Animal Recognition using Joint Visual Vocabulary

Heydar Maboudi Afkham

Master of Science Thesis
Stockholm, Sweden 2009
Animal Recognition using Joint Visual Vocabulary

Heydar Maboudi Afkham

Master’s Thesis in Computer Science (30 ECTS credits) at the Scientific Computing International Master Program
Royal Institute of Technology year 2009
Supervisor at CSC was Alireza Tavakoli Targhi
Examiner was Stefan Carlsson

TRITA-CSC-E 2009:012
ISRN-KTH/CSC/E--09/012--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication

KTH CSC
SE-100 44 Stockholm, Sweden

URL: www.csc.kth.se
Abstract

This thesis presents a series of experiments on recognizing animals in complex scenes. Unlike usual objects used for the recognition task (cars, airplanes, ...) animals appear in a variety of poses and shapes in outdoor images. To perform this task a dataset of outdoor images should be provided. Among the available datasets there are some animal classes but as discussed in this thesis these datasets do not capture the necessary variations needed for realistic analysis. To overcome this problem a new extensive dataset, KTH-animals, containing realistic images of animals in complex natural environments. The methods designed on the other datasets do not preform well on the animals dataset due to the larger variations in this dataset. One of the methods that showed promising results on one of these datasets on the animals dataset was applied on KTH-animals and showed how it failed to encode the large variations in this dataset.

To familiarize the reader with the concept of computer vision and the mathematics backgrounds a chapter of this thesis is dedicated to this matter. This section presents a brief review of the texture descriptors and several classification methods together with mathematical and statistical algorithms needed by them.

To analyze the images of the dataset two different methodologies are introduced in this thesis. In the first methodology fuzzy classifiers we analyze the images solely based on the animals skin texture of the animals. To do so an accurate manual segmentation of the images is provided. Here the skin texture is judged using many different features and the results are combined with each other with fuzzy classifiers. Since the assumption of neglecting the background information in unrealistic the joint visual vocabularies are introduced. Joint visual vocabularies is a method for visual object categorization based on encoding the joint textural information in objects and the surrounding background, and requiring no segmentation during recognition. The framework can be used together with various learning techniques and model representations. Here we use this framework with simple probabilistic models and more complex representations obtained using Support Vector Machines. We prove that our approach provides good recognition performance for complex problems for which some of the existing methods have difficulties.

The achievements of this thesis are a challenging database for animal recognition. A review of the previous work and related mathematical background. Texture feature evaluation on the "KTH-animal" dataset. Introduction a method for object recognition based on joint statistics over the image. Applying different model representation of different complexity within the same classification framework, simple probabilistic models and more complex ones based on Support Vector Machines.
## Contents

1 Introduction 1
   1.1 Introduction .................................................. 1
   1.2 Thesis Outline ................................................. 4

I Backgrounds 5

2 Related Work 7
   2.1 Object recognition ............................................. 7
   2.2 Animal Recognition ........................................... 9

3 Computer Vision and Mathematical Background 11
   3.1 Feature Extraction .............................................. 11
   3.2 K-means Algorithm ............................................. 12
   3.3 K-Nearest Neighborhood Classification ....................... 14
   3.4 Single Histogram Class Models for Image Segmentation ....... 15

II Dataset 19

4 Dataset 21
   4.1 MSRC Dataset ................................................. 21
   4.2 KTH-Animal Dataset ........................................... 22
   4.3 Dataset Setting ................................................ 23
   4.4 Database Generator Software ................................ 24

III Methodologies 27

5 Fuzzy Classification 29
   5.1 Method .......................................................... 29
   5.2 Results .......................................................... 31

6 Joint Visual Vocabulary 35
Chapter 1

Introduction

1.1 Introduction

In this thesis the aim is to develop a method for recognizing animals based on their skin texture and color. Object recognition is referred to the task of finding objects in images or video sequences. Humans have a good robustness in this task due to their advanced visual system. One of the aims of the computer vision is to develop algorithms for object recognition with a robustness close to human vision. In order to train the computers to recognize objects various images datasets are provided by different groups [32, 12, 8]. These datasets are usually divided into a test and train subset. Mathematical model build over the train subset and their evaluation is done over the test subset. There are several parameters that makes the task of the object recognition a challenging problem and at the same time an interesting task. A few of these problems are listed below.

• The object appearing in different images may appear in different view points such rare and side view of the cars.

• The object is partially covered by other objects like a chair partially hidden by a table. or

• The object appearing in different illuminations like two pictures of a building taking in different times of a day.

• The object appearing in different scales.

One may realize that the mathematical model made for recognizing an object should be robust enough to perform well under different conditions. It should be mentioned that there are different difficulties in detecting various objects. For example objects like cars which usually appear in a rigid shape are recognized easier from animals that may appear in different shapes and positions. In figure 1.1, some sample animal classes are shown. An image can be described by different features such as texture, color, shape. One of the most important tasks in objects recognition problem is to extract features that can describe objects in a distinctive way. Using such features will lead us to better object classifica-
CHAPTER 1. INTRODUCTION

Figure 1.1. Different difficulties that can be faced when dealing with the animal classification problem. The images in each row present some sample images of one class. These variations in viewpoint, shape, illumination, color and etc. make the task of animal recognition a challenging task.

Animals as objects have significance in content-based image and video retrieval as they carry a lot of semantic information about natural scenes. Unfortunately, they are also difficult to recognize due to occlusion with natural scenes, having deformable bodies with varying shape, complex background in natural. Additionally, animals usually appear in images of outdoor environments under different illumination conditions, view points and scales which has been a challenge for general object recognition. Due to the fact that the problem is known to be difficult, many authors have evaluated the performance of their recognition methods on images of animals. At the same time, few attempts have focused directly on animal recognition. Furthermore, the existing databases are usually limited to simple settings and contain small number of animal classes. In the existing literature, the performance of recognition or segmentation algorithms is usually shown on just a few classes e.g. zebra, cheetah and giraffe [11] or cow, sheep, bird, cat and dog (along with 16 non-animal classes) [25].

The object recognition task is usually performed on objects with distinctive shapes such as cars, bikes and airplanes [12]. The shape and the local regions of these objects do not vary much from image to image when compared to images of animals. Additionally, animals usually appear in images of outdoor environments under different illumination conditions, view points and scales. The difficulties one will face when dealing with such images can be divided into two major groups: technical difficulties and natural variations. Technical difficulties mainly occur due to low quality of images combined with variations
1.1. INTRODUCTION

![Figure 1.2. Examples of intra-class similarities and inter-class differences in the database. The two images on the left show two different goats that vary in both texture and color, while the two images on the right show a female lion and a cougar which look similar with respect to both texture and color.](image)

in scale or inaccurate segmentation of the image regions containing the animals. Several well known segmentation and classification algorithms with good overall performance do not provide a sufficient level of robustness for animal classes due to the complexity and variability of the background and the objects themselves. Furthermore, low quality of the images or compression artifacts affect the performance of such methods. The variability due to changing illumination conditions, scales and view points give rise to further difficulties. Natural differences occur often between the animals belonging to the same class due to varying physiological attributes or gender. In both cases, the animals can have completely different visual appearance. An example of such a problem is illustrated in Figure 1.2 in which an American goat appears to be completely different from an Asian goat with respect to both skin texture and color. At the same time, a female lion looks similar to a cougar in both texture and color.

Studies on object recognition and detection in indoor and outdoor environments [26, 27] have shown that the use of the contextual information in images leads to a better recognition and detection. However, the contextual information is usually omitted intentionally and treated as a noise. As a result, most of the previous models were designed to avoid the background information [32, 25, 23]. In this work, we take advantage of the contextual information of an object to obtain a better recognition performance. In case of animal recognition it is not difficult to justify the use of context as different animals usually live in different environments. The chance of correctly classifying an animal increases when the relevant context exists and is used. The method introduced in this thesis learns the relation between color and texture information of the animal region and it’s context in an unsupervised manner to aid the classification process. Additionally, due to the ability to encode complex visual structures in both the categorized objects and the surrounding background, no segmentation is required during recognition.

As it will be shown through experiments, many of existing methods, which is shown promising results for recognition problem, cannot properly represent the diversity of animal classed with complex background that occur in natural scenes. Therefore, to be able to better represent the intra-class variability of the animal classes and the different context in which the animals can appear, more complex internal representations are required.

In this thesis two different methodologies for animal recognition are presented. In the first method it is assumed that an accurate segmentation of the objects is provided while in the other method the assumption is that no information regarding the localization of the
CHAPTER 1. INTRODUCTION

object is provided. Each of these methodologies require a different approach for creating models. As a result, the main contributions of this paper can be summarized as follows:

- A review of the available literature in objects recognition and classification and the mathematical background related to the scope of this thesis.
- Introducing our database "KTH-animal" of animal images which captures a broad range of natural variations and common technical difficulties.
- Texture feature evaluation on the "KTH-animal" dataset.
- Introduction a method for object recognition based on joint statistics over the image.
- Applying different model representation of different complexity within the same classification framework, simple probabilistic models and more complex ones based on Support Vector Machines.
- Experimental evaluation on the proposed database.

1.2 Thesis Outline

The outline of this thesis is as follow. In chapter 2 the related work done in the scope of this thesis is given. In chapter 3 the computer vision and mathematical background for familiarizing the reader with concepts of the thesis is given. The KTH-animal database is introduced in chapter 4 also this chapter contains a review of some available databases. The first methodology of this thesis is introduced in 5. This experiment assumes that the location of the animal is known by the system. The experimental results of this method can also be seen in this chapter. The second methodology is introduced in 6. In this experiment only the textural information of the image is required for the recognition task. The results of this experiment can be found in the same chapter. Finally in chapter 7 conclusions of the thesis and it’s future work is discussed.
Part I

Backgrounds
Chapter 2

Related Work

The purpose of this chapter is to review the related literature to the methodologies introduced in this thesis. In this thesis the aim is to develop a method for recognizing animals based on their skin texture and color. For this reason it is necessary to review some of the most related methods that tend to recognize objects based on their textural appearance. It is also needed to review some of the previous attempts focused on animal recognition regardless of the features used for their recognition purpose.

2.1 Object recognition

There is a host of approaches to visual object recognition. The two most widely used are

- sparse methods which find distinctive regions in the image [18], calculate features for these regions, [17] and classify the objects. Examples of such an approach can be found in [7, 28];

- dense methods which segment the images and classify each resulting region into one of the object classes [1, 4].

An accurate segmentation can be a great aid for the aforementioned recognition methods. Since segmentation is not a well-defined problem [16], segmentations produced using different algorithms might not be similar. For challenging databases, it is hard to find a method which behaves well on every image of the database. Furthermore, segmentation algorithms have the same complexity as object recognition algorithms with their own problems. Segmentation algorithms such as normalized cuts [15] or graph cuts [3], highly depend on their parameters. These methods were applied on our animal database and failed due to the large complexity in the background. These segmentation methods are sensitive to the edges appearing in the image and they require the number of image segments as an input argument. On one hand the problem of using these algorithms for animal recognition is that natural skin texture or sudden changes of illumination may result strong edges on the animal skin. This fact will force the segmentation method to divide the animal region into several regions. On the other hand when we are dealing with complex pictures and we wish to segment all the images into a certain number of segments, segmentation fails.
This is because each image has its own segmentation setting. This brings up the question how to find the best segmentation for each image? Instead of answering this question the focus is to introduce a method that can perform the recognition task without any need of the segmentation information.

Recently many classification methods are built on top of different sparse features. These features are computed much faster than the dense features. This fact makes them a good candidate for realtime object detection and classification. There features highly depend on the region detector they are based on and the region detector usually extract information from the sharp edges. Such edges are usually remain unchanged with changes in illumination and viewpoint. One of the most used sparse methods is the SIFT [14] descriptor. Due to fast computation of the SIFT features they have been used in many applications. Another class of sparse feature descriptors are the affine invariant region descriptors. These features have shown a good performance in detection and recognition in the literature introduced by S. Lazebnik et al. [11]. Their work is mainly focused on texture classification based on such features. Among the datasets they used for evaluating their features they used a three class animals dataset. This approach was applied on three different animal classes zebra, cheetah and giraffe. All these animals are among the classes of animals with complex skin texture. Because of this fact these classes made a perfect choices for their approach.

When dealing with an complex image database, it is important to use a method capable of equally capturing information from the different classes. Sparse methods depend on the region detector used and how complex the structure and texture of object and background are. Like segmentation algorithms region detectors are sensitive to edges. As a result using such methods may fail, especially when dealing with animal classes with smooth skin texture and complex background. This is mainly due to the lack of enough interest regions on the animal’s body. To solve this problem, either a segmentation algorithm or a dense method which uses the information of all the pixels of the image should be applied.

The first step of creating good appearance models is extracting robust features from the images. Due to the problems discussed in the previous paragraph dense methods are used for feature extraction in this thesis. Two often used dense methods are MRF [30] and maximum response filter banks [29]. These methods use the exact intensity of the image or segmented region to extract textural information from the image. These feature extractors will be discussed in more details in the up coming chapter. The data gathered from the images is then clustered using the k-means[9] algorithm. Each the cluster center is considered as a visual word in the visual vocabulary. After creating the visual vocabulary the features from each region are labeled with the closest word in the dictionary and the normalized histogram of the distribution of visual words is then assign to it. Later, a test image or a region is classified as the training image with the closest distribution histogram of visual words. The closeness is determined using a distance meter such as $\chi^2$ distance. Such methods were originally applied to textile and texture database in which the whole image consisted of one uniform texture. This methodology usually does not preform well on the outdoor images without an accurate segmentation of the objects in the image provided.

With the need for recognizing objects in the outdoor environments different databases has been introduced. One of the most used databases in this field is the MSRC database
2.2. ANIMAL RECOGNITION

[32], introduced by J. Winn et al. The authors proposed a method for classifying different regions of the image and it showed good performance. In their approach, the image is divided into several sub-regions and each region is classified separately depending on its distribution of visual words. They reported a 93.4% classification rate. Later, S. Savarese et al. [23] defined a model for the appearance and the shape of an object class by finding correlations between different visual words. The performance of 93.8 classification rate on MSRC database is reported. These approaches focus on recognizing the different regions of the image, which brings the need for using a segmentation algorithms before performing the recognition task. To avoid the use of segmentation algorithms, Schroff et al. introduced the single-histogram class models [25]. Single-histogram provides both recognition and segmentation of the test images using the first order statistics over the image. This approach uses a histogram of the visual words per class as the model. The methods provide a pixel-wise segmentation of the image into object class regions to give simultaneous segmentation and recognition. In order to achieve this goal, the single-histogram class models are compared with histogram of local distributions of visual words using Kullback-Leibler divergence. The single-histogram models are much simpler than the models used in the previously mentioned methods, still classification performance reported was comparable to the performance of other methods with 93.43% of the classification rate. The segmentation accuracy of this method was reported to be 75.07%. The single-histogram model is not rich for capturing large variations, but sufficient enough for the available databases[19].

The method of most relevance to use is the single-histogram class model [25]. This is because to perform the recognition task no extra information such as segmentation is required. For this reason a special attention is given to this method.

2.2 Animal Recognition

Not many attempts solely focused on the task of animal recognition. One of the earliest methods applied to an animal database was proposed done by C. Schmid [24]. The author constructed models for content-based image retrieval using Gabor-like filters. The method was tested only on four different classes zebra, cheetah, giraffe and a class of human face. All animals used in this work had complex skin texture. Later, D. Ramanan et al. introduced methods to detect textured animals using the shape and texture information in video sequences [21, 22]. In an application for searching images on the internet Tamara L. Berg and David A. Forsyth [2] used four cues: nearby texts on the web pages, color, texture and shape to re-rank the images retrieved by Google image search. They reported that animals are among the hardest classes of objects for recognition in computer vision. They also stated that, animals with complex textures, which have been subject for object recognition in many experiments [24, 21, 22, 11], are easier to detect than the animals with a smooth skin pattern.
Chapter 3

Computer Vision and Mathematical Background

To do texture based object recognition in computer vision one needs to do review several related topics in both mathematics and computer vision. The purpose of this chapter is to introduce these literature to the reader. In this chapter first we explain different relevant feature extraction method and later we describe the mathematics behind making "visual dictionaries" used for object recognition. Afterwards a few machine learning and classification methods are discussed. Finally a more detailed discussion is given regarding the single histogram class modes [25] for both object recognition and classification.

3.1 Feature Extraction

To be able to automatically learn and recognize object it is needed to describe these objects in a way understandable by computers. Different approaches have been created on mathematical image description and feature extraction depending on the problems. Due to the discussions in the related work, chapter 2, the aim is to extract dense features from every pixel of the image. Also due to the nature of this problem we are interested in employing different cues of information such as texture and color. Two main texture methods that can provide such information are the maximum response filter banks [29] and MRF [30] methods. In this section a review of these method and their usage and applications is given.

The feature vectors extracted using the maximum response filter banks introduced by M. Varma et al. [29] are invariant to viewpoint and illumination changes. To extract these features several filters are convolved with the image. A visualization of MR8 filter bank can be seen in figure 3.1. This filter bank consists of two rotationally symmetric filters and two edge and bar filters at 6 different rotations and scales. In the MR8 filter bank only 8 responses are recorded. To obtain this at each scale the maximal value over the different rotations in considered.

Computation of maximum response filter banks are really expensive due to the number of convolutions in their computation. For this reasons the same authors ran another set of experiments on using the exact intensity if the image pixels for extracting features from
the images. Their experiments resulted a new series of feature descriptors known as MRF descriptor [30]. This descriptor does not capture any invariance in viewpoint or illumination. Still the experiments have shown promising results comparable to maximum response filter banks. The MRF descriptor simply records the intensity of a local neighborhood of the pixels in a certain order. In there experiments the classification using MRF features outperformed the performance of MR8 descriptor. Also due to easy computation of MRF features and their easy adoption for being applied on color images, they were used in many applications [32, 23, 25].

These feature descriptors were traditionally used on textile dataset. Figure 3.2, gives a visualization of such datasets. This dataset were used in experiments evaluating MRF [30] and MR8 [29] feature descriptors. This dataset only deals with illumination and rotation change of certain textures. Another similar but more complete dataset is the KTH-TIPS dataset. This dataset also includes images of the materials in different scales.

One characteristic of such dataset is that either the whole image contains one certain texture or they all have a constant background easily detected. Also all these images are taken in controlled environment and less noise is involved when models are created over these datasets.

To describe the procedure of how models are made from these extracted features we first need to describe the k-means algorithm. K-means is an essential tool for making different vocabularies from the data.

3.2 K-means Algorithm

In many scientific fields researches need to deal with large statistical data gathered from different sources. In order to analyze this data in applicable way, there is a need for identification of the repeating patterns and reduce the data size without losing significant information. One of the most used algorithms for finding such patterns is the k-means algorithm.

The objective of the k-means algorithm is to partition a set of feature vectors, \( \{x_i : i = 1..n, x_i \in \mathbb{R}^m\} \), into \( k \) clusters. The value of \( k \) is assumed to be known. Each cluster is identified by a cluster center \( c \). The feature vector \( x_i \) is assumed to belong to the cluster \( j \) if \( d(x_i, c_j) < d(x_i, c_l) \) for every \( l \neq j \) with respect to the metric distance \( d(.) \). The set of
3.2. K-MEANS ALGORITHM

![Figure 3.2](image)

Figure 3.2. Pictures from the Columbia-Urtech dataset. This dataset were used in experiments evaluating MRF [30] and MR8 [29] feature descriptors.

The optimum cluster centers \( \{c_1, c_2, \ldots, c_k\} \) minimizes the cost function

\[
\Phi(\text{clusters}, \text{data}) = \left\{ \sum_{j=1}^{k} \sum_{x_i \in \text{cluster}_j} d(x_i, c_j) \right\}. \tag{3.1}
\]

The set of cluster centers found by the k-means algorithm gives a local minimum to the cost function 3.1. The procedure the k-means follows to find this set is as follows,

- Initially choose the \( k \) cluster centers from the feature vectors randomly.
- Until the cluster centers don’t move,
  - Assign each feature vector to the closest cluster center.
  - Replace the cluster centers with the mean of the feature vectors assigned to them.

The distance function used in the k-means algorithm is usually chosen to be the euclidean distance defined as \( d(x, y) = (x - y)^T(x - y) \). Often a weighted distance needs to be used to implement the weighted distance between two points one might use the weight matrix \( A_{n \times n} \) and define the weight function \( d_A(x, y) = (x - y)^T A(x - y) \).

The set of the cluster centers is often referred as a dictionary and each cluster center is called a word. Usually in computer vision each object is described with a set of feature vectors. Using a dictionary we can assign each feature vector to the closest word in the
dictionary. Finally the normalized histogram of the words used in the object describes the object. The dimension of this descriptor vector is the number of words available in the dictionary.

3.3 K-Nearest Neighborhood Classification

After defining the dictionary and descriptor vectors, each object can be represented using one or several feature vectors. These features are then labeled as the nearest word in the dictionary and a normalized histogram of visual words is then assigned to the object. Let's assume that the set

\[ A = \{(x_i, y_i) : x_i \in \mathbb{R}^n, |x_i| = 1, y_i \in [1..m]\} \]

is the set of all descriptor vectors extracted from a database with \( m \) different classes. The purpose of a statistical learning procedures is to introduce a function \( f \) which minimizes the cost

\[ \sum ||f(x_i) - y_i||. \] (3.2)

Usually to create such a function the set \( A \) is divided into two subsets, \( A^{train} \) and \( A^{test} \), for training and test of the model. The function \( f \) is designed over the training subset and its performance is measured using the test subset. There are several ways to introduce the function \( f \). Here we focus on the first nearest neighborhood classification which is widely used for textile classification.

In this methodology the whole train set, \( A^{train} \), is considered to be the model. If we look closer at the elements of \( A \) we realize each element is actually a probability distribution assigned to a class. For an element \( (x_i, y_i) \in A \) the element at the \( j^{th} \) position of the vector \( x_i \) can be seen as the probability of existence of word \( j \) in the object \( x_i \).

The k-NN classifiers use this presentation of the elements of \( A \) classifies an unknown distribution \( x \in \mathbb{R}^n \) to the \( k^{th} \) closest distribution from the distribution provided in \( A^{train} \). In this case the function \( f \) return the index of the class to which the closest distribution belongs to.

The closest distribution can be found using different probability distribution distances. One of the most famous distances used in different literature is the Euclidean distance defined as,

\[ d(x, y) = \sqrt{\sum_{i=1}^{n} (x^{(i)} - y^{(i)})^2}. \]

This distance acts poorly in capturing the similarities between two different distributions when in some dimensions the difference between \( x^{(i)} \) and \( y^{(i)} \) is large. To overcome this problem the distance measure \( \chi^2 \) was introduced. The \( \chi^2 \) distance can also be seen as the normalized Euclidean and it is defined as

\[ d(x, y) = \sum_{i=1}^{n} \frac{(x^{(i)} - y^{(i)})^2}{x^{(i)} + y^{(i)}}. \]
3.4 SINGLE HISTOGRAM CLASS MODELS FOR IMAGE SEGMENTATION

To measure the similarity between to distributions other distance measures have also been used. In this literature we will mainly use the $\chi^2$ distance for this purpose.

When the number of elements in $A^{train}$ grows large using the k-NN models becomes inefficient since each unknown vectors has to be compared with all the vectors in $A^{train}$. Also due to existence of noise individual distributions might be corrupted.

Figure 3.4 shows the procedure of making visual dictionaries and creating models from the training dataset and recognition of test images in textile datasets.

One benefit of working with textile images is that the whole images contains one pattern and the problem is classification of different images. When looking at outdoor images different objects may appear in one image. For these datasets the models can not be created over the entire image, since depending on the objects appearing in the image the histogram of the visual words appearing in the image will change. Also when dealing with outdoor images the objects appear in complex background while the distribution of the visual words appearing in the background will result a noisy description of the objects which makes the task of object recognition almost impossible.

To solve this problem many approaches have been taken. As mentioned in many approaches a there is a need for providing a segmentation of the image. Using a segmentation one can decrease the noise of the object model in classification stage. Here the intension is to construct a model which can recognize the objects in the image without the aid of a segmentation process. The focus is only on single histogram class models [25] which provide a reasonable recognition rate without the aid of any segmentation algorithms.

3.4 Single Histogram Class Models for Image Segmentation

The single histogram classes models [25], introduced by F. Schroff et al. aim to obtain a segmentation of the image through pixel wise classification of the image. The best performance of this method was measured using the MRF descriptors built on $5 \times 5$ patches, on the MSRC dataset [32].The idea behind the single histogram class models is that with having the average distribution histogram of each class it is possible to classify local regions of the test images based on the distribution of the visual words in a special metric space. Here we discuss the training and testing process of single histograms individually.

To build visual words dictionary 30% of total feature vectors extracted from the dataset were considered. Using this dictionary a normalized histogram of visual words is assigned to each interest region. Lets assume that $p^j$ denotes the normalized histogram of the $j^{th}$ exemplar region of class $c$. The single histogram models of this class is defined as,

$$\hat{q} := \frac{\sum_j n^j p^j}{\sum_j n^j}.$$ 

Here $n^j$ denotes the number of pixels in $j^{th}$ exemplar region. $n^j$ is a weighting factor of different exemplars if it is set to one then all the exemplars are treated equally in making the model. Figure 3.3 is a visualization of the training and testing procedures of the single-histogram models.
During the testing process the method classifies the normalized histogram $p$ of a test region or a neighborhood defined using a sliding window technique, to the closest single histogram model. The closeness was defined by two different metric meters. First the Euclidean distance between the two histograms $D_{L2}(a, b) = \sum_{i=1}^{n} (a_i - b_i)^2$, and the Kullback-Leibler divergence defined as,

$$D_{KL}(a, b) = \sum_{i=1}^{N} a_i \log \frac{a_i}{b_i}.$$

As argued the results of classification had significant improvement when KL divergence were used.

One of the interesting techniques used in this paper is the sliding window technique which enables us to provide a pixel wise classification of the image. Here at each window position the pixel at the center of the window is classified due to the distribution of the visual words appearing in the window. This way the method can provide a segmentation of the images. Also the KL divergence distance function ensures that the result of the segmentation is smooth. Later we will use a similar technique to introduce a method which obtains a rank based recognition of the objects in the scene. Also as mentioned in the chapter 2 it is important to employ databases containing realistic data [19]. The MSRC database [32] does not have many characteristics of such database.
3.4. SINGLE HISTOGRAM CLASS MODELS FOR IMAGE SEGMENTATION

Figure 3.4. The visual description of dictionary extraction, model training and test image classification on textile datasets.
Part II

Dataset
Chapter 4

Dataset

As concluded in the previous chapter a realistic dataset is needed to measure how realistic the analysis are. Recent studies [19] show that most of the available databases lack many features required for realistic analysis. Most of these datasets contain objects with similar viewpoints and orientations, normalized sizes and having a little or not background clutter. To build a method for animal recognition a large dataset of different animals was required. After an initial study of the existing animal classes in the available datasets the need for gathering a new animals dataset got stronger. These datasets contained too few classes of the animals. Also the images available in these classes were not capturing the realistic pose and illumination changes in the outdoor images taken from the animals.

4.1 MSRC Dataset

MSRC dataset [32] is one of the most used datasets for testing recognition methods provided. Among the different classes of this database there are a few animal classes available. In figure 4.1 some sample images of the images of this database are shown. By taking a closer look at these images and the other animal images in the dataset one can see that most of the objects have a unique pose and they appear in an almost constant illumination. The context information of all these images is the same, since most of these images were taken from the animals living in urban area’s. Also this database contains too limited instances of objects per class which makes the training of complex models hard on this dataset. These models usually require more data for achieving better robustness. In table 4.1 number of images in the animal classes seen in second version of the MSRC database.

For these reasons the animal classes appearing in this database does not make a good choice for animal recognition. Also for most of the animals, due to living in specific regions of the world and appearing in a special environment, it is natural to get aid from their surrounding context for recognition. It has to be mentioned that other datasets have similar problems with the animal classes.
CHAPTER 4. DATASET

<table>
<thead>
<tr>
<th>Class Name</th>
<th># Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird</td>
<td>30</td>
</tr>
<tr>
<td>Cat</td>
<td>23</td>
</tr>
<tr>
<td>Cow</td>
<td>42</td>
</tr>
<tr>
<td>Dog</td>
<td>27</td>
</tr>
<tr>
<td>Sheep</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 4.1. Detailed information of the classes in second version of MSRC database.

Figure 4.1. Some sample images available in the MSRC database [32].

4.2 KTH-Animal Dataset

Animals usually appear in different poses and different illuminations in nature. Also when trying to recognize animals different difficulties appear. For instance the difficulties among the objects of one class. Often animals with different shape and texture are categorized as one class. When encountering such an animal it is necessary to have the images of different appearances of the animal in the dataset. Also the number of images in each class should be large enough so that the training and testing of the data would be possible.

Due to non-existence of a comprehensive annotated animal image database which captures all of these variations, a database containing 1740 images in nineteen classes of different animals is created which captures all possible natural conditions and variations. The detailed information regarding the number of the images in the animals dataset can be seen in table 4.2. The images are low quality JPEG images like most of the images found on the internet. The images were gathered from images of the Corel database [13], combined with images gained from Yahoo internet image search results. The images were segmented manually into foreground and background regions for the learning purpose. A sample of the animals in the images of the database, which appear in a variety of positions, scales, viewpoints and illuminations, can be seen in Figure 4.2. The purpose of gathering such database is to introduce a database which captures most of the necessary natural difficulties for creating a realistic model for animal recognition. For example the database captures intra-class variations resulting from physiological differences. It is intra-class differences and inter-class similarities that make the task of animal recognition especially challenging.
### 4.3 Dataset Setting

In each image of KTH-animal dataset only one animal class is presented in each image. Also for each image a manual segmentation is provided. In the manual segmentation the pixels belonging to each object in the image are labeled as the index of the object and the pixels belonging to the background hold the value equal to the number of the objects in the image plus one. In figure 4.3 some images of the database and their manual segmentation are shown.

The following matlab code can help accessing the images and their segment files in the dataset.

```matlab
function [ im , seg_im ] = image_loader(path,class,image_index)

    orig_path = [path '/' class '/original/'];
    seg_path = [ path '/' class '/segment/'];

    orig_files = dir( [ orig_path ‘*.jpg’] );
    filename = orig_files(image_index).name;

    im = imread( orig_path + filename );
    im = im2double(im);

    seg = imread( seg_path + filename );
    seg = im2double(seg);
```

---

<table>
<thead>
<tr>
<th>Class Name</th>
<th># Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear</td>
<td>105</td>
</tr>
<tr>
<td>Cougar</td>
<td>100</td>
</tr>
<tr>
<td>Cow</td>
<td>97</td>
</tr>
<tr>
<td>Coyote</td>
<td>100</td>
</tr>
<tr>
<td>Deer</td>
<td>100</td>
</tr>
<tr>
<td>Elephant</td>
<td>100</td>
</tr>
<tr>
<td>Giraffe</td>
<td>84</td>
</tr>
<tr>
<td>Horse</td>
<td>100</td>
</tr>
<tr>
<td>Kangaroo</td>
<td>90</td>
</tr>
<tr>
<td>Leopard</td>
<td>100</td>
</tr>
<tr>
<td>Lion</td>
<td>98</td>
</tr>
<tr>
<td>Panda</td>
<td>97</td>
</tr>
<tr>
<td>Penguin</td>
<td>80</td>
</tr>
<tr>
<td>Sheep</td>
<td>68</td>
</tr>
<tr>
<td>Skunk</td>
<td>62</td>
</tr>
<tr>
<td>Tiger</td>
<td>100</td>
</tr>
<tr>
<td>Zebra</td>
<td>77</td>
</tr>
</tbody>
</table>

Table 4.2. Detailed information of the classes in KTH-animal dataset.

Similar but less varied animal image databases that have been used for feature or classifier evaluation include: a butterfly database containing 619 images of seven different classes of butterflies [11] and a bird database containing 600 images (100 samples each) of six different classes of birds (100 images per each class) [10].
In this dataset the images of different classes are stored in different directories and each class directory holds two subdirectories containing the original and segment images. The filename of the original images are the same but the original images are stored as JPG files and the segment images are stored as PNG files.

### 4.4 Database Generator Software

When dealing with huge datasets depending on the application one might require to work with a subset of the dataset. Usually selecting this subset is a hard and time consuming task. In order to ease this process a software for viewing these images and the manual segmentation is provided. A screen shot of this software can be seen in figure 4.4.

This software can perform some other tasks beside creating a subdataset of a given dataset. Here is a summary of the applications of this software and the way to use them.
4.4. DATABASE GENERATOR SOFTWARE

![Sample images from animals database and their manual segmentation. In this figure the values of manual segmentation are changed for visualization.](image)

- **Database Generator**: When a dataset is loaded using the software there is variable assigned to each image of the dataset called *Copy Image*. When the button *DatabaseGenerator* is pressed the images which are selected to be copied are copied into a new location with the same structure as the old dataset.

- **Creating A Dataset Log**: Some times there is a need for ranking the images of the dataset or adding comment to some images. To make this task easy the application is made in a way to hold this information for different images and the information is written into a file in the following format. This option is useful for editing the database or when we wish to group some images in different classes together.
Figure 4.4. A screenshot of the database generator application.
Part III

Methodologies
Chapter 5

Fuzzy Classification

Here the aim is to develop a method which can recognize the animals based on their skin texture. A natural way to start the analysis is to consider the dataset as a textile dataset. In this case we neglect the background information and we only process the region marked by the ground truth mask. This assumption might be unrealistic but will provide us with essential information about textural behavior of the dataset. Also with employing a segmentation algorithm a similar method can be used for recognizing the objects appearing in different segments but this problem is not in the scope of this thesis.

The purpose of the analysis was to find a feature that can best represent the textural information of all different animal classes. To achieve this goal two different texture descriptors \{MR8, MRF\} and different color systems \{RGB, Opponent, HSV\} were used. As will be argued in this chapter it is interesting to see how the behavior of different classes was different regarding different texture descriptors and color spaces. This behavior was so different that made it almost impossible to find a special setting for the texture descriptors which describes all the classes well.

To overcome this problem a method is needed that can use all these information and obtain a recognition rate close to the best recognition rate obtained by the best classifier of each class. The solution given in this chapter uses a voting method to eliminated some of the possible classes the object might have in different stages. Each stage is connected to a pool of classifiers. The classifiers in each pool analyze the object regarding a certain aspect. In figure 5.1, a demonstration of how different stages are connected to each other is given.

5.1 Method

For a feature descriptor with certain parameters, say \( \delta \), and a set of training images \( I_1^{train}, I_2^{train}, \ldots, I_n^{train} \) the feature vectors for each pixel in the training images are clustered using k-means clustering. For a set of cluster centers \( c_1^{\delta}, c_2^{\delta}, \ldots, c_m^{\delta} \), each image \( I \) is represented by a \( m \) dimensional normalized frequency histogram \( n(c^{\delta} | I) \). We define the classifier \( \phi_\delta \) as a function, which classifies every test image \( I' \) with respect to descriptor \( \delta \) as

\[
\phi_\delta(I') := \gamma(\argmin_i d(n(c^{\delta} | I'), n(c^{\delta} | I_i^{train}))), \tag{5.1}
\]
where $d(\cdot)$ denotes the $\chi^2$ distance between the histograms and $\gamma$ is the index function which maps the index of every training image to the index of the class containing it.

To solve the problem and achieve a classifier which acts close to the best classifiers for every class, we introduce a fuzzy classifier which uses the information of different classifiers to classify the test image. Let’s assume that $P = \{\phi^{\delta_1}, ..., \phi^{\delta_n}\}$ is a pool of classifiers and $(A, m_P)$ is the fuzzy set of probable classes for the test image $I^t$, with fuzzy value defined as

$$m_P(j) := \left| \left\{ i : \phi^{\delta_i}(I^t) = j \right\} \right| / |P|, \quad j \in A. \quad (5.2)$$

A fuzzy classifier, $\Phi_P$, is defined on the pool $P$ as

$$\Phi_P(I^t) := \arg\max_j \{m_P(j) : j \in A\}. \quad (5.3)$$

In this equation the test image is classified to class with the highest fuzzy value in the set $A$.

Here two ways of combining the information of different descriptors is considered. First is to make one pool of classifiers containing all the classifiers. This way we can define a fuzzy classifier on top of this pool.

Second is to consider different pools of classifiers for certain types of descriptors. These pools are connected to each other in a chain. As the fuzzy set $(A, m)$ moves forward on the chain least probable classes for test image are eliminated by having their fuzzy value equal to zero. To define this chain of pools of classifiers first it is needed to define a restriction over classifier defined in equation 5.1. Let $(A, \alpha)$ be a fuzzy set, we define the classifier $\phi^{\delta}$ with restriction to $(A, \alpha)$

$$\phi^{\delta}(I^t) \big|_{(A, \alpha)} := \gamma(\arg\min_i \left\{ d(n(\delta^i | I^t), n(\delta^i | I^{train}_i)), \alpha(\gamma(i)) > 0 \right\}). \quad (5.4)$$

Here the image $I^t$ is being classified only to the elements of $A$ with their fuzzy value greater than zero. Using this definition we can define other restrictions such as $m_P(i) \big|_{(A, \alpha)}$ and $\Phi_P(I^t) \big|_{(A, \alpha)}$, where $P$ is a classifier pool. This means that we neglect any element of $A$ with its fuzzy value equal to zero.
5.2. RESULTS

Assume that $P_1, P_2, ..., P_n$ are different pools of classifiers and $(A, m_{\Phi_0})$ is a fuzzy set of probable classes with $m_{\Phi_0}(i) = 1/|A|$ for every $i \in A$. The chain of classifier pools is described for the test image, $I^t$, as

$$(A, m_{\Phi_0}) \xrightarrow{P_1} (A, m_{P_1}) \xrightarrow{P_2} ... \xrightarrow{P_n} (A, m_{P_n}),$$

where at each step $m_{P_i}$ is defined on classifier pool $P_i$ with restriction to the fuzzy set $(A, m_{P_{i-1}})$. Finally at the last stage the test image is classified to

$$\Phi_{P_n}(I^t) = \operatorname{argmax}_j\{m_{\Phi_n}(j) : j \in A\}.$$  \hfill (5.6)

The main idea behind this method is when we are employing a large number of classifiers and no classifier votes on a class the probability of that class being the true class of the object tends to zero. This way setting the fuzzy value of the class to zero and neglecting it in further computations seems reasonable. This way the further stages of the classifier pools in the chain have to decide between fewer number of classes and this make the task of object recognition easier.

5.2 Results

In this experiment we use a subset of animals dataset containing the animal classes \{Cougar, Coyote, Deer, Elephant, Goat, Horse, Leopard, Lion, Tiger\}. Animal classes appearing in this subset contain all different texture complexities and also this subset most of the natural and technical difficulties of animal recognition task.

To perform the recognition task 50% of the images was randomly taken as the train set and the rest were considered as the testing set. In order to obtain results independent from the test and train set the experiments were repeated several times and the average classification rates is reported.

Tree different groups of features descriptors were used in these experiments. The first set of these descriptors are the MRF descriptors [30] with different parameters. These descriptors were applied to grayscale images. Each descriptor was applied on both normalized and not normalized views of the object. Here the normalization is denoted to normalizing the gray level intensity of pixels of the objects by by subtracting them from their mean and dividing by their standard deviation. The patch sizes MRF descriptors captured the textural information from patch sizes 3, 5, 7, 9, 11. This way it is ensured that the objects in analyzed in different scale levels. The MR8 descriptor [29] used filters with different maximum $\sigma$ values on a $40 \times 40$ patch. The maximum $\sigma$ value was changed in the values 2.5, 5, 7, 5, 10. The final set of descriptors is the color descriptor. The color features were extracted from HSV, RGB, Opponents color systems. To make dictionaries from the color information. The three dimensional vectors presenting the color of different pixels of the objects were clustered using k-means algorithm.

Using these classifier we present different classifier pools and we present three different experiments in this section. In the first two experiments we analyze the dataset based on texture and color individually. It can be seen from tables 5.1 and 5.2 that the behavior
CHAPTER 5. FUZZY CLASSIFICATION

Table 5.1. The best color spaces for classification of each class.

<table>
<thead>
<tr>
<th>Class</th>
<th>Classification Rate</th>
<th>Color system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Cougar</td>
<td>0.49 ± 0.05</td>
<td>HSV</td>
</tr>
<tr>
<td>2 - Coyote</td>
<td>0.66 ± 0.06</td>
<td>Opponent</td>
</tr>
<tr>
<td>3 - Deer</td>
<td>0.35 ± 0.05</td>
<td>HSV</td>
</tr>
<tr>
<td>4 - Elephant</td>
<td>0.64 ± 0.05</td>
<td>Opponent</td>
</tr>
<tr>
<td>5 - Goat</td>
<td>0.62 ± 0.07</td>
<td>Opponent</td>
</tr>
<tr>
<td>6 - Horse</td>
<td>0.92 ± 0.03</td>
<td>RGB</td>
</tr>
<tr>
<td>7 - Leopard</td>
<td>0.84 ± 0.05</td>
<td>RGB</td>
</tr>
<tr>
<td>8 - Lion</td>
<td>0.72 ± 0.06</td>
<td>RGB</td>
</tr>
<tr>
<td>9 - Tiger</td>
<td>0.82 ± 0.05</td>
<td>Opponent</td>
</tr>
</tbody>
</table>

Table 5.2. Classification using different fuzzy classifiers.

<table>
<thead>
<tr>
<th>Class</th>
<th>MRF</th>
<th>MR8</th>
<th>MRF+MR8</th>
<th>Color</th>
<th>MRF+MR8+Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Cougar</td>
<td>0.36</td>
<td>0.22</td>
<td>0.37</td>
<td>0.47</td>
<td>0.39 ± 0.03</td>
</tr>
<tr>
<td>2 - Coyote</td>
<td>0.57</td>
<td>0.63</td>
<td>0.68</td>
<td>0.66</td>
<td>0.62 ± 0.04</td>
</tr>
<tr>
<td>3 - Deer</td>
<td>0.33</td>
<td>0.25</td>
<td>0.34</td>
<td>0.33</td>
<td>0.36 ± 0.03</td>
</tr>
<tr>
<td>4 - Elephant</td>
<td>0.71</td>
<td>0.80</td>
<td>0.81</td>
<td>0.63</td>
<td>0.74 ± 0.04</td>
</tr>
<tr>
<td>5 - Goat</td>
<td>0.52</td>
<td>0.49</td>
<td>0.57</td>
<td>0.61</td>
<td>0.57 ± 0.06</td>
</tr>
<tr>
<td>6 - Horse</td>
<td>0.70</td>
<td>0.71</td>
<td>0.75</td>
<td>0.91</td>
<td>0.75 ± 0.07</td>
</tr>
<tr>
<td>7 - Leopard</td>
<td>0.92</td>
<td>0.88</td>
<td>0.93</td>
<td>0.85</td>
<td>0.94 ± 0.01</td>
</tr>
<tr>
<td>8 - Lion</td>
<td>0.67</td>
<td>0.75</td>
<td>0.77</td>
<td>0.70</td>
<td>0.71 ± 0.04</td>
</tr>
<tr>
<td>9 - Tiger</td>
<td>0.78</td>
<td>0.75</td>
<td>0.80</td>
<td>0.79</td>
<td>0.82 ± 0.06</td>
</tr>
</tbody>
</table>

of different classes is different regarding different descriptors. In the first experiment it is important to see that if we truly wish to find a feature that best classifies the animals it has capture information form all different color spaces. It can also be seen that the classification rates only based on color are comparable to the classification rate of other descriptors. This experiment was done to analyze how the color features behave. So far the features pools have not been employed.

In the second experiment several feature pools are defined and the classifiers, classify the test objects to the class with maximum vote. It is natural to classify the object to the class with the minimum distance to the models when the number of votes are equal. The classifier pools defined in this experiment are as follow.

- **MRF**, containing all MRF descriptors and classifiers.
- **MR8**, containing all MR8 descriptors and classifiers.
- **MRF + MR8**, containing both MRF and MR8 descriptors and classifiers.
- **COLOR**, containing all color descriptors and classifiers.
- **MRF + MR8 + Color**, containing all the classifiers used in this experiment.
5.2. RESULTS

<table>
<thead>
<tr>
<th>Class</th>
<th>Chain 1</th>
<th>Chain 2</th>
<th>Chain 3</th>
<th>Chain 4</th>
<th>Chain 5</th>
<th>Chain 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Cougar</td>
<td>0.28</td>
<td>0.49</td>
<td>0.32</td>
<td>0.51</td>
<td>0.49</td>
<td>0.51 ± 0.06</td>
</tr>
<tr>
<td>2 - Coyote</td>
<td>0.67</td>
<td>0.73</td>
<td>0.70</td>
<td>0.72</td>
<td>0.76</td>
<td>0.73 ± 0.05</td>
</tr>
<tr>
<td>3 - Deer</td>
<td>0.31</td>
<td>0.35</td>
<td>0.31</td>
<td>0.37</td>
<td>0.37</td>
<td>0.36 ± 0.05</td>
</tr>
<tr>
<td>4 - Elephant</td>
<td>0.83</td>
<td>0.72</td>
<td>0.82</td>
<td>0.71</td>
<td>0.83</td>
<td>0.72 ± 0.05</td>
</tr>
<tr>
<td>5 - Goat</td>
<td>0.58</td>
<td>0.67</td>
<td>0.54</td>
<td>0.66</td>
<td>0.69</td>
<td>0.61 ± 0.05</td>
</tr>
<tr>
<td>6 - Horse</td>
<td>0.75</td>
<td>0.92</td>
<td>0.75</td>
<td>0.91</td>
<td>0.89</td>
<td>0.92 ± 0.04</td>
</tr>
<tr>
<td>7 - Leopard</td>
<td>0.93</td>
<td>0.86</td>
<td>0.89</td>
<td>0.86</td>
<td>0.93</td>
<td>0.86 ± 0.05</td>
</tr>
<tr>
<td>8 - Lion</td>
<td>0.81</td>
<td>0.78</td>
<td>0.81</td>
<td>0.75</td>
<td>0.81</td>
<td>0.78 ± 0.06</td>
</tr>
<tr>
<td>9 - Tiger</td>
<td>0.81</td>
<td>0.84</td>
<td>0.78</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84 ± 0.04</td>
</tr>
</tbody>
</table>


The fuzzy classifier used in this stage contained only one stage and the results were generated based on a voting of the classifiers of the classifier pool. Several interesting properties of the classifiers can be seen by taking a close look at the table 5.2. It is interesting to see that MRF and MR8 descriptors seem capture different information and by their combination a higher classification rate is achieved in every class. It is also interesting to see that the classification rates obtained from color classifier is higher all other combinations.

In the third experiment we wish to combine different classifier pools in fuzzy classifier with different stages to obtain an order of fuzzy classifiers which gives a performance close to the best performance on all the animal classes. In this experiment six different fuzzy classifier chains were used.

1. **MR8 → MRF → COLOR**
2. **COLOR → MR8 → MRF**
3. **MRF → MR8 → COLOR**
4. **COLOR → MRF → MR8**
5. **TEXTURE → COLOR**
6. **COLOR → TEXTURE**

Here the TEXTURE classifier pool indicates the union of MRF and MR8 classifier pools. It can be clearly seen in table 5.3 the results achieved using the last fuzzy classifier chain are the highest classification rates achieved in these set of experiments.
Figure 5.2. A demonstration of how different stages of fuzzy classifiers are connected to each other. At each stage the number of possible classes for the image is reduced.
Chapter 6

Joint Visual Vocabulary

So far the analysis were mostly focused on feature evolution and extraction. Different features were extracted from the image and a voting method was introduced for combining these features to obtain a stronger universal classifier. In this part of the thesis another methodology is being followed. A limitation of the method introduced in chapter 5 and many other methods is that their performance depends on the performance of different segmentation algorithms.

To capture the texture and color information, we use the dense MRF texture descriptor and use an approach similar to the single-histogram models [25]. Histograms of visual words have been used previously for region or image level classification for both dense and sparse methods. To build the vocabulary from training images, for every pixel of the image in the training set, a feature vector is created from a small patch, here $5 \times 5$ in the CIE-LAB color space. These feature vectors are then clustered using k-means algorithm. Each of the cluster centers generated using this method is called a visual word or a texton. Based on these set of cluster centers, we assign each pixel in the training images with the closest visual word in the vocabulary. Then, we compute histograms of visual words for each of the training patches. Therefore each training example is modeled by a histogram. The idea of the single-histogram [25] method is, to combine these histograms and represent each object category by a single one. The advantage of this method is the compactness of the model; however, it might be to simple to capture all variations in the training data. Then, the test image is converted into its corresponding texton map. Based on this map, a histograms of visual words is computed for patches within the image. For classification e.g. the k-nearest-neighbor classification (k-NN) approach can be used to find the closest histogram within the training sets.

To avoid the need for segmenting the region or object from the background, a sliding window is slid across the image to generate a histogram of visual words for each position. The center pixel of the window is then classified based on the obtained histogram. Therefore, an image can be classified into several object classes such as sky, grass, or animal. This can be obtained by making an assumption that all regions and objects are represented by separate classes. In other words, there is no generic background class. This is a limitation when dealing with object classification within an image with complex background
such as those of natural scenes in which animals are usually found. Furthermore, as it was already mentioned, there exist large intra-class variability and inter-class similarities. It is not a surprise that extremely simple models such as the single-histogram models might not be able to capture all these variations. Obviously, there is a need for more complex representations, which will also be proven through the experiments.

In our model, we start with the single histogram and instead of considering each visual word individually, we look at the relation between the visual words within a region. This is inspired by text processing principles, where the relation of central words with other word in a specific text or paragraph (region) is considered. Therefore here we calculate probability of occurrence of a visual word conditioned on the visual word in the center of the patch. More specifically we perform $\Pr(t_j | t_i)$ when $t_j$ is the visual word in the center of a patch and $t_i$ is any other visual words in the vocabulary. With this second statistical model, we capture the structure of the object when the region contains only information of the object. It even captures more information when the region contains some information of the object context or background. Therefore against of many other recognition approach, such as single histogram, which ignore the background, we take the advantage of available information from the context of the object to describe it. We will evaluate this hypothesis in our experiments.

Different animals appear in different contexts. The appearance such as texture and structures of the surrounding the animal can help recognizing it. The method proposed in this thesis both captures the joint distribution of the visual words appearing in the object also when an almost accurate ground truth segmentation provided, it can capture the joint distribution between the visual words appear in the images and those appearing the neighborhood background of the object. In this manner the contextual information around the animal is being included in the models in an unsupervised way. It is shown in figure 6.1, increasing the amount of joint contextual information effects the overall classification.

### 6.1 Joint Probabilities

The idea of using the joint probabilities is to capture the joint distribution of the visual words, in order to obtain better features for classification. Joint distribution of the textons captures the probability of different visual words appearing in a neighborhood of each other, in different classes. In this application the neighboring visual words were determined using a sliding window technique. It can be easily shown that the maximum likelihood estimate of appearing the visual word $t_j$ in a neighborhood of the visual word $t_i$ with respect to the class $c$ is calculated as

$$\Pr(t_j | t_i, c) = \frac{\sum_{q \in N_{(c,t_i)}} (T(q) == t_j)}{\sum_{q \in N_{(c,t_i)}} 1}$$

(6.1)

where, $N_{(c,t_i)}$ denotes the union of local neighborhoods of the pixels labeled as the visual word $t_i$ in the training region of class $c$ and $T(q)$ returns the visual word positioned at the pixel $q$.  

36
6.1. JOINT PROBABILITIES

The model learnt for each class is \( M_c = [m_{c}^i]_{K \times K} \) with \( m_{c}^{i,j} = \Pr(t_j | t_i, c) \). The \( i^{th} \) row of this matrix is the visual words distribution around the visual word \( t_i \). This vector is denoted by \( m^i_c = [m^i_{c,j}]_{1 \times K} \).

With having the visual words dictionary and models for classes, \( C_1, C_2, \ldots, C_n \) and the background class \( C_B \), we wish to define the probabilities \( \Pr(C_i | I_{\text{test}}) \) for every test image \( I_{\text{test}} \). Every visual word in the test region is then classified according it’s neighboring visual word distribution. Assume that \( n(N|T) \) is the normalized histogram of the visual words within the neighborhood \( N \) with the center visual word \( t_i \), positioned at pixel \( p \).

Using these information this visual word is classified as

\[
    c^*(p) = \arg\min_{c \in \{C_1, \ldots, C_n, C_B\}} \{d(n(N|T), m^i_c)\},
\]

where, \( d(\ldots) \) denotes the \( \chi^2 \) distance between the histograms. Finally the probability of occurrence of each class within the test image for \( c \in \{C_1, \ldots, C_n\} \) is defined as

\[
    \Pr(c | I_{\text{test}}) = \frac{|p \in I_{\text{test}} : c^*(p) = c|}{\sum_{j=1}^{n} |p \in I_{\text{test}} : c^*(p) = C_j|}.
\]

6.1.1 Support Vector Machine Ensemble

We used SVMs [6] models as an alternative representation of the distributions of visual words over the images of different animal classes. As described previously, we conditioned the probability of occurrence of an animal class \( c \) in the center of the given image window \( w_p \) on the visual word that was assigned to the center pixel of the window \( t(p) \). This means creating separate models, for each of the visual words, trained on a subset of the training data. Creating such an ensemble of classifiers might increase performance, but also divides the problem into smaller subproblems. This is particularly important in case of SVMs for which the complexity of the training process is approximately quadratic in the number of training samples. In our experiments, this allowed to perform training on all histograms extracted from windows centered at all the pixels in the training regions (18.6 million histograms). Consequently, we trained \( K \) multi-class SVM classifiers \([6, 20]\). \( \{F_{t_i}(x)\}_{i=1}^{K} \), one for each of the center visual words. Each classifier was trained on a labeled subset of all the training visual word histograms.

\[
    \{(n(w_p|T), c) : p \in Q_c \cap t(p) = t_i \}_{c \in \{C_1, \ldots, C_n, C_B\}}.
\]

In this paper we will use the \( \chi^2 \) kernel [5] \( K(x, y) = \exp\{-\gamma \chi^2(x - y)^2\} \), which has shown to give good performances for histogram-like features in several domains [5, 20].

In this work, we used the pairwise approach, in case of which \( N(N - 1)/2 \) two-class machines are trained for each pair of classes. The final decision is then obtained by considering as output of each classifier the class label and counting votes for each of the classes. The test sample is then assigned to the class which receives most votes.

During recognition, each pixel of the test image was labeled based on the visual word histogram extracted from the surrounding window.

\[
    c(p) = F_{t(p)}(n(w_p|T)).
\]

37
Note that appropriate classifier was selected depending on the visual word in the center of the window. The final decision as well as the ranking of hypotheses was then calculated according to Eq. 3.

6.2 Results

This section presents the experimental setup we used, the procedure we followed, and the results we obtained during the evaluation of our approach based on both types of models and the single histogram technique. All the experiments are performed on the animal image database being one of the contributions of this paper.

6.2.1 Experimental Setup and Procedure

There are several parameters affecting the performance of the proposed method. One of such parameters is the size of the sliding window. This parameter has a significant influence on the performance and will be discussed in more details later. Another value which determines how well the images can be modeled is the size of the visual words dictionary. The size of the visual words dictionary should be large enough to capture the variations of the dataset. However, when the number of words grows, the efficiency of the algorithms drops. Therefore in the experiments we used a visual words dictionary of size 1500. The kernel and training parameters for the experiments with SVM were the same for all models ($\gamma = 1$ and $C = 100$) and selected based on a small set of preliminary experiments.

In order to perform the experiments, the database was randomly split into a test training sets. Each set contained 50% of the images. The training set was first used to build the dictionary of visual words and then to train the models. In case of the experiments with the single histogram and joint probability methods, the experiments were performed several times for different random splittings. Since, the obtained results were stable over the splittings, the experiments with SVM were conducted for one pair of training and test sets only.

6.2.2 Experimental Results

We performed two types of experiments in order to find the proper parameters of the methods and evaluate their performance. First, we investigated the influence of the sliding window size on the overall results. One of the problems addressed in this work was whether the use of contextual information can facilitate the recognition. The size of the window is the main parameter which determines how much contextual information is being captured. In our experiment we varied the size of the window from $31 \times 31$ pixels to windows with $211 \times 211$ pixels and measured the recognition rate for the joint probability approach. The size of most of the images used in the experiments was approximately $384 \times 384$ pixels. The results are presented in Figure 6.1. On one hand, when the window is too small the classification rate is low since too little information is captured about the object. On the other hand, when the window is too large the classification rate drops, since the models models contain more information about the background than the object itself. The best
6.2. RESULTS

![Figure 6.1. Dependency between the classification rate and the size of sliding window. In this experiment a visual words dictionary of size 1500 were used. The size of most of the images in the experiments was equal to $384 \times 384$ pixels.](image)

A classification rate was achieved for the sliding window of size $121 \times 121$ pixels, which, when compared to the size of the images, clearly shows that a lot of contextual information was available for the model.

In the second set of experiments, we evaluated the performance of our method based on two types of models (simple probabilistic model and SVM) and compared it to the one achieved by the single histogram technique. The overall results for each approach are given in Figure 6.1. It is apparent that the model consisting of a single histogram for each class was unable to encode the complex dependencies in the data. As the more sophisticated methods are employed, the classification rate increases by 26% in case of the joint probability model and another 10% in case of the SVMs. Methods like single histogram [25] were not designed to find objects in complex unknown backgrounds. Usually these methods are applied on datasets where almost every pixel of an image belongs to a class and for the classification the method is forced to choose between one of these classes. In the case of animals dataset a background class is defined and the pixels classified to the background classes are neglected when the probability of appearance of different animal classes is being computed. This performance is maximized when the learning method can best distinguish between background and animal visual words. Naturally one expects the background class contains large variations and due to performance of the single histogram class models it can be easily seen that these models fail to encode the complex background.

When it comes to image search applications, more than one hypotheses can be considered. Most of the web based image search engines find images based on the relative words appearing in images. When employing a computer vision algorithm for aiding such systems we always expect to get the result among several highest ranked images. This way a better decision can be made by combining the results of different sources of information. Table 6.1 shows the percentages of correct classification among the first two or three hypotheses. It is clear that, the applied recognition framework can output not only a single
CHAPTER 6. JOINT VISUAL VOCABULARY

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Rate 1</th>
<th>Rate 2</th>
<th>Rate 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-hist. [25]</td>
<td>0.39±.01</td>
<td>0.60±.01</td>
<td>0.70±.01</td>
</tr>
<tr>
<td>Joint Prob.</td>
<td>0.65±.02</td>
<td>0.78±.01</td>
<td>0.84±.02</td>
</tr>
<tr>
<td>SVM</td>
<td>0.75</td>
<td>0.84</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 6.1. The classification rates obtained by the introduced methods and the single-histogram method. Average values with standard deviations are presented for the single histogram and joint probability techniques. The rates 1, 2, and 3 correspond to classification rates when one, two, or three best hypotheses were taken into account.

<table>
<thead>
<tr>
<th>Single-Hist</th>
<th>Joint</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rate 1</td>
<td>Rate 1</td>
</tr>
<tr>
<td>Rate 2</td>
<td>Rate 2</td>
<td>Rate 3</td>
</tr>
<tr>
<td>1-bear</td>
<td>0.07</td>
<td>0.42</td>
</tr>
<tr>
<td>2-cougar</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td>3-coyote</td>
<td>0.38</td>
<td>0.52</td>
</tr>
<tr>
<td>4-elephant</td>
<td>0.32</td>
<td>0.64</td>
</tr>
<tr>
<td>5-giraffe</td>
<td>0.27</td>
<td>0.33</td>
</tr>
<tr>
<td>6-goat</td>
<td>0.18</td>
<td>0.60</td>
</tr>
<tr>
<td>7-horse</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>8-leopard</td>
<td>0.35</td>
<td>0.78</td>
</tr>
<tr>
<td>9-lion</td>
<td>0.31</td>
<td>0.88</td>
</tr>
<tr>
<td>10-panda</td>
<td>0.50</td>
<td>0.67</td>
</tr>
<tr>
<td>11-penguin</td>
<td>0.46</td>
<td>0.68</td>
</tr>
<tr>
<td>12-tiger</td>
<td>0.18</td>
<td>0.88</td>
</tr>
<tr>
<td>13-zebra</td>
<td>0.54</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 6.2. A detailed comparison of performance of the three method for each of the single classes. Classification rates for two and three best hypotheses are also presented for the method based on the Support Vector Machines.

decision, but is also able to provide a meaningful ranking of hypotheses.

The obtained results can be further analysed. By plotting the results for each single class, we can see if there are difficult classes that pose a problem for the recognition system. Indeed, it can be seen from Table 6.2 that such animals as giraffe, bear or cougar are particularly difficult to recognize using the evaluated methods. Still, it can be observed that the correct classification is usually among the first two or three hypotheses, and the classification rate quickly improves when more than single decision is considered.

To complete the analysis the joint probabilities method with the best parameters obtained from the previous analysis was ran on all the 19 classes of animals. The results of this experiment can be seen in table 6.3.

40
6.3 Joint Visual Vocabulary for General Object Recognition

In the mathematical details of the methods described in chapters 5 and 6 there is no restriction on the visual properties of the objects these methods learn. Despite this fact, so far these methods had only been applied on the KTH-animal. In this section the focus is on the behavior of the Joint Visual Vocabulary method on several non-animal image categories taken from the MSRC database [32]. In this section it is important to show that the Joint Visual Vocabulary method shows a reasonable behavior on non-animal classes in this database.

In this experiment we used a subset of the MSRC database [32]. This subset contains the classes \{Building, Sky, Plane, Face, Car, Bicycle\}. This subset contains several visually different objects rather than the usual challenges we discussed in animal recognition. Usually most of the objects appearing in this subset have a rigid shape and differences between different objects is due to the viewpoint and and their color. Because of an almost accurate manual segmentation provided, this database makes a good candidate for our test.

For the experiments in this section the same methodology were used in 6.1. Due to change of the dataset some small modification was made into the method. In this dataset there are images with multiple objects within each image. Also the number of images is too few to create a robust background class model. For these reasons here we only analyze the region selected in the ground truth to identify the object. This correction makes our work more comparable to the recognition part done in Single-Histogram class models [25].

### Table 6.3. 19 class classification rate.

<table>
<thead>
<tr>
<th>Animal</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear</td>
<td>0.40</td>
</tr>
<tr>
<td>Cow</td>
<td>0.18</td>
</tr>
<tr>
<td>Giraffe</td>
<td>0.21</td>
</tr>
<tr>
<td>Gorilla</td>
<td>0.52</td>
</tr>
<tr>
<td>Kangaroo</td>
<td>0.22</td>
</tr>
<tr>
<td>Panda</td>
<td>0.59</td>
</tr>
<tr>
<td>Penguin</td>
<td>0.42</td>
</tr>
<tr>
<td>Sheep</td>
<td>0.18</td>
</tr>
<tr>
<td>Skunk</td>
<td>0.39</td>
</tr>
<tr>
<td>cougar</td>
<td>0.36</td>
</tr>
<tr>
<td>Coyote</td>
<td>0.54</td>
</tr>
<tr>
<td>deer</td>
<td>0.12</td>
</tr>
<tr>
<td>elephant</td>
<td>0.60</td>
</tr>
<tr>
<td>goat</td>
<td>0.56</td>
</tr>
<tr>
<td>horse</td>
<td>0.88</td>
</tr>
<tr>
<td>leopard</td>
<td>0.76</td>
</tr>
<tr>
<td>lion</td>
<td>0.84</td>
</tr>
<tr>
<td>tiger</td>
<td>0.84</td>
</tr>
<tr>
<td>Zebra</td>
<td>0.87</td>
</tr>
</tbody>
</table>

6.3.1 Results

Parameters such as the size of the sliding window and the size of the visual words dictionary are usually database dependent. These parameters usually depend on the variation of the database. These parameters were obtained through heuristic analysis. In our analysis we obtained the optimum size of the visual words dictionary as 100 and the size of the sliding window 25 pixels. It is clearly seen that the size of the dictionary used here is much smaller than the size of the dictionary used in single-histogram class models. This shows that the information captured by our method is indeed different than the information captured using the single-histogram class models.

Since it is required to generate recognition results independent from the training set 10 random partitions were created over the MSRC database. The average recognition rate
Table 6.4. Confusion matrix for classification on six classes \{Building, Sky, Plane, Face, Car, Bicycle\} of the MSRC[32].

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-building</td>
<td><strong>0.75</strong></td>
<td>0.15</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>2-sky</td>
<td>0.00</td>
<td><strong>0.98</strong></td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3-plane</td>
<td>0.06</td>
<td>0.00</td>
<td><strong>0.94</strong></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4-face</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td><strong>1.00</strong></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5-car</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td><strong>1.00</strong></td>
<td>0.00</td>
</tr>
<tr>
<td>6-bicycle</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.06</td>
<td><strong>0.94</strong></td>
</tr>
</tbody>
</table>

on these partitions using the setting described above is $0.88 \pm 0.02$. This recognition rate is lower than the recognition rate achieved by the single-histogram models. By taking a close look at the table 6.4 one can see that the only class acting poorly using these models is the building class and other classes have a reasonably high classification rate. Usually buildings consists of different parts with different texture information. As mentioned previously MSRC database does not contain enough instances of different objects. There for it is possible that the models are not completely filled for the building class to capture all variations.
Part IV

Conclusions and Future Work
Chapter 7

Conclusions and Future Work

7.1 Conclusions

Animals have proven to one the most difficult objects for the object detection task. The challenges one faces in the task of animal recognition are not reflected in most of the available datasets. Moreover most of the method created on these datasets are incapable of capturing large variations of a realistic dataset. Also most of the recognition methods based on these datasets need to have a segmentation information of the images as prior knowledge. In chapters 1 and 2 a review of some of these existing methods is given.

The first achievement of this thesis is gathering and annotating a challenging database of animal images. This database is introduced in chapter 4. The number of images available in each class is large enough for training and testing complex models. Also as it can be seen in chapter 4 the images of the dataset gather most of the difficulties a realistic animal database should have. These difficulties are argued in chapter 1.

With having a new database it was necessary to do a feature evaluation test to see how different features behave on this dataset. In chapter 5 several textural analysis is done on the animal dataset with the unrealistic assumption of having perfect segmentation and knowing the location of the animal. With this assumption it was possible to analyze the animals dataset only based on their skin texture. These analysis showed that different cues of information play a vital role in the classification and by correctly combining them one can achieve higher classification rate.

There is a need for recognizing objects within images without having the segmentation information of the image. To satisfy this need the joint visual vocabularies were introduced in chapter 6. The models build on joint probabilities capture the relation between the visual words of the objects. To encode the contextual information they also capture the relation between the visual words of the objects and the visual words appearing in the surrounding of the objects. This process is done in an unsupervised manner. To train these models a database with manual segmentation of the objects is required. In the test process the images can be classified based on the joint distribution in the local regions of the image. This experiments showed that contextual information can be a great aid when dealing with outdoor images. Also it was showed that how such a method can be employed in a rank
CHAPTER 7. CONCLUSIONS AND FUTURE WORK

Most of the analysis were done on the KTH-animal dataset. To show that the proposed method is capable of describing general objects it was applied to 5 non-animal classes from the MSRC database. Reasonable results were achieved through these experiments and it show that the method is capable of capturing information from general object categories.

7.2 Future Work

In the proposed these only textural information of the images were used. As discussed it is possible to combine different cues of information to achieve a better and more robust classifier. One of the ways to continue the work of this thesis is to create joint visual vocabularies from different cues of information such as shape, color and texture and combine them together. Moreover creating multi-scale models can also aid the recognition rate. These models can be made both on dense and sparse features. Depending on the database the performance of these features might be different.
Bibliography

BIBLIOGRAPHY


