belong to an image. This set of probabilities is assumed to provide the context information, which is then fed to a meta annotator to finalize the word assignment process to an unknown image.

In this thesis, we express the image by different aspects of visual information, using a variety of feature spaces extracted from the whole image, avoiding segmentation. Then, we formulate the unsupervised annotation problem as a kind of supervised classification problem, where the class labels correspond to the words in the document vocabulary. Proposed approach avoids the clustering of the visually similar features, corresponding to image segments as in [6], [2]. This approach employs the content and context information at the

same time, ranking the word assignment process by the class membership values obtained during the evaluation of the probability mass functions of words. Instead of modeling the relation between the image segments and annotation words as proposed in [3], [7], [11], [4], our approach utilizes general outlook of the image content from different perspectives. Using more than one feature space, enables us to relate the words carrying different visual information to the same image. One may use the result of annotation as an input to an intelligent segmentation to further improve the result of the initial annotation obtained from the holistic method proposed in this study. This combination of top down and bottom up approaches is quite consistent to the human visual system. Furthermore, it is already known that segmentation is a problematic issue in image processing and a good segmentation can be obtained only by means of using priori information about the image context.

In this thesis, we also aim to reveal the problems in the performance measures of the annotation systems. Although several measures are proposed in the literature, their usage in annotation problem should involve careful inspection of all criterion. So, in this thesis, these performance measures and related criteria are discussed thoroughly.

In this chapter, first the image representation model and description methods are presented. Then, the system architecture is described and the annotation scheme is explained. At the end of the chapter, brief information about the realization of the system is provided.

The notation used in the description of system is provided at Table 3.1.

3.2 Image Representation

In an annotation problem, image is represented by two major information levels: Firstly, image content is represented by low level visual features and secondly, the high level semantic information is represented by a set of words, called document.

Considering the first level of information, one may easily realize that the most critical step is the feature extraction to represent the raw image data. This is the most crucial step in defining the low level description space. Our experience shows that, current low level description spaces are far from describing a generic scene in a meaningful way, consistent with the human visual system. Especially, when a single description method is used only one aspect of the image can be represented. For this reason, we decide to use a set of low level features to describe various properties of color, shape and texture of images. However, using more than one descriptor requires detailed inspection of the description spaces and behavior of system at each space. Studies that use several descriptors, usually, describe the image with

a vector incurred by the concatenation of features extracted by these descriptors [3], [4], [9], [11], [17]. However, the concatenation of features from distinct description spaces necessitate the normalization of the obtained vector so that all the features are scaled to values in a certain range. Consequently, the original description space is deformed causing information loss. Moreover, the concatenation results in a large vector containing several features and this arises the curse of dimensionality problem. As the dimension of the description space increases, the number of samples may become statistically insufficient to describe the images in the high dimensional description space. In order to avoid these drawbacks, we construct separate annotation systems for each description space and make use of the results from all these annotation systems to assign keywords to a given image.

For the second level of information, semantic representation of the image is constituted by the annotation words. Each annotation word is considered as a class label and each image has a membership value for the words in the dataset. In other words, the system assign keywords to a given image by means of evaluating the word membership values of that image.

3.2.1 Low Level Visual Descriptors for Content Information

A popular set of descriptors can be found in MPEG-7 descriptors. MPEG-7 is a multimedia content description standard developed by MPEG (Moving Picture Experts Group) [19]. It is used for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code. It provides fast and efficient searching and content identification. Also, it describes the main issue about the content of a multimedia material and it can be used to index a big range of applications.

In this study, a subset of MPEG-7 Visual Descriptors are used in content representation of images. The selection of descriptors depends on the visual content of the images in the dataset. We use a set of images from the Corel Stock Photo Collection in this study and it is well known that color layout, color structure, scalable color, homogenous texture and edge histogram descriptors from MPEG-7 successfully represent the images in the Corel dataset [2]. An empirical study verifies the selection of the description space, as the description power of the descriptors varies depending on the problem domain.

Let us briefly explain the descriptors used in this study. Further details can be found in [20].

- **Color Layout:** It effectively represents the spatial distribution of color of visual signals in YCrCb color space by using discrete-cosine transformation (DCT). Its advantage is that, it does not depend on image format, resolution or bit-depths.
- **Color Structure:** It is a color feature descriptor that captures both color content and information about the structure of this content by sliding a structuring element of 8x8 pixels over the image. Its main functionality is image-to-image matching and its intended use is for still image retrieval. It is able to distinguish between two images in which a given color is present in identical amount but structure of the group of pixels having that color is different.
- **Scalable Color:** It is a color histogram in HSV color space, which is encoded by a Haar transform. Its binary representation is scalable in terms of bin numbers and bit representation accuracy over a broad range of data rates. It is useful for image-to-image matching and retrieval based on color feature. Retrieval accuracy increases with the number of bits used in the representation.
- **Homogenous Texture:** Homogenous Texture features are extracted using Gabor Filters in five different scales and six different directions. It provides a precise quantitative description of a texture that can be used for accurate search and retrieval.
- **Edge Histogram:** It represents the spatial distribution of five types of edges, namely four directional edges and one non-directional edge. It can retrieve images with similar semantic meaning. Thus, it primarily targets image-to-image matching, especially for natural images with non-uniform edge distribution.

Let, the number of descriptors used in the representation of images be D and the number of images be N . The i^{th} image is represented as I_i and the j^{th} descriptor is represented as Δ_j . For a given image, low level feature vectors are extracted for each descriptor Δ_j . Feature vector extracted from image I_i using the descriptor Δ_j is represented as δ_{ij} . Therefore, for each image I_i , for $i = 1, 2, \dots, N$, its description for all descriptors Δ_j , for $j = 1, 2, \dots, D$, is extracted and D different descriptions are obtained for the same image.

3.2.2 Semantic Words for Context Information

Context information is represented by the words, related to the image. It is assumed that low level content information and high level context information is somehow related to each other. Alas, in most of the practical problems there is a serious gap between the two, which

complicates the image annotation problem. This gap is referred as semantic gap problem in the CBIR literature and the proposed systems have been trying to bridge the gap between the complex semantic information and the simple visual information [1]. So far, researchers have not been able to create satisfactory solutions for bridging the content and context information.

In the proposed hierarchical architecture, high level description of image consists of its annotation words. The hierarchical architecture treats the annotation words as class labels and assumes that an image may belong to one or more classes.

High level descriptors, that is the words are represented in the hierarchical architecture as follows; the number of words in the data set is L , the l^{th} word in the dataset is represented as w_l . The document of an image, that is the words of an image I_i is represented as T_i where $T_i = \{w_{i1}, \dots, w_{im}\}$, and the j^{th} word of image I_i is represented as T_{ij} . Each image is described with at least one word and at most M words, $1 \leq m \leq M$.

3.3 Mathematical Definition of Image Annotation Problem

Annotation problem can be defined mathematically as follows; a training set S consisting of N images in set $I = \{I_i\}_{i=1}^N$ and their associated text documents in set $T = \{T_i\}_{i=1}^N$ such that, $S = \{(I_1, T_1), (I_2, T_2), \dots, (I_N, T_N)\}$, is given. Each image in the dataset is described by a set of visual descriptors, $I_i = \{\delta_{i1}, \delta_{i2}, \dots, \delta_{iD}\}$ where δ_{ij} is the feature vector representing the i^{th} image in the j^{th} description space. Each text document T_i consists of a set of words, $T_i = \{w_{i1}, w_{i2}, \dots, w_{im}\}$, where w_{im} corresponds to the m^{th} word of the i^{th} image, and $w_{im} \in W$ where $W = \{w_1, w_2, \dots, w_L\}$, L is the number of words in the dataset. Given a test image Q , problem is to assign a document A , which is obtained from the elements of W , to Q .

Each word in W is considered as a class label. Each image in I is associated to a set of images in the vocabulary W . While searching for the best set of words for a given image, each word is assigned a membership value to the image. This membership value, $p_{l,i}$, is referred as the word membership value and indicates the level of association between the image I_i and the word w_l .

Table 3.1: Notation for the description of the hierarchical architecture.

N :	number of images
D :	number of descriptors
L :	number of words
M :	max number of words assigned to an image
δ_{ij} :	feature vector extracted from the i^{th} image by using the j^{th} descriptor.
I_i :	the i^{th} image
w_l :	the l^{th} word
$p_{l,i,j}$:	the membership of the l^{th} word for the i^{th} image in the j^{th} description space
$p_{l,i}$:	the membership of the l^{th} word for the i^{th} image
\underline{P}_{ij} :	vector containing all the word membership values for the i^{th} image in the j^{th} description space such that $\underline{P}_{ij}=[p_{1,i,j} \dots p_{l,i,j} \dots p_{L,i,j}]$
\underline{P}_i :	vector containing all the word membership values for the i^{th} image obtained using the decisions of a set of description spaces $\underline{P}_i=[p_{1,i} \dots p_{l,i} \dots p_{L,i}]$
A_j :	the j^{th} level-0 annotator
T_i :	the document of the i^{th} image where $T_i = \{w_{i1}, \dots, w_{im}\}$, $1 \leq m \leq M$
w_{im} :	the m^{th} word of the i^{th} image where $m = 1, 2, \dots, M$
$Meta - A(\cdot)$:	the meta-level annotator

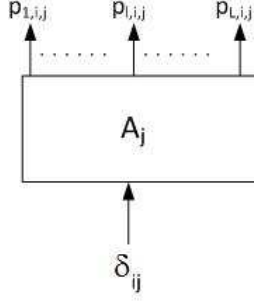


Figure 3.1: Level-0 Annotator.

3.4 Hierarchical Architecture for Annotation

We propose to solve the annotation problem defined above by means of a hierarchical learning architecture which consists of two layers. In the first layer, called level-0, information from all visual description spaces are processed separately and candidate annotation words and their membership values are estimated for a given image. These candidate words are assumed to provide context information for the given image. In the second layer, called meta-level, information provided by level-0, is considered and most probable words are assigned to an unknown image.

3.4.1 Level-0 Annotators

Level-0 consists annotators, which assign a membership value to each word in the vocabulary based on distinct low level visual features and the high level context information provided by the annotation words. Therefore, we have a set of descriptors at level-0. An annotator in level-0 is depicted in Figure 3.1 where, each annotator is shown as A_j , and $j = 1, 2, \dots, D$. For a given image I_i , an annotator A_j takes as input the low level description d_{ij} of the image I_i and gives as output the word-membership values of that image for all words $w_{l,j}$, for $l = 1, 2, \dots, L$. The output $p_{l,i,j}$ refers to the membership value of image I_i for word w_l under the description of the j^{th} visual description space. All membership values for the image I_i provided by annotator A_j is represented by $\underline{P}_{i,j}$ which is a vector constructed as follows:

$$\underline{P}_{i,j} = [p_{1,i,j} p_{2,i,j} \dots p_{l,i,j} \dots p_{L,i,j}] \quad (3.1)$$

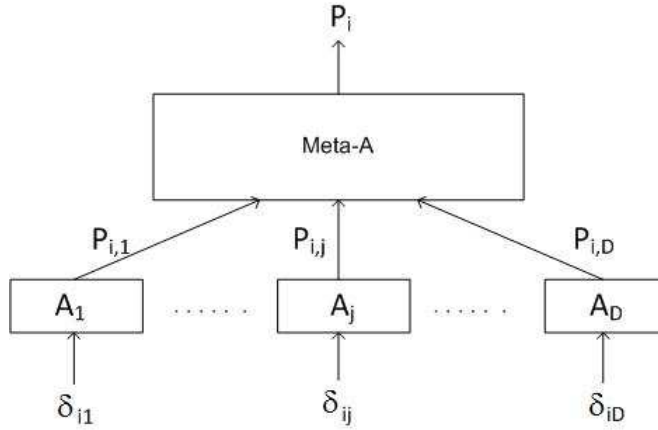


Figure 3.2: System Architecture.

3.4.2 Meta-Level Annotator

In the meta-level, a set of final annotation words is selected by aggregating the results of level-0. In other words, meta-level processes the output of all level-0 annotators. Meta-level is depicted in Figure 3.2. For a given image I_i , it receives the word membership values $\underline{P}_{i,j}$ produced by the level-0 annotators and outputs the final membership value \underline{P}_i of the annotation words for the image I_i .

3.5 Annotation Training

In a typical classification problem, given a set of labeled training data and a set of unlabeled test data, the aim is to find the labels of the test data using the training data. The basic problem defined in pattern classification is the two-class classification problem where each sample is assigned to one of priori known two classes. Then, multi-class classification problem, in which the number of priori known classes is larger than two and each sample is assigned to one of the classes, is defined. As the two-class classification problem is much easier, many researchers propose using two-class classifiers to solve the multi-class classification problem [21].

In the case of image annotation, it is possible to model the problem as a variant of multi-class classification. Given a set of observations I , representing the images and a set of words W in a dataset, annotation is defined as a multivariate transformation $T : \Delta \rightarrow W$. An image is represented by its visual features $\delta_{ij} \in \Delta$. In other words, annotation can be considered as a multivariate mapping from the low level visual space to the high level semantic space, where each image can be assigned to more than one class corresponding to

a word.

When considered in the view of the classical pattern classification methods, the above multivariate mapping can be formulated as a multi-class classification problem, where a visual descriptor representing a sample image belongs to one or more classes. Therefore, one of the well-known classifiers can be employed for each descriptor to train the system at level-0. Each annotator at level-0 is trained for a single visual description space by means of a classifier. In the meta-level, result of all level-0 annotators are aggregated to obtain the final classification of a given sample, where the sample is assigned to one or more classes.

3.6 Automatic Annotation

Now, let us explain how the automatic annotation of an unknown image is accomplished with the proposed architecture. First, visual features of the image is extracted for all low level description spaces. Then, each of the obtained feature vector is fed to a distinct level-0 annotator where the word membership values are estimated for all words. Estimation of the membership values can be based on several criteria and a wide range of algorithms can be applied to estimate the membership values. After the membership values are evaluated by all level-0 annotators, they are fed to the meta-level to be processed to incur the final word membership values. In the meta-level the most straightforward approach to combine the results of level-0 annotators is to sum up their results for all words individually. Several other combination methods can be applied which will be discusses in the following sections. Once the final word membership values are obtained from the meta-level, a set of words with the highest membership values is selected as the keywords of the given image. The annotation process is described in Algorithm 3.

3.7 Performance Measures

Annotation problem has aroused during the studies on CBIR and it is related to CBIR, as we have already mentioned in Chapter 1. As a consequence, the performance measures for image annotation are adapted from the CBIR performance measures. The most widely used performance criterion are precision and recall. In the annotation system, these criterion are evaluated on per word basis. For a word w , **precision** is the number of correct annotations with this word divided by the number of annotations with this word. Equation 3.2 is the formula evaluating the precision of word w , where $\#w_{correct}$ denotes the number of correct

Algorithm 2 Automatic annotation scheme of HANOLISTIC

Require: An image I_i

Ensure: Assign annotation words to the given image I_i

for each description space j (from 1 to D) **do**

 Extract feature vector, d_{ij}

 Feed d_{ij} to the level-0 annotators A_j and get the membership values \underline{P}_{ij}

end for

Feed the results of the level-0 to the meta-level annotator $meta - A(\cdot)$ to obtain the final memberships \underline{P}_i

Select words in \underline{P}_i with the highest membership values

annotations with w and $\#w_{annotations}$ denotes the total number of annotations with w .

$$w_{precision} = \frac{\#w_{correct}}{\#w_{annotations}} \quad (3.2)$$

While, **recall** of a word w is the number of correct annotations with this word divided by the number of annotations with this word in the ground truth. The formula for recall of w is provided in Equation 3.3, where $\#w_{correct}$ denotes the number of correct annotations done with w and $\#w_{ground}$ denotes the total number of annotations with w in the ground truth.

$$w_{recall} = \frac{\#w_{correct}}{\#w_{ground}} \quad (3.3)$$

Precision and recall values are evaluated per word and the mean-per-word values are computed to give the system performance.

Another criteria used in the evaluation of annotation performance is the **mean average precision**. This is proposed based on the idea that we do not know exactly how many words should be used in the annotation of an image. So, instead of annotating with a certain number of words, a ranking is provided. Mean average precision emphasizes returning the related results as early as possible. For a given word, images are ranked according to their membership to this word and Equation 3.4 is applied to this ranking to find the average precision of this word, where r is the rank, N is the number retrieved, $rel()$ is a binary function on the relevance of a given rank and $P()$ is the precision at a given cut-off rank. After the average precision is evaluated for all words, their mean is computed to give the mean average precision.

$$AveP = \frac{\sum_{r=1}^N (P(r) \cdot rel(r))}{numberofrelevantdocuments} \quad (3.4)$$

Coverage percentage is another criteria used as a performance measure in the automatic image annotation systems. First, each image is annotated with certain number of words then for each image the percentage of true annotation is evaluated. For a given image I , the coverage percentage CP is computed with Equation 3.5, where I_{true} refers to number of true annotations for image I and I_{all} refers to the number of all annotations for image I . This measure does not penalize wrong annotations. Consequently, as the number of annotation words the system returns increases, coverage percentage increases too.

$$CP = \frac{I_{true}}{I_{all}} \quad (3.5)$$

F-measure (F-score) is the weighted harmonic mean of precision and recall and is evaluated with the following formula:

$$F = \frac{2 \cdot precision \cdot recall}{(precision + recall)} \quad (3.6)$$

F-measure is useful when a single performance measure is needed. It is more reliable than the coverage-percentage as coverage-percentage is insufficient to take false annotations into consideration.

3.8 Discussion on the Pros and Cons of HANOLISTIC

The hierarchical architecture has the following principles:

1. **Simplicity:** Most of the state of the art automatic image annotation methods such as Probabilistic Latent Semantic Analysis [5] and Cross Media Relevance Model [4] involves expensive computation of the joint probability density functions of image regions and annotation words. During these computations many assumptions on the statistical independence among the regions and words are done. The proposed system avoids these type of assumptions and complex computations of algorithms. Most of the studies in the literature are based on the segmental approach, which segments the image and tries to find relations between the image regions and semantic words. These studies, not only assume that the segmentation correctly partitions the image into semantically meaningful regions [3], but they also assume that observing the regions in an image and the semantic words, which express the image, are mutually independent

[4]. These assumptions introduce uncontrolled and unknown error to the system from the initial step of segmentation through the last step of statistical analysis of data.

2. **Simultaneous Processing of Content and Context Information:** The critical issue in annotation problem, is making use of semantic and visual information at the same time to overcome the semantic gap problem. In the state of the art image annotation systems, first low level visual information is extracted and then clustered to generate a visual code-book. Then, the semantic words are associated to the visual words via some probabilistic techniques [3], [4], [11]. However, this scheme causes propagation of error which aroused during the visual information processing. For this reason, Sayar and Yarman-Vural [22] proposed using semantic information as a constraint while processing the visual information. In HANOLISTIC, we integrate the semantic information as a supervision tool for learning in the very first stages of computation so that the relation between the low level and high level information is established. This approach enables us to avoid unsupervised clustering of visual data, which assumes no correlation to the semantic words at the initial phase of the annotation.
3. **Holistic View of Image Through Different Perspectives:** Most of the state of the art annotation systems, describe a given image by means of a set of features extracted from a number of image regions. In this method, certain image properties are considered by means of a single description method. However, this approach lacks the ability of describing the visual properties of an image from different perspectives. For this reason, Akbaş and Yarman Vural [2] proposed using ensemble of descriptors for representing various texture, color, and shape properties of images. In our system, we also describe the whole images by ensemble of descriptors.

3.9 Realization of the System

HANOLISTIC is realized by using very popular fuzzy k-nn algorithm [23]. Systems are fed by the popular MPEG-7 descriptors [20].

3.9.1 Realization of the Level-0 Annotators

It is well known that nearest neighbor approaches such as [2] and [10], perform considerably well in many pattern recognition problems. Tang, in [10], uses the nearest-neighbor approach,

considering the neighbors' words while annotating a given image. On the other hand, Akbas in UEVD [2] uses fuzzy k-nearest neighbor approach for clustering the training images and finds the corresponding cluster for a given image then selects a subset of the words of that cluster. Observing the success of systems, we employ the supervised version of fuzzy k-nearest neighbor algorithm. In our approach, labels to a sample image are assigned by looking at its k-neighbors' labels together with the distance of the neighbor from the sample image. Algorithm assigns high probability to words that appear in close neighborhood. This approach has two advantages:

1. **Fuzzy Logic and the Principle of Least Commitment:** The Principle of Least Commitment [18] is proposed by David Marr for the development of intelligent computer vision algorithms. The idea is stated as 'Don't do something that may later have to be undone.' In this view, before making a crisp decision, utilize degree of membership as long as possible. The adaptation of this principle follows directly for each image to belong to one or more classes (words). Hence, the images can be considered as having membership values for words. In order to annotate an image, words with the highest membership values is selected. And this selection takes place at the end of the processing.
2. **Supervised Learning vs. Unsupervised Learning:** Instead of directly clustering images and then selecting words, supervision is integrated to the learning system for assigning the class labels. That is, annotation is formulated as a supervised classification problem. We believe that this probabilistic approach and the supervision plays an important role in the superior results obtained during our experiments.

Another reason for choosing the k-nearest neighbor algorithm for the level-0 annotators is, its simplicity and fast computation relative to other approaches.

The Fuzzy k-Nearest Neighbors Rule

Fuzzy k-NN is proposed by Keller *et al.* in 1985 [23]. The idea behind the algorithm lies in Marr's principle of least commitment, which is also confirmed by Keller in [24]. In the crisp k-nn algorithm, membership information is lost once a sample is assigned to a certain class, while in the fuzzy version all information from k-neighbors are combined. Especially, in an annotation problem, one needs to find the membership of a sample for all words. For this particular problem, the fuzzy logic gains more importance than a classical classification problem.

In the proposed hierarchical architecture; given a test image I_i , the membership values of the image for all words are evaluated as follows:

1. Find the k -nearest-neighbors of the image
2. Apply the formula in Equation 3.7 for a word corresponding to a class l , to find the membership of image for each word. In this equation, $p_{l,i,j}$ refers to the membership value of image I_i for class l in the j^{th} description space.

$$p_{l,i,j} = \frac{\sum_{k=1}^K p_{l,k,j} \left(\frac{1}{\|\delta_{ij} - \delta_{kj}\|^{\frac{2}{m-1}}} \right)}{\sum_{k=1}^K \left(\frac{1}{\|\delta_{ij} - \delta_{k\text{æ}V\text{ert}}\|^{\frac{2}{m-1}}} \right)} \quad (3.7)$$

The denominator in equation 3.7 is a normalization factor. And m is a scaling factor used to scale the distance between the images I_i and I_k . One of the disadvantages of this approach is that m is arbitrary. However, there are two meaningful ranges for the values of m .

- (a) If $m < 1$ then as m gets smaller the influence of distant samples on the annotation of I is increased.
- (b) If $m > 1$ then as m gets greater the influence of distant samples on the annotation of I is decreased.
- (c) As m approaches $\pm\infty$ then the annotation result approaches to the annotation by crisp k-nn.

Note that, for each descriptor δ_{ij} extracted from image I_i , $i=1,2,\dots,N$, we obtain a set of membership values $\underline{P}_{L,i}$ for words $W = \{w_1, w_2, \dots, w_L\}$, yielding total of $(D \times L)$ membership values.

In two steps, membership value of I for all words are assigned. For those words that do not appear in the neighborhood of k has the membership value equal to 0. This step is repeated for all level-0 is annotators.

3.9.2 Realization of the Meta-Level Annotator

Several algorithms can be used at the meta-level annotator, where Annotation is performed based on word membership values \underline{P}_{ij} , with $\underline{P}_{ij} = [p_{1,i,j} p_{2,i,j} \dots p_{l,i,j} \dots p_{L,i,j}]$, coming from level-0 annotator. Meta-level is the final annotator, which outputs the word membership values \underline{P}_i for image I_i .

Summation Annotator

Since the level-0 annotator outputs a set of independent membership values assuming that the reliability of annotators are all equal, summation of word membership values is a suitable approach for the meta-level of the annotation system. Therefore, a straightforward approach is to simply add the membership values. Then, we assign the top M words with highest membership values to the unknown image. Mathematically, for i^{th} image the word membership values are evaluated using Equation 3.8, where $\underline{P}_{i,j}$ is the word membership vector for image I_i in the j^{th} description space. Intuitively, \underline{P}_i represents an overall score for a word to belong image I_i .

$$\underline{P}_i = \sum_{j=1}^D \underline{P}_{i,j} \quad (3.8)$$

There is a close relation between the summation annotator and the well-known majority voting approach. What we apply here is not directly voting but still the idea of democracy is adapted by the summation annotator. For a typical combination of classifiers problem, Kittler in [25] states that using the sum as the combination rule outperforms other classifier combination schemes. Similarly, in our empirical study, we examined the advantage of summation over the other combination schemes.

Weighted Summation Annotator

An alternative to the summation annotator is weighted summation annotator, where each annotator is assigned reliability values and these values weight the result of annotators while summing up the results. If weighted summation is to be used at the meta-level, then a weighting principle should be decided. The idea that follows directly is to evaluate the performance of each level-0 annotator and determine weights accordingly. For this purpose, system needs to be trained by the cross-validation technique to learn the weights. A subset of training images, consisting of 500 images, is selected randomly to evaluate the weights. At this point two approaches can be used:

1. **Overall performance based:** In this method, for each level-0 annotator, the precision and recall values are evaluated over the validation set for all words. Then, mean precision and mean recall are computed over all words for that annotator. Using these values, the F-score is computed by using Equation 3.6. F-score is used as the weighting factor for the annotator. Weighting factors, computed for all level-0 annotators are used in the meta-level to find the final membership values with the following equation.

$$\underline{P}_i = \sum_{j=1}^D w_j \underline{P}_{i,j} \quad (3.9)$$

2. **Per word performance based:** It is, also, possible to weight each level-0 annotator for each word. In this case, F-score values of a level-0 annotator, for each word is computed using a validation set. Then, the final word membership value w_l of an image I_i is found, using Equation 3.10. Similarly, word membership values for all words are evaluated for image I_i , as

$$p_{l,i} = \sum_{j=1}^D w_{l,j} p_{l,i,j} \quad (3.10)$$

where, $p_{l,i}$ is the membership value of the l^{th} word for the i^{th} image, $w_{l,j}$ is the weighting factor of the j^{th} level-0 annotator for the l^{th} word and $p_{l,i,j}$ is the membership value of the l^{th} word for the i^{th} image in the j^{th} description space.

Selection of Maximum

One alternative to the voting scheme for annotation is the selection of maximum membership value for a word among the outputs of all level-0 annotators. In this case, membership values, coming from level-0 annotators are considered for each word. The maximum membership value for a given image is selected. The formula for the maximum membership selection is as,

$$p_{l,i} = \max_j p_{l,i,j} \quad (3.11)$$

where $j = 1, 2, \dots, D$.

Whichever voting scheme is used, to finalize the annotation, words with the highest membership values are selected as the annotation words.

3.9.3 Automatic Image Annotation by HANOLISTIC

Let us explain automatic image annotation process in HANOLISTIC for the described realization techniques, which is fuzzy-knn for the level-0 annotators and summation annotator for the meta-level. Given an image I_i , it is processed by level-0 annotators separately. For each description space, its features δ_{ij} are extracted, and then fed to a level-0 annotator. Level-0 annotator finds the K nearest neighbors of the image based on the Euclidean distance between the feature vectors of images. After the neighbors are determined, for each word

w_l a membership value $p_{l,i,j}$ is computed using equation 3.7, where $p_{l,i,j}$ is the membership value of l^{th} word for the i^{th} image in the j^{th} description space. Word membership values for image I_i in the j^{th} description space is referred as $\underline{P}_{i,j}$. At the meta level, assuming that a summation annotator is implemented, output of level-0 annotators are summed up by equation 3.8. In the end, a final word membership vector \underline{P}_i is obtained and M words with the highest membership values are selected as the annotation words of the image I_i . The algorithm of HANOLISTIC for automatic image annotation is provided in Algorithm 3. The algorithm can be modified to apply weighted summation annotator or the selection of maximum method by changing the formula 3.13.

Algorithm 3 Automatic image annotation by HANOLISTIC

Require: An image I_i

Ensure: Assign annotation words to the given image I_i

for each description space j (from 1 to D) **do**

 Extract feature vector, δ_{ij}

 find the K nearest neighbors of I_i in the j^{th} description space

for each word w_l (from 1 to L) **do**

 evaluate the word membership value $p_{l,i,j}$ by

$$p_{l,i,j} = \frac{\sum_{k=1}^K p_{l,k,j} \left(\frac{1}{\|\delta_{ij} - \delta_{kj}\|^{\frac{2}{m-1}}} \right)}{\sum_{k=1}^K \left(\frac{1}{\|\delta_{ij} - \delta_{kj}\|^{\frac{2}{m-1}}} \right)} \quad (3.12)$$

end for

$\underline{P}_{ij} = [p_{1,i,j} \dots p_{l,i,j} \dots p_{L,i,j}]$ is obtained

end for

evaluate the final word membership values \underline{P}_i at the meta-level

$$\underline{P}_i = \sum_{j=1}^D \underline{P}_{i,j} \quad (3.13)$$

Select M words in \underline{P}_i with the highest membership values

CHAPTER 4

EMPIRICAL STUDY

In this chapter empirical studies to show the power of HANOLISTIC is explained. The system is examined thoroughly, over several parameters and the results are reported. The hierarchical and the holistic approach is compared to the other automatic annotation systems in the literature. Experiments are conducted on MATLAB (R2007A) environment.

4.1 Experimental Setup

4.1.1 Data Set

A subset of Corel Draw Photo Collection is used in the experiments, which is the same dataset as in [2], [3], [4], [9], [11] and [26] to be able to compare the performances. In this dataset, there are 5000 images each annotated by a set of words, where the number of annotations for the images varies from one word to five words. There are 374 distinct words in the dataset. Some images and their annotations are illustrated in Table 4.1, Table 4.2 and Table 4.3.

The dataset is partitioned into two, with 4500 training images and 500 test images. 500 images are selected from the 4500 training images to be used for validation purposes. The number of words in the test set is 263, and 260 of them also take place in the training set. Thus, ideally it is possible to annotate only 260 of the words. The number of annotation words associated to each image varies between one and five. Therefore, how many words are required to annotate a test image is known precisely. Although this allows a flexibility for the number of word in annotation, it brings a bias to the precision, recall and coverage percentage while measuring the performance.

The frequency of the words varies greatly, which brings another bias to the annotation problem. Those words, which appear more frequently tend to dominate the result over less

Table 4.1: Sample images from the 5000 image Corel Dataset whose annotations are directly related to the image content.



Table 4.2: Sample images from the 5000 image Corel Dataset whose annotations involve abstract level of information.

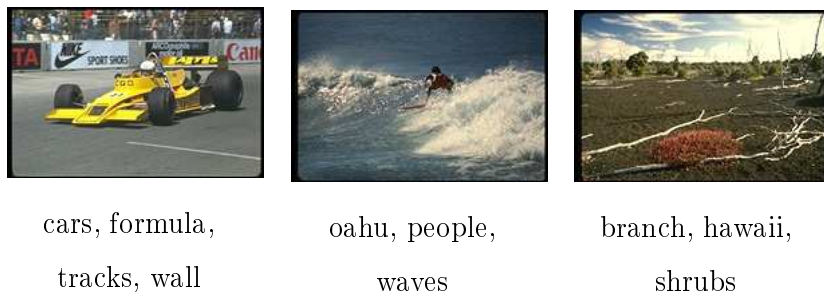


Table 4.3: Sample images from the 5000 image Corel Dataset whose annotations are not directly related to the image content.

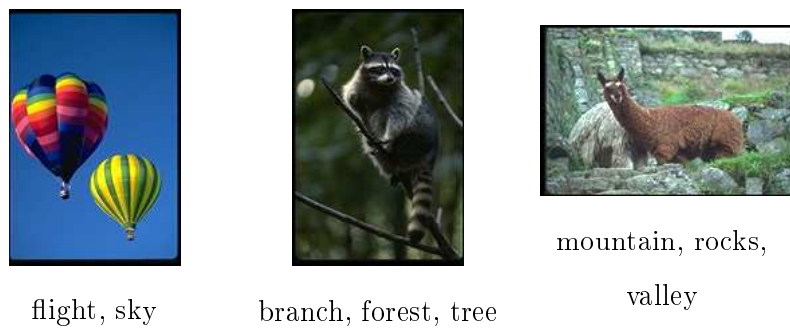


Table 4.4: The best k -values determined on the test set.

	Color Layout	Color Structure	Scalable Color	Homogenous Texture	Edge Histogram
k	9	2	5	17	15

frequent words. Consequently, the systems usually, miss the infrequent words in annotation. In order to overcome this problem, several techniques have been proposed, such as the rare words scheme of Wang [6]. We suppress the dominance of frequent words by constraining the neighborhood determined by the k -value and avoid to make decisions based on large clusters. The effect of increasing k -value is further described in Section 4.2.1.

4.1.2 Selection of Descriptors

The selection of the descriptors should represent various visual properties of Corel Dataset. It is well-known that the dataset mostly carries color and texture information. Considering this fact, five MPEG-7 visual descriptors, given in Chapter 3 are employed, namely; color layout, color structure, scalable color, homogenous texture and edge histogram. MPEG-7 Visual Descriptor features are extracted using the XM(eXperimentation Model) software which is made available on the web by Stephan Herrmann[27]. After the features are extracted, a level-0 annotator is constructed for each of the visual description spaces. In this system, we make use of the power of independent descriptors to represent the visual information.

4.1.3 Estimating the Best k -Values for Fuzzy-knn

The best k -values are estimated by running the system for a set of k -values on the test samples. In this experiment, we examine the performance of individual descriptors. Various k -values are tested and the k -values which maximize the performance is selected as shown in Table 4.4.

4.1.4 Exploring the Performances of Individual Descriptors

Let us, explore the performance of individual descriptors. In our hierarchical annotation system this corresponds to observing the performance of level-0 annotators. Once all the test images are annotated by all level-0 annotators, we examine the annotations of each level-0 annotator individually. For this purpose, we compute the precision and recall value

Table 4.5: Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by color layout features.



pair, fox, **den**,
rocks



den, fox, tree



arctic, fox, **den**,
grass



water, **whales**



beach, water,
whales



water, **whales**



bulls, fields,
moose, water



antlers, **moose**



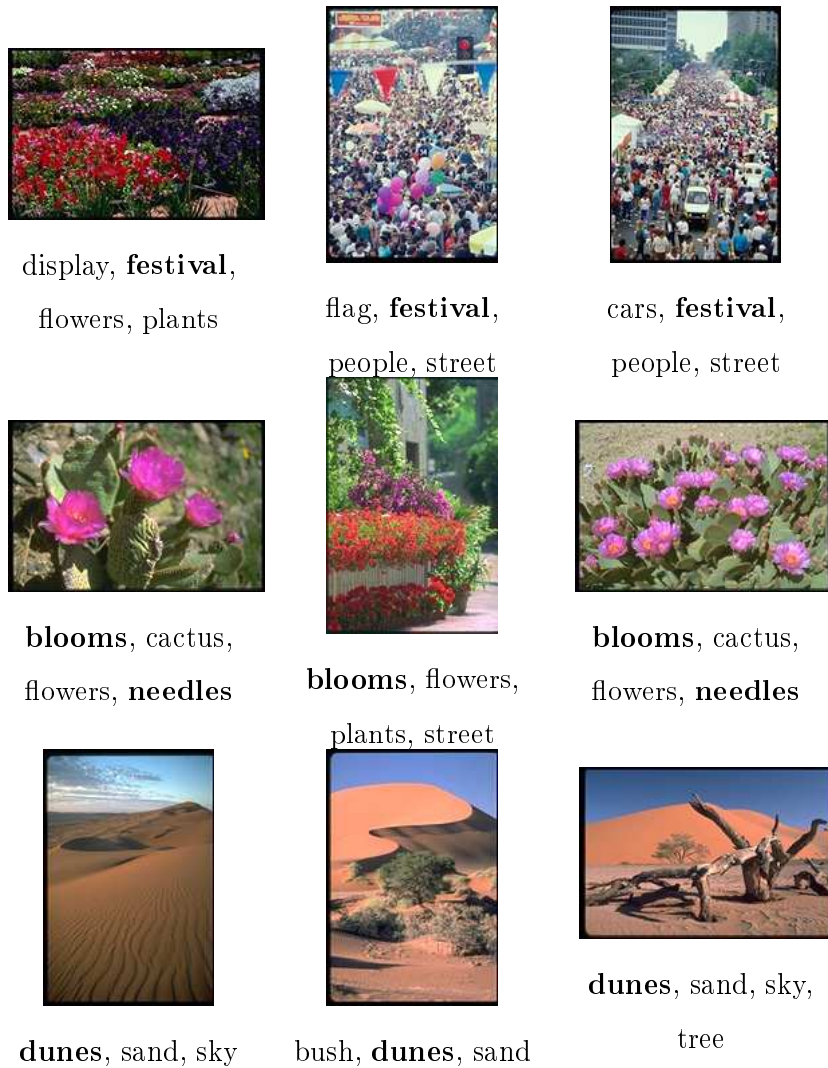
antlers, fields,
moose

of each annotator for each word and then select a set of words with highest precision and recall values. We noticed that each descriptor has a discriminative power over a subset of the annotation words. This is quite natural, because the visual properties which dominates the image may change for each word.

Annotation with **Color Layout** features, has the highest precision and recall values for the words; 'den', 'relief', 'moose', 'formula', 'whales', 'nest', 'pillar' and 'mosque'. Among these words 'den', 'relief', 'moose' can be correctly used in annotation only by color layout features. A close look at the images with these words indicates that these words can be discriminated by the spatial distribution of colors.

Annotation with **Color Structure** features, has the highest precision and recall values for the words; 'festival', 'needles', 'blooms', 'dunes', 'sphinx', 'mare' and 'foals'. Among these words 'festival', 'needles', 'blooms', 'dunes' and 'sphinx' can be correctly used in annotation

Table 4.6: Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by color structure features.



only by the color structure features. These words basically correspond to the color textured regions in the images.

Annotation with **Scalable Color** features, has the highest precision and recall values for the words; 'whales', 'dance', 'outside', 'mare', 'foals', 'sun', 'jet', 'plane', 'formula', 'horses', 'swimmers'.

Annotation with **Homogenous Texture** features, has the highest precision and recall values for the words; 'flight', 'tracks', 'formula', 'zebra', 'polar', 'jet', 'reefs'

Annotation with **Edge Histogram** features, has the highest precision and recall values for the words; 'formula', 'tracks', 'turn', 'runway', 'sky', 'water'. Word that can be used in

Table 4.7: Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by scalable color features.

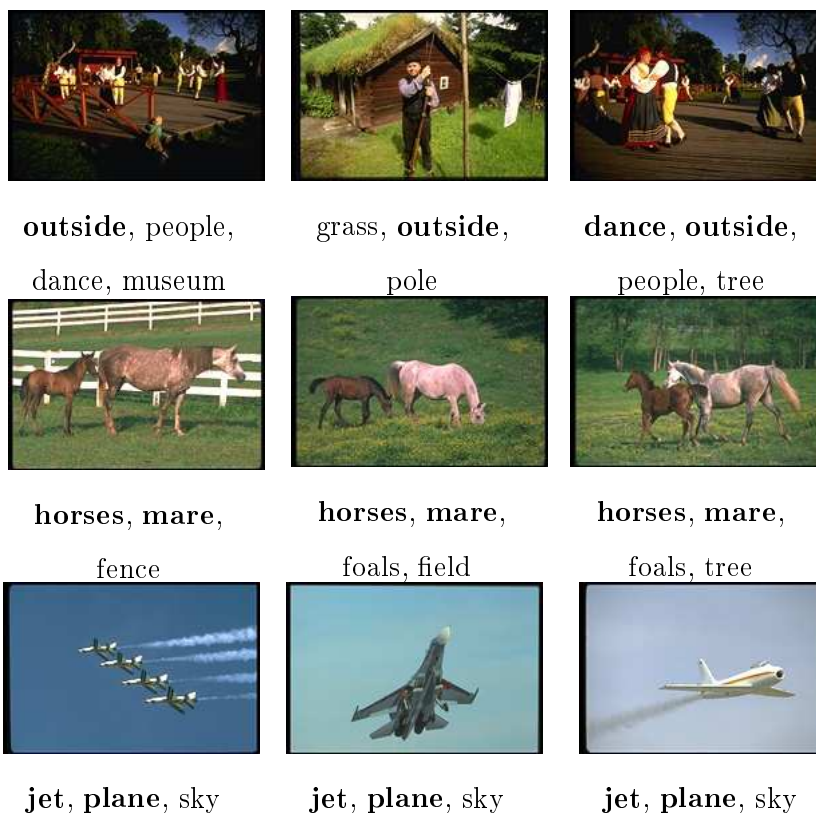
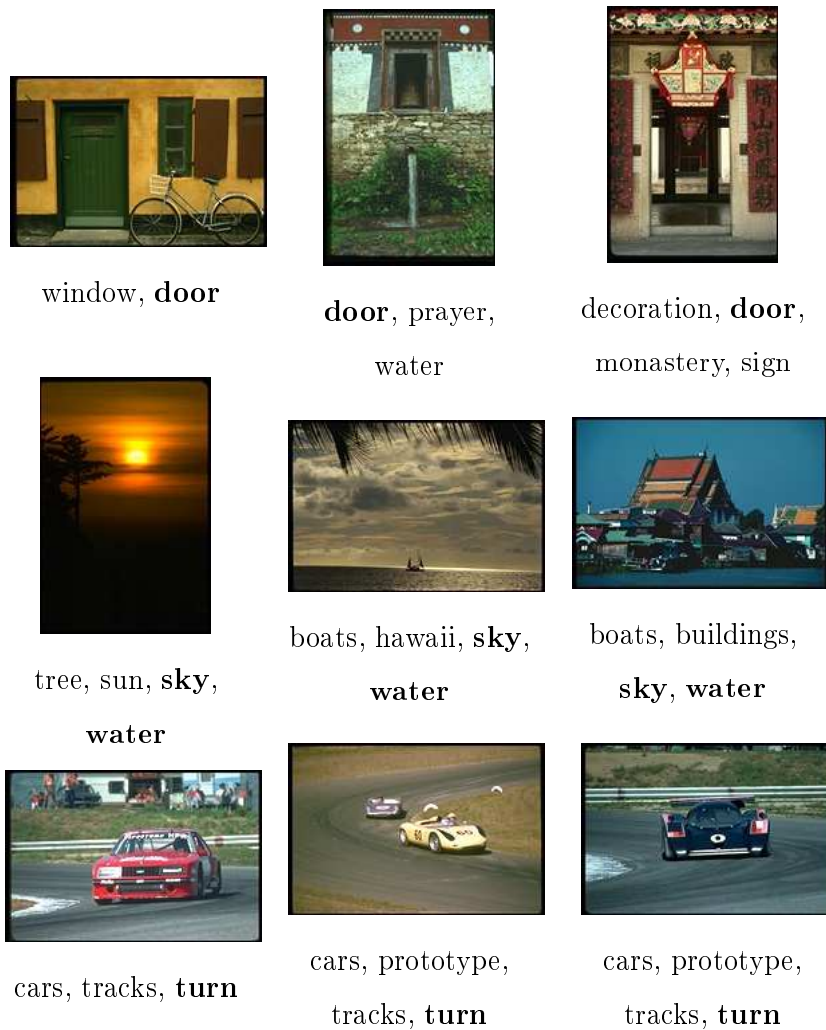


Table 4.8: Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by homogenous texture features.



Table 4.9: Sample images, first two columns from the training set and the last column from the test set, which can be successfully discriminated by edge histogram features.



annotation only by edge histogram is 'door' but its precision is 0.5 implying that only half of the annotations with 'door' is correct.

In conclusion, the annotators with different description of images share and gain the expertise to recognize different set of words. This is one of the main reasons for the high generalization performance of HANOLISTIC.

4.1.5 Exploring the Overall Performance of Descriptors in HANOLISTIC

It is observed that, the proposed hierarchical annotation system has the highest precision and recall values for the words; 'pool', 'bengal', 'den', 'needles', 'blooms', 'dunes', 'mosque', 'sphinx', 'whales', 'dance' and 'formula'. These words have high precision and recall values

Table 4.10: Sample words with relatively high precision and recall values for level-0 annotators based on the five descriptors and for HANOLISTIC with summation annotator at the meta-level.

Descriptors	words							
	'pool'		'bengal'		'mosque'		'formula'	
	rec.	prec.	rec.	prec.	rec.	prec.	rec.	prec.
CL	0.54	0.54	0	0	1	0.50	0.75	0.75
CS	0.81	0.81	1	0.66	1	0.50	1	0.80
SC	0.54	0.46	0.16	0.33	0	0	0.75	0.50
HT	0	0	0	0	0	0	1	0.80
EH	0.09	0.25	0	0	0	0	1	0.66
HANOLISTIC	0.90	0.83	0.83	1	1	1	1	1

in at least one of the level-0 annotator. Among these words, the most interesting ones and their precision and recall values are shown in Table 4.10. These words show the performance increase at the meta-level or due to expertise of the distinct classifiers. For example, some of the word 'mosque' is correctly annotated by the level-0 annotator fed with the color layout description and some other "mosques" are annotated by the level-0 annotator fed with the color structure description. At the meta-level the precision and recall become 1, which indicates that all the images annotated as 'mosque' in the dataset is correctly annotated by HANOLISTIC. Another interesting example is the word 'pool', whose precision and recall increase by combining the results of level-0 annotators at meta-level.

4.2 Automatic Annotation Performance

In this section, we discuss estimation of parameters for the level-0 and meta-level annotators.

4.2.1 Setting Parameters for Fuzzy-knn at Level-0

First, the **neighborhood parameter** k of fuzzy-knn is set by means of leave-one-out cross validation technique. In this method, each training image is treated as a test image and the system is tested for several k -values. After all the training images are considered, the k -value, which yields the highest precision and recall is determined. The k -values evaluated

Table 4.11: The parameter k determined by cross-validation for each descriptor.

	Color Layout	Color Structure	Scalable Color	Homogenous Texture	Edge Histogram
k	6	4	4	4	4

Table 4.12: Performance of HANOLISTIC for a function of increasing k -values.

k	Mean-per-word precision	Mean-per-word recall	# words rec>0
5	0.37	0.25	112
10	0.44	0.22	101
15	0.44	0.20	96
20	0.46	0.18	88

by this method are shown in Table 4.11.

We run the system for increasing values of k and observed the behavior of precision and recall. In this experiment, at each step k parameter is set to a constant value for all descriptors and system is run under the same k -values for all annotators. Table 4.12 shows the performances of annotators observed for different k -values. As it is shown in the table, recall decreases as k increases. This is not surprising because, as the number of neighboring images increases, the frequent words are observed in a wide-neighborhood, which dominates the annotation.

Second, the **scaling parameter** m is set by considering two properties stated in Chapter 3. If $m < 1$ then distant samples have greater affect on the annotation of an image. On the other hand, if $m > 1$ then distant samples have smaller affect on the annotation of unknown an image. We run the system for several values of m which are depicted in Table 4.13 and decided that the best performance is obtained for $m = 2$. Therefore, in our experiments we set m equal to 2.

4.2.2 Performance of Level-0 Annotators

System is run with the k -values determined by cross-validation and individual performances of level-0 annotators is observed as depicted in table Table 4.14. In the table, annotator

Table 4.13: Performance of HANOLISTIC with summation annotator method at the meta-level for several values of m .

m	Mean per-word precision	Mean per-word recall	# words rec>0	F-score
0.5	0.37	0.19	97	0.25
1.5	0.35	0.23	105	0.27
2	0.38	0.22	104	0.28
2.5	0.39	0.21	103	0.28
3	0.39	0.21	103	0.28

Table 4.14: Performance of Level-0 annotators.

Descriptor fed to the Level-0 Annotator	Mean per-word precision	Mean per-word recall
Color Layout	0.14	0.14
Color Structure	0.27	0.28
Scalable Color	0.16	0.15
Homogenous Texture	0.09	0.10
Edge Histogram	0.11	0.12

working on color structure features has the highest performance. This is due to the fact that, color features are very important features for the Corel Dataset.

After observing the individual results, we decided to explore the system with several combinations whose results are shown Table 4.15. Results in this table, imply that the system performs even better without the edge histogram descriptor. On the other hand, color layout, scalable color and color structures have significant contribution while the homogenous texture has little contribution which is still important for the overall system performance.

4.2.3 Exploring the Performance of Meta-Level

Summation Annotator is a straightforward solution for the meta-level. Given an unknown image, it applies Equation 3.8 to the results of level-0 annotators and then five words with the highest membership values are assigned to the image as annotation words. Most of the automatic annotation systems assign five words to each image [11], [4], [2], [3], [10], [9]. In

Table 4.15: Performances for several descriptor combinations.

Descriptors	Mean per-word precision	Mean per-word recall	# words rec>0	F-score
CL,CS,SC,HT	0.37	0.24	111	0.29
CL,CS,SC,EH	0.34	0.22	109	0.27
CL,CS,EH,HT	0.38	0.22	104	0.28
CL,SC,EH,HT	0.30	0.18	89	0.22
CS,SC,EH,HT	0.34	0.22	107	0.27
CL,CS,SC	0.31	0.25	113	0.27
CL,CS,SC,EH,HT	0.38	0.22	104	0.28

order to compare the performance of HANOLISTIC with those systems, we also assign five words to each image. When the results are explored for this method, it is observed that, the number of words that is used in annotation is 147 and the number of words that at least once annotate an image correctly is 103 (number of words with recall > 0). This implies that 44 words are never used in correctly annotating the images. Performance measures evaluated for the hierarchical architecture with summation annotator is provided in Table 4.16.

In the case of weighted summation annotator we use two methods: The first method is called as the **overall performance based** weighting. It assigns weights to each level-0 annotator based on the F-scores of the individual annotators. In this method, F-score of the annotators are evaluated over a validation set of images selected randomly among the training images. In the meta-level, Equation 3.9 is used to combine the results of level-0 annotators and five words with the highest membership values are selected for the annotation. Since the validation set is selected randomly, weights and consequently performance measures varies slightly at each run of the system. An instance of the results is provided at Table 4.16. In this, method on the average 167 number of words are used in annotation and 113 of them are used correctly at least once.

Second method, which is referred as **per word performance based** weighting, evaluates the f-score values of each annotator for each word and employs the Equation 3.10 to the output of level-0. Then, five words with the highest membership values are selected as annotation words. Again a validation set of images is selected randomly among the training images. An instance of the results for this method is, also, provided in Table 4.16. 155 words

Table 4.16: Performance of HANOLISTIC over 263 words for several meta-level techniques.

Technique	Mean Prec.	Mean Rec.	# words recall>0	F-score
summation annotator	0.39	0.22	103	0.28
overall weighting	0.35	0.24	113	0.29
per-word weighting	0.32	0.20	98	0.25
max. selection	0.26	0.20	97	0.22

are used in annotation and among these words, 98 are used correctly at least once.

The last method, that is experimented in meta-level is the **maximum selection**, which selects the maximum membership value for each word among the outputs of level-0 annotators and obtain a membership vector including membership values for each word. Finally, five words with the highest membership values are selected. This method uses 159 words in annotation among which 97 have recall > 0 . Performance measures are presented in Table 4.16.

4.2.4 Exploring the Overall System Performance

Let us investigate the overall performance of HANOLISTIC as a function of the number of annotation words, and find answer to the question: How do precision and recall change as the number of annotation words increases? While comparing the system with state-of-the-art annotation systems, we assign each image five words, which has become a standard in the comparison of the proposed methods in the literature. Let us now, observe the response of coverage percentage to the change in the number of words assigned to test images. We already know that the coverage percentage increases as the number of assigned words increase. We need to see to what extend, the coverage percentage can be increased without assigning too many words. For this purpose, number of words assigned to an unknown image is set as 5, 10, 25, 50 and 100 and the observations are reported in Table 4.17. This result is quite impressive in the sense that we may reduce the vocabulary from 374 to 100, yet get a coverage percentage of 0.85. One may employ this result to segment the images under the supervision of the annotation word. This segmentation is expected to yield semantically more meaningful regions compared to an initial segmentation followed by clustering for the generation of a code-book.

Table 4.17: Effect of increasing the number of assigned words on the performance of per-word based weighted summation annotator.

# words	mean-prec.	mean-rec.	# words recall>0	cov-per.
5	0.35	0.24	113	0.52
10	0.21	0.35	141	0.62
25	0.08	0.48	174	0.73
50	0.04	0.56	186	0.80
100	0.04	0.65	200	0.85

Table 4.18: Comparison of HANOLISTIC with other systems in the literature

Model	Mean Per-word Precision	Mean Per-word Recall	# words with recall>0	F-score
Co-occurrence [15]	0.03	0.02	19	0.02
Translation Model [3]	0.06	0.04	49	0.05
CMRM [4]	0.10	0.09	66	0.09
Max. Entropy [26]	0.09	0.12	-	0.10
CRM [9]	0.16	0.19	107	0.17
CRM-Rectangles [11]	0.22	0.23	119	0.22
MBRM [11]	0.24	0.25	122	0.24
CSD-prop [10]	0.20	0.27	130	0.23
HANOLISTIC	0.35	0.24	113	0.28

4.2.5 Result

We concluded that the best performance of the hierarchical architecture is obtained by combining the output of level-0 annotators by means of overall performance based weighted summation annotator in the meta-level. The comparison of the system performance with other systems in the literature is presented in Table 4.18.

4.3 Automatic Annotation Examples

Sample annotations by the hierarchical architecture with overall weighted summation annotator for a set of images from the test set is provided in Table 4.19 and Table 4.20. It is

observed that HANOLISTIC annotates the images quite meaningfully. Moreover, our subjective analysis on the annotation results of HANOLISTIC indicates that, it assigns words which are not originally in the manual annotation but are still related to the content of the image. For example, for the first image in Table 4.19 which is manually annotated as 'jet, plane, sky', HANOLISTIC assigns additional words 'clouds' and 'smoke' which are really related to the image content. Similarly, for the second image, HANOLISTIC assigns additional words 'clouds' and 'buildings'. Other examples can be seen in Table 4.19 and Table 4.20.

Table 4.19: Sample annotations from the test set by HANOLISTIC with overall weighted summation annotator at the meta-level.












	<p><u>HANOLISTIC:</u> plane, jet, sky, clouds, smoke <u>Manual:</u> jet, plane, sky</p>		<p><u>HANOLISTIC:</u> sun, water, clouds, build- ings, city <u>Manual:</u> city, sun, water</p>
	<p><u>HANOLISTIC:</u> sun, sea, birds, waves, beach <u>Manual:</u> birds, clouds, sun, water</p>		<p><u>HANOLISTIC:</u> tree, clouds, sun, water, sunset <u>Manual:</u> hawaii, sky, sunset, tree</p>
	<p><u>HANOLISTIC:</u> tree, sky, beach, people, palm <u>Manual:</u> tree, peo- ple, palm, beach</p>		<p><u>HANOLISTIC:</u> people, buildings, tree, street, city <u>Manual:</u> mountain, people, road</p>
	<p><u>HANOLISTIC:</u> wa- ter, sunset, sun, city, boats <u>Manual:</u> light, shore</p>		<p><u>HANOLISTIC:</u> sky, water, buildings, ruins, tree <u>Manual:</u> hill, shore, water</p>
	<p><u>HANOLISTIC:</u> wa- ter, tree, garden, grass, field <u>Manual:</u> coral, fish, ocean</p>		<p><u>HANOLISTIC:</u> garden, flowers, tree, grass, cot- tage <u>Manual:</u> fence, flowers, grass, vines</p>

Table 4.20: Sample annotations from the test set by HANOLISTIC with overall weighted summation annotator at the meta-level.

	<p><u>HANOLISTIC:</u> water, sky, grass, temple, waves <u>Manual:</u> buildings, sky, water, waves</p>		<p><u>HANOLISTIC:</u> sky, water, boats, harbor, house <u>Manual:</u> boats, buildings, sky, water</p>
	<p><u>HANOLISTIC:</u> buildings, statue, night, people, light <u>Manual:</u> light, night, statue</p>		<p><u>HANOLISTIC:</u> water, iguana, lizard, marine, rocks <u>Manual:</u> iguana, lizard, marine, rocks</p>
	<p><u>HANOLISTIC:</u> horses, field, mare, foals, flow- ers <u>Manual:</u> field, foals, horses, mare</p>		<p><u>HANOLISTIC:</u> tree, flowers, tulip, sky, garden <u>Manual:</u> flowers, sky, tree, tulip</p>
	<p><u>HANOLISTIC:</u> grass, leaf, plants, close-up, bear <u>Manual:</u> leaf, pots</p>		<p><u>HANOLISTIC:</u> stone, pillar, road, temple, sculpture <u>Manual:</u> pillar, sculpture, statue, stone</p>
	<p><u>HANOLISTIC:</u> birds, nest, branch, leaf, grass <u>Manual:</u> birds, nest, tree</p>		<p><u>HANOLISTIC:</u> grass, bear, tun- dra, polar, tress <u>Manual:</u> bear, grass, polar, tundra</p>

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis a novel approach called HANOLISTIC (Hierarchical Automatic Image Annotation System Using Holistic Approach) for the automatic image annotation problem is proposed. In the literature, annotation is either considered as a supervised problem [6], where each image belongs to a class and each class is assigned a set of annotation words or it is considered as an unsupervised problem [2], where each image is assigned a set of words without any class information. HANOLISTIC proposes a solution for the unsupervised annotation problem by employing a supervised training method, where each word in the dataset is considered as a class label. Hence, the problem is quite similar to a multi-class classification problem. However, in the annotation problem an image belongs to more than one class. We model the relations between the low level visual image features with the high level semantic information by means of estimating the word membership values to determine the most probable words of an image.

While modeling the relation between the visual features and the semantic features, we employ a holistic approach, which extracts visual features from the whole image using a set of descriptors. In this way, an image is represented in several perspectives. Most of the studies in the literature, represents the image by extracting features from image regions in which case an image is represented with several feature vectors each of which is obtained from a different part of the image. In this scheme, relation between image regions and image words remains ambiguous, since it is not clear which image region give rise to which word. Therefore, the segmental approach is prone to error. However, in the HANOLISTIC approach an image is represented in several perspectives but for all perspectives features are extracted from the whole image. Hence, all image features are associated to all image words.

Hierarchical architectures for image annotation is already proposed in the literature for the unsupervised annotation problem. However, the scheme followed in those approaches first

clusters the images based on low level visual information. Then, a word selection technique is applied to annotate the image. This approach is too naive to model the relations between the image regions and annotation words. On the other hand, HANOLISTIC uses the semantic information as supervision rule while processing the low level visual information at the very first stages of annotation.

The hierarchy proposed in this thesis can be considered to provide improvement in the annotation performance at the meta-level. Initially, candidate words, which have high membership values at the end of level-0, are suggested by level-0 annotators. Each set of candidate words produced at level-0 can be considered as a single result for the annotation of an image. Instead of using level-0 annotations directly, the obtained information is processed further by the meta-level annotator to improve the annotation result.

The system has some superiorities compared to the available annotation systems in the literature. First, it makes use of several independent visual descriptors, for color, shape and texture properties to express the content information of an image. Considering an image in several perspectives enables the system to capture relation of the image with different semantic words. For example, it may be the case that a semantic word is properly represented by the color information while another semantic word may not be represented well by color, but, may be represented successfully by texture features. Hence, representing the image in several perspectives enables the system to associate each image with several annotation words.

Second superiority of HANOLISTIC is that, it applies fuzzy algorithms until the very last decision step. This ensures minimum information loss. For example, it may be the case that, a word, which does not have very high word membership value at any of the level-0 annotators, has a high membership value at the end of the meta-level. Another advantage of the system over the available systems is that, HANOLISTIC avoids segmentation which is a very problematic process introducing error into the system. Moreover, the applied algorithm, fuzzy k-nn is a simple and computationally inexpensive method compared to the available systems such as [5], [4]. HANOLISTIC process visual and semantic information simultaneously which is very crucial in terms of eliminating the semantic gap problem. If first, images are clustered based on visual information then semantic information is associated to the clusters, error introduced in clustering propagates until the end of the annotation process. Moreover, at the end of the clustering it is expected that visually similar images forming the clusters are also semantically similar. Unfortunately, in most of the practical problems this is not necessarily the case. Besides these advantages stated, it is empirically verified that

the proposed system is superior to state of the art annotation systems.

Let us now, discuss the major drawbacks of the proposed annotation system. First of all, we face the well-known general difficulties in automatic image annotation problem which originates from the human subjectivity of the manual annotation. Available datasets contain semantic information, that is words, which are not directly related to the image content. For example, an image annotated with word 'hawaii' has very little or no visual information related to that word. Similarly, there are images in the dataset, annotated with words having higher abstraction levels such as 'dance'.

Another problem in automatic image annotation systems comes from the performance measures. Although there are several performance criteria proposed in the literature, non of them has the power of objectively measuring the quality of automatic image annotation by itself. Therefore, one needs to take into account several measurements while comparing the performances of systems. The most widely used criteria are mean precision and mean recall. One may choose to consider the F-score in comparison while it carries the information from both recall and mean. On the other hand, considering coverage percentage would be misleading as it lacks the ability to penalize wrong annotations. In our study, we consider as many of the performance criteria as possible to be able to analyze the system thoroughly.

For the future work, we can further improve the annotation process of the current systems by combining the proposed approach with the existing segmental approaches. The holistic approach, proposed in this study, can be further employed in a segmentation process by relating the regions to the annotation words through the classical annotation systems, such as, CMRM [4], MBRM [11]. This task can be achieved by allowing a large number of annotation words for each image in our HANOLISTIC system. Then, apply, a semi-supervised segmentation approach which is constrained by the annotation words belonging the image. Therefore, segmentation and annotation are somehow supervised by the results obtained from the HANOLISTIC system. The top-down approach of the proposed system, combined with the bottom-up approaches of the current annotation systems is expected to improve the quality of segmentation together with annotation.

Apart from moving from an holistic approach to a segmental approach, it is also possible that alternative algorithms can be tested on level-0 and meta-level. Note that, one should employ annotators which outputs membership values of each word, rather than crisp decisions about the labels or annotation words of images.

Although several visual descriptors are used in HANOLISTIC, further study can be conducted on other description methods, extracting either global features or local features,

and several generalization methods can be examined on these features including the features proposed in this study to inspect the behavior of distinct descriptors.

REFERENCES

- [1] 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*, vol. 22, no. 12, pp. 1349–1380, 2000.
- [2] E. Akbaş, “Automatic image annotation by ensemble of visual descriptors,” Master’s thesis, Middle East Technical University, Ankara, Turkey, 2006.
- [3] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” in *ECCV ’02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, (London, UK), pp. 97–112, Springer-Verlag, 2002.
- [4] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *SIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 119–126, ACM Press, 2003.
- [5] F. Monay and D. Gatica-Perez, “Plsa-based image auto-annotation: constraining the latent space,” in *MULTIMEDIA ’04: Proceedings of the 12th annual ACM international conference on Multimedia*, (New York, NY, USA), pp. 348–351, ACM, 2004.
- [6] J. Li and J. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1075 – 1088, 2003.
- [7] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, “Matching words and pictures,” *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

- [8] F. T. Y. Akbas, Emre; Vural, “Automatic image annotation by ensemble of visual descriptors,” *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, 17-22 June 2007.
- [9] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf, eds.), Cambridge, MA: MIT Press, 2004.
- [10] J. Tang and P. H. Lewis, “Image auto-annotation using ‘easy’ and ‘more challenging’ training sets,” in *7th International Workshop on Image Analysis for Multimedia Interactive Services*, (<http://eprints.ecs.soton.ac.uk/12477/>), pp. 121–124, Korea Information Science Society, ["/lib/utills:month_12477" not defined] 2006.
- [11] S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 02, pp. 1002–1009, 2004.
- [12] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, pp. 241– 259.
- [13] R. W. Picard and T. P. Minka, “Vision texture for annotation,” *Multimedia Syst.*, vol. 3, no. 1, pp. 3–14, 1995.
- [14] F. Monay and D. Gatica-Perez, “On image auto-annotation with latent space models,” in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, (New York, NY, USA), pp. 275–278, ACM Press, 2003.
- [15] Y. Mori, H. Takahashi, and R. Oka, “Image-to-word transformation based on dividing and vector quantizing images with words,” in *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
- [16] D. M. Blei and M. I. Jordan, “Modeling annotated data,” in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (New York, NY, USA), pp. 127–134, ACM Press, 2003.
- [17] D. Monay, Florent; Gatica-Perez, “Modeling semantic aspects for cross-media image indexing,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, Oct. 2007.
- [18] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: WH Freeman.

- [19] MPEG (Moving Picture Experts Group), “MPEG-7 overview.”
- [20] International Organization for Standardisation: Coding of Moving Pictures and Audio, “Multimedia content description interface, part 3 visual,” Technical Report ISO/IEC JTC1/SC29/WG11/N4062, 2001.
- [21] R. Tax, D.M.J.; Duin, “Using two-class classifiers for multiclass classification,” *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2, pp. 124–127 vol.2, 2002.
- [22] A. Sayar and F. T. Y. Vural, “Image annotation by semi-supervised clustering,” *ICIAR, 2008*.
- [23] J. Keller, M. Gray, and J. Givens, “A fuzzy k-nearest neighbor algorithm,” *IEEE Transaction on Systems, Man and Cybernetics*, vol. 15, no. 4, p. 580, 1985.
- [24] P. Keller, J.M.; Gader, “Fuzzy logic and the principle of least commitment in computer vision,” *Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century., IEEE International Conference on*, vol. 5, pp. 4621–4625 vol.5, 22-25 Oct 1995.
- [25] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, “On combining classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [26] J. Jeon and R. Manmatha, “Using maximum entropy for automatic image annotation.,” in *CIVR*, pp. 24–32, 2004.
- [27] S. Herrmann, “MPEG-7 eXperimentation Model (XM).”