Universität Karlsruhe (TH)
Fakultät für Informatik
Institut für Theoretische Informatik
Prof. Dr. rer. nat. Alex Waibel

Diplomarbeit

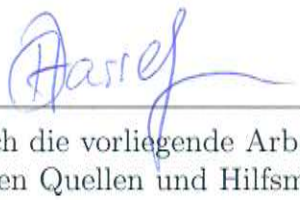
Bridging Global and Local Features in Pattern Analysis

with Application to Car Manufacturers' Logo Recognition

Dennis Harres

30. Juni 2006

Prof. Dr. Alex Waibel
Dr. Rainer Stiefelhagen
Jie Yang, Ph.D.
Datong Chen, Ph.D.



Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 30. Juni 2006

Abstract

Human vision is of remarkable nature. We barely notice the amazing variety of (visual) conditions, under which it performs day-to-day, and the high complexity behind our perception. As self-evident and trivial it seems for a human being, it challenges researchers around the world even more in their goal to achieve comparable results in the field of pattern analysis, as a part of artificial intelligence. Thus, the recent development addresses more complex problems with increasing relaxation of restrictions in application's conditions. This leads to reinforced allusion to human vision (i.e. [42]).

In the scope of this work, we propose a new representation approach which depicts a complex pattern, based on local descriptors [30] and geometrical constraints between them. This achieves bridging local representation and global representation in order to obtain a non-rigid appearance perception as an imitation of human vision. We present and examine the simplest such representation: a pair of local descriptors. In terms of textual categorization, we call the feature pattern a "visual word" [27] and assign a set of visual words to "describe" the target class in the best way. In human language we would call them "keywords". For finding the best keywords we present a feature selection method based on feature relevance definitions [26] and document frequency for ranking.

As application for our approach, we chose car manufacturer's logo recognition. This scenario features many classic challenges like image transformations and environmental challenges in the corresponding data (car images). The variety of logo appearances represent a good basis for measurement and comparison of representation methods and their impact in recognition process, rather than that of classifiers' performance.

The experimental results show partially significant improvement of recognition performance when utilizing the proposed representation method. Further, according to the achieved recognition rates our representation method exhibits less sensitive behavior to object appearance and thus captures the overall object "structure" in a better manner, but still distinctive enough for recognition process.

We conclude on the interest on more complex patterns beyond the pair representation, where a single feature vector can represent a single object as in global representation, and where it can have a correlation with the words of human language like "face", "chair", "window" in terms of categorization problems.

Acknowledgements

This work was conducted at the *Interactive System Labs* (ISL) of Carnegie Mellon University, USA, and the *Institut für Logik, Komplexität und Deduktionssyteme* at the Universität Karlsruhe (TH), Germany. I would like to thank Prof. Dr. Waibel for the InterACT student exchange program, which gave me the opportunity to do my research in Pittsburgh, PA, provided an insight into American culture and enabled a memorable, unique experience.

I am grateful to my advisors Jie Yang, Ph.D. for his support of my work and the open-hearted help with the organization of my stay and beyond research, as well as Datong Chen, Ph.D. for the conduction, discussions and suggestions to my work. I would like to thank Dr. Rainer Stiefelhagen for the support in Karlsruhe, Germany.

I would like to thank Dr. Thomas Schaaf, Linda Hager, Celine Carraux, and Kristen Messinger for their help in administration and organization of my stay at CMU, Jan Niehues and Kay Rottmann for their hospitality during the first days, and all other fellow students and ISL colleagues for the great time in Pittsburgh.

Contents

List of Figures

List of Tables

1. Introduction

Pattern analysis has been under continuous and intense exploration since the beginning of artificial intelligence. However, there are still many unsatisfactorily solved problems and high number of challenging conditions – but generic approaches are barely available [3]. Human vision and cognitive capabilities are still unreachable for any algorithmic solution in terms of robustness and accuracy, especially, under permanently changing conditions and with application in manifold environments.

The recent development in the field of pattern analysis [10, 17, 21, 30, 40, 41, 42] reveals a broad examination of new challenging tasks. Moreover, despite the classic application fields, a new bias toward the human-related recognition (visual categorization [27, 33, 35, 40, 41], traffic signs [18, 19, 20, 21, 22, 23, 24], etc.) can be observed. This enhances the aspects of human-computer interaction: recognition of key objects which create communication basis both for vision and speech. In the everyday life we often use trademarks and brand names for depicting and distinguishing a specific object out of its kind. This information characterizes this object in the same way color, size, and shape already do in such situations. This "label" extends the precision of our communication and often facilitates it by reducing the needed information amount. Hence, the recognition of trademarks and brands in form of a logo has been already of interest but rather limited to the document domain [1, 2, 3, 4, 5, 17].

In this work we are going to discuss a novel problem of car manufacturer's logo recognition, as an example application setting for our proposal of a new representation method for challenging conditions and uncontrolled environment.

1.1. Goal of this Research

Although logo recognition is an already well-known and in the past often explored problem, our extent to a new out-of-text domain of car manufacturers meets a

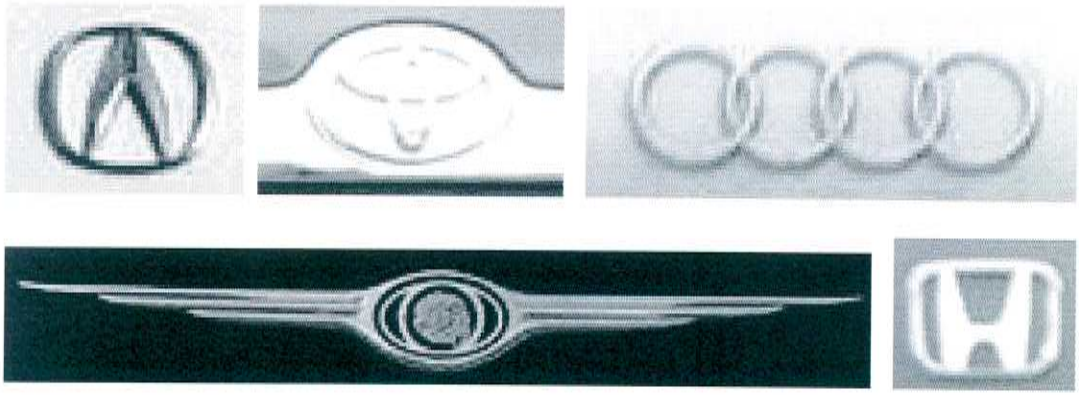


Figure 1.1.: Visual examples of encountered challenges in car logo recognition scenario.

new challenge of the three dimensional real life world. With the new setting, new constraints are to be defined and new challenges ought to be met.

In addition to well-known "classic" challenges, like translation, rotation, and scale, as well as the environmental ones – due to the illumination and weather changes, we see following issues, as shown in figure 1.1, requiring special attention. First of all, car logos in particular are not planar structures and change their appearance with varying light direction by casting shadows. Further, low contrast images are common due to similar or shiny background which makes luminance information barely useful. Both luminance and color information lose, further, their importance as robust representation by reason of partially significant appearance changes among the samples of one class (compare figure 6.1). As some objects tend to have the same shape form or dimensions, this is *not* true for car logos, too – they reveal many variations in size and partial resemblance to other structures, i.e. letter shapes on license plates.

In terms of a recognition task, there exist multiple fields we can tackle to meet the above-named challenges. We want to consider the most early stage of the recognition process: representation phase and explore the additional impact of preprocessing. Representation proposals of previous work can be coarsely divided in two groups: global representation and local representation. While global approaches are common in many appearance-based pattern recognition tasks, they often fail with even small changes in lighting conditions and poses. Local description, on the other side, is more insensitive to common image deformations [28] and has gained more attention in recent years [8, 27, 28, 30, 33, 35, 40, 41].

In this work we are going to propose and a novel representation approach: *bigram local feature* and discuss it within the introduced scenario of car logo recognition

by using its challenging conditions. We see this approach as the first step in bridging global and local features - from an individual local feature to creation of a local "pair". Extending this to more complex and/or sequential pattern eventually leads to global representation, but rather based on local(!) appearance. Hence, these complex patterns incorporate the structural perception, rather than the captured appearance.

1.2. Possible Fields of Application

Car manufacturer's logo recognition can be used in multiple scenarios with versatile goals, which can be obviously separated in application as a primary or supporting task.

In the field of parking lot surveillance, car model recognition can be used to create an intelligent environment. Often, car owners leave their vehicles without noticing its correct location or forgetting it by the time of pick up. Only some of them would remember the license plate number, but surely most of them would know the car manufacturer and the color of their vehicle. These details would be enough to locate the few possible candidates on the whole parking lot area.

Another application field with primary priority to car manufacturer's recognition – as a part of an automated system – would be car registration. Either such system would be employed for toll-paid highways or for car identification in a regular inspection.

The field of intelligent vehicles presents a major candidate for car make recognition in a supporting role. First, a human-understandable identification of surrounding and approaching vehicles facilitates the interaction with the information and warning systems of an intelligent vehicle. On the other hand, logo recognition in common sense can be employed for interactive navigation systems using buildings of general known fast-food and coffee franchising chains as landmarks.

Besides the car related logo recognition, environment or scene representation benefits in general from logo recognition for a more precise object description and characterization.

1.3. Outline

The structure and contents of the next chapters summarizes our system as application independent initialization (3), our approach (4) with feature selection (5),

and application dependent setup (6) (compare figure 3.1).

In chapter 2 we present a survey on recent work. Due to wide range of related approaches on the side of our proposed method and the new application scenario, it covers the fields of symbol and shape recognition, face recognition, local representation, textual categorization, and feature selection.

Chapter 3 describes some *initialization* steps and details of the implemented system for car logo recognition. We show how the approaches of local representation and of document categorization are combined for recognition task, as found in recent work.

We present the proposal of our *bigram local feature* in chapter 4. The advantages of local and global representation are discussed and represent the motivation for bridging both representation manners. Further, we describe the construction of our bigram local feature vector out of a pair of local descriptors and its utilization as a "visual word".

In chapter 5 we explain our linear *feature selection* filter. It is based on three rules of feature relevance. They aggressively adjust the computed document frequency ranking accordingly, in a class-wise relation in order to select some number of (class-)distinctive features. This number is identified by the subset selection function.

The results of the conducted experiments are discussed in chapter 6. We explain the *application dependent* system setup, collected data, and composition of data sets. Then we define the comparison set: different representation methods and evaluate the achieved performance of each in the car logo recognition scenario.

Finally, in chapter 7 we conclude the results of this work and discuss future improvements to our approach and further application fields.

2. Related Work

In first place, car manufacturer's logo recognition essentially belongs to a superior class of classic symbol and shape recognition. Thus, it has a wide range of common problems within this field. On the other hand, it features some specific challenges, only found beyond the classic two dimensional approaches like analysis of textual symbols and shapes. Therefore, an abstraction to application in real world environment and problem extension to the third dimension are needed. Appropriate candidates for such problems are the related fields of traffic sign and license plate recognition. They share a large subset of environmental challenges we defined in chapter 1. Still, even if used in 3D environment they exhibit only two dimensional or planar nature of patterns. Specific to car manufacturer logos, the 3D structure of the objects creates new challenges within the pattern itself. Thus, we are going to reference some basic approaches of face recognition to explore their relevance and impact in our scenario.

2.1. Symbol and Shape Recognition

2.1.1. Textual Logos and Shapes

Although, in recent time the focus in the document analysis has moved from the re-engineering problems to indexing and information retrieval [6], symbol recognition is a still interesting research domain. As for discussion on invariance and robustness of current approaches under real-world conditions, problems of geometric invariance of logos in the document domain has been of interest for a while [4]. Further, utilization of local context within shape matching task has gained attention some years later [13]. This represents the first stage of using local context (descriptors, in terms of recent research) in the neighborhood of the interest points. Indeed, in the recent update [8] to [13], a comparison to the SIFT local descriptor [30] is discussed. Finally, exploration of further new challenging problems like occlusion and noise [8, 17] has become an interesting research subject in more recent time.

An overview on recognition of symbols in documents is given in [1]. Even some statistics may be a little outdated, the major categories of textual symbol and shape recognition problems are: technical and facility drawings, maps of various types, musical scores, logos and others. Classification phase of the investigated approaches often uses template matching. Neuronal nets seem to be not well represented by the time of the survey, although they can be variably used for logo recognition beyond controlled environment and conditions [5, 17]. Nevertheless, many of proposals in this survey deal with more document specific or algorithmic solutions (e.g. [7, 25]). This survey makes clear the distance between the document logo recognition and application in uncontrolled environment – as in case of car maker recognition – because the proposed methods in the document domain address different set of problems.

As summarized in a further, more recent survey on symbol recognition [3], a dominant symbol recognition technique is missing among the many approaches available. A generic symbol recognition remains a challenge because of the domain dependent design and knowledge.

2.1.2. License Plates

In the domain of license plate recognition, the change of priorities is evident. Developing a stable system for uncontrolled environment becomes more interesting than a novel matching solution. The range of such problems varies from expanding dynamic range [14], which is a well-known hardware limitation, to utilization of tracking methods [15, 16] to achieve more robust and accurate detection and recognition.

In [9] a coarse-to-fine strategy for multi-class shape detection is presented. The search for instances from multiple classes is followed by arrangement of the detected subsets to a global interpretation. The recognition of alphanumerical characters on license plates showed very good results. Again, we meet a combination of local representation, in this work edge segments, and global representation, which was achieved by a structural interpretation of edge segments' combinations.

2.1.3. Traffic Signs

In the same scenario of scene interpretation or environment perception, traffic sign recognition plays a major role. Despite of its higher relevance in road traffic understanding, as well as its navigation status, traffic signs share the same setting and environmental conditions as our logo scenario. Nevertheless, the similarity in recognition can be only narrowed down to the common aspects in environmental

challenges. To their advantage, traffic signs usually have a distinctive set of well comprehensible and elementary rules, like shape, color, size, and placement in the environment [18, 19] which are almost not transferable to logos in a sufficient way.

In general, traffic sign recognition is meant to have real-time capability [18, 21] and thus such systems are limited in computing power and apply fast, well-known methods [18], i.e. Haar wavelets [21] and other [24]. It is usual to separate the detection and recognition part. In this case, the detection phase attracts more attention due to distinctive traffic sign content. Thus, examined methods for detection vary from more sophisticated techniques like genetic algorithms [19] to the simpler horizontal and vertical (color) projections [20, 22]. Further, some systems capture a closer view at the detected sign either with the same camera [24] or using another telephoto camera [18] to obtain better recognition rates. For the same reason some proposals exploit the advantage of in this field common video systems and apply tracking, i.e. [21].

2.2. Face Recognition

A problem of three-dimensional lighting invariance gained a general importance in the field of face recognition. Although these problems are of a more complex and import nature than the one in our scenario, some essential and simple approaches, in terms of computational complexity, could be of interest for our work.

Hence, there are two major categories of recent proposals: two dimensional image processing and three dimensional model mapping. Obviously, any of the latter approaches would be exaggerative for this scenario, but a possibly interesting approach from the first category came to our attention: self quotient image [23]. A simple yet promising algorithm for a lighting invariant representation represents an in-between stage of a gray-scale image and an edge map, eliminating any equal intensity blobs and smoothing luminance changes. The achieved representation reveals additional parameter to the edges: thickness or intensity. While this method is successfully used for global representation approaches, it is new to apply it in local descriptors.

Further, an application of Gabor wavelets or filters has gained much attention in recent research, i.e. [42], in the field of face recognition. Gabor kernels exhibit a two-dimensional receptive field profiles similar to those of the mammalian cortical simple cells. Gabor wavelets have the desired attributes to capture the object with scale, orientation and illumination invariance if combined to a multi-orientation

and -scale feature vector. Typically, Gabor features are used in global representation approaches. We are going to utilize them for the baseline system with global representation as well.

Well known and often applied in the field of face recognition *PCA* [12] and *LDA* methods [11] attack similar lighting challenges as in our setting. While LDA is sensitive to the number of class representatives in order to achieve robust discriminative representation, we abstain from utilizing it in the developed system. For sake of dimensionality reduction, as a possible problem of global representation, we examine the impact of PCA for the logo recognition problem.

2.3. Local Representation

In order to solve occlusion and image transformation issues of the already explored and upcoming problems, methods of local representation eventually gained attention over years. According to [30], the first attempt of matching by using local interest points can be traced back to Moravec and 1981. With improved algorithm by Harris "corner detectors" were broadly accepted for image matching tasks after 1992. But only in 1997 the local descriptors approaches could be extended so far, that general image recognition was possible. Since then, local interest points have gained a very high research attention.

According to Mikolajczyk and Schmid in their performance evaluation study [31], SIFT [30] is the most resistant approach utilizing local descriptors on images with real geometric and photometric transformations.

In general, local representation is done in two stages. First, some points of interest must be detected. Typically, they have to be located on such spots, which are likely to remain stable over transformations. Then, a description of the region around those points has to be created. This last step has the major impact on recognition performance [31]. That fact encouraged to search for further improvements on the descriptors, for instance PCA-SIFT [28].

Local representation is successfully used for unsupervised learning [10], where an attempt of global representation through a joint probability density function on the shape of the constellation of local features is described. In this case, only the positions of the local descriptors are concerned for the learning of the statistical shape model.

2.4. Textual Categorization

Recently, some computer vision approaches have successfully dealt with adaptation of text categorization problems for visual pattern recognition and categorization [27, 40, 41]. Indeed, the problems of recognition and categorization have a close relation to each other. While the first one engages the *identification* problem among the object within the *same* class or group, the latter one concerns only the *distinguishing* those groups. Obviously, both of them are similar in their approaches, but differ in the domain space and parameter significance. Thus, for identification purposes we would utilize the approach with *distinctive* features within *one* target class. On the other side, a categorization approach would still utilize the distinctive features, but in the *intra-class* sense of multiple target classes, and hence, requiring a more *general* representation within *each* class.

The importance of adopted approaches for computer vision is motivated by the "semantics-oriented" results [35] as known in the text domain.

2.5. Feature Selection

With the parallels in the textual categorization new problems arise in managing of large vocabularies. This motivated us to extend the related work survey to the field of feature selection, as we expect to deal with large number of features when creating pairs out of a set of individual features.

Initially, the term of feature *relevance* should be defined, as contributed in [26]. The authors suggest to concern three relevance definitions: (1) strongly relevant, (2) weakly relevant and (3) irrelevant features. Obviously, the prediction accuracy is not affected by removing the irrelevant features. More interesting is the difference of weak and strong relevance. While weak relevant features may contribute to prediction accuracy, the loss of strong relevant features implies worse performance.

Further in this work [26], two models of feature selection are discussed. The simpler *filter* model can be regarded as a preprocessing step, where some designed function assigns weights to the features according to some a priori assumptions. With the *wrapper* model one has a more sophisticated way for relevance assignment by analyzing the feedback of the induction algorithm in an iterative process, which can be already of high computational complexity even with simple heuristic search algorithms, i.e. backward elimination and forward selection.

On the practical side, in [29] a performance comparison of common feature selection methods in text categorization was presented. While *information gain* (IG)

and χ^2 achieved most effective results, a correlation of the both methods with *document frequency* (DF) was observed. The minor drawback of DF was the less aggressive term removal, but on the side of advantages, significant simplicity of the approach and linear computational complexity. Thus, document frequency proved to be a reliable measure for selecting informative features.

2.6. Conclusions on previous work

With our proposal and application scenario, we cover a wide range of previous work. Obviously, recognition approaches of the document domain address different problems oppose to our application scenario and concentrate more on versatile shape representation and recognition techniques. In the field of license plates and traffic signs, which share the same environment with our scenario, the content plays significantly less important role due to more distinctive class of objects (and their appearance). Here, the environmental challenges and detection are more often discussed. That is, fusion between these fields is needed for our problem.

Even face recognition is from a entirely different domain, we keep trace of this field which features the most intense exploration of illumination and three-dimensional appearance problems. Their influence in recognition process is less significant in our approach, though. Further, to solve three-dimensional problems and affine image transformations we deploy local representation to achieve non-rigid object description.

Textual categorization and document retrieval have already motivated others to implement the techniques from that field in computer vision problems. We are going to use "visual word" representation and search for feature selection approaches in order to find relevant ones.

With coverage and fusion of many different domains and problems we hope to comply – at least partially – with the demand for generic symbol/pattern recognition technique [3].

3. Details of the Developed System

Before we introduce the main approach of this work, a short overview of steps prior to our method will be given to create the *context* for it. The content of this chapter resembles the common part of some other local feature based recognition methods found in recent related work. Thus, we discuss some implementation details as used in our system. In general, at the end of this chapter other researchers complete the representation phase. Our contribution is going to be discussed separately in the next chapter.

A visual outline of our system is shown in figure 3.1. This chapter explains the application independent initialization process for our approach and leaves the application dependent setup, like data sets, labeling, preprocessing, and classification methods, for chapter 6.

Like in any other approach, some principal design decisions determine the choice of the supplementary methods like mentioned above preprocessing, etc. For our problem this major decision is to utilize local representation according to prevailing conditions, in order to be able to meet the major challenges [41].

3.1. Local Representation

In general, local representation phase runs through two stages:

3.1.1. Key point detection

It is claimed in [31] that the performance of the local descriptors does not depend on the choice of key point detector, but rather on the accuracy and repeatability of the specific detection method. Therefore, we avoid utilizing some proposed point detectors [30, 32, 33, 34] in order to minimize the error on such early stage of the recognition process, as the algorithmic details of proposed detectors could reveal some possible weakness within our challenges. Instead, we will use feature selection, which will be discussed in chapter 5, at the end of the representation process. This way we can create a relative big number of potential key points

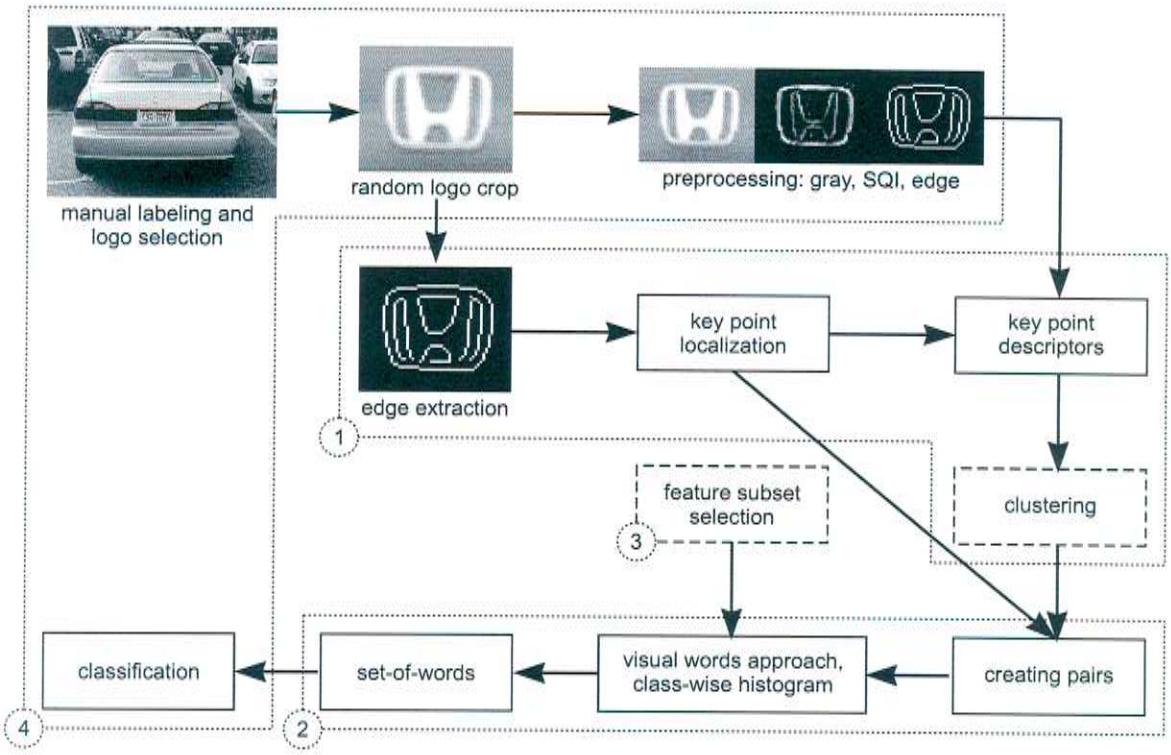


Figure 3.1.: A schematic view of our approach. This diagram presents the outline of our system for car manufacturer's logo recognition. During testing the dashed lined steps provide the results from the training phase. (1) Chapter 3, application independent initialization; (2) Chapter 4, our approach; (3) Chapter 5, feature selection method; (4) Chapter 6, application dependent system setup.

and thus, a high density coverage of the future point relationships – leaving the decision making of their relevance as the last step to be *class-wise*(!).

For key point localization we apply the Canny filter for edge extraction. All the coordinates of the edge points are mapped into a list of (x, y). However, we limit the list's maximal size and take only the first 300 coordinates of spatially uniform distributed grid points.

3.1.2. Region description

In the second stage of local representation, a robust region description around the estimated key points is to be found. Characteristics like invariance to affine image transformations and lighting independent representation are desired. In

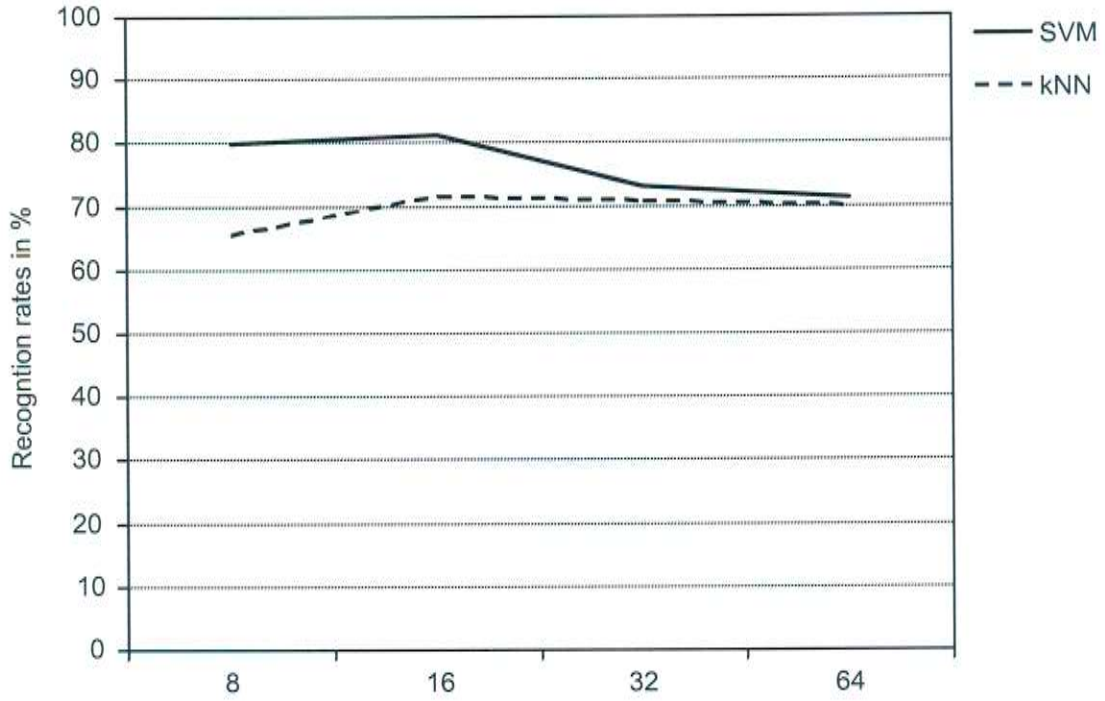


Figure 3.2.: Estimation of the number of clusters for the "pair" feature.

our system we utilize the description phase from SIFT approach [30], which is described in appendix A.1.

For each key point of previously created list a SIFT description vector of 128 dimension is created. This step is iteratively done for all input images.

3.2. Codebook

In the next step we partition the extracted SIFT descriptors across all classes of training images to reduce them to a some number of distinctive general representatives. The number of clusters was determined experimentally. Our feature based on 16 clusters delivered the best performance as shown in figure 3.2.

For the clustering task we utilize one of the simplest partitioning methods: k-means [37]. After initialization of predefined k number of cluster centers, k-means iteratively assigns the points to the closest center and then recalculates new center points of the updated clusters. The iteration stops when all the points stay in the previously assigned cluster and thus, their centers don't drift.

There are some open issues about k-means: its convergence to local optima, initialization dependence and lack of knowledge about the parameter k. As a matter of fact, even with those problems k-means algorithm is most successfully used on large data sets, mostly because of simple implementation and computational attractiveness [36].

Afterward, we assign each SIFT descriptor its closest cluster center and thus, reduce the dimensionality from 128 to one.

4. Pattern Analysis via Ordering Local Features

Since some problems of pattern analysis still reveal non-trivial challenges and the proposed approaches exhibit qualifications to solve a *dedicated* kind of problems, an effort on their fusion and combination has been always of interest. In general, this effort leaned toward combination of rather different approaches with a weighting function for prediction evaluation (i.e. [2]) or by a two-stage recognition processes, like a coarse-to-fine strategy in [9] or joint probability density function on the shape of the constellation of local features [10].

In this chapter we are going to present, to our best knowledge, a novel approach of bridging global and local representation by rather designing a new feature than utilizing existing ones with additional, feature-extern relational knowledge.

4.1. Main Idea

The abstract goal of our approach is to design a new representation approach which depicts a complex pattern based on local descriptors. In this work we examine the simplest such representation: a pair of local descriptors. In terms of textual categorization we call such pair a "visual word". In the training phase each class is assigned a set of visual words which describe the target class in the best way. In human language we call them "keywords". During recognition we match the detected visual words to our trained vocabulary and predict the labels in the test data set.

First, let's begin with advantages and disadvantages of global representation – the initial point for the motivation. Then, we examine the local representation for recognition with its pros and cons. At last we speak about bridging them together in a bigram local feature.

4.1.1. Global Representation

This method of representation is based on *global appearance* of the objects. In general, it doesn't matter which feature space we select to represent the images, but rather that we use the whole object for that task. That is, PCA, LDA [11] or any other feature space transform remain global representation if we apply them on the entire image, for example, faces.

The biggest advantage of global features is, probably, the ease of application. Representation phase can be proceeded with minimal computational and algorithmic efforts. Most of the common methods are also fast and reliable under some controlled conditions.

Beyond the controlled conditions, and thus in real world applications, begin the disadvantages. The famous is certainly the lighting, as global appearance suffer much when the source or direction of light changes. The second major contra argument is the lack of or insufficient invariance to common image deformations and pose, depending on the method. Furthermore, most of the global methods require a precise detection algorithm, which is usually independent of the representation approach and makes the recognition process more error-prone.

4.1.2. Local Representation

This kind of representation was of high interest in recent research. Typically, some smaller regions of the global image and relations between them create the basis for recognition. In this case, the representation has no information about global object appearance and "knows" only some local image patches.

With this comes the biggest advantage of local representation: robustness on changing appearance. Though it can be extended only to a certain degree of perspective image deformations, they achieve good results on affine image changes [31]. Also the lightning invariance is easier to achieve due to small size of affected image regions and their independent processing. Further, the detection process of interest points is an integral part of representation phase. Key point detection for local representation delivers robust performance [31] with different detection algorithms [30, 32, 33, 34].

It seems, global and local features exchange their advantages and disadvantages, which is also true for the fact of application simplicity. In this case, local representation requires additional effort – a grouping or arranging constraint often coupled with classification method.

Especially, we want to keep and use all advantages, with candidate like in SIFT [30], for our approach.

A new approach in the scenario of local representation has been borrowed from textual categorization. "Visual words" [27] are local features used for visual recognition problems with methods from the textual field. For this, interest point descriptors are usually matched by clustering in small partitions and thus, relatively high cluster number. This quantization step creates a vocabulary which all the point descriptors are assigned to in order to become a "word". Analog to text categorization problems they are used without any spatial context in *bag-of-words* [27, 35, 40, 41]. The distribution or histogram of the words is used for classification.

While geometric relations between the local descriptors may be unimportant for categorization problems with a limited class number [27], we expect to prove their high relevance in a recognition problem like ours in chapter 6.

With visual words a simple yet robust method is given for employing local representation.

4.2. Bigram Local Feature

We are going to construct feature vectors from clustered (analog to [35]) SIFT image descriptors with cluster numbers k_i for i-th descriptor from the codebook (see section 3.2) and *additionally* with their key point attributes:

- absolute image coordinates x_i and y_i
- scale s_i
- orientation θ_i

We define our bigram local feature as a pair of descriptors i and j and their *relative* spatial information. We choose such descriptor to be a "start" point, that $k_i \leq k_j$. Then, our feature vector is

$$f_{i|j} = (\partial s_{i,j}, \Delta\theta_{i,j}, \Delta x_{i,j}, \Delta y_{i,j}, k_i, k_j)^T \tag{4.1}$$

This representation keeps the invariance attributes of individual features. This is obvious for the first two elements: $\partial s_{i,j}$ and $\Delta\theta_{i,j}$. The $\Delta x_{i,j}$ and $\Delta y_{i,j}$ are normalized to θ_i and thus, rotation/translation invariant[1]. The scale invariance will be discussed some paragraphs later.

[1]By design this is true if both descriptors i and j have those invariance attributes.

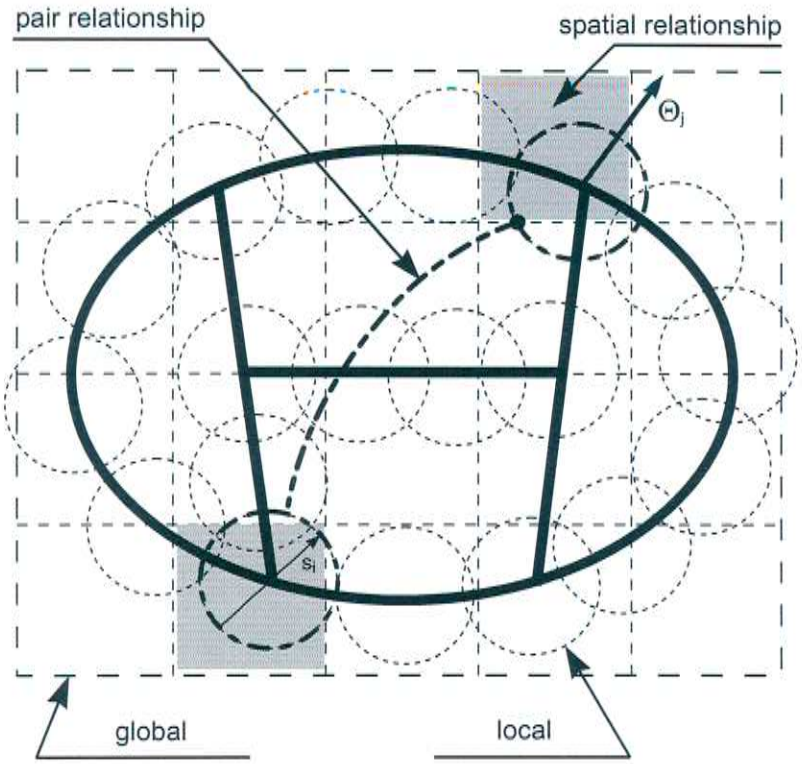


Figure 4.1.: A graphical overview of the bigram local feature. Bridging global and local features: a pair relationship of two individual local descriptors. Aside the image descriptor, each individual local feature owns coordinates (x_i, y_i), scale s_i and orientation θ_i.

At this point the choice of classification type determines further procedure, as we have the first four continuous and the next two dimensions discrete values.

4.2.1. Visual Words

In our setting we chose to utilize the visual words' matching as discussed above. Therefore, discretization of the first feature vector elements is needed to make them comparable by each dimension – a "visual character"[2]. This is, a quantization step as required for this kind of representation [27], even our feature consists of more than one dimension[3]. Aside the possibility to employ a clustering algorithm one more time and to take some drawbacks of these methods into account,

[2]Indeed, our bigram feature can be treated as "word" in the broad sense, because of its 6 degrees of freedom. Those can be regarded as letter positions in a string. The assignment of this positions would be regarded as a character which can be grouped to an "alphabet".

[3]This leads us to a "string matching" for visual words.

we decide to create a small number of bins, depending on the dimension of $f_{i|j}$ and its "meaning".

Thus, often used number of eight bins with centers in $\left\{\frac{k\pi}{4}, k = 0 \dots 7\right\}$ were predefined for the orientation dimension, similar to [30].

The scale ratio has less predictable values and would fit the logarithmic scale the best. Depending on scenario, five bins with centers at $\frac{1}{4}, \frac{1}{2}, 1, 2, 4$ must be sufficient for accurate results.

The remaining Δx and Δy are separated in G bins, G is a constant depending on the size of input images, the scale of image descriptors and objects' structure. Paired with the maximum distance of the available individual features, we normalize Δx and Δy relative to G, making them scale invariant.

4.2.2. Pairs creation

In respect to the application scenario, an algorithmic decision must to be made, in which way the local descriptors are merged to a pair. Even with a small number of local descriptors[4] the number of pairs could increase significantly. In our system we implement the pair creation as follows:

Let L be the distance matrix for all $i, j \in \{1, \dots, N\}$, where N is the number of interest points:

$$L_{ij} = \left\| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} x_j \\ y_j \end{pmatrix} \right\|_2 \tag{4.2}$$

Maximal distance for two local descriptors in a pair is defined as

$$r_{max} = \frac{1}{2} \cdot \max_{i,j=1\dots N} L_{ij} \tag{4.3}$$

We limit the maximum distance by the factor of $\frac{1}{2}$. First of all, we obtain a smaller list of possible pairs. Second advantage is the reduction of "outlier"-pairs across the global appearance of an object, which are possibly sensitive to registration errors. Instead, the key points in the margin areas of the images can be still "connected" by multiple intermediate stops or "bridges" (that is, by a sequence of pairs).

[4] Within the scope of this work we chose to create more key points in each input sample, as it would be necessary in a real world application. Thus, we obtain a high number of pairs and determine their relevance later (see chapter 5). This procedure allows to control the error before feature creation and to create comparable experimental settings.

In respect, minimal distance is defined as:

$$r_{min} = \frac{1}{G} \cdot 2r_{max} \tag{4.4}$$

The number of bins, G, virtually creates a $G \times G$ grid over the image (see figure 4.1). The coordinates of all key points are correlated to the cells of this grid by the quantization step. Thus, we set the minimal distance to 1 grid unit.

Then, let F be the pairs set, such that

$$F = \left\{ f_{i|j} : \forall i : \forall j \neq i : \begin{array}{ll} k_i < k_j, & L_{ij} < r_{max} \\ k_i = k_j, & r_{min} < L_{ij} < r_{max} \end{array} \right\} \tag{4.5}$$

where $i, j \in \{1, \ldots, N\}$. In other words, we select such pairs with the maximum distance of r_{max} between the key points if the key points correspond to descriptors of different clusters. For pairs with descriptors from the same cluster we apply an additional minimal distance requirement to limit the number of "uninteresting"[5] pairs, as direct neighbor key points are most likely to have similar descriptors and thus, to belong the same descriptor cluster.

4.2.3. Set-of-words

A subset selection on the above bigram local features is required to determine the distinctive ones before we can use them for recognition. For that we create an occurrence matrix of our features regarding them as visual words (i.e. by string matching or a hashing function) and apply the method presented in chapter 5. As result a subset for each class with distinctive visual words is returned.

Similar to the popular bag-of-words representation [27, 35, 40, 41] in context of text categorization, we utilize "set-of-words" representation for the classification. While bag-of-words has been proved to work with context-free features, it is not of interest in our case, where visual words stand for spatial *context* and furthermore, are of distinctive nature through the feature selection process. These two steps replace the *statistical* measure of bag-of-words on individual feature representation. Even worse, the statistical aspect would introduce more error on the confidence of the distinctive visual words by making them dependent on the number of similar pairs, and thus, dependent on the key points – which number is in respect sensitive to the method of key point localization.

[5] In terms of information theory, i.e. feature vector $(1, 0, 0, 0, k_i, k_i)^T$ represents only k_i - which is an individual feature.

5. Keyword Selection

In this chapter we present an approach for feature selection. We combine the definitions on feature *relevance* [26] and *document frequency* measure from text categorization field to create a simple, yet promising linear scoring function for discriminative features.

In general, there exist two main ideas which describe features' role in classification [11]. The first one gives the importance to the most common attributes, where the classification depends on partitionability of feature values, for instance, simple classification problem square vs. rectangle with "height" and "width" as attributes. The second idea concentrates on the rare or discriminative features. In this case, one or more dimensions belong to a specific class and build a class specific sub-space. For the above problem, such attribute could be (a fuzzy definition of) "equal sides". In our setting of car manufacturer's logo recognition, a distinctive feature set seems more desirable as car logo classes resemble each other in many aspects of appearance and geometric constraints.

Further, two major types of feature selection are presented in [26]. Their important difference lies in the utilizing the *wrapper mechanism* for decision evaluation using the classification results with the last selected feature subset. The wrapper model delivers better results as there is no need in defining principal factors, responsible for the relevance of features. With the help of an induction function the previous feature subset is modified and tested on some data set. Test results influence the induction algorithm to find a better feature subset. The big disadvantage of a wrapper method in contrast to its counterpart *filter method* is the iterative approach and therefore, higher computational complexity. In a high dimensional problem like ours, we decide to employ a filter method.

In general, our problem has a major characteristic, or even challenge, of text categorization: the high dimensionality of the feature space, which needs an aggressive dimensionality reduction. Similar to text categorization, it is our task to eliminate the non-informative terms and to select those terms which describe a specific category in the best way. On the other hand, our "terms" may be uncorrelated in their nature unlike the human language and its words. This behavior is interesting enough to be kept traced in the future.

5.1. Scoring Function

As shown in previous chapter, in our setting the transition from feature vectors to visual words is likely to reveal a big number of keywords, due to the wide range of possible assignments – unique words – in a bigram local feature vector. Therefore a hard decision function is needed to reduce the set to only few of them while assigning well distinguishable scores. The score should decrease fast enough from 1 to 0 in order to eliminate the mass of irrelevant keywords, and thus, to achieve an aggressive dimensionality reduction.

As shown in [29], document frequency proved to be an adequate measurement of term goodness and delivered almost the same results in comparison to more complex methods of feature selection. On the other hand, since the document frequency could not perform an aggressive dimensionality reduction, we don't intend to use it for feature selection itself in a direct way, but rather utilize it for the initialization of our method to rank visual words.

The main idea is to construct a set of rules with compliance to definitions of feature relevance found in [26]. However, in terms of the more complex wrapper model, those definitions are based on prediction accuracy. In order to utilize the simpler filter model, we use the document frequency as an estimation function for recognition accuracy. Indeed, this assumption is not groundless if we examine the document frequency in a class-wise manner. Obviously, a feature present in the majority of instances of class i would contribute to its recognition accuracy if it doesn't occur in classes $j, j \neq i$.

Following rules were picked to define the major behavior of the scoring function:

1. *ignore* words occurring only in few documents of this class – **irrelevance**
2. *prefer* words with high ranks only in this class – **strong relevance**
3. *degrade* words with low ranks in other classes – **weak relevance**

Let D_c be a set of documents belonging to one class c:

$$D_c = \{d_i : d_i \in c\} \tag{5.1}$$

and

$$df_{cw} = \frac{|D_c \supset x_w|}{|D_c|} \tag{5.2}$$

be a class specific document frequency for word x_w, normalized by the number of documents of the corresponding class c. Previous approaches didn't consider the class-wise aspect and sum across all classes.

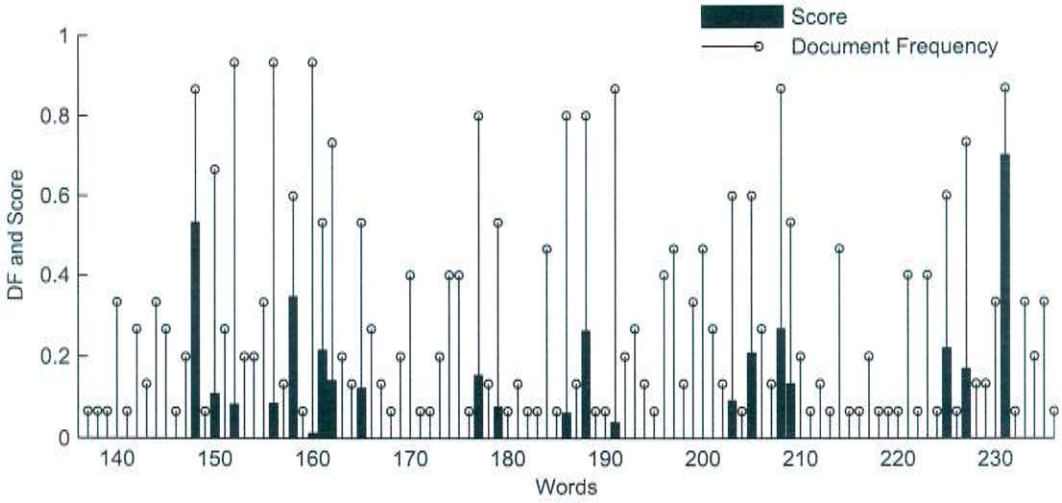


Figure 5.1.: A comparison of given document frequencies of a small number of words of one specific class and their importance (scores) considering a cross-class correlation.

We further define two thresholds t_r and t_n which values are to be determined empirically. The first threshold t_r describes the document frequency at which a word is good enough to be considered as *relevant*. t_n stands for the tolerance of some possible noise for a specific word across other classes. The document frequency below t_n will be ignored and considered to be zero. Therefore the document frequency between t_n and t_r will mean *weak relevance*[1].

Using them we create a scoring function

$$s_{cw} = |\{df_{cw} : df_{cw} > t_r\}| \cdot \frac{df_{cw}}{\sum_i |\{df_{iw} : df_{iw} > t_r\}|} \cdot \frac{\sum_i |\{df_{iw} : df_{iw} \leq t_n\}|}{C} \quad (5.3)$$

where $i \in \{1, \ldots, C\}$ and where C is the number of all classes. All the terms in this equation represent the above rules in the same order as we defined above. The score matrix $S \in [0, 1]^{C \times W}$ consists of the elements s_{cw}, where W is the number of all words.

The returned result and the functionality should be understood in the following way. The score is 1.0 if the word occurs in all of the documents in only one class with no occurrences in any other class (except for noise). Any other *significant* presence ($> t_r$) of this word in another class cuts its score in half (for further

[1] in this case, less degradation in score

classes $\frac{1}{3}, \frac{1}{4}, ...$) and *less relevant* occurrence reduces the score only marginal for the first foreign class, but with increasing penalty for further classes.

As a rule of thumb, we can conclude that every score above 0.5 belongs to a class-specific distinctive visual word.

An example of the effect on document frequencies after applying the scoring function can be found in figure 5.1. Here we can see that only some of the words with high document frequency in a specific class are able to keep their scores after the cross-class comparison. In other words, most of them have high document frequencies in other classes making them useless as a *distinctive* feature. Similar to text categorization, we have such "stop words" - but in our abstraction they just mean general appearance or structural constraint. That is, all classes of round logos would have visual words describing the geometry of a circle.

5.2. Subset Selection

After having computed the score matrix S, we have to choose a "good" subset of features for each class. However, the number of distinctive visual words is going to be various depending on the logo class and its similarity to remaining classes. To reduce the computational overhead, we estimate the number of top highest scores to be considered:

$$N = \frac{1}{C} \cdot \sum_{i=1}^{C} |\{s_{iw} : s_{iw} > 0, \forall w = 1, \ldots, W\}| \tag{5.4}$$

Further, an average score for each class of the top N words is defined by

$$a_c = \frac{1}{N} \cdot \sum_{i=1}^{N} s_{ci} \tag{5.5}$$

where such i is

$$i \leftarrow s_{ck_i} > s_{ck_j}, \forall j > i, k_i = 1, \ldots, W. \tag{5.6}$$

Let v_c be a partial set of visual word labels (indexes) for class c

$$v_c = \{w_1, \ldots, w_n\}, s_{cw_i} > a_c \tag{5.7}$$

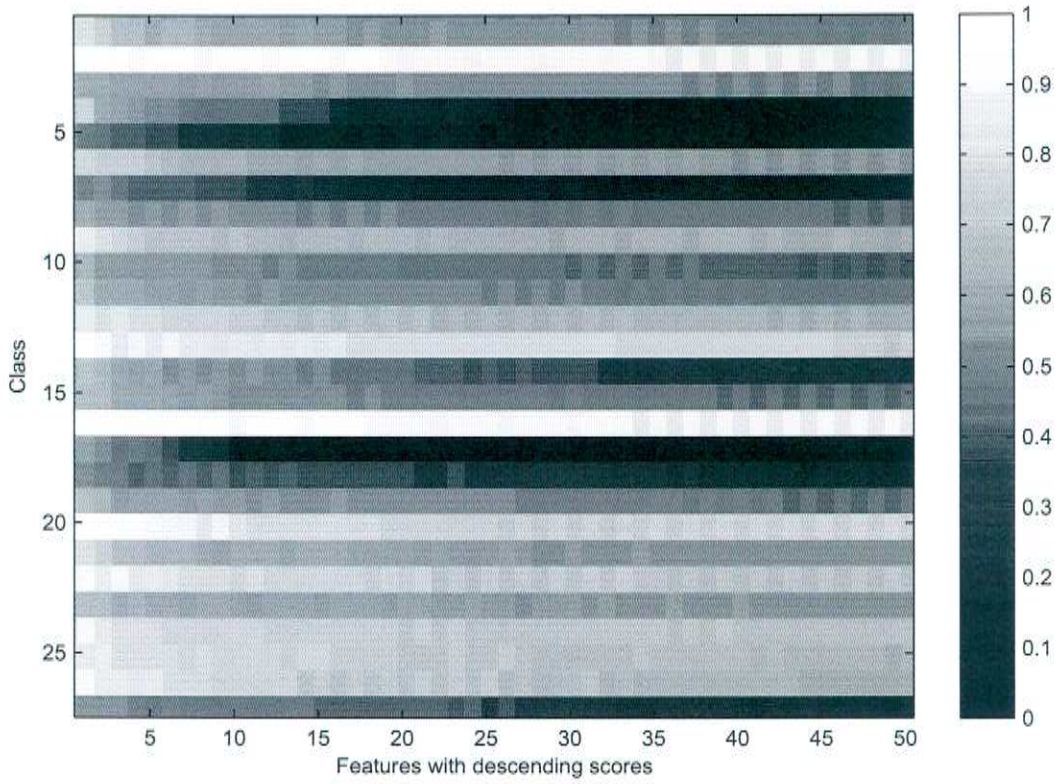


Figure 5.2.: Scores among all classes. This heatmap reveals the distribution of the scores for all classes and for the best 50 keywords.

then the resulting visual words subset V is

$$V = \bigcup_{i=1}^{C} v_i \tag{5.8}$$

5.3. Conclusion

In figure 5.2 a score matrix for the first 50 best words is presented. Truly, not all of the classes could gain the highest score of 1.0. That is, there are no perfectly distinctive visual words or bigram local features available. There are some evident reasons for this behavior.

At first, if we consider, for instance, the class with highest number of distinctive keywords number 2 and 16, which are Audi and Mercedes-Benz respectively, the reason is self-explanatory: their form and shape differ in significant way from

those of other car logos. Meantime, some poor candidates like class number 4, 5, 7, 17 and 27 – Buick, Chevrolet, Dodge, Mitsubishi and Pontiac – confirm the above assumption on geometric appearance on the negative side. Especially logos of the last three ones represent challenging shapes of square[2], three triangles and a single vertically stretched triangle. For such cases, a single bigram local feature is not distinctive enough. We expect improvement with extension to more complex feature structures. Further, the high variance in logo samples applies to all of the negative list, but significantly to Buick and Chevrolet (compare figure 6.1). This results in trouble finding common keywords for major portion of the class data set, yet distinctive in contrast to similar looking logo classes. However, it is likely to have diminished influence of the effect – as for pair representation – on classification accuracy due to multi-dimensional (multiple visual words) representation. Higher complexity grade than a pair of the proposed bigram local feature is required for better results when distinguishing similar, elementary shapes.

[2]details of Dodge's bullhead are almost always lost, so there is only the boundary shape available which is almost rectangular.

6. Experiments and Evaluation

After we have presented our approaches, a set of experiments need to be done to do a comparison to other baseline systems in terms of recognition performance. It is of high interest to explore the performance improvement between global and local representation. The latter one is separated into individual feature representation in comparison to our bigram local feature. Furthermore, some variables ought to be empirically estimated for an appropriate system initialization. While presenting the results we are going to discuss them and to make an evaluation in their deviation. But let's describe the system setup first.

6.1. System Setup

6.1.1. Data Sets

For the upcoming experiments pictures of cars were taken on an outdoor parking lot. In different sessions we were able to to capture various weather conditions, as sunshine, overcast, rain, and snow. With the high number of cars and the frequent alteration in their art (manufacturer, make and color, with additional changing weather conditions), almost all of the captured samples reveal a unique shot of a vehicle.

As this work emphasizes only the feature extraction and classification part of a recognition process, we assume an a priori known position of a logo on each image, and skip the detection part in our standalone system – for which the detection is not less challenging as the classification problem itself. But in any proposed scenario for car manufacturers' logo recognition (see section 1.2), the logo position will be either known or easy predictable from the higher knowledge of the main system. For instance, knowing the position of the corresponding vehicle or possessing knowledge of the image segmentation can deliver high probability regions for the logo position. A local detection in this areas of interest will be a much smaller effort in comparison to detection on the complete image. Thus, we select the logo position in the images manually and crop them with randomized size and horizontal/vertical shift in position to simulate an inaccurate detection.

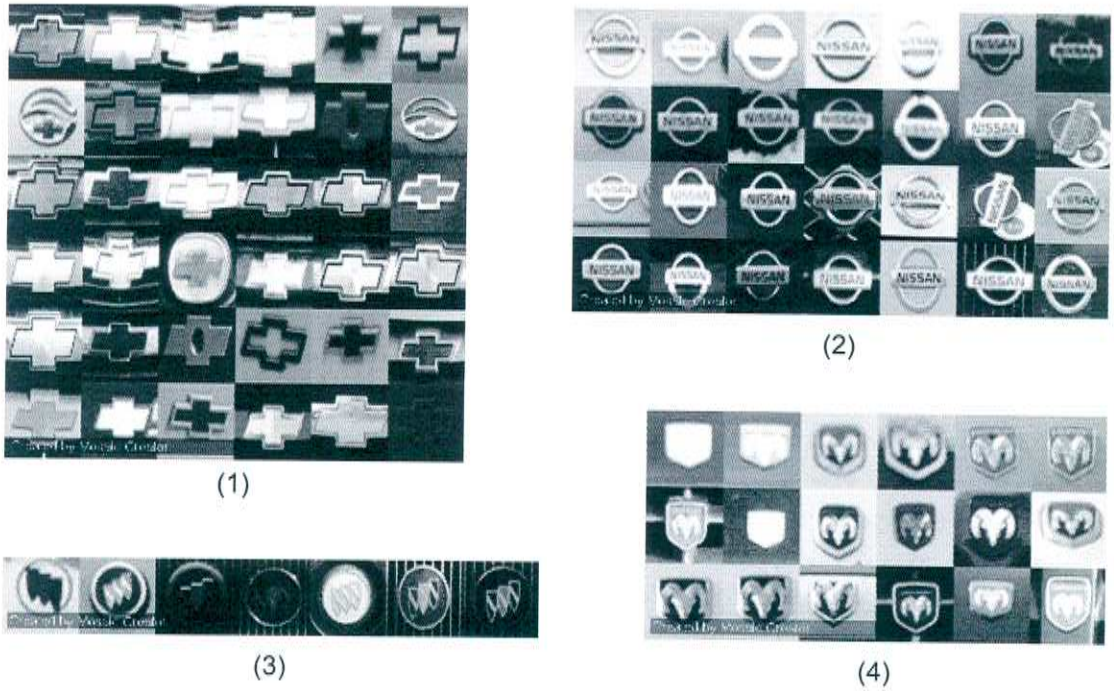


Figure 6.1.: Sample images from the database. Actual images used for recognition with obvious challenging appearance: (1) Chevrolet, (2) Nissan, (3) Buick, (4) Dodge.

All images were manually labeled into 27 classes of car manufacturers. Each class was divided in a training and a testing set with a proportion of 50%-50%. For most of the experiments a twofold cross-validation is used to evaluate the results to eliminate the random factor favoring better scores depending on sample selection.

6.1.2. Feature Sets

To compare the proposed feature, we created several feature sets. All of them base on the *same* data sets, that is, the features were built on exactly the same logo samples in order to be able to measure and compare feature performance in a credible way. Following three sets were defined:

- *Global features:* Often found in appearance-based approaches. We use the whole logo crop normalized to the size of 16×16 pixels as input vector for the utilized classifiers. However, the dimensionality of the input vectors is rather high with 256. For comparison, we additionally apply principal component

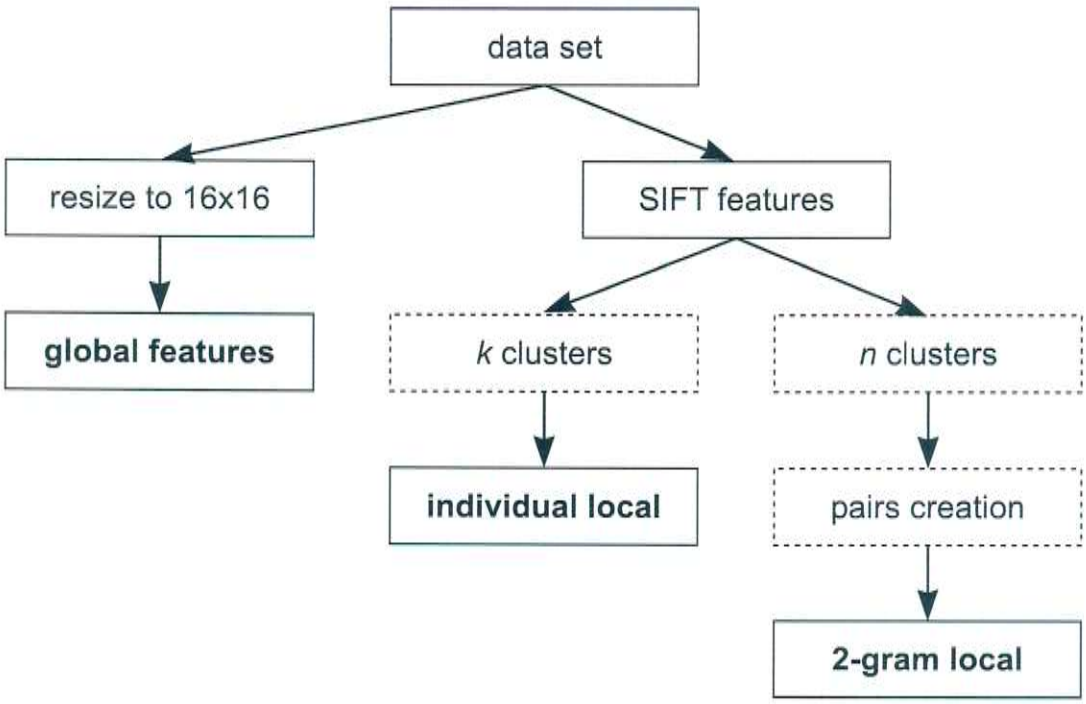


Figure 6.2.: Feature set creation and inheritance. This diagram shows the schematic way from data sample to feature vector. We use identical data samples for feature creation. Thus, a comparison on the *feature* level is possible.

analysis (PCA [12]) for dimension reduction. Further, we extract Gabor feature for an alternate global representation.

- *Individual local features:* Those are local features as used in broad sense of this term in previous work. With clustering the local image descriptors we obtain "visual words" and utilize the bag-of-words representation for classification. For a plausible comparison to the proposed pair feature, the local descriptors were extracted with the same SIFT approach ([30] and introduction in appendix A.1) on the same set of interest points.
- *Pairs features:* This feature set is based on in the chapter 4 presented proposal. We used the same database of local features but clustered the image descriptors differently to create our bigram local feature. Visual words approach was applied on them as well. By reason of high dimensionality we utilized the own feature selection method (see chapter 5) and chose the set-of-words representation (see section 4.2.3 and figure 3.1).

An overview and relations between the feature sets are shown in the figure 6.2.

6.1.3. Preprocessing

In section 1.1 presented challenges require a careful choice of general, yet distinctive feature base for successful recognition process. According to our scenario such classic information like color, size, or shape can be easily disregarded. With shadow impact and alternating fore- and background luminance intensity even gray-scale representation could result in unsatisfactory recognition rates. Thus, we decide to choose the most robust representation under the given conditions: *edges*.

SIFT features are known to be sensitive to non-linear lighting changes [30], i.e. light source and direction alteration for 3D objects. Therefore, we are going to apply multiple preprocessing methods for the input data and to evaluate the impact on the results. We chose the following preprocessing arts:

- *Gray*: no preprocessing (except for obligatory normalization). Gray-scale images are usually used for SIFT extraction.
- *Edges*: edges representation, achieved with Canny filter.
- *SQI*: self quotient image approach [23] applied for lighting invariant representation as known in face recognition.

With these three methods we want to do a basic exploration of preprocessing impact on all feature sets described above. We expect to see different results on the challenging conditions of the collected data.

6.1.4. Training Classifiers

Recognition performance is known to be dependent on the choice of the classification method. Thus, two classifiers will be used for the experiments to reveal the importance of representation and feature extraction.

- *kNN*: is broadly used for classification problems and therefore well-known for its simplicity, fast performance and stable recognition rates. A summary on this classifier is presented in appendix B.2.
- *SVM*: support vector machine algorithm proved to deliver state-of-the-art high recognition performance even on complex distributions in high-dimensional feature space. For optimal performance we referred to [39] and utilized radial basis function (RBF) $K(x, y) = e^{-\gamma\|x-y\|^2}, \gamma > 0$ as kernel. Then, we searched for best parameters for RBF kernel, C and γ, by cross-validating the test data set. This ensured the optimal recognition performance of support vector machine, which is known to be extremely parameter

dependent [39]. An introduction to SVM classification is given in appendix B.1.

All classification experiments are conducted as two-fold cross-validation by exchanging the test and training data sets. This way we are going to minimize the dependence of recognition performance on the randomness factor of data separation. Further, we concern 50%-50% data sets to be more comparable to a real world application, as opposed to, sometimes too often used, ten (or higher) fold cross-validation. Obviously, with our data separation it is easier to judge on the features' role in the recognition task by stressing the classifiers under more challenging conditions, in terms of amount of test and training data.

6.2. Performance Comparison

For performance comparison we are going to use the overall recognition rate:

$$R = \frac{\sum_{i=1}^{n} |D_i| \, M_{ii}}{\sum_{i=1}^{n} |D_i|} \tag{6.1}$$

where M is a confusion matrix

$$M_{ij} = \frac{|\{d_k \in D_j : h(d_k) = i\}|}{|D_j|} \tag{6.2}$$

where D_j is the set of documents/images from class j, and $h(d_k)$ is the predicted class label after the classification.

Further, we exchange the test and training data sets and rerun the experiments to obtain a two-fold cross validation in order to get data set selection unbiased results.

The experimental results are presented in figures 6.3 for visual comprehension and in table 6.1 for numerical reference.

6.2.1. Global Features

Evidently, global representation methods fall back behind other two approaches – regardless the classification method. While "gray" and "edge" preprocessed inputs show low recognition accuracy, a 256-dimensional "SQI" [23] preprocessed feature achieves equal recognition rates with 32-dimensional Gabor [42] representation. To

*k*NN	Edges	Gray	SQI	Gabor	Ø
Global	25.6	40.7	**52.1**	51.7	39.5
Indiv. local	**67.6**	64.1	55.1	-	62.3
Bigram local	**82.8**	78.8	76.2	-	79.3

SVM	Edges	Gray	SQI	Gabor	Ø
Global	31.4	47.9	**51.8**	**51.8**	43.7
Indiv. Local	**85.7**	79.3	70.5	-	78.5
Bigram local	**87.4**	83.8	81.8	-	84.3

Table 6.1.: Summary on recognition rates (2-fold CV in %) for all preprocessing methods. Gabor as comparison to global features, not in the average.

investigate this, we further applied PCA [12] to all three 256-dimensional global features. The results are presented in table 6.2 and show the same behavior analog to Gabor filter. Thus, it was impossible to improve this score in a significant manner. Then, we enlarged the crop size to 32×32 pixels to examine if it would improve the recognition performance. As shown in the same table 6.2, it was not the case. This is true for both classifiers. Obviously, neither further improvement is able to replace the lack of representation information, as global features fail to capture it in a class-wise close manner.

The impact of preprocessing method is high, especially for the *k*NN classifier. Unprocessed, trivial gray-scale image has been confirmed to be unsatisfactory as a feature. Though it was evident, that an edge map wouldn't make much sense for the given configuration, we used it in the global representation for the sake of completeness.

*k*NN	SQI	Gray	Edge	Ø
Global	52.1	40.7	25.6	39.4
+ PCA 32dim	46.6	41.4	26.3	38.1
Improvement	-5.5	0.8	0.7	-1.3

Gray 32×32	Improvement
40.9	0.2
40.9	-0.5

SVM	SQI	Gray	Edge	Ø
Global	51.8	47.9	31.4	43.7
+ PCA 32dim	48.9	54.8	33.0	45.6
Improvement	-2.9	6.9	1.6	1.9

Gray 32×32	Improvement
45.1	-2.8
54.4	-0.4

Table 6.2.: Dimension reduction with PCA and 32×32pix input (2-fold CV in %).

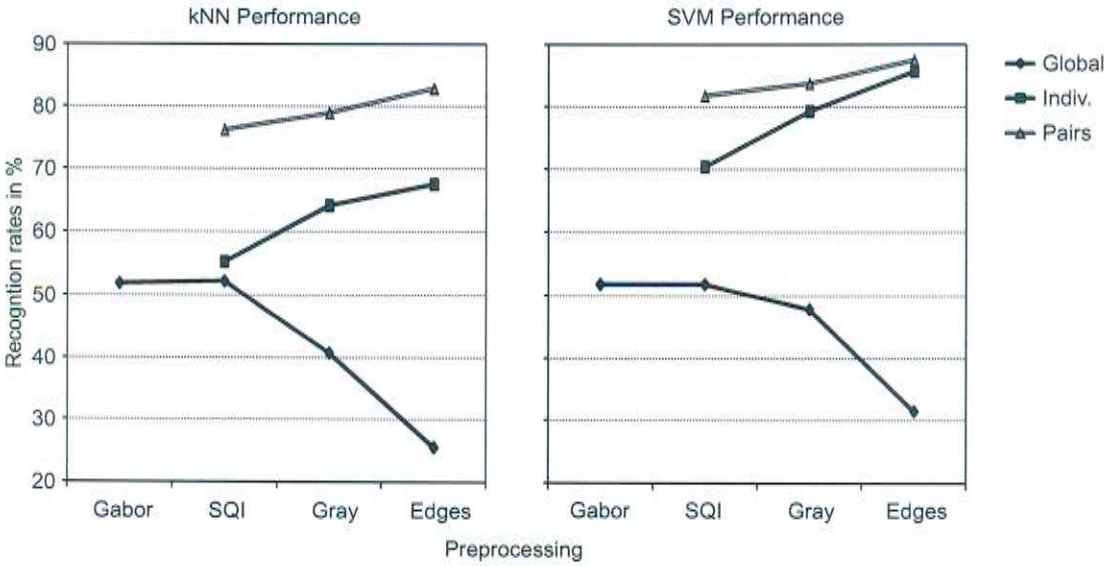


Figure 6.3.: Experimental results. Comparison between applied preprocessing methods for three feature sets and two classification methods kNN and SVM.

6.2.2. Individual Local Features

While a comparison of our feature to global representation would be less fair, this is the real baseline system, in terms of similar configuration and closer relation to our approach.

First of all, our assumption on edge preprocessing has been proved to outperform the common gray-scale appearance for SIFT region patches for this scenario.

However, SQI falls back behind on the local features, even by featuring the edge-like appearance. We assume, that this behavior can be traced back to the "incompatibility" with SIFT local image description so far, that the important distinctive edges of SQI are weaker or of low intensity in comparison to general high intensity edges, which the SIFT image description is sensitive to.

If we compare the recognition rates of both classifiers, we see an overall improvement of 15% for the SVM classification method. In our opinion, this an evidence of challenging and probably non-linear distribution of feature vector projections in the feature space. Thus, the bag-of-words representation of individual local features is not optimal and has much room for improvement.

The advantage of local representation is obvious for our setting. The improvement in prediction accuracy is very significant with $\approx 15\%$ for kNN and $\approx 34\%$ for SVM

kNN	ΔR best/worst	rel. to worst	rel. to best	std. dev. σ
Global	26.5	+103.5	-50.7	13.3
Individual	12.5	+22.7	-18.6	6.4
Bigram local	6.6	+8.7	-8.0	3.3

SVM	ΔR best/worst	rel. to worst	rel. to best	std. dev. σ
Global	20.5	+65.6	-39.6	10.8
Individual	15.2	+21.6	-17.7	7.6
Bigram local	5.6	+6.8	-6.4	2.8

Table 6.3.: Stability of features over different preprocessing methods (in %). It is evident, that global features need a most careful choice of image preprocessing. The proposed bigram local feature is the most insensitive in this comparison.

classification, for best results in the group accordingly.

Still, the results exhibit a high preprocessing dependent behavior, where kNN has a large gap of $\approx 12.5\%$ between the best and worst preprocessing method, as well as SVM with $\approx 15\%$ difference. This makes it clear, that this individual feature representation is sensitive to image preprocessing methods as shown in the table 6.3. Despite the fact, that the proposed bigram feature is the most insensitive to preprocessing methods, its robustness lies in the design as both our feature and individual local feature use the SIFT approach for interest points description.

6.2.3. Pairs Features

The proposed bigram local feature showed best performance during all experiments. This behavior can be observed regardless the choice of preprocessing method. Further, the improvement was confirmed with both classification methods.

Before we compare the pair feature to other approaches, there are some points of interest in the experimental results. Hence, the striking steadiness of the results (in contrast to other representation approaches) across the different preprocessing methods is summarized in the table 6.3. The differences between the maximal and minimal recognition rate average out to around 6% for each classification method. That is, the design of the proposed bigram local feature is less *appearance dependent* in general. The information gain through incorporated spatial constraints

makes it possible. On the other hand, the proposed edge preprocessing for image descriptors enhances the prediction significantly relative to other experiments.

The comparison of the classifiers performances leads to a conclusion of almost optimal representation for the given problem of logo recognition. The kNN method closes on the SVM classification, which has in average 5% better accuracy in each preprocessing method. We see, that there is still some room for improvement for our features and thus, to extend the work to a more complex designs than pairs.

A major advantage is, on the other side, that the proposed bigram local feature representation enabled us to achieve a comparable prediction accuracy with a much simpler and, in terms computational complexity, cheaper algorithm as kNN, which is significant.

If we compare this representation method to the baseline individual features, an overall improvement is evident. On the side of SVM classification it is less with 4.9% in average and 1.7% for the best edge preprocessing. With kNN the proposed representation achieves outstanding 17% average gain in prediction accuracy with 15.2% again for the best edge preprocessing.

The question of *appearance independent* design of the proposed bigram local feature can be further confirmed through the comparison to the individual local representation. Thus, both of approaches use the SIFT image descriptors for the interest points, which is known to deliver most invariant representation [31], but obviously they gain much more insensitivity to appearance when paired with our geometric constraint.

6.2.4. Evaluation Summary

The set of experiments on car manufacturer's logo recognition has confirmed the advantage of local representation in contrast to global one under challenging conditions. Among the two local approaches, the proposed pair feature has been proved to capture the variety in the data in a better way, with less sensitivity to appearance. Further, the feature was more generic such that both utilized classification methods achieved comparable performance.

During the experiments an interesting behavior of prediction accuracy in relation to image preprocessing was revealed. They correlated in such manner, that any preprocessing method, which enhanced the recognition rate for global representation, showed worse results in local approaches and vice versa. Even this behavior seems to be not have been of interest before and was discovered by accident in this setting, it is of believable nature, as both of the global and local representation have oppositional purposes and design attributes.

7. Conclusions and Future Work

The goal of this work was to design a new representation approach which depicts a complex pattern or sequence based on local descriptors. We examined the simplest of such representation: a pair of local descriptors. In terms of textual categorization we called such pair "visual word" [27]. In the training phase each class was assigned a set of visual words which describe the target class in the best way – "keywords" in terms of human language. During recognition we matched the detected visual words to our trained vocabulary in order to predict the labels in the test data set.

In the scope of this work we applied our approach to the car manufacturer's logo recognition in a scenario with extreme challenging conditions. Among the two local approaches, the proposed pair feature has been proved to capture the variety in the data in a better way, with less sensitivity to appearance. Further, the feature was more generalizable such that both utilized classification methods achieved comparable performance. A significant prediction accuracy improvement was observed for a simple, yet very popular kNN classifier in relation to both the global and baseline local approaches.

We still see much future work. The visual words representation for proposed bigram local feature proved to be good enough in the context of the pair relationship. With extension to more complex relations and increasing number feature members, new representation and classification ways are to be examined.

In a final state of this research we see a standalone – as usual for global representation – n-gram local feature for distinctive object description and recognition. In the context of human language we can imagine a complete meaning for "visual word" as correlation to an object or a class of objects.

On the other hand, it is of interest to employ further optimizations on the feature sub-level like different local description approaches. For instance, in [28] a new PCA descriptor for the SIFT approach [30] is presented with even more robustness to common image deformations.

While we utilized a simple key point detection for local descriptors with aggressive feature selection in this work, an application of proposed methods like *Harris* [32], *Harris-Laplace* [33], *Difference of Gaussian* [30], and *Harris-Affine* [34] for point

detection would be of interest to evaluate how stable the detection process will work in order not to disturb the geometric constraint of the proposed feature.

Finally, we presented an approach for feature selection. We combined the definitions of feature *relevance* [26] and the *document frequency* measure from text categorization field to create a simple yet promising linear scoring function for discriminative feature selection. Obviously, the method is not limited to our application scenario and should be applied to classic feature selection benchmarks.

While we utilized our representation approach for a recognition problem, an application in detection approaches is to be explored in the future. With structural relations within the proposed feature detection of objects becomes more convenient. For instance, face detection seems to be an attractive field of application.

Similar to detection task, we can imagine that our method by design can be utilized for solving categorizations problems, though some modifications in parameters are to be expected.

A. Representation Methods

A.1. Scale Invariant Feature Transform

As described in [30], scale-invariant feature transform (SIFT) mainly consist of four stages:

1. Scale-space peak selection
2. Key point localization
3. Orientation assignment
4. Key point description

Specific to our setting and a different application manner of SIFT, we utilized a different approach in key point localization. Thus, we concentrate on introduction to only the last two stages.

A.1.1. Orientation assignment

The third phase determines the major orientation of the image region around the proposed key point. With this, any method of image representation of underlying image patch obtains rotation invariance if the image patch is processed relatively to its orientation.

The scale of the key point identifies the scale of the Gaussian smoothed image, L, and makes the computation scale-invariant. For each image sample, $L(x, y)$ the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are computed as following:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad \text{(A.1)}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right) \quad \text{(A.2)}$$

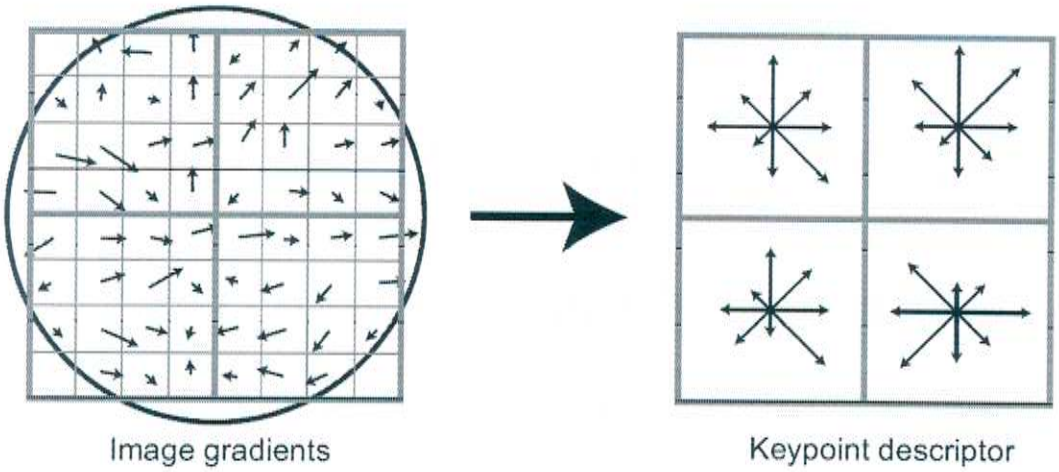


Figure A.1.: Local image descriptors of the SIFT approach as in [30].

In the next step an orientation histogram is built based on θ with 36 bins covering the whole range of 360 degrees. Each sample is weighted by the gradient magnitude and by a Gaussian-weighted circular window (σ is 1.5 times the scale of the key point). The dominant directions are identified by the peaks of the histogram. With multiple peaks multiple key points are created with difference in the corresponding orientation.

A.1.2. Key point description

In the final stage, using the orientation and scale from the previous step, a descriptor is built with the property to remain invariant to affine image deformations, based upon the image gradients in the neighborhood of the key point [28].

First, the gradient image around the key point is smoothed by Gaussian blur depending on the key point scale. Magnitudes and orientations of this image are then computed relatively to the key point orientation (by rotating the descriptor coordinates and gradient orientations). Small arrows in figure A.1 represent the orientation of each location.

Further, the magnitudes are weighted by a Gaussian window around the key point in order to prevent high emphasis on the edge regions which are more sensitive to misregistration errors and transformations.

The descriptor (illustrated on the right side of figure A.1) consists of 4×4[1] ori-

[1] In the figure a 2×2 subregions are presented for better visualization purposes.

entation histograms. This allows significant shift in gradient positions by still contributing to the same orientation histogram.

To avoid all boundary effects like abrupt histogram changes due to sample shift from one to another histogram, trilinear interpolation is used to map each gradient into the adjacent histogram bins. That is, each entry is weighted by the distance to the central value of the bin.

The histogram values over all histograms create the descriptor vector, which has 128 dimensions with 8 orientations within the 4×4 array of histograms. This representation is brightness invariant as it is computed on the image gradient. To prevent contrast dependence, the feature vector is normalized to unit length (as contrast affects the gradient magnitudes).

As to non-linear illumination changes, the author of [30] proposes a threshold of empirically estimated value of 0.2 to reduce the influence of large gradients. Those magnitude but not the orientations are most likely to be affected by not trivial illumination changes.

A.2. Self Quotient Image

In our system we utilized the self quotient image (SQI) approach [23] as a simple case of lighting independent representation. The method is based on smoothing filtering, for example Gaussian filter.

The self quotient image Q of an image I is defined as following:

$$Q = \frac{I}{\hat{I}} = \frac{I}{F * I} \tag{A.3}$$

where $\hat{I}$ is the smoothed version of I, F is the smoothing kernel. The division is done pixel-wise.

The major advantage of this representation and its simplicity is the lack of necessity of any training process of the target classes as it uses the original image itself to obtain this representation.

Indeed, the result Q depends much on the choice of F. Thus, with a small kernel size we approximate to one and with too large values halo effects to be expected. This influence is minimized by utilizing multiple kernel sizes:

$$\frac{1}{N} \sum_{\Omega} WG = 1 \tag{A.4}$$

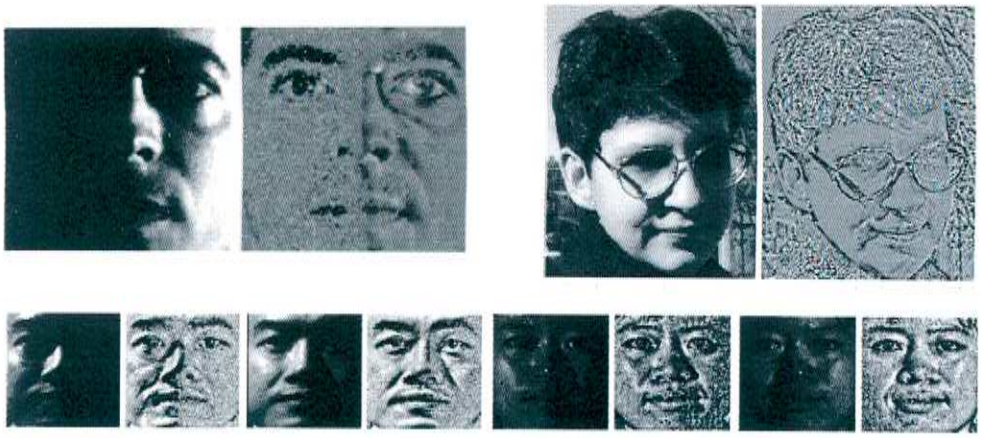


Figure A.2.: Sample images with SQI preprocessing from [23].

where W is the weight, G the Gaussian kernel, N normalization factor and Ω is the convolution kernel size. This creates an anisotropic smoothing filter.

The implementation of the SQI approach is summarized below:

1. Choose some smoothing kernels $G_1, G_2, \ldots, G_n$ and compute corresponding weights $W_1, W_2, \ldots, W_n$, then smooth I by each weighed anisotropic filter:

$$\hat{I}_k = I \oplus \frac{1}{N} W_k G_k, k = 1, 2, \ldots, n \tag{A.5}$$

2. Compute self-quotient image

$$Q_k = \frac{I}{\hat{I}_k} \tag{A.6}$$

3. Transfer it with a non-linear function T to compress the dynamic range for recognition results (i.e. log, arctangent or sigmoid):

$$D_k = T(Q_k) \tag{A.7}$$

4. Sum up the transferred results

$$Q = \sum_{k=1}^{n} m_k D_k \tag{A.8}$$

where m_k are the weights for the filters and can be set to 1.

In figure A.2 some examples of SQI are presented.

B. Classification Methods

B.1. Support Vector Machine

The support vector machine (SVM) classifier finds a linear hyperplane which separates two-class data with maximal *margin* [38]. The margin is defined as the distance of the closest training point to the separating hyperplane. For given observations $x_i \in R^n$, and corresponding labels $y_i \in \{-1, 1\}^m$, one finds a classification function:

$$f(x) = sign(w^T x + b) \tag{B.1}$$

where w, b represents the parameters of the hyperplane.

Data sets are not always linearly separable. The SVM takes two approaches to this problem. Firstly it introduces an error weighting constant C which penalizes misclassification of samples in proportion to their distance from the classification boundary. Secondly a mapping Φ is made from the original data space of X to another feature space. This second feature space may have a high or even infinite dimension. One of the advantages of the SVM is that it can be formulated entirely in terms of scalar products in the second feature space, by introducing the *kernel*

$$K(u, v) = \Phi(u) \cdot \Phi(v) \tag{B.2}$$

Both the kernel K and penalty C are problem dependent and need to be determined by the user.

As there exist some kernel functions, a radial basis function

$$K(u, v) = e^{-\gamma \|u-v\|^2}, \gamma > 0 \tag{B.3}$$

can be considered as most reliable in terms of classification performance [39].

In the kernel formulation, the decision function can be expressed as

$$f(x) = sign\left(\sum_i y_i \alpha_i K(x, x_i) + b\right) \tag{B.4}$$

where x_i are the training features from data space X and y_i is the label of x_i. Here the parameters α_i are typically zero for most i. Equivalently, the sum can be taken only over a select few of the x_i . These feature vectors are known as *support vectors*. It can be shown that the support vectors are those feature vectors lying nearest to the separating hyperplane.

In order to apply the SVM to multi-class problems it is useful to utilize the one-against-all approach. Given an m-class problem, m SVM's are trained, each distinguishes images from some category i from images from all the other $m-1$ categories j, $j \neq i$. The given query image is then assigned to the class with the largest SVM output.

B.2. k-Nearest Neighbor

The kNN classification method belongs to the simplest and yet well performing algorithms for many classification problems. The ranking in kNN is based on the labels assigned to the k nearest training samples of the input. The similarity to the neighbors is measured by the (Euclidean) distance between the two feature vectors.

Thus, for a query y from the feature space K^n and m training features $x_i \in K^n, i = 1 \ldots m$, the k closest neighbors are defined by

$$t = \arg\min_{i=1\ldots m} (\|x_i - y\|_2) \tag{B.5}$$

and $L(t)$ is the corresponding label.

The prediction result R is usually the label of the majority of the neighbors, such as

$$R = L(\arg\max_{t_i, i=1\ldots k} L(t_i)) \tag{B.6}$$

Bibliography

[1] L. P. Cordella and M. Vento. Symbol and Shape Recognition. In *Lecture Notes in Computer Science*, Vol. 1941, pp. 167-182, Jan 2000.

[2] J. Neumann, H. Samet and A. Soffer. Integration of Local and Global Shape Analysis for Logo Classification. In *Lecture Notes in Computer Science*, Vol. 2059, p. 769, Jan 2001.

[3] J. Lladós, Er. Valveny, G. Sánchez and Enric Martí. Symbol Recognition: Current Advances and Perspectives. In *Lecture Notes in Computer Science*, Vol. 2390, pp. 104-127, Jan 2002.

[4] D.S. Doermann, E. Rivlin, and I. Weiss. Logo recognition using geometric invariants. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pp. 894-897, Oct. 1993.

[5] E. Francesconi, P. Frasconi, M. Gori, S. Marinai, J. Q. Sheng, G. Soda, and A. Sperduti. Logo Recognition by Recursive Neural Networks. In *Lecture Notes in Computer Science*, Vol. 1389, pp. 104-117, 1997.

[6] K. Tombre and B. Lamiroy. Graphics recognition - from re-engineering to retrieval. In *Proceedings of the Seventh International Conference on Document Analysis*, pp. 148-155, Aug 2003.

[7] J. Llados, E. Marti, and J.J. Villanueva. Symbol recognition by error-tolerant subgraph matching betweenregion adjacency graphs. In *EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23 (10), pp. 1137-1143, 2001.

[8] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. In *EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27 (11), pp. 1832-1837, 2005.

[9] M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *Lecture Notes in Computer Science*, Vol. 1842, pp. 18-32, Jan. 2000.

[10] Y. Amit, D. Geman, and X. Fan. A coarse-to-fine strategy for multiclass shape detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26(12), pp. 1606-1621, Dec. 2004.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19(7), pp. 711–720, 1997.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. In *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–96, 1991.

[13] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *Eighth International Conference on Computer Vision*, p. 454, 2001.

[14] T. Naito, T. Tsukada, K. Yamada, K. Kozuka, and S. Yamamoto. Robust recognition methods for inclined license plates under various illumination conditions outdoors. In *Proceedings of the 1999 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems.*, pp. 697-702, 1999.

[15] Y. Yanamura, M. Goto, D. Nishiyama, M. Soga, H. Nakatani, and H. Saji. Extraction and tracking of the license plate using Hough transform and voted block matching. In *Proceedings of the 2003 IEEE Intelligent Vehicles Symposium.*, pp. 243-246, 2003.

[16] V. Kamat and S. Ganesan. An efficient implementation of the Hough transform for detectingvehicle license plates using DSP's. In *Proceedings of the 1995 Real-Time Technology and Applications Symposium.*, pp. 58-59, 1995.

[17] M. Gori, M. Maggini, S. Marinai, J.Q. Sheng, and G. Soda. Edge-backpropagation for noisy logo recognition. In *Pattern Recognition*, vol. 36 (1), pp. 103-110. 2003.

[18] J. Miura, T. Kanda, and Y. Shirai. An active vision system for real-time traffic sign recognition. In *Proceedings of the 2000 IEEE Intelligent Transportation Systems*, pp. 52-57, 2000.

[19] A. de la Escalera, J.M. Armingol, and M. Mata. Traffic sign recognition and analysis for intelligent vehicles. In *Image and Vision Computing*, vol. 21(3), pp. 247-258, 2003.

[20] M.A. Garcia, M.A. Sotelo, and E.M. Gorostiza. Traffic sign detection in static images using Matlab. In *IEEE Conference on Emerging Technologies and Factory Automation*, vol. 2, pp. 212-215, 2003.

[21] C. Bahlmann, Y. Zhu, V. Ramesh, M Pellkofer, and T. Koehler. A system for traffic sign detection, tracking, and recognition using color, shape, and motion information. In *Proceedings of the 2005 IEEE Intelligent Vehicles Symposium*, pp. 255-260, 2005.

[22] J.-C. Hsien and S.-Y. Chen. Road Sign Detection and Recognition Using Markov Model. In *14th Workshop on Object-Oriented Technology and Applications*, pp. 529–536, 2003.

[23] H. Wang, S. Z Li, and Y. Wang. Face recognition under varying lighting conditions using self quotient image. In *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[24] S.-H. Hsu and C.-L. Huang. Road sign detection and recognition using matching pursuit method. In *Image and Vision Computing*, vol. 19, pp. 119-129, 2001.

[25] C.V. Jawahar and A.K. Ray. Fuzzy statistics of digital images. In *IEEE Signal Processing Letters*, vol. 3 (8), pp. 225-227, 1996.

[26] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pp. 121-129, 1994.

[27] G. Csurka, C. Dance, C. Bray, and L. Fan. Visual categorization with bags of key points. In *Proceedings Workshop on Statistical Learning in Computer Vision*, 2004.

[28] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 506-513, Volume 2, 2004

[29] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412-420, 1997.

[30] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, 2004.

[31] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of Computer Vision and Pattern Recognition*, June 2003.

[32] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pp. 147-151, 1988.

[33] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, pp. 525–531, 2001.

[34] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, pp. 128–142, 2002.

[35] L. Zhu, A. Rao, and A. Zhang. Theory of keyblock-based image retrieval. In *ACM Transactions on Information Systems*, pp. 224-257, 2002.

[36] A. K. Jain, M. N. Murty and P. J. Flynn. Data clustering: a review. In *ACM Computing Surveys*, Volume 31, (3), pp. 264-323, 1999.

[37] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* 1973, John Wiley and Sons, Inc., New York, NY.

[38] V. N. Vapnik. *The nature of statistical learning theory*, Wiley, New York, 1998.

[39] CW. Hsu, CC. Chang, and CJ. Lin A practical guide to support vector classification.

[40] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, W.T. Freeman. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 370-377, 2005.

[41] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 883-890, 2005.

[42] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. In *IEEE Transactions on Image Processing*, Vol. 11, no. 4, pp. 467-476. Apr. 2002.