FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Novel Visual Media Summarization Methods for Retrieval Applications

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Markus Hörhan, B.Sc. M.Sc.
Matrikelnummer 0400625

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Ao.Univ.Prof. Mag. Dr. Horst Eidenberger

Diese Dissertation haben begutachtet:

Prof. Dr. Harald Kosch	Ao.Prof .Dr. Mathias Lux

Wien, 15. Juni 2020

Markus Hörhan

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Novel Summerization Methods for Visual Retrieval Applications

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Markus Hörhan, B.Sc. M.Sc.
Registration Number 0400625

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Mag. Dr. Horst Eidenberger

The dissertation has been reviewed by:

Prof. Dr. Harald Kosch	Ao.Prof .Dr. Mathias Lux

Vienna, 15th June, 2020

Markus Hörhan

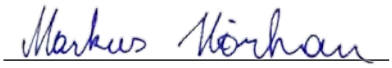
Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Markus Hörhan, B.Sc. M.Sc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. Juni 2020



Markus Hörhan

Danksagung

Ein herzliches Dankeschön an meinen Betreuer Horst Eidenberger, der mich exzellent betreut hat, mir Denkanstöße, Motivation und äußerst hilfreiche Ideen für diese Arbeit gegeben hat.

Weiters möchte ich mich bei meiner Frau Ksenia Skorokhodova für die vielen Ideen zur Arbeit über Make-up robuste Gesichtserkennung und für die Hilfe bei der Erstellung des Daten-Sets bedanken. Außerdem hat sie mich stets motiviert und auch in schwierigen Zeiten dazu ermutigt weiter zu machen. Ein besonderer Dank gilt auch meiner Mutter und meiner Familie. Zusätzlich danke ich auch Robert Sorschag, der mich speziell zu Beginn der Arbeit mit wertvollen Informationen unterstützt hat.

Zuletzt möchte ich mich bei Spreitzer Johannes und Jeff Crowder für das Korrekturlesen bedanken.

Acknowledgements

A heartfelt thank you to my supervisor Horst Eidenberger, who gave me excellent support, fruitful thoughts, motivation and extremely helpful ideas for this work.

I would also like to thank my wife Ksenia Skorokhodova for the many ideas on working on make-up robust face recognition and helping with the creation of the dataset. She has also always motivated and encouraged me to continue working in difficult times. Special thanks also go to my mother and my family. In addition, I also thank Robert Sorschag, who especially supported me with valuable information at the beginning of my work.

Lastly, I would like to thank Spreitzer Johannes and Jeff Crowder for proofreading.

Kurzfassung

Aufgrund der stetig wachsenden Datenmengen, die weltweit gesammelt werden, ist es von großer praktischer Relevanz, über effiziente Methoden zur automatischen Zusammenfassung dieser Informationen zu verfügen. Auf einer sehr hohen Abstraktionsebene versucht die automatische Zusammenfassung von Daten, Teilmengen von Objekten wie Sätze, Bilder oder Videos zu finden, die die gesamte Datenmenge möglichst gut repräsentieren.

In dieser Arbeit werden neuartige Methoden zur Zusammenfassung von visuellen Daten vorgestellt, die auf der visuellen Wahrnehmung des Menschen, insbesondere auf den Gestaltgesetzen beruhen. Diese Gesetze definieren Regeln darüber, wie Menschen die Welt um sich herum wahrnehmen und wie sie visuelle Reize vereinfachen, ohne dass diese Reize an Bedeutung verlieren. Nach unserem Kenntnisstand beschränken sich viele in der Vergangenheit entwickelte Bildverarbeitungsmethoden auf rein technische Aspekte und lassen psychologische Theorien wie die Gestalttheorie außer Acht. Mit dieser Arbeit wollen wir einen Beitrag leisten, um dieser Entwicklung entgegenzuwirken. In mehreren Experimenten haben wir unsere neuartigen visuellen Methoden zur Zusammenfassung von Daten auf verschiedene Computer-Vision-Probleme wie die Kategorisierung von Videoszenen, die gewichtsinvariante Gesichtserkennung, die Bildklassifizierung, die Identifizierung von Autos in Bildern und die Make-up-robuste Gesichtserkennung angewendet. Darüber hinaus haben wir unsere experimentellen Ergebnisse mit modernsten Methoden verglichen. Für diese Vergleiche haben wir einerseits öffentlich zugängliche Testdatensätze verwendet und andererseits selbst Testdatensätze zusammengestellt.

Die wahrscheinlich wichtigste Methode, die während dieser Dissertation entwickelt wurde und die zur Zusammenfassung von Daten verwendet werden kann, ist der Gestalt-Interest-Points-Algorithmus (GIP). Der Algorithmus ist schnell und sehr effektiv, da er nur sehr wenige, aber gut ausgewählte Bildinformationen extrahiert und dadurch sehr kompakte semantische Zusammenfassungen von Bildern erstellt. Er basiert auf den Gestaltgesetzen der Schließung und Kontinuität, d. H. der Idee, dass im Gegensatz zu anderen lokalen Bildbeschreibungsmethoden bestimmte schwächere Kandidaten - zusätzlich zu den lokalen Extrema - auch als Merkmale nützlich sein könnten. Der GIP-Algorithmus war die Grundlage für die Gestalt-Regions-of-Interest-Methode (GROI). Das Trainieren eines CNNs durch GROI-Bilder übertrifft deutlich die Genauigkeit eines CNNs, das aus rohen Pixelbildern für den Bereich der makeup-robusten Gesichtserkennung trainiert wurde. Darüber hinaus ist unsere vorgestellte Methode robuster gegen Überanpassung als der

herkömmliche Ansatz, bei dem ein CNN anhand von Rohpixelbildern trainiert wird. Der größte Vorteil der GROI-Methode ist, dass der semantische Inhalt von Bildern kompakter zusammengefasst werden kann als bei ganzen Bildern. Dies ist ein sehr wichtiges Argument in Big-Data-Anwendungen wie der Gesichtserkennung.

In dieser Arbeit haben wir gezeigt, dass der Computer mithilfe unserer neuartigen Methoden zur Zusammenfassung von visuellen Daten, die auf der menschlichen Wahrnehmung basieren, mehrere praktisch relevante Computer-Vision-Probleme zuverlässig lösen kann. Darüber hinaus haben wir unsere experimentellen Ergebnisse mit modernsten Methoden verglichen und festgestellt, dass unsere Ansätze sehr vielversprechend sind. Für die Zukunft sind weitere Studien erforderlich, um die entwickelten Methoden auf andere Probleme und größere Datenmengen anzuwenden.

Abstract

Due to the steadily increasing amount of data that are collected worldwide, it is of great practical relevance to have efficient methods for automatic data summarization. At a very high abstraction level, automatic data summarization attempts to find subsets of objects such as sentences or visual data like images or videos that cover information about the entire set.

In this work we present novel visual summarization methods, which are based on the visual perception of human beings, in particular on the Gestalt Laws. These laws define theories about how people perceive the world around them and the simplification of the visual stimuli without loss of meaning. To the best of our knowledge, many computer vision methods developed in the past are limited to purely technical aspects and omit psychological theories, such as Gestalt theory. With this work, we want to contribute to counteracting this fact.

In several experiments, we applied our novel visual summarization methods on various computer vision problems like video scene categorization, weight-invariant face recognition, image classification, identifying cars in images and makeup-robust face recognition. Furthermore, we compared our experimental results to state-of-the-art methods. In order to make these comparisons, we used publicly available data sets on the one hand and on the other hand we compiled data sets ourselves.

Probably the most important method that was developed during this dissertation and that can be used for summarizing data is the Gestalt Interest Points (GIP) algorithm. The algorithm is fast and highly effective because it extracts very little but well-selected image information and thereby creates very compact semantic summaries of images. It is based on the Gestalt Laws of Closure and Continuity, i.e. the idea that, unlike in other local image description methods, certain weaker candidates may – in addition to the local extrema – also be useful as interest points. The GIP algorithm was the foundation for the Gestalt Regions of Interest (GROI) method. With the GROI images we improved the accuracy of a CNN for the domain of makeup-robust face recognition. Training a CNN with GROI images clearly outperforms the accuracy of a CNN trained with raw pixel images for the domain of makeup-robust face recognition. Additionally, our presented method is more robust against over-fitting than the conventional approach, training a CNN from raw pixel images. The biggest advantage of the GROI method is that it is possible to summarize the semantic content of images more compactly than from whole

images. This is a very important argument in particular in big data domains such as face recognition.

In this work, we have demonstrated that the computer can robustly solve several practically relevant computer vision problems by using our novel visual summarization methods based on human perception. Furthermore, we compared our experimental results to state-of-the-art methods and found that our approaches are highly competitive. Further studies are needed to apply the developed methods to other problems and larger data sets.

Contents

CHAPTER 1

Introduction

1.1 Automatic Data Summarization

The main idea of automatic data summarization [Ahm19] is to find a subset of data which gives as much information about the entire set as possible. Such techniques are widely used in industry today. Search engines are one example; others include summarization of documents, image collections, and videos. Document summarization [YWX17] tries to create a representative summary or abstract of the entire document, by finding the most informative sentences, while in image summarization [SKA+15] the system finds the most representative and important (i.e. salient) images. For surveillance videos, one might want to separate the important events from the uneventful context [PHK+18]. At a very high level, summarization algorithms try to find subsets of objects (like a set of sentences, or a set of images), which cover information about the entire set.

Image collection summarization [SS20] is another application example of automatic summarization. It is often performed by selecting a representative set from a larger set of images. A summary in this context is useful to show the most representative images of results in an image retrieval system. Video summarization [SPP16] is a related domain, where the system automatically creates a trailer from a long video. This also has applications in consumer or personal videos, where one might want to skip the uneventful scenes. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the redundant frames captured.

In this work we present the utilization of novel visual summarization methods for various application domains, e.g. the separation of action scenes from non-action scenes through summarizing a set of videos into these two categories. Classifying a set of images into food and non-food images or the categorization of horse and non-horse images are other application domains also presented in this work. Another example is the differentiation of real violence from martial arts videos. We also present the recognition of faces of people

who have experienced a significant change in weight. To accomplish this recognition task, we summarized different face images of one person to learn a generalized representation of this person's face. The assumption was that it would be possible to recognize a person whose weight has changed exclusively from the previously learned generalized representation of this person's face. A similar example are our experiments with makeup-robust face recognition, although a different approach was chosen to solve the problem. Learning a generalized representation of cars by summarizing vehicle images is also an application that is presented in this work. This generalized representation is then used to detect cars in images. All the novel visual summarization methods utilized to solve the above-mentioned problems are presented in detail later in this work.

The remainder of this work is organized as follows. The subsequent elements of the introduction provide important background knowledge for the following chapters, present the state-of-the-art in the related fields and finally conclude all the contributions. We first present a brief introduction to computer vision in Section 1.2 because we utilized techniques and methods of computer vision to summarize visual information. To gain a clearer understanding of the fundamental issues with which we were confronted during this work Section 1.3 describes some major challenges of computer vision. Because the deep learning revolution had a significant impact on our work, we give a short introduction to this topic in Section 1.4. The presented methods are based on the Gestalt theory and therefore we will have a closer look at this theory in Section 1.5. The state-of-the-art works related to our work are presented in Section 1.6. This introduction will be finished in Section 1.7 with an overview of all the contributions of the already published works presented in the following Chapters 2 to 8.

1.2 Computer Vision

To summarize visual information, we employed techniques and methods from computer vision. Computer vision [SSKG20] is a research discipline that aims at extracting semantic information from visual data sources to help computers "see" and understand the content of digital images such as photographs and videos. Usually, the first stage is feature extraction [NA19], which results in a summarized description of the visual data source. A good descriptor should have the ability to handle intensity, rotation, scale and affine variations to some extent. Feature extraction is motivated by reducing the size of a visual data source as well as by the elimination of redundancy and possibly noisiness.

Finally, to categorize the visual descriptions into classes a categorization method is needed, e.g. [CDF+04]. Generally, many methods can be employed for the categorization of multimedia descriptions though some methods are more frequently used in one area than another. The Support Vector Machine, Neural Networks, Decision Trees or Nearest Neighbor methods are some examples for categorizing visual descriptions.

There are many potential application domains for computer vision [Sze19]. For this reason, it has been an active research topic over decades. Innovative methods and improvements of existing approaches are constantly required. This dissertation has been

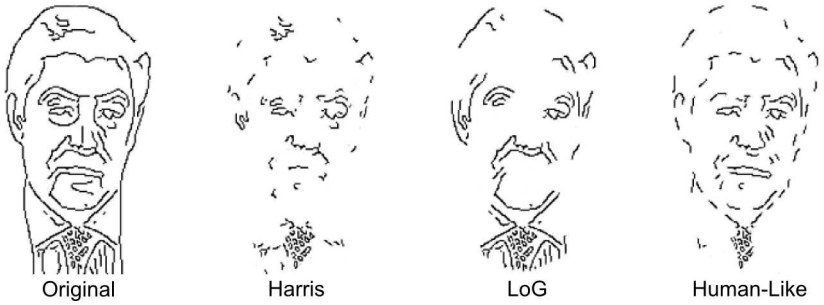


Figure 1.1: The left edge map of a face is represented by points from a Harris corner detector and the second a Laplacian of Gaussian (LoG) operator. Many of the important face features (e.g. the eyes) are thereby lost. The rightmost image shows the GIP description. It preserves the perceptual features of the original stimulus well, while not producing a longer description than the LoG operator [Eid11].

going on for several years. It contributes to this research area and introduces innovative computer vision approaches. Discovering novel visual features and the development of innovative methods to extract them is an important part of this dissertation. It was expected that integrating principles from Gestalt psychology [Kof35] into the proposed methods should improve the results. The visual perception of human beings is based on these principles and they make it highly effective. Therefore, the machine may also benefit from Gestalt principles. The Gestalt law of closure states that the perception of individuals fills in visual gaps in incomplete shapes. For example, humans are able to recognize a whole circle, even if there are gaps in its contour. For our approach this means that due to the Gestalt law of closure it is still possible to recognize what an image depicts, only by considering its local representation. This effect is shown in Figure 1.1. Obviously, such interest point sets are more useful for media understanding than points from which humans cannot identify the semantic content of an image. If the user cannot reconstruct the object from the interest points, how should the machine? The experimental results are highly encouraging: The novel features perform exceptionally well, and the classification process delivers practically acceptable recall and precision values.

As we already mentioned above, the summarization of object images for object identification and localization is also a part of this dissertation and according to [LOW+18] one of the most fundamental and challenging problems in computer vision. There are publicly available datasets, which we used for our experiments. The proposed methods were tested, refined and compared to existing state-of-the-art algorithms through extended experiments. To gain a clearer understanding of the fundamental issues with which we were confronted during this work, the next section describes some major challenges of computer vision.

Figure 1.2: We have no difficulty seeing that this image shows two apples, although there are big differences between them, e.g. color, texture, shape, size, orientation and lighting. In addition, the front apple partially covers the back.

1.3 Challenges of Computer Vision

According to [Sej18a] computer vision seems easy, perhaps because it is so effortless for humans. Initially, it was believed to be a trivially simple problem. However, after decades of research, computer vision still cannot reach the capabilities of human vision. One reason for this is that people have not yet understood how their visual perception works. Sejnowski et al. stated that studying biological vision requires an understanding of the perception organs like the eyes, as well as the interpretation of the perception within the brain. Much progress has been made but there is still a long way to go. Another reason why it is such a challenging problem is because of the complexity inherent in the visual world. A given object may be seen from any orientation, under any lighting conditions, with any type of occlusion from other objects, and so on. According to Sejnowski et al., a true vision system must be able to "see" in any of an infinite number of scenes and still extract something meaningful.

Furthermore, Sejnowski et al. [Sej18a] stated that in the 1960s, it was assumed that

computer vision was an easy problem to solve and researchers completely underestimated it. In the early days of computer vision, one tried to match object templates with objects in images to realize object recognition. Figure 1.2 shows why this approach did not work. For us humans it is easy to see that there are two apples in the image but not for the machine. The apples are very different in their properties and the front apple partially covers the one in the back. According to Sejnowski et al. the solution to this problem was not to compare a template pixel by pixel with an object in the image, but to use an abstract version of the template, i.e. extracting features from the images.

Nixon et al. [NA19] stated that a good feature is one that makes an object clearly distinguishable. For example, there is one apple variety whose apples are green and another whose apples are red. If you have an apple from these two varieties and you want to know which of the two varieties it belongs to, you just have to look at the color feature. If one feature is not enough to make an object clearly distinguishable, it can also be the combination of several features. For example, there are a lot of apple varieties whose apples are of the same color. Due to the color alone, no clear statement could be made. But an apple could then be clearly assigned to an apple variety, for example, by its color, size and texture. A photo of an object contains a lot of details, which in no way contribute to distinctness. But if you made a schematic drawing of the photo, you could for example highlight only the clearly distinguishable features. For a long time, a challenge of computer vision has been to find out or automatically determine these distinguishable features of a specific application domain. This challenge is also a very important core theme of this work.

In 2016, research in the field of computer vision was strongly influenced by the deep learning revolution. Since this revolution also had a significant impact on our work, we will give a brief introduction to this topic in the next section.

1.4 The Deep Learning Revolution

This dissertation has been going on over several years because we wanted to investigate all the aspects presented in this work in depth. Moreover, the research topic dealt with is important and of the highest practical relevance. Around the year 2016, the deep learning revolution [Sej18b] and its leap forward in performance had a big impact on our work. We had to rethink our methods because deep learning set entirely new standards in terms of solving computer vision problems.

According to Sejnowski et al. [Sej18b] the historic timeline of the deep learning milestones until the deep learning revolution began in the year 2012 when Geoffrey Hinton et al. published a paper [KSH12] on object recognition in images that used deep learning to train AlexNet, a deep convolutional network. Using the ImageNet database of more than 15 million labeled high-resolution images in more than 22,000 categories as a benchmark, AlexNet achieved a 18 percent reduction in the error rate. This enormous increase in performance led to the development of ever larger networks which now reach human levels of performance. In 2015, Kaiming He and colleagues developed a deep learning

network [HZRS16] that is strongly inspired by the visual cortex, reducing the error rate in the ImageNet database to 3.6 percent. Sejnowski et al. pointed out that in 2016 the Go-playing program AlphaGo that used deep learning networks to evaluate board positions and possible moves won against the Korean 18-time world Go champion Lee Sedol. Even DeepMind, the company that had developed AlphaGo, was surprised by the great success which was due to using Deep Learning.

Until 2016, the results we obtained in this dissertation were highly competitive compared to the prevailing baseline methods. But that changed very quickly as a result of the deep learning revolution in 2016. Suddenly the baseline was much higher. The solution was not to beat deep learning for our application domains. Because of the outstanding performance that seemed impossible. Instead, we found a way to combine our method with deep learning. More details about this topic and extensive experimental results are presented later in this work.

Since the presented methods in this work are based on Gestalt theory, we will have a closer look at this theory in the next section. This is important background knowledge in order to understand the described approaches more easily.

1.5 Gestalt Theory

The fundamental algorithm presented in this work was strongly inspired by elements of Gestalt theory. Already in 1923, there was one of the most important publications of the psychologist Max Wertheimer [Wer23], which dealt with the topic of Gestalt. It was published long before computers were available. Gestalt theory is about perceptive organization of patterns and parts. The so-called Gestalt Laws define theories about how people perceive the world around them and about the simplification of the visual stimuli without loss of meaning.

According to Michaelsen et al. [MM19], the most interesting visual stimuli for a human observer are other human subjects. Human observers will concentrate most of their attention on the faces. Michaelsen et al. described that this could be quantitatively verified by the use of eye trackers. They also stated that according to the Gestalt law of symmetry, symmetrical arrangements attract the viewer's attention noticeably and faces in frontal view are one example for reflection symmetry. So, it seemed obvious to us to apply the Gestalt-based algorithm we developed on face recognition. The results are presented later in this work.

A key element of Gestalt theory [Wer23] is that people tend to group objects with certain similar properties. A group is then perceived as a new larger object called *Gestalt*. For example, buildings are something man-made and are not found in nature. However, people can recognize buildings instantly, even though they exist in thousands of different ways. According to the theory, the laws of Gestalt grouping greatly assist the analysis of such objects. In the book of Michaelsen et al. [MM19] the authors show the occurrence of various Gestalt Laws in different example images. According to their findings, the

following Gestalt Laws can be observed in the facade of a building such as the one depicted in Figure 1.3.

- *Law of Similarity:* The windows of the building are similar e.g. in color, shape and size. We tend to perceive rows of windows as we group the windows based on the similarity. In some image processing algorithms, the similarity of objects is measured by color, texture, object shape, aspect ratio, properties of detected interest points, and so on.
- *Law of Proximity*: An example of this law are the exhaust pipes at the top of the roof. Some of them stand alone on the roof and others stand side by side with no gap. On the far left are five pipes close together and in contrast, the rightmost exhaust pipe is all alone. The closely spaced pipes are perceived as groups by humans. In some image processing algorithms this law is implemented as the pixel spacing between objects.
- *Law of Continuity*: According to this law, lines are always seen as following the simplest path. The black and white dashed line on the road is usually perceived as a solid line rather than as a collection of single short black and white lines.
- *Law of Closure*: Someone can clearly see cars behind the fence, though they are not completely visible through the fence. This is due to the Gestalt law of closure which describes that our brain fills in the missing gaps in incomplete shapes to produce meaningful information. We use this law in our algorithms to reduce the information contained in pictures to the essentials. This makes it possible to abstract the information in these images and process the images more efficiently. The abstraction of image information is usually a prerequisite, e.g. for object and face recognition. However, this will be described later in this work.
- *Law of Good Figure*: According to this law, people prefer to perceive objects which have a simple and therefore memorable structure. The tree in the foreground on the left side is perceived as a tree and we usually do not focus our attention on every single leaf of this tree. There are obvious similarities with image processing algorithms, which in many cases also simplify objects through abstraction.
- *Law of Common Fate*: This law states that people perceive visual objects that move at the same speed and/or into the same direction as parts of a group. In our opinion, this law does not appear in our example image, although one could perceive the cars driving one behind the other as a motorcade and therefore as a group of cars.

According to [Sej18c] great progress has been made with deep learning, but people often classify objects better and faster than algorithms. The authors argued that even partially visible information can be fully recognized by humans through automatic completion in the brain. We believe that knowledge about human perception should be more widely

Figure 1.3: Picture of a house facade in Vienna. Some of the gestalt laws can be identified in this image.

incorporated into the development of computer vision algorithms than has been done so far. Sejnowski et al. stress that many methods developed in the past are limited to purely technical aspects and omit the psychological theories, such as Gestalt theory. With this work we want to contribute to counteracting this fact. The state-of-the-art works related to our work are presented in the next Section.

1.6 State-of-the-Art

In recent years, the field of computer vision shifted from statistical methods to deep learning neural network methods. There are still many challenging problems to solve in computer vision. To demonstrate the effectiveness of the novel visual summarization methods presented in this dissertation, they are applied to the following general computer vision problems:

- Video Scene Classification
- Face Recognition under different Conditions
- Object Detection
- Image Classification

In the following sections we show how images can be described using feature extraction and we present selected works on Gestalt-inspired computer vision because both topics are prerequisites for our approaches. Because one main application domain in this work is face recognition under different conditions, we present the current state-of-the-art in this field and other relevant background information below.

1.6.1 Image Description by Feature Extraction

Since feature extraction and local descriptors play a central role in this work, a few examples of selected techniques follow. In [LZL+19], many common local feature descriptors and feature extraction techniques for image matching are summarized. However, these descriptors and techniques are not limited to the field of image matching alone but are also used in many other areas of computer vision. The authors divide the local feature description techniques into six categories:

1. Gradient-based methods
2. Intensity-based methods
3. Spatial frequency-based methods
4. Moment and probability-based methods
5. Learning-based methods
6. Convolutional neural network-based methods

In the work of Zheng et al. [ZYT18] a detailed overview of methods for image description and retrieval is given, from the early 1990s until now. According to the authors mainly global descriptors were used to describe images until the early 2000s. Gradually, the global

image descriptors were replaced by local descriptors because they partially overcome the invariance limitations of global methods. In early 2000, the well-known SIFT and the Bag-of-Words (BoW) [CDF+04] model were introduced, which since then have been used for more than ten years to solve many computer vision problems. Zheng et al. stated that since 2012, the deep learning-based methods became dominant, as they achieved higher accuracies in many areas than the present methods. Since then, computer vision researchers have focused more and more on deep learning methods, especially CNNs.

Zheng et al. [ZYT18] divide the image description methods broadly into SIFT-based and CNN-based methods. The SIFT-based methods are further organized by the authors into methods using large, medium, and small codebooks for encoding, e.g. BoW. The CNN-based methods are subdivided by the authors into methods that use pre-trained or fine-tuned models, as well as hybrid methods. Zheng et al. described the pipeline of SIFT-based retrieval as follows: 1) interest point detection and description of the region surrounding the points, 2) codebook training to partition the descriptors into visual words, 3) feature encoding is the process of mapping local descriptors to the visual words generated by the codebook training, e.g. by k-means clustering [LLH16]. As mentioned earlier, we now introduce the SIFT-based methods according to Zheng et al. [ZYT18].

- SIFT-based methods using small codebooks

  A small codebook consists of no more than a few thousand visual words. The computational complexity for clustering a codebook strongly depends on the codebook size and is moderate in the case of a small codebook. Some common methods for codebook generation and encoding are BoW [CDF+04], vector of locally aggregated descriptors (VLAD) [JDSP10], and Fisher vector (FV) [PSM10]. In works where these methods are used, codebook sizes are usually small, e.g. 64, 128 or 256 visual words. The computational complexity for all three methods is similarly high but this does not matter much due to the small codebook size. A big advantage of these methods is that relatively little of the original information is discarded during encoding. In order to process the high-dimensional output of VLAD / FV quickly during image retrieval, e.g. approximate nearest neighbor (ANN) methods are used [ML14]. Another possibility is the dimensionality reduction by principle component analysis (PCA), which even improves the retrieval accuracy [JC12].

- SIFT-based methods using medium-sized codebooks

  Medium-sized codebooks are compounded of 10k-200k visual words. Compared to small codebooks, the computational effort does not increase dramatically. Therefore, flat k-means can be used for the codebook generation [TAJ13] and nearest neighbor search or ANN methods for encoding. In order to loose less information during encoding, Hamming Embeddings (HE) [JDS08] are often used together with medium-sized codebooks. HE makes visual words more discriminative by taking the Hamming distance between the HE signatures of local features into account.

- SIFT-based methods using large codebooks

Large codebooks consist of at least one million visual words. To assign the data to a large number of clusters, hierarchical k-means (HKM) [NS06] and approximate k-means (AKM) [PCI+07] were used in many works. Since the memory requirements of large codebooks are high, in [CTYH12] local features are discarded if their distance to the nearest visual word is above a threshold. In some works, weights are assigned to the visual words. With visual word weighting, burstiness is a considerable problem. Repeating structures in an image can negatively affect the retrieval process, which is called burstiness. While in many works trying to eliminate burstiness as something unwanted, this phenomenon is used in [TSOP15] as a feature.

As mentioned above, we now introduce the CNN-based methods, which have become very popular in recent years. Zheng et al. [ZYT18] divide these methods into three categories: 1) methods using pre-trained CNN models, 2) methods using fine-tuned CNN models and 3) hybrid methods.

- CNN-based methods using pre-trained CNN models

  Generating a CNN model from data is usually a lengthy process. Therefore, pre-trained models already exist, e.g. ResNet [HZRS16], which can be used for a retrieval task. However, the use of a pre-trained model also has disadvantages that are referred to in various current works [ZZW+16] as transfer effect. Experiments showed that the deeper layers of a CNN exhibit worse generalization ability than the shallower layers. In addition, the retrieval accuracy depends heavily on the data with which the transferred model was trained. The more similar these data are to the data to which the pre-trained model is to be applied, the higher the accuracy will be. In several works feature descriptors are extracted from the fully connected (FC) layers. However, as the filters of the underlying layers have already been applied to the input image, the FC descriptors can be considered as global features. Numerous other methods [NYD15] extract local descriptors in the intermediate layers and therefore have the invariance advantages known from the SIFT-based local descriptors. After applying the convolution operations in the low layers of the CNN, the resulting activation maps can be interpreted as column vectors. These column features can be encoded using the techniques already known from the SIFT-based methods, e.g. VLAD, FV and BoW. An alternative to encoding is pooling [TSJ15] that discards some of the input information. In many cases Max-Pooling is used, whereby for every m x n square neurons of a convolutional layer only the activity of the most active neuron is preserved for the further calculation steps. Despite the data reduction, the performance of the network is generally not reduced by pooling, but even offers some advantages, such as increased calculation speed and reduction of over-fitting.

- CNN-based methods using fine-tuned CNN models

Besides methods that use pre-trained CNN models, there are also methods that utilize fine-tuned CNN models. The purpose of fine tuning is to enable the CNN model to deal with data that it was not originally trained for. So, a pre-trained CNN model is retrained or, in other words, refined to apply it to new data. In recent years, for the fine tuning mainly datasets have been used which consist of buildings and general objects. One important work regarding fine tuning is [BSCL14].

- Hybrid CNN-based methods

  The third category of CNN-based methods are the hybrid methods, which select image patches from the input image using different techniques and these patches are then fed into a CNN. In some methods, these patches are additionally transformed, e.g. by feature extraction before further processing by the CNN. These methods are called "hybrid" because, like the SIFT-based methods, interesting regions are selected in an image before they are processed by the CNN. Different techniques [ZYT18] are used in the literature to select regions: 1) the input image is divided into uniform patches, 2) a sliding window strategy is applied 3) keypoint / region detectors are used 4) region proposal strategies are applied to suggest potential objects. The hybrid methods use the same feature encoding techniques that are already known from the SIFT-based methods, e.g. VLAD, FV and BoW. On the other hand, several works that deal with large codebooks exploit the inverted index on the patch-based CNN features [LKZT16]. In Chapter 8, we present a novel visual–perception inspired local–description approach, which is related to hybrid methods.

The authors of [ZYT18] made the following main observations in the course of their work as they compared the method categories described above with respect to their accuracy. For all methods, their retrieval accuracy decreases sharply as soon as the feature vector dimensionality falls below 256 or 128 bins. With SIFT-based methods, the use of a medium-sized codebook leads to the highest accuracies. HE methods in combination with medium-sized codebooks improve the trade-off between recall and precision even further. If small codebooks are used, the result is a high recall, but the precision is not very high and therefore the distinctness of the visual words is low. In CNN-based methods, as expected, the fine-tuned CNNs perform particularly well when applied to data with a distribution similar to the training data. Interestingly, however, the training data do not always have to be similar to achieve high accuracies, as the authors showed.

Additionally, the time and memory requirements of the methods are compared and discussed in [ZYT18]. The authors found out that with the SIFT-based methods, the bottleneck is the feature computation time. CNN methods are fast if GPUs are used but using GPUs for SIFT-based extraction would also greatly increase the performance. Regarding training time, according to the authors, the bigger the codebook, the longer the training takes. If a CNN is used, the training can take hours or even days and usually takes longer than the SIFT-based methods. But training can be shortened by pooling or small codebooks, albeit often by loss of accuracy. On the one hand, the authors showed

that CNN-based methods work well on all data sets to which they have applied them. These methods provide high accuracies if enough training data are available. On the other hand, SIFT-based methods also have their advantages. These methods work on gray-scale images and are therefore not dependent on color information. The authors stress that the absence of color information in the data can be a disadvantage for CNN-based methods. But too many different colors in the data can also have negative effects on CNN-based methods. If small objects are to be detected or objects are mostly occluded, the SIFT features can also work better.

1.6.2 Gestalt-inspired Computer Vision

In this Section we describe selected works dealing with Gestalt-inspired computer vision methods. The areas of application and problems to be solved are very different for such methods. Table 1.1 summarizes the methods regarding to their application domain, the applied Gestalt Laws and features.

In [BFL+18], a computer vision approach is presented, which performs medical image segmentation by using Gestalt principles. The authors introduced a Gestalt psychology-based abdominal multi-object segmentation method for soft tissues. Medical image segmentation is a very important topic as it supports experts in computer-aided diagnosis, image-guided surgery and other medical image applications. Their method works as follows. Firstly, they equally divide the input image into square patches. After that, a big surrounding patch is stretched around the patches to include neighboring context information for further processing. In the following clustering step, the minimum gradient pixel of a square patch is selected as the initial cluster center. The two Gestalt principles proximity and similarity are applied as the condition to cluster the pixels inside the corresponding surrounding patch. The spatial proximity is calculated by the Euclidean distance, and the intensity similarity by the Manhattan distance. The method used above creates irregular visual patches, which are now classified in soft tissues. For the classification, the Gray-level co-occurrence matrix (GLCM) [SQX+09] feature is extracted, which provides information about the texture of the visual patches. Subsequently, a KNN classification strategy is applied, which additionally uses medical expertise. The authors showed that their method delivers a better performance than two state-of-the-art methods [WCM+13] and [OLH+12] that do not use knowledge about human perception.

To locate and recognize Ship License Numbers (SLNs), the authors of [LSD+17] use some of the Gestalt Laws in their approach. Recognizing SLNs is important because it allows ships to be quickly identified. According to the authors, that's the first work that deals with the detection of SLNs. However, because the application domain is similar to the recognition of car license plates, they compare their method to car number plate recognition methods [DISB13]. Because recognizing SLNs presents different challenges than recognizing car license plates, i.e. SLNs exist in different colors, sizes, textures and aspect ratios, it is not possible to just use the same methods. Therefore, they proposed a new method for this specific application domain.

Work	Task	Gestalt laws	Features
[BFL+18]	tissue segmentation	similarity proximity	GLCM texture
[LSD+17]	ship license number detection	similarity proximity continuity	region intensity mean, region color mean, outer boundary intensity mean, outer boundary color mean, stroke width, gradient magnitude mean on border, region centers coordinates, aspect ratio
[Zen17]	object reification	closure	edges key vertices
[ZZH16]	detect crowd groups and their moving directions	similarity proximity common fate closure	tracklets
[YXGS16]	salient object detection	similarity proximity closure	color texture size location shape boundary structure
[EG18]	image representation by curve features	similarity proximity continuity	edges curve partitioning points SIFT

Table 1.1: An overview of selected state-of-the-art Gestalt-inspired computer vision approaches.

Their SLN recognition procedure consists of four main steps which in turn are subdivided: (1) Coarse text extraction; (2) Fine SLN location; (3) Fake-SLN elimination; (4) Missed-character compensation. They use the Gestalt Laws in different steps of their method. In the text-region coarse extraction stage [GK13] they build an MSER tree consisting of potential text regions. From this tree several possible text-group hypotheses are formed by character similarity features including geometrical features, the intensity and color mean of the region, the intensity and color mean of the outer boundary, and the stroke width and gradient magnitude mean on the border. The proximity features are mainly spatial information, i.e., x, y coordinates of the region centers. Assuming that characters of a SLN should have approximately the same aspect ratio they use the similarity of aspect ratios for the following precise localization of the SLNs. Although the authors do not explicitly mention it, they also use the Gestalt law of continuity because in

their method they assume that the characters of the same SLN are usually arranged in an approximately horizontal line. The authors showed that their method is highly competitive in comparison to five other state-of-the-art car license–plate recognition methods. Their approach is another example of the fact that the utilization of Gestalt Laws leads to high efficiency in various application domains of computer vision.

In [Zen17] a method for object reification based on the Gestalt law of closure is proposed. Object reification is the creation of objects of which only their incomplete or partly occluded shape is known. The issue of object completion has been studied in many works [GWM14], [HOS15b], [HOS15a]. Basically, the method consists of four phases. In the first phase, edges are detected in the binary images. Then key vertices are determined on the edges whose curvature is above a defined threshold. Subsequently, the edges are connected by approximated lines, ellipses, and circles. The experiments presented in the work show that their method accurately completes the shape of objects. Since the dataset used is limited to only three images, only the basic concept of the method is presented. An evaluation of the approach using larger datasets would be desirable and interesting for future work.

In the work of Zhao et al. [ZZH16] several Gestalt Laws are applied to detect crowd groups and in which directions these groups move. The algorithm is based on the clustering of so-called tracklets in video scenes. According to Zhao et al. there are other works on tracklets clustering based on the Gestalt Laws of grouping, e.g. [OMB14], [BM10], [FZS12]. A tracklet is a fragment of the track followed by a moving object. For more reasonable tracklet clustering, the authors utilize the Gestalt law of proximity, the law of similarity and the law of closure. More specifically, the authors assume that spatio-temporal adjacent tracklets should be part of the same cluster, which corresponds to the law of proximity. The next assumption is that tracklets with similar lifespans and moving direction should belong to the same cluster, which is related to the Gestalt law of similarity. The inclusion of the tracklets' moving directions also utilizes the Gestalt law of common fate, although it is not explicitly mentioned in the paper. Third, the authors consider a semantic region as a complete group according to the law of closure. As already mentioned, the proposed method detects not only groups of crowds but also their main directions of movement. This additionally makes it possible to detect abnormal moving directions of individual subjects, e.g. important for surveillance. Tracklets which are not part of the main path could be considered as abnormal behavior. The authors compared their method to four outstanding and two state-of-the-art approaches. The experiments showed that their Gestalt-inspired method is very competitive, not least because the approach is inspired by how people are likely to perceive groups and their directions of movement.

Yu et al. [YXGS16] present a novel computational model for salient object detection in complex scenes, which incorporates the Gestalt Laws of proximity, similarity, and closure into saliency computation. They have found that many of the existing salient object detection works cannot deal satisfactorily with complex scenes, i.e. scenes with complicated shaped objects and heavily textured backgrounds. The paper argues that there are far

too few salient object detection approaches that incorporate human perception and, in particular, the Gestalt Laws. This is another example that strengthens our hypothesis that theories about human perception are neglected in many computer vision areas. According to the authors, the Gestalt Laws of proximity and similarity are included in some existing salient object detection works, but other laws such as closure, symmetry and continuation are barely used. A problem with the less used laws could be according to the authors, that one would need information in advance about the objects to be detected, which is a chicken-and-egg dilemma. The authors assume in their model that the two factors of attention and perceptual grouping are fundamentally important for salient object detection. They base this assumption on the so-called *sensory enhancement theory* [DD95] of neuroscience. To get the attention information for their model, they use the already existing eye-fixation prediction model GBVS [HKP07]. From the fixation map generated in this way, different Gestalt features are extracted, and these are in turn used to generate a Gestalt graph on which the so-called Personalized Power Iteration Clustering Algorithm is applied.

The Gestalt features, which are measured between neighboring super pixels are represented as color similarity, texture similarity, size similarity, location similarity and shape similarity. To overcome the chicken-and-egg dilemma mentioned above, the authors proposed the following approach which relies on object proposals [YUG+13]. In order to incorporate the Gestalt law of closure first object proposals (hypotheses) which could include the actual objects are created. Thereafter, the closure of these objects is evaluated based on their boundary structures and this information is combined to obtain the overall closure map. According to the authors, such an approach could also be used to extract other Gestalt Laws, e.g. symmetry and continuation. The mentioned Gestalt graph carries the Gestalt proximity information because only super pixels are connected by edges when they are located within the second order neighborhood. In the extensive experiments of the work, the authors' method is compared to eleven state-of-the-art methods. The experiments show that the proposed method outperforms several of the state-of-the-art methods and that although the method is unsupervised.

Eventually, the authors of [EG18] introduced a key point and local descriptor-based method for describing images. Additionally, the selection of key points is improved by the Gestalt Laws of proximity, similarity, and continuity. We agree with the authors that while CNNs achieve human-like performance, limited training data and computing power are available in many applications. Therefore, there is still a need for local key point-based methods. The authors' method works as follows and mostly relies on object shapes because they are important descriptive features for humans too. In the first step, by horizontally and vertically scanning the image, the pixels with gradients higher than a predefined threshold are selected as starting points for the subsequent edge detection. For these starting points, neighboring points are selected according to certain rules. The neighboring points are new starting points for their neighborhood. This edge tracking process gradually creates an edge map. For each edge point, a gradient-based feature is calculated to determine if the point is a so-called curve partitioning point

(CPP) candidate. From these CPP candidates, only certain CPPs are selected, using the Gestalt Laws. According to the law of proximity, CPPs that are close together are grouped into one CPP. Edges with similar slope and curvature are grouped according to the law of similarity and the resulting redundant CPPs are removed. According to the law of continuity CPPs that do not separate different classes of edges are removed. After detection of the CPPs, bags are created by k-means clustering [LLH16] to remove redundant CPPs and make the resulting image representation more meaningful with respect to the semantic content. The k-means algorithm is fed with SIFT descriptors that describe the regions around the CPPs. In order to code the frequencies of occurrence of the CPPs within the bags, the vector quantization technique is used. The resulting histograms do not contain any location information of the CPPs and therefore, finally, the spatial pyramid matching technique is applied to integrate this information into the image representation as well.

The authors applied their approach to various publicly available datasets and divide their experiments into multi- and single-label classification problems. Their approach was compared with two other methods based on ORB [RRKB11] and SIFT [Low04] and in most cases it outperforms the other techniques for the given application domains. According to the authors, the method works particularly well for objects made by humans, such as e.g. airplanes, bottles, and buses. In future work, the authors also want to include other Gestalt Laws to further improve performance. And as in our work, they also want to use CNN based image representation methods in future experiments. The authors as well as we think that the Gestalt Laws combined with a CNN work very well to replicate human perception for content rich image representations.

1.6.3 Face Recognition under different Conditions

Face recognition has been an important research topic for more than thirty years. Since humans can recognize faces very well, there is still no intelligent system that can measure to human performance. Researchers from various disciplines, e.g. psychology, pattern recognition, and computer vision are working to solve the problem, and many articles have been published on the subject. Face recognition is commonly used to verify a face against faces in a database or to identify it in a set of face images. It is of high practical relevance in many fields, e.g. automated border control (ABC) or video-based surveillance. According to [GPK18] various aspects make facial recognition even more difficult. The authors subdivide these aspects further into intrinsic aspects like expression, weight, age or plastic surgeries and extrinsic aspects such as pose, illumination, occlusion, image quality, morphing, spoofing, or makeup. An overview of research on face recognition accompanied by the approaches concerned with the various aspects that influence face recognition can be found in [GPK18]. Below we list important aspects of these approaches.

- Image morphing

  Image morphing is used to merge biometric information of two or more faces into an artificially created face. The misuse of this technique in the context of so-called face

morphing attacks may be a major obstacle for facial recognition methods. Suppose a criminal and his accomplice look alike then they can create a morphed face image and apply for a passport. With this passport they could both go through ABC gates and if the morphed face image is well done, it can even delude people. One can easily obtain software for face morphing and the culprit does not have to be an expert in the field. A comprehensive overview over this topic and the published literature can be found in [SRM+19].

- Face spoofing

  Face spoofing or presentation attack is a serious security threat and a major obstacle for face recognition systems. If the attacker is e.g. wearing a mask or holding a face photo from a social network in front of a surveillance camera to fake a different identity, this is called a presentation attack. Another form of a presentation attack is video in which a face is seen. These videos can be played in front of a surveillance camera via tablet or smartphone. With this spoofing technique it becomes possible to fake physiological signs of life, such as eye blinking, facial expressions, and movements in the head and mouth. Meanwhile, many anti-spoofing methods have been published to make face recognition systems robust against spoofing attacks, but it remains a difficult problem to solve. In [SOPP18] Souza et al. provide a comprehensive overview on face spoofing detection. The authors see the lack of large data sets in real-world scenarios as one of the biggest obstacles to the further development of face spoofing detection methods.

- Body weight variations

  Body weight variations can have a big impact on the appearance of a face. This could be a challenge for face recognition systems when it comes to recognizing a person whose weight has changed since the face was captured in the database. It is often observed that with age variations, the weight of an individual also changes and the combination of these two factors makes face recognition even more difficult. To the best of our knowledge, there are currently not many works that deal with this topic, though it is a serious problem of crucial practical relevance. To contribute to the solution of this problem we present an approach for weight-invariant face recognition in Chapter 4.

- Makeup

  Makeup can also change a face to a greater or lesser extent and thus could be a major challenge for face recognition systems. In some approaches, it is attempted to remove the makeup before face recognition, such as in [LLH18]. Li et al. [LSW+18] proposed an algorithm for makeup-invariant face verification by introducing a bi-level adversarial network (BLAN). BLAN reduces the sensing gap between makeup and non-makeup images and can additionally remove makeup in facial images by creating synthesized faces. Because of the practical relevance, we have explored this topic intensively throughout this work and present our makeup-robust face recognition approach in chapters 7 and 8.

1.7 Contributions

In Chapter 2 we present an approach for summarizing video content to detect action scenes. Action scenes usually contain higher motion activity than other scenes in feature films while showing events like fights, gun shots, and car crashes. This work investigates motion and event detection to separate action scenes from non-action scenes. In contrast to existing work, the proposed system does not consider the shot structure of video. The approach uses SVMs to classify GIST-based global motion features, SIFT-based local motion features, and bag of MPEG-7 Color Layout features. Two test sets of movies and user–generated action movies are used to evaluate the system. The results of a frame-level evaluation indicate that especially, the global motion approach represents a good trade-off between accuracy and speed. A scene-level evaluation shows that the combined system compares well to existing shot-based approaches.

Chapter 3 shows our approach to summarizing videos into violent content and martial arts content. Real violence is unwanted content in video portals but it is forensically relevant in video surveillance systems. Naturally, both domains have to deal with mass data which makes the detection of violence by hand an impossible task. We introduce one component of a system for automated violence detection of video content: the differentiation of real violence from martial arts videos. In particular, we introduce two new feature transformations for jitter detection and local interest point detection based on Gestalt Laws. Descriptions are classified in a two-step machine learning process. The experimental results are highly encouraging: The novel features perform exceptionally well, and the classification process delivers practically acceptable recall and precision values.

In Chapter 4 we present a method for weight-invariant face recognition. We propose two improvements of the Gestalt Interest Points (GIP) algorithm for the recognition of faces of people that have underwent significant weight change. The basic assumption is that some interest points contribute more to the description of such objects than others. We assume that we can eliminate certain interest points to make the whole method more efficient while retaining our classification results. To find out which GIP can be eliminated, we did experiments concerning contrast and orientation of face features. Furthermore, we investigated the robustness of GIP against image rotation. The experiments show that our method is rotation-invariant and in this practically relevant forensic domain outperforms methods such as SIFT, SURF, ORB and FREAK.

In Chapter 5 we propose an efficient algorithm that utilizes the combination of discrete cosine transform (DCT) and phase correlation (PC) for fast object detection. To test the algorithm's classification performance and computational complexity we developed a prototype and conducted several experiments with a publicly available car dataset. Furthermore, we compared our experimental results to a state-of-the-art object detection method. The proposed method uses the energy compaction property of DCT and requires a smaller number of coefficients than fast Fourier transformation (FFT)-based techniques to compute PC. The computational complexity and memory requirements are significantly

reduced using this method. According to our results, the proposed algorithm outperforms the baseline method with respect to training time and classification accuracy.

In Chapter 6 we introduce the novel Inter-GIP Distances (IGD) feature and its integration into the Gestalt Interest Points (GIP) image descriptor. With the ongoing growth of visual data, efficient image descriptor methods are becoming more and more important. Several local point-based description methods have been defined in the past decades. Accuracy and descriptor size are important factors when selecting the appropriate method for a given retrieval problem. The method presented in this work describes images by only a few very compact descriptors. To test our descriptor, we developed an image classification prototype and conducted several experiments with a publicly available horses dataset and a food dataset. Our experiments show that only a few of the very compact GIP image descriptors are necessary to quickly classify the images from the datasets with high accuracy. Furthermore, we compared our experimental results to state-of-the-art local point-based description methods and found that our method is highly competitive.

In Chapter 7 we propose a novel approach for the domain of makeup-robust face recognition. Most face recognition schemes usually fail to generalize well on such data where there is a large difference between the training and testing sets, e.g., makeup changes. Our method focuses on the problem of determining whether face images before and after makeup refer to the same identity. The work on this fundamental research topic benefits various real-world applications, for example automated passport control, security in general, and surveillance. Experiments show that our method is highly effective in comparison to state-of-the-art methods.

Finally, in Chapter 8 we present a novel visual perception–inspired local description approach as a pre-processing step for deep learning. With the ongoing growth of publicly available visual data, efficient image descriptor methods are becoming more and more important. Several local point-based description methods were defined in the past decades before the highly accurate and popular deep learning methods such as Convolutional Neural Networks (CNNs) emerged. The method presented in this work, combines a novel local description approach inspired by the Gestalt Laws with deep learning and thereby it benefits from both worlds. To test our method, we conducted several experiments on different datasets of various forensic application domains, including makeup-robust face recognition. Our results show that the proposed approach is robust against overfitting and only little image information is necessary to classify the image content with high accuracy. Furthermore, we compared our experimental results to state-of-the-art description methods and found that our method is highly competitive. For example, it outperforms a conventional CNN in terms of accuracy in the domain of make-up robust face recognition.

CHAPTER 2

Action Scene Detection from Motion and Events[1]

2.1 Introduction

In this work we investigated approaches to detect action scenes in feature films and user generated video. The detection of video scenes that contain specific content or events is interesting for various applications like video search, summarization, classification, and navigation. Content-based analysis methods for automatic scene detection mainly differ in the scene types of interest and the analyzed content.

Well-defined rules exist for feature films to classify scenes by considering entities like time, place, and story line. One rule states that action scenes contain a series of shots with high motion activity and fast edits. This creates tense atmosphere and a sense of kinetic action and speed. However, such rules usually do not apply to user-generated content where action scenes have to be formulated in a more intuitive way. Therefore, action scenes are only supposed to contain specific events like explosions, car crashes, and gun shots in opposite to non-action scenes.

Existing scene classification systems roughly apply the following approach: (1) use shot boundary detection (2) to segment different scenes by a clustering of similar, nearby shots and (3) assign each scene to one of the given classes. In opposite to such approaches, this work investigates if action detection is possible without shot and scene segmentation. This would introduce some advantages. On the one hand, this is more flexible as it allows the beginning and end of scenes to lie within shots instead of forcing them to be on shot boundaries. Thus, it is especially interesting for user-generated videos with no or only few shot boundaries. On the other hand, it can be used to classify very short video sequences with a length of just a few seconds.

[1] Published in: 18th International Conference on Image Processing. IEEE, 2011. [SH11]

Our system works on video frame basis where each frame is classified independently by global motion, local motion, and visual event detection. A fusion step is then performed to generate larger video segments that are classified as action or non-action scenes.

2.2 Related Work

A few visual analysis systems for action scene detection have been proposed so far. All of them work on a scene level using motion and shot structure statistics. In the work of [LOM04], [CSWL09], and [LLZ+08] motion vectors of MPEG-1 video are used directly to capture motion statistics. In [LLZ+08] additionally, shot-based MPEG-7 motion activity descriptors are employed while [GXYF06] uses dominant motion direction histogram of shots. Statistics about the shot change rate and shot length are applied in [CSWL09] and [GXYF06]. The actual action scene classification is performed with a state machine in [LOM04], with a SVM in [CSWL09], and with thresholding in [LLZ+08] and [GXYF06]. The approaches of [LLZ+08] and [GXYF06] use additional audio features to improve their system. The latter work further classifies detected action scenes into the two categories, chase scenes and fight scenes, by motion and inter-shot similarity measures. All approaches above have been solely evaluated on professional feature films.

Another research direction that is related to our work is violence detection. For instance, [LW09] presented an audio-visual approach to detect fire or explosions, blood, gun shots, screams and shouts that indicate violent film shots using motion statistics and dominant colors. Recently, [SCVJA10] evaluated the bag of visual feature approach using SIFT features versus local spatio-temporal features to classify short video clips into violence and non-violence. This approach shares some similarities with our work, where local features are also used as bag of features to detect visual events.

2.3 Action Scene Detection

In order to find action scenes, motion and event detection are first performed independently of each other before a fusion and smoothing step generates a combined classification. Thereby, motion detection can be seen as the foundation of the system while event detection is used to enhance the results.

2.3.1 Motion Detection

Two different motion detection approaches are investigated in this work. The general idea of both approaches is to capture motion using the difference between a video frame and frames of the following second. This difference is small if no or little motion exists between the investigated frames and it is high if high motion or shot boundaries exist. In this way, a single motion feature is generated for each video frame. The length of one second was chosen to avoid that multiple shot boundaries are situated within an interval. In consequence, the shot structure of a movie is implicitly captured without shot

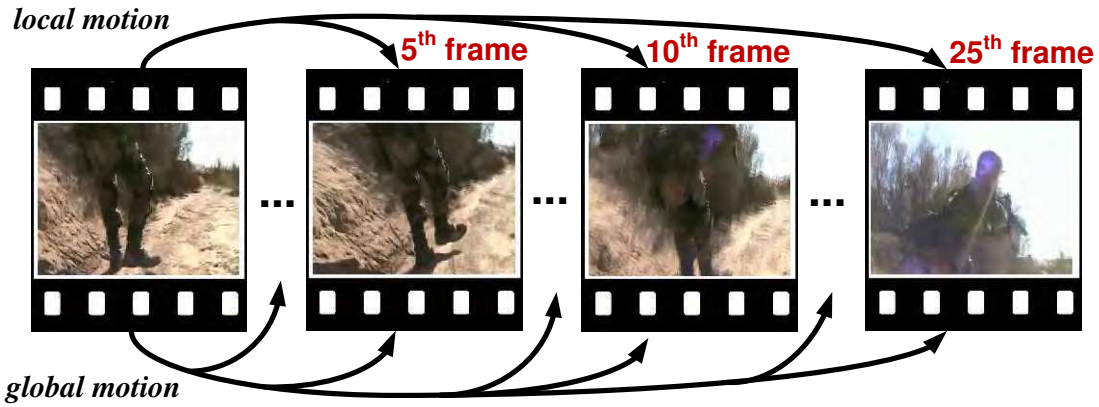


Figure 2.1: Motion features extraction. The difference of an investigated frame (leftmost one) to the frames of the following second are captured by global gist (lower arrows) and local SIFT features (upper arrows).

boundary detection and only a few frames are affected by shot boundaries in scenes with little motion. The classification of each frame is done in the same way for both approaches. Thereby, the respective motion features represent the input for SVM classification with a radial basis function (RBF) kernel that was chosen as in [CSWL09]. The parameters of the kernel are optimized with a grid search strategy. Training is performed with a set of 100 features from action and non-action sequences, extracted from one movie of the test set.

Global Motion

First, a global GIST feature [OT06] is extracted from every frame of the video. This feature gathers the image gradients of sub-regions in 16 orientation histograms with 8 bins each. A simple motion feature is then generated for every video frame by comparison with the features of the 25 successive frames using Euclidean distance. The arrows in the bottom part of Figure 1 show this task for the leftmost frame in the film strip. The resulting motion features have 25 dimensions.

Local Motion

While the global motion is suitable to gather the general amount of change in a scene, local motion can be used to describe what is going on in a scene on a much finer level. It collects information about the motion of different objects in relation to the camera movement. In contrast to global motion, we extract local motion characteristics of one second by comparison of the starting frame to the 5th, 10th, and 25th following frames instead of using every frame. This is shown in the upper part of Figure 2.1. Since such a sparse frame matching would hardly work with classical tracking or optical flow

approaches, local SIFT descriptors are used instead. In our experiments the given frame intervals have performed best.

We extract SIFT descriptors from Difference of Gaussian (DoG) points [Low04] in every frame to compute local motion. A filtering step is used to limit the maximum number of DoG points per frame to 350. Thereby, points with low contrast to their scale space neighborhood are rejected because they tend to be unstable over time compared to high contrast points. Frames with less than 10 descriptors (e.g. black frames) are thereby rejected. Similar to the original SIFT approach, a SIFT descriptor of the investigated frame is re-detected in a following frame when the relative Euclidean distance between the nearest neighbor and the second nearest neighbor is higher than 60 percent. Finally, the number of re-detections between a frame and its 5th, 10th, and 25th successor are used as first 3 dimensions of the resulting motion vector. This vector further includes the ratio between the first dimension and the other ones (the number of re-detections of the 10th and 25th frame compared to the 5th frame). In early experiments we also investigated more complex statistics to describe local motion considering the geometric correspondences of matches and the actual motion change (translation, scale, orientation), but the results indicated that the simple statistics perform better.

2.3.2 Event Detection

We further use state-of-the-art object classification methods to find a set of events that indicate violent activities and action scenes. These events are represented by images labeled with blood, emergency, fire, armed forces, police or weapons where event-related objects are shown. These images have been downloaded from Google image search and Flickr for each of these labels and for an additional non action event class. We used about 100 images per class.

Event detection is performed with the popular bag of visual features approach [SCVJA10] using MPEG-7 ColorLayout (CL) features that are densely sampled from about 300 uniform image regions considering 3 different scales. These CL features consist of the 12 values, extracted from the first coefficients of a discrete cosine transformation. We use CL features because the events of interest are better described by color than by texture or shape.

All CL features extracted from the downloaded images are clustered with K-Means to generate a codebook. A codebook dimension of 250 was chosen because of the faster computation and similar results compared to the dimensions 500 and 1000 on a single test movie. For training and prediction CL features are extracted from images of each class and from each video frame, respectively. Finally, each CL votes for up to 3 codebook bins (cluster centers) using a nearest neighbor search with Euclidean distance and a distance threshold of 0.5.

Classification is done with a cascaded SVM with RBF kernel, where the 6 event classes as well as the non-action event class are learned. The best matching class together with a posterior probability estimate is used as classification output.

2.3.3 Fusion and Smoothing

Motion classification provides the basis for combined results. If the global and local motion results agree on the same class for a frame, this result is adopted without considering the results of event detection. Otherwise, the event detection results are used to decide if a frame belongs to action or non-action. For applications where only global motion or local motion is used in combination with event detection we propose an alternative fusion. In this case, event detection overrules the motion classification of those frames where the posterior probability estimate of an event class is higher than 80 percent.

As next step, we smooth the fused results to generate longer sequences of the same scene type. Thereby, a frame is classified as action or non-action scene when at least 150 of the next 250 frames vote for action or non-action, respectively. Frames are not classified when fewer votes are given for both classes which happened to be the case for less than 5 percent of all frames in our experiments. The size of 250 frames was chosen according to the minimal action sequence length (10 seconds) in the ground truth.

2.4 Results

We have evaluated our system on two different datasets, one with Hollywood action movies and another, smaller one, with user-generated action movies. Furthermore, the proposed approaches have been evaluated on frame-level and scene-level to allow for better comparison with existing approaches.

2.4.1 Datasets

The first dataset includes the 10 feature films listed in Table 2.1. The length, number of action scenes, and percentage of action frames of these movies are given in Table 2.2. The second dataset also consists of 10 action movies but generated by amateur filmmakers. These movies were taken from YouTube and have a length between 4 and 11 minutes. Although most of these movies follow a straight story line and consist of different shots, there are considerable differences to the movies of the first dataset. For instance, bad lighting and stabilization effects often occur together with video artifacts, no expensive special effects are employed and toy guns with imitated shot sounds are frequently used. In order to generate the ground truth for both datasets we followed the approach in [LOM04] and annotated scenes as action scenes when at least one of the following events occurs: fire or explosion; violence like fighting, gun shots, robberies, shouts and screams, car chases or crashes, and sounds like alarm or breaking glass. Since there are no open evaluation sets for action scene detection so far, we made this ground truth publicly available under [Sor].

2.4.2 Evaluation

Table 2.1 shows the frame-level results of the individual modules for the Hollywood dataset. The recall values (left sub-columns) state the percentage of correctly classified

Film Title	Global Motion		Local Motion		Event Detection	
Crank 1	58.1	67.4	63.6	71.9	16.1	67.5
Crank 2	58.2	67.7	59.8	66.8	15.0	71.6
Boondock St.	59.9	68.7	61.5	67.6	14.8	73.5
The Hunted	69.8	80.0	71.8	78.7	24.5	84.9
John Rambo	67.6	77.9	68.5	77.2	25.3	70.9
Public Enem.	71.1	82.0	75.9	83.1	33.1	77.0
Shoot Em Up	61.0	70.9	61.3	68.9	31.8	56.3
Smokin Aces	67.7	75.4	73.2	79.0	32.9	74.4
Transporter	56.5	72.0	62.5	75.3	12.1	58.9
Wanted	62.5	74.3	64.9	71.8	12.8	40.2
Overall	**64.2**	**74.3**	**67.9**	**75.0**	**22.3**	**67.8**

Table 2.1: Recall (left) and precision (right) of frame-level evaluation for motion and event detection.

#	Length	Action Frames	Frame Level		Action Scenes	Scene Level	
1	84 min	33%	63.8	70.0	15	66.7	68.8
2	96 min	47%	63.3	68.8	26	88.5	71.9
3	108 min	35%	65.8	68.9	23	69.6	69.6
4	91 min	23%	78.9	83.5	19	63.2	63.2
5	87 min	30%	74.5	79.0	20	77.3	68.0
6	143 min	18%	82.6	87.1	16	93.8	83.4
7	87 min	48%	63.7	70.2	19	57.9	68.8
8	104 min	22%	78.7	82.1	23	56.6	68.4
9	95 min	34%	61.7	76.5	13	69.3	64.3
10	95 min	42%	71.8	71.3	23	62.6	72.0
$\sum$	**972**	**33%**	**71.6**	**76.7**	**198**	**70.6**	**70.9**

Table 2.2: Results of the combined system.

frames while the precision values (right sub-columns) state the number of correctly classified frames divided by the number of all classified frames. With an overall recall of 64.2% and a precision of 74.3% global motion performs only slightly worse than local motion. The overall precision of the event detection module, 67.8%, is also promising although its recall is only 22.3%. This low recall follows the fact that only those frames (32.9%) with a posterior probability estimate above 80% are classified as action or non-action scenes for this evaluation, see Section 2.3.3.

The results of the combined system are shown in Table 2.2 for a frame-level and a scene-level evaluation. On frame-level, the combined system achieved a much higher

Global Motion	Local Motion	Event Detection
125 fps	4 fps	14 fps

Table 2.3: Frames analyzed per second.

recall than all individual modules (+5% compared to local motion) and a slightly higher precision. The scene-level results are computed according to [LOM04]. Although the overall results of both evaluations are similar, there exist high variations for some movies. Especially Movie 4 (The Hunted) and Movie 8 (Smokin Aces) have much better results for the frame-based evaluation than for the scene-based one. The main reason for this is the low action frame percentage of these movies (see column 'Action Frames' of Table 2.2) as correctly classified non-action frames count higher for the frame-based evaluation. The achieved recall of 70.6% is lower than the recall of the related shot-based systems of Section 2.2 (74% - 96%) while the precision of 70.9% is similar or better (62% - 71%). However, most of these related systems are evaluated on smaller datasets of only 4 movies.

A further frame-level evaluation was done for the amateur action movie dataset. The achieved overall recall of 68.2% and precision of 71.6% are just a few percent lower than for professional material. A closer examination of the separated motion and event detectors indicates that local motion and event detection works fine for this content while the global motion performance decreases compared to the results of Table 2.1. We conclude that large motion changes are captured from unstable handheld cameras even for scenes without significant movement. However, as the dataset size is very small with a total length of just 67 minutes it is unclear if these results can be generalized to different kinds of user generated content.

2.4.3 Run time

All videos have a frame rate of 25 fps and a frame size of 480x320 pixels. The motion and event detectors were developed in C++ and for classification the LIBSVM library was used. The performance evaluation was done on a single core of an Intel 2.66 GHz quad core machine. Table 2.3 shows the number of frames that are analyzed per second including image IO and SVM classification.

2.5 Conclusions and Future Work

The main contribution of this work is an action scene detection system that works fine without using the underlying video structures of feature films, namely shots and scenes. For this task, we propose the combination of several state-of-the-art content analysis approaches that lead to different trade-offs between accuracy and speed. The proposed global motion approach appears to be the best general choice considering the high frame rate and good results. The overall system shows comparable results to shot-based approaches on professional feature films and promising results on user-generated action

movies. To facilitate future research, we further made the ground truth of our test set publicly available [Sor].

We plan to extend our system by the additional use of audio event detection because even user-generated action scenes contain typical sounds like gun shots, screams, and explosions. Furthermore, we want to use human-centered motion and event detection in order to detect events like people running, laying down, hitting and kicking or the appearances of blood on face or body, or a weapon in hand.

CHAPTER 3

New Content-Based Features for the Distinction of Violent Videos and Martial Arts[1]

3.1 Introduction

In this paper we described a content-based solution for the detection of violent video content. More specifically, we focused on one step in a greater plan: the differentiation of user-generated violent videos (which are often objectionable content on video sharing websites) from martial arts videos (which are not). The novel methods are two general-purpose feature transformations and a categorization scheme that balances over- and under-fitting.

The greater plan is a three-step process for violence detection: First, retrieving everything from a source that is potentially violent, secondly, filtering out martial arts and similar content, and thirdly, classifying the remaining videos as violent or not. Potential applications include the forensic analysis of video surveillance content and automated blocking of unwanted content on video sharing websites. Both applications are of highest relevance today: For example, in the Vienna underground approx. 480MB video content is produced per second. Currently, without knowing the exact train/station number and time, it is not possible to retrieve forensically potentially relevant content. An automated process with fair precision would improve this situation drastically.

In many works on violence detection, the classification task is applied on highly discriminative film genres, e.g. horror and romance. In contrast, we propose a method,

[1] Published in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. IEEE Press, 2013.[HE13]

which automatically distinguishes between user-generated violent videos and martial arts videos. Obviously, this problem requires a more sophisticated approach since a lot of content similarities do exist between these genres. For this end, we present the first implementation of a novel feature for video genre identification which is based on high-level perceptual Gestalt principles. Theoretically, this feature was first described in [Eid11].

In the remainder of the paper, we describe the novel feature transformations, the categorization approach, the ground truth and the experimental results. In addition, the next Section summarizes relevant related work.

3.2 Related Work

Only a few works about content-based violence detection can be found in the literature so far. In [GMK+10] a combination of two individual kNN classifiers (one for audio features and one for visual features) is used to distinguish between violence and non-violence video segments. The approach of [CHWS11] utilizes motion, blood detection, face detection and some film production rules together with an SVM to detect violence in movies.

The second research direction that is related to our work is video genre classification. For instance, in [YLYH07] semantic features and text features derived from the title, tags, and video description are used to perform web video genre classification. In the work of [HSH07] average shot length, color variance, motion, lighting key and visual effects are combined to categorize action, drama and thriller films.

3.3 Proposed Approach

Below, we describe the employed content-based features, classifiers and the ground truth. The test videos were segmented automatically into shots using the free tool Shotdetect [Mat]. Shots are decomposed into frames, of which the novel *Gestalt Interest Points (GIP)*, jitter descriptions, color and SIFT are extracted. For extracting GIP, color and SIFT features, each 25th frame of a shot serves as a key frame; for jitter detection, the first n frames of each shot are taken as key frames. A detailed description of the descriptions and their semantic interpretation by categorization follows.

3.3.1 Gestalt Interest Points

To describe the local information in an image, we use the novel GIP and SIFT. GIP is based on the Gestalt law of closure [Kof35] and the idea that, unlike other local methods, some weaker points are also useful as interest points in addition to the local extrema.

The Gestalt law of closure states that the perception of individuals fills in visual gaps in incomplete shapes. For example, humans are able to recognize a whole circle, even if there are gaps in its contour. For our approach this means that due to the Gestalt law of closure it is still possible to recognize what an image depicts, only by considering its

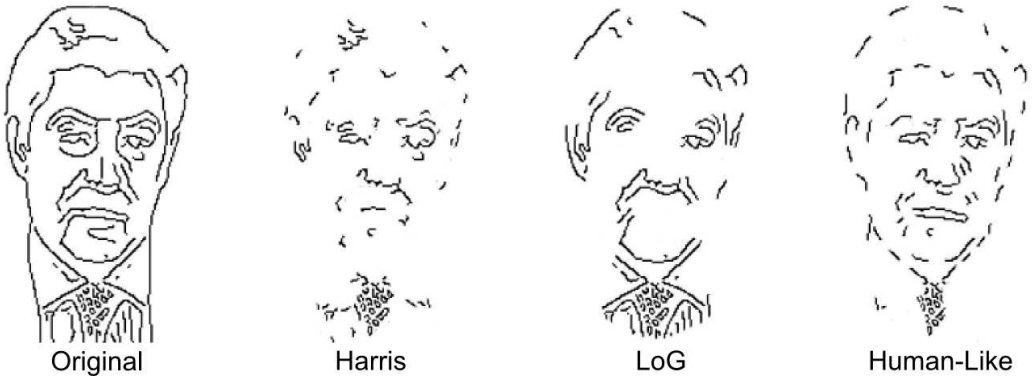


Figure 3.1: The left edge map of a face is represented by points from a Harris corner detector and a Laplacian of Gaussian (LoG) operator. Many of the important face features (e.g. the eyes) are thereby lost. The rightmost image shows the GIP description. It preserves the perceptual features of the original stimulus well but does not produce a longer description than the LoG operator. [Eid11]

local representation. This effect is shown in Figure 3.1. Obviously, such interest point sets are more useful for media understanding than points from which humans cannot identify the semantic content of an image. If the user cannot reconstruct the object from the interest points, how should the machine?

We developed several different methods for the implementation of GIP. It turned out that the following procedure leads to the best classification results for capturing body movement: Every n-th frame is converted to a gray scale image and then convoluted with an edge operator (e.g. Sobel) to get the gradient vectors and gradient vector magnitudes for each image location. The resulting gradient image is split into $m - by - n$ (e.g. 10x10) macro-blocks. For each block, we identify the three largest gradient magnitudes. These magnitudes are used to construct the image descriptor. It is composed of the three positions, the three orientations and the average of the three above selected magnitudes of each image block. Among the advantages of this straightforward scale-less implementation are the guarantee that the visual object shape is preserved in the description, and that heaps of high-curvature interest points are avoided: compared to SIFT, SURF and related methods the local description is more evenly distributed over the entire input signal without ending up in a global description.

3.3.2 Jitter

User generated videos often contain a considerable amount of jitter because such videos are often captured with hand-held cameras. In our work, jitter is a very useful descriptor for user-generated video identification. To extract the jitter descriptors we use an adapted version of Matlab's video stabilization algorithm. For our purpose, we employ only the first part of this algorithm. To accomplish jitter detection, only the first 25 frames of each

shot are taken into account. Experiments showed that this magic number represents a well-performing trade-off between accuracy and speed in the given domain. The following procedure is applied to the key frames.

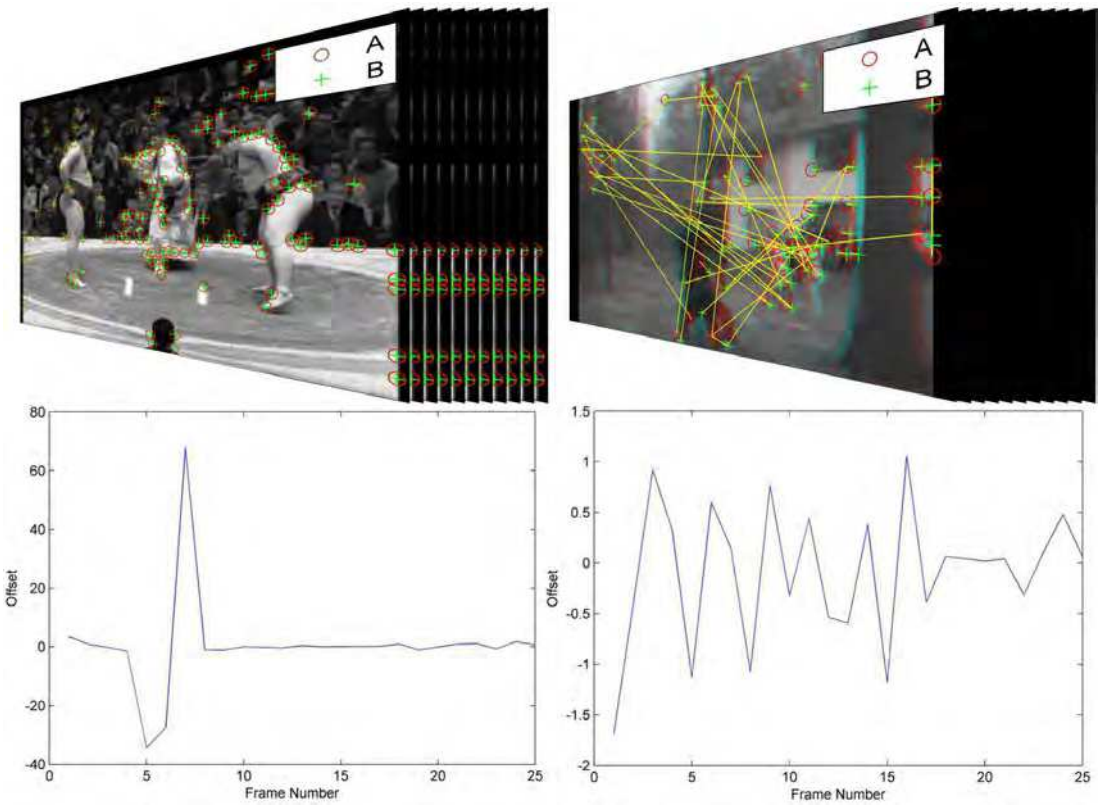


Figure 3.2: *Left-Top*: Interest point correspondences between consecutive frames of a martial arts video. *Right-Top*: Interest point correspondences between consecutive frames of a violent video. *Left-Bottom*: Interest point displacements over 25 frames for a martial arts video. *Right-Bottom*: Interest point displacements over 25 frames for a violent video.

For the extraction of the jitter descriptor the displacements of selected interest points from a frame A to its successor B are measured. In both frames a corner detection algorithm selects interest points around salient image regions such as corners. In the next step, the selected points in frame A and frame B are matched using correspondences in their neighborhood. Figure 3.2 shows the matched points with circles indicating points of frame A and plus symbols indicating points of frame B. The lines connecting points of frame A with points of frame B in the right top of Figure 3.2 represent the displacement vectors. In the example, they indicate a video that contains a significant amount of jitter. For performance reasons, we compute not more than 150 displacement vectors between each pair of key frames. The 150 x- and 150 y-components of the vectors are concatenated to form a 300-dimensional vector. Additionally, the mean is taken from these 300 values. For the 25 shot key frames, we get 25 mean values. Arranged in the

temporal order of the frames they form a signal; cf. bottom of Figure 3.2. Means and variances of the interval lengths between neighboring zero crossings are then calculated from this signal. Our experiments show the value of this description type in the sketched domain.

3.3.3 Color

Color is a fast and effective way for describing visual media. We assume that color characteristics differ significantly between martial arts videos and user generated violent videos due to quality differences in the recording devices and the fact that professional videos undergo color correction in post-production. Therefore, the video quality of consumer videos should not come up to professionally produced content. Our color feature serves as a baseline for jitter descriptions. We utilize the HSV color space to describe each key frame. The feature vector is composed of mean and variance of each of the three HSV color channels. In early experiments we also investigated RGB and CIE XY color space, but the results indicated that the HSV color space performs better for the given task.

3.3.4 SIFT

Inspired by [SCVJA10], we use SIFT combined with the Bag of Features (BoF) approach for our prototype. SIFT descriptors are extracted per key frame and transformed into histograms. We employ the SIFT-BoF implementation of the freely available CORI framework [Sor11] to compute this feature that serves as a baseline for the GIP features in the prototype.

3.3.5 Classification

Our goal is to group the shots of the test videos into *two classes*: *martial arts shots* and *user generated violent shots*. To accomplish the task, we use a separate Support Vector Machine (SVM) [SS02] for each of the four features and finally fuse the classifier decisions by applying a combination of a decision rule and majority voting. Combining SVM categorization together with rule-based method generalizes well for the given domain, because the SVM is a rigid method that avoids over-fitting on the feature data. Feeding the SVM output into a decision rule, however, opens a new degree of freedom which allows to adapt to the semantics in the ground truth. The classification algorithm was composed based on experiments in Weka and is designed as follows.

Jitter detection plays a fundamental role in the classification process. If the jitter-SVM judges a shot with confidence score above an empirically defined threshold as a user-generated violent shot, the other classifiers are not considered anymore. The confidence score is computed from the relative number of frames of a shot classified as user generated violence. The judgments of the classifiers for the other features are taken into account if the confidence score of jitter detection is below the confidence threshold. In this case the final decision is derived by majority voting of the three non-jitter classifiers.

	Martial arts			User gen. violence		
	Rec.	Prec.	F1	Rec.	Prec.	F1
Jitter-based classification	51	77	61	95	85	90
Color-based classification	72	41	52	64	87	74
GIP-based classification	81	64	72	85	93	89
SIFT-based classification	81	44	57	65	91	76
Fused classification	83	96	89	99	94	97

Table 3.1: Recall, precision and f1-score of the proposed method for both video categories.

3.4 Results

3.4.1 Dataset

We assembled a dataset that contains videos of the dataset from [SCVJA10] and videos taken from youtube.com and gorillafights.com. The dataset is composed of 214 videos (10 hours in total), 107 for each category. The martial arts category consists of wrestling, sumo, boxing, kick boxing, karate and cache fight videos; the user-generated violent part is composed of indoor and outdoor videos that show people who fight against each other. In order to generate the ground truth for the dataset, we labeled each video by hand. As mentioned above, some videos of one category are very similar to videos of the other and, therefore, they can easily be misclassified.

3.4.2 Evaluation

For the evaluation of the proposed methods, the dataset was divided into two parts: one-third for training and the rest for testing. In the first step of the experiment, the categorization process was conducted for each of the four features separately. Finally, the algorithm for fusing the classification outputs, as described in Section 3.3.5, was applied. Table 3.1 presents the results.

The f1-scores of the color-based classification are the lowest among all features. Nevertheless, this proves that color and also lighting conditions are in average to some degree different between the two categories. These differences can also be observed by the naked eye. An advantage of color is the low computational effort required for feature extraction, which is the lowest of all four considered features. The achieved f1-scores of the jitter are significantly higher than the f1-scores of the color feature. This justifies the statement above, that color (and SIFT) serves as baseline features for the two new approaches. Some scenes in martial arts videos are produced with handheld cameras and therefore the occurrence of jitter is not limited to user-generated violent videos. This leads to confusion and errors in the classification process. Some intended camera movements may also be misinterpreted as jitter.

The evaluation results for SIFT-based classification are slightly better than color-based

classification. However, the GIP-based categorization clearly outperforms SIFT. This is a remarkable result, since both features are general-purpose methods applied on the same domain and ground truth. The result supports our thesis that weaker interest points are also useful features and not only the local extrema: Sometimes it makes sense to give up stronger points for isolated weaker ones that satisfy the Gestalt rules.

As we can see from the experimental results, we proved that the GIP are an effective feature in classification of video genres. Besides, jitter is very useful to discriminate user generated videos from professionally produced videos. The last row of Table 3.1 makes clear that all four described features in combination with the proposed classification algorithm are highly effective for distinguishing between martial arts videos and user generated violent videos.

3.5 Conclusions and Future Work

Violent videos are often objectionable content on video-sharing websites, but martial arts videos are not. Motivated by this problem, we developed a system which automatically discriminates between these video genres. For this task, we proposed the novel GIP feature and a novel method for jitter detection. Experiments showed that these features together with Color information and SIFT are well suited for the given task. Classification is performed using one SVM for each description type. A combination of a decision rule and majority voting fuses the classifier results.

We are currently working on a violence detection system that integrates the proposed algorithm. In a preprocessing step the system will separate violent videos from other video types. The resulting set of violent videos is processed to filter out martial arts videos. Furthermore, since the results have been exceptionally good, we will investigate the GIP feature and jitter detection in greater detail and for other domains. We are positive that both offer great potential for solving various problems of content-based retrieval.

CHAPTER 4

Gestalt Interest Points for Image Description in Weight-Invariant Face Recognition[1]

4.1 Introduction

In this work we suggested a novel approach for the description of face images that outperforms the state-of-the-art methods for the domain of significant change of person weight. Face recognition has been an active topic of scientific research for decades now. The process requires the description of face images and the classification of the descriptions (e.g. by machine learning). Methods for description include holistic approaches (e.g. Eigenfaces) as well as local methods. Applications include image annotation for content-based search, automated video surveillance and ex-post forensic face identification. In the latter area, months and even years may lie between two images. This makes the association of a suspect with a proof (e.g. an image from a surveillance camera) a difficult task – in particular if the suspect has experienced significant weight change in the meantime. In court the authors have seen cases of *in dubio pro reo* acquittal due to insufficient biometric methodology for the association of face images.

This led us to test our GIP description algorithm on the domain. To improve its performance, we added two modifications that are described in the subsequent Section. Both are targeted at typical properties of face images: On one hand, face features are often distinguished by high contrast which is to a certain degree due to morphology of the human skull. On the other, face features tend to have a clear orientation. Both aspects are influenced by weight change: Weight gain reduces the availability and contrast of face features which also influences their orientation. The investigation of the reasonability

[1] Published in: Visual Communications Proceedings. SPIE, 2014. [HE14]

of these assumptions and their implementation is – next to the identification of the best-performing algorithm – a second target. Furthermore, we hope to improve not just the recognition performance of the GIP method but as well the efficiency of the description data.

Below, we describe the GIP algorithm, its modifications (next Section), the test dataset (Subsection 4.3.1), the evaluation process and the baseline features to which GIP is compared (Subsection 4.3.2) and the results (Subsection 4.3.3). It turns out that on the given domain the modified GIP algorithm outperforms state-of-the-art description methods such as SIFT, SURF, ORB and others. It dominates them both in terms of recognition accuracy and of description compactness. In summary, the method described in this paper produces shorter description that contain more weight-invariant face information.

4.2 Proposed Approach

Below, we briefly describe the Gestalt Interest Points feature (GIP) that are used for image description in the proposed algorithm [HE13]. After that, we explain algorithm enhancements and modifications that were specifically developed to improve the quality of the GIP features for overweight person recognition.

4.2.1 Gestalt Interest Points

The Gestalt Interest Points (GIP) are based on the Gestalt Laws of closure and continuity, i.e. the idea that, unlike in other local image description methods, certain weaker candidates are – in addition to the local extrema – also useful as interest points. The algorithm works as follows: The input image is converted to gray scale and then point-wise convoluted over the Sobel edge operator to get the gradient vectors and gradient vector magnitudes of the image. The gradient image is split into m by n (e.g. 16x16) macro blocks. For each block, the three largest gradient magnitudes $m_{1,2,3}$ are identified. The descriptor of each GIP point is composed of the three magnitude values, the three positions of m and the three orientations of m. Extended experiments have shown that this simple recipe results in descriptions that satisfy the major Gestalt Laws.[HE13]

Elimination of Low-Contrast Macro Blocks

For the visual perception and recognition of human faces edges appear to carry far more of the important image semantics than areas with low contrast. According to this assumption, we assume that interest points in low contrast image macro blocks may sometimes be discarded for the benefit of better edge description elsewhere. Hence, below we experiment with a simple scheme for the elimination of low-contrast points: For each image macro block we calculate the variance of gray values. If the variance of a block is below a certain threshold t, then the three interest points of this macro block are discarded. During our experiments we investigated the influence of t on recognition

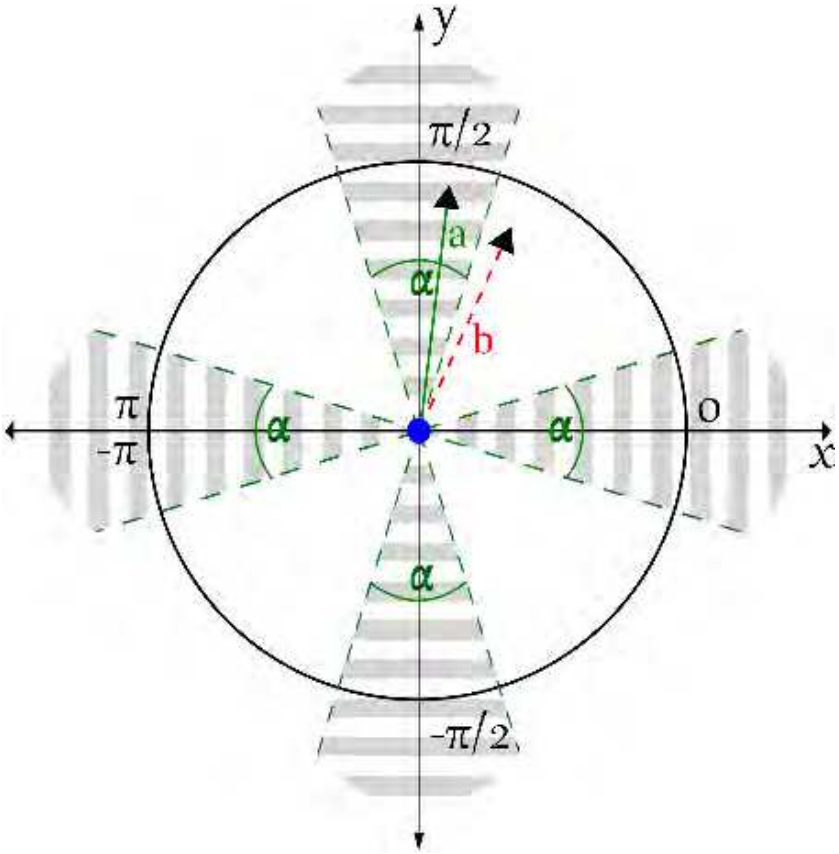


Figure 4.1: The point in the origin indicates a GIP and vector **a** its gradient, which is within one of the four circle segments. In this case the GIP will be accepted as interest point. If vector **b** was the gradient of this GIP, the GIP would be discarded because its underlying edge has diagonal orientation.

accuracy. Results are depicted in the Figures 4.4, 4.5. Explanations are given in the Results Section.

Elimination of Interest Points on Diagonal Edges

The similarity grouping experiments of Olson and Attneave [OA70] showed that human beings are significantly faster in grouping horizontal or vertical lines than of diagonals or other patterns. As an explanation for this they assumed that significantly larger parts of the receptive field are oriented horizontally and vertically than diagonally. This idea inspired us to experiment with discarding interest points that are not on horizontal or vertical edges, as these might be less important for the description and recognition of faces. To prove the hypothesis we integrated this concept into the GIP feature extraction algorithm.

Figure 4.1 depicts the basic idea of the implementation of this modification of the GIP

algorithm. The adjustable inclination angle α defines circle segments. We apply the inverse tangent function on the gradient vectors of each GIP to get the directions in which the gradients point. If one gradient vector of a GIP doesn't point in a direction within one of the circle segments, the GIP is discarded. In this case the underlying edge is considered too diagonal and therefore its interest points are considered to be of too limited use for the face recognition process.

Combination of Both Concepts

In our variation of the original GIP algorithm we integrated both modifications introduced above. In the first step we eliminate all image macro blocks which are underneath threshold t. The contrast of these macro blocks is considered as too low and therefore we discard the whole block and hence the GIP points within the block. In the second step, the GIP in the remaining macro blocks are tested for being sufficiently straight (horizontal, vertical). If not, they are also discarded. These two steps should eliminate a considerable number of GIP points which – so the hypothesis – do not contribute to the face recognition process to a sufficient degree. Hence, the modifications should make the GIP extraction algorithm more efficient for the problem at hand. Moreover, describing an image with less information should have a positive effect on resource usage and the performance of the recognition process.

4.3 Results

In this Section, the used dataset and the experimental results are presented and discussed.

4.3.1 Dataset and Evaluation Task

To our knowledge, a standardized dataset for the recognition of faces of overweight people is currently not available. The commonly used databases (FERET, UMIT, etc.) do not include such material. This is unfortunate as the problem is of high practical relevance, in particular in the forensic application of face recognition.

Therefore, we had to compile a dataset for our experiments. It turned out that pairs of face images with significant weight gain/loss in-between are hard to gain. Eventually, we succeeded in assembling a dataset of face photos for a group of fifteen persons who underwent significant weight change (at least twenty kilograms) in less than one year. The majority of the photos were taken from a diet web forum.[Red] Others were provided by acquaintances of the authors.

For the experiments below, without loss of generality we employ the face images with lower weight as the training set. The test set consists of the face photos that show the higher weight. Figure 4.2 shows three example images and descriptions extracted by the GIP algorithm. The evaluation task is to associate each test image with the corresponding training image. Success is measured as accuracy, i.e. here the number of true positives. The ground truth is provided by the authors.

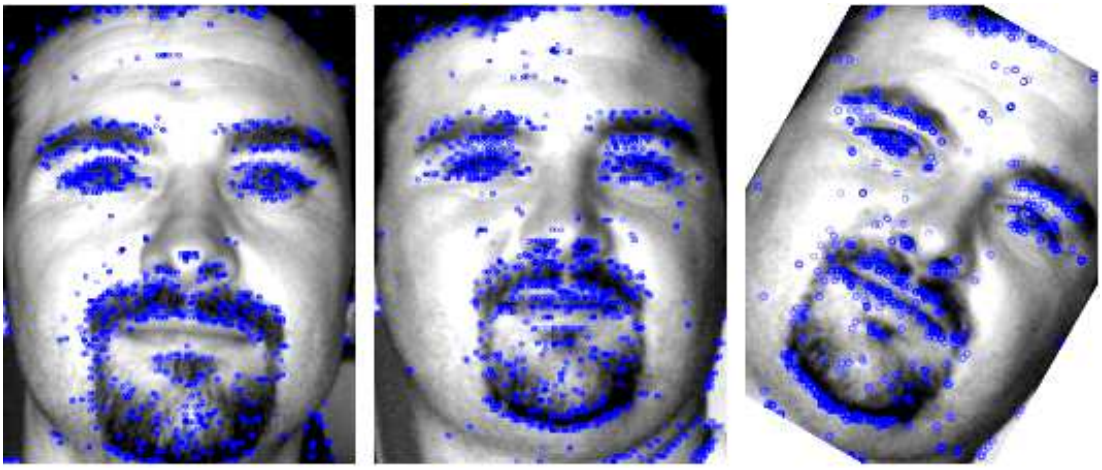


Figure 4.2: Our novel variations of the GIP-algorithm were applied on pictures such as these examples. GIPs which are within low-contrast macro blocks and many of the GIPs on diagonal edges were discarded. *Left*: Shows a normal weight person and the detected GIP points, indicated as dark blue circles. *Middle*: Shows the same person after 30 kilograms of weight gain and the detected GIP points. *Right*: Shows the person image rotated by 30 degrees and the detected GIP points.

Remark: In practical forensic application, pictures of suspects (e.g. taken by a surveillance camera) are typically of highly variable quality. To evaluate how well our and the state-of-the-art local description algorithms can deal with this aspect, the photos in the dataset are left in their original resolutions, ranging from 201x285 to 508x728 pixels. However, the contrasts of the test images were adapted to the contrasts of the training images using histogram equalization because this step improves the general classification performance without limiting the generality of the experiment.

4.3.2 Baseline Features

Five commonly used local feature description methods were chosen to be compared with the GIP feature: SIFT [Low04], SURF [BETVG08], MSER [MCUP02], FREAK [Ort12] and ORB [RRKB11]. We quantized/trained all of them with the popular BoVW-algorithm [CDF+04]. The description methods can be characterized as follows.

- *SIFT:* The Scale-Invariant Feature Transform algorithm [Low04] is very popular for detecting and describing local features in images. We employ the *OpenCV* [Bra00] implementation to compute this feature. Each interest point is described by a 128 elements vector.
- *SURF:* The Speeded Up Robust Features algorithm [BETVG08] works similar to SIFT but is several times faster than SIFT. We use the *OpenCV* [Bra00] implementation for our work. Each point is again described by a 128 elements vector.

Table 4.1: Overall Results.

Method	Accuracy	Accuracy 30° rotated	Average Number of Values per Face
BoVW+SIFT	20%	13.3%	25,309
BoVW+SURF	33%	13.3%	57,984
BoVW+MSER	6.7%	6.7%	132
BoVW+FREAK	20%	20%	59,473
BoVW+ORB	13.3%	13.3%	8,883
BoVW+GIP	53.3%	53.3%	52,536
BoVW+GIP minus low-contrast IPs	53.3%	53.3%	23,086
BoVW+GIP minus diagonal IPs	46.7%	46.7%	23,101
BoVW+GIP minus both IP types	46.7%	46.7%	10,425

- *MSER:* The Maximally Stable Extremal Regions algorithm [MCUP02] detects blobs in images. Each blob is described by a four elements vector. For our work we use the *VlFeat* [VF08] implementation of MSER.

- *FREAK:* The Fast Retina Keypoint algorithm [Ort12] uses binary descriptors to describe keypoints. It employs a pseudo- human-like manner of capturing visual information – coarse in peripheral regions of the retina and fine in the central fovea region. FREAK is a particularly fast image description algorithm. We use the *dovgalecs*[Kor] implementation for our work. Each interest point is described by a 64 elements vector.

- *ORB:* The Oriented FAST and Rotated BRIEF algorithm [RRKB11] is basically a fusion of two other algorithms (FAST keypoint detector and BRIEF descriptor) with many modifications (e.g. rotation invariance) to enhance the performance and add functionality. We use the *OpenCV* implementation for our work. Each point is described by a 32 elements vector.

For classification of the features we simply use the Euclidean distance. Hence, all standard descriptors as well as our approach are employed in exactly the same way. This is a mandatory requirement for comparing the description performance for the recognition problem at hand.

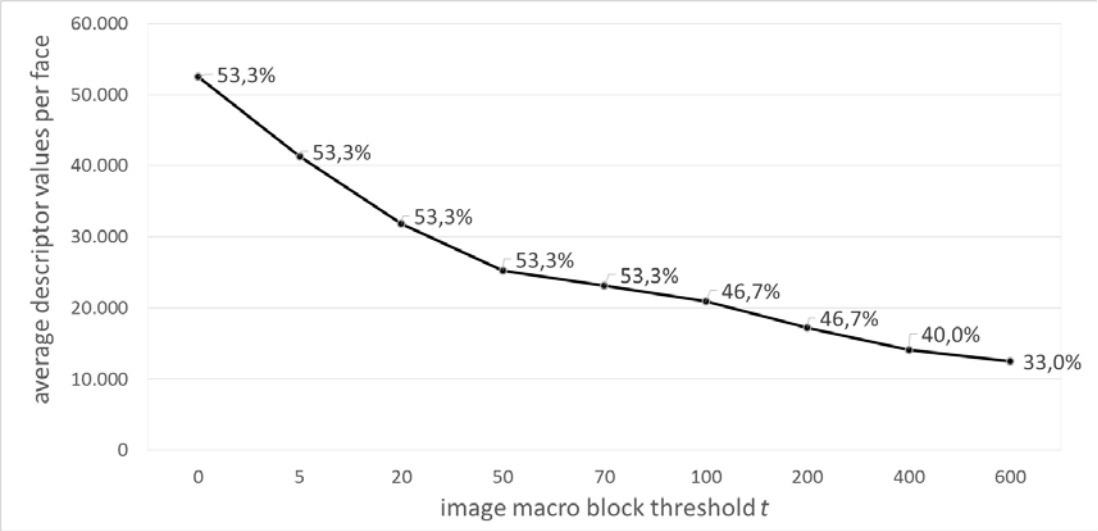


Figure 4.3: The average number of description values per face and categorization accuracy in percent as a function of image macro block variance threshold t. The variance of an image macro block has to be above t. A value of $t = 0$ means that no image macro blocks are discarded (equivalent to the original GIP algorithm).

4.3.3 Evaluation

The second column of Table 4.1 shows that using the five state-of-the-art features SIFT, SURF, MSER, FREAK and ORB to identify the faces of people who experienced significant weight change delivers only moderate classification accuracies. Of all five features, with 33 percent SURF provides the best results. However, to obtain this result in average 57,984 description values per face (last column) are necessary. This number is calculated as the product of the average number of description vectors per face (453) by the size of one description vector (128). Using the MSER feature in average only 132 description values per face are required. In return, the classification accuracy is only 6.7 percent which is far below an acceptable rate for practical application.

Compared to the five popular features above, the GIP description algorithm is in its original form with 53.3 percent by far more accurate. This performance is 20 percent ahead of the SURF algorithm that leads the state-of-the-art description methods. As Figure 4.3 shows, discarding interest points that lie within low-contrast macro blocks does not affect the accuracy until the threshold reaches $t = 70$. At the same time, the average number of description values per face goes down from 52,536 to 23,086. That is, by discarding low-contrast blocks we maintain the original accuracy of the GIP algorithm but reduce the amount of data to just 44%. The required information for this result is even less than the information SIFT, SURF and FREAK need to achieve their lower performance. Hence, we consider it just to say that for the given domain, the GIP approach clearly dominates the state-of-the-art algorithms.

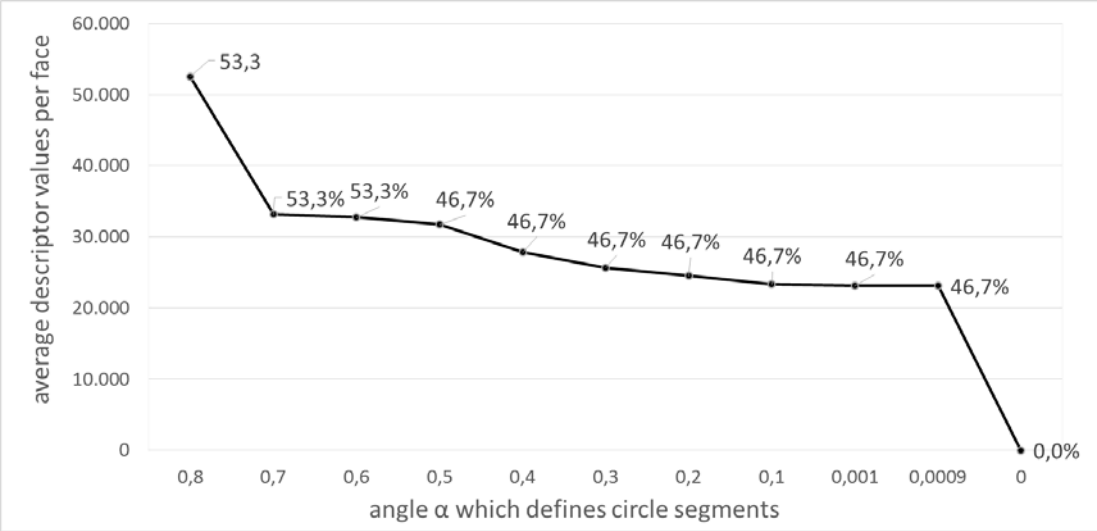


Figure 4.4: The average description values per face and categorization accuracy in percent as a function of circle segment angle *alpha*. A value of $alpha = 0.8$ radians means that no GIPs are discarded (equivalent to the original GIP algorithm). If there are no perfect straight lines in the face image dataset then for $alpha = 0$ the accuracy will drop to zero.

Figure 4.4 shows that the elimination of diagonal edges in the GIP algorithm does not affect the accuracy until an $alpha = 0.64$ is reached, but this modification reduces the average number of description values per face drastically. With $\alpha = 0.0009$ and 23,101 description values the algorithm still reaches an accuracy of 46.7 percent. A classification accuracy of 46.7 percent is still significantly higher than the results reached by the commonly used local feature transformations. In comparison to the original GIP algorithm less than half of the information is enough to achieve results that are only slightly inferior. We find these results encouraging to employ these modifications also in other application domains.

The combination of discarding interest points with low contrast and of points on diagonal edges leads to a significant reduction of the average number of description values required per face to 10,425 values. Figure 4.5 illustrates the behavior of the algorithm. An accuracy of 46.7 percent is still significantly higher than the results of the baseline features. Therefore, the combination of both methods appears to be a well-performing trade-off between performance and accuracy. This result supports our hypothesis that interest points on almost horizontal or vertical edges are more useful for face description than other points. Furthermore, it indicates that our hypothesis (certain interest points in low contrast areas can be neglected) has empirical substance. There appears to exist a trade-off between Gestalt perception and the focusing on salient points.

Eventually, we evaluated the sensitivity against rotation of our approach. All baseline feature extraction methods are to a certain degree rotation-invariant. To find out how robust the GIP approach is against rotation, we conducted an experiment with all test

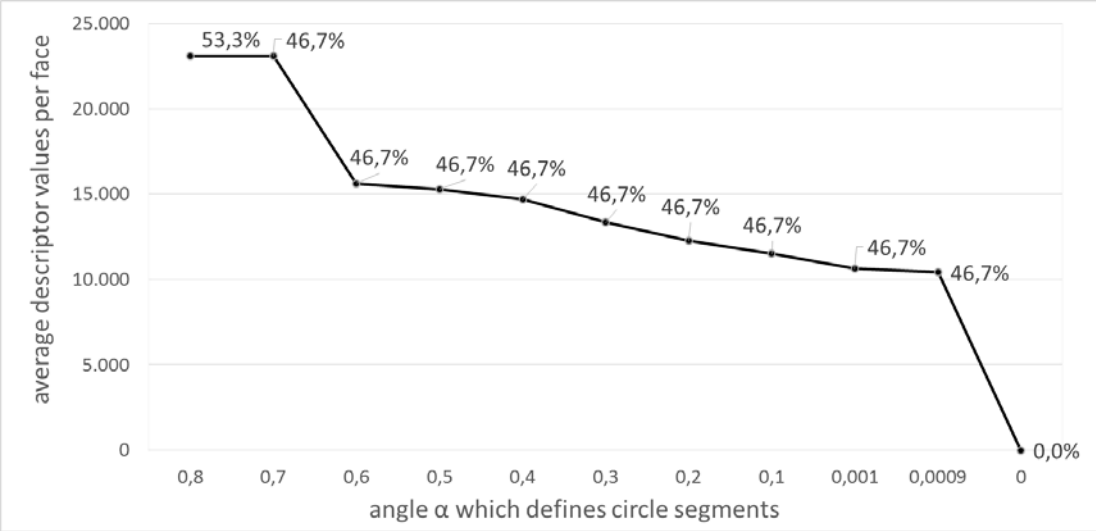


Figure 4.5: The average number of description values per face and categorization accuracy in percent as a function of circle segment angle *alpha* with $t = 70$.

images rotated by 30 degrees. The third column of Table 4.1 depicts the outcome. As can be seen, both SIFT and SURF deliver lower accuracy for rotated images. The accuracy of MSER, FREAK and ORB remains constant. Likewise, GIP is not affected by rotation: The performance remains constant. Hence, we consider it fair to conclude that the modified GIP approach is a highly competitive local description approach for the problem under consideration.

4.4 Conclusions and Future Work

We have introduced a novel approach for the description of face information for recognition. The algorithm is based on our Gestalt Interest Points approach and modified in two respects: Certain low-contrast points are eliminated, and diagonal points are neglected. The approach was tested empirically. The results are twofold:

1. The modified GIP approach describes faces significantly better than the state-of-the-art methods do. Its accuracy is at least 20% better than of the first competitor (SURF). As assumed, the relative completeness of Gestalt interest points makes a huge difference in recognition performance.

2. GIP descriptions are more compact than most other descriptions and they are rotation-invariant. That is, we need less disk space and processing power for description storage and evaluation. This is an important advantage in a big data domain such as face recognition. The rotation invariance is a simple requirement satisfied by most – yet not all – algorithms.

For the future, we plan the following enhancements:

1. One major theme is the provision of an industry ready algorithm. For that we require an automated procedure for the optimization of threshold t. We plan the implementation of a heuristic scheme that is based on state-of-the-art methods from operations research. In this process, we will continue to refine and enlarge our image database by collecting more pairs of face images with significant weight change.

2. To improve the perceptual level of GIP, we are developing an algorithm that fuses it with the FREAK approach. There, detected Gestalt points serve as input for the FREAK algorithm. Hence, FREAK is employed to describe GIP points.

These modifications should widen the applicability of the approach while preserving the central idea of description in accordance to the Gestalt rules.

CHAPTER 5

An Efficient DCT template-based Object Detection Method using Phase Correlation[1]

5.1 Introduction

In we suggested a novel template-based object detection approach that outperforms the state-of-the-art methods for the domain of vehicle detection. Object detection has been an active topic of scientific research for decades now and various approaches (e.g. viola jones method) emerged. Applications include image annotation for content-based search, automated video surveillance and pedestrian detection.

Many psychological works [Pyl02, Kos94] assume that human object recognition is based upon object templates stored in the brain or on simplified representations of them. It is therefore not surprising that this concept was also employed for computers. The traditional theory of template models [New92] assumes that image-like representations of different views of an object are stored in the brain. This theory goes hand in hand with high computational complexity and considerable storage requirements. The templates are probably transformed in a more efficient representation to compensate these drawbacks. We adopted this assumption for the proposed algorithm and reduced the object templates to their essential aspects.

In visual template matching, the goal is to find the region in one image that matches a specific template. According to [Eid12a] template matching falls into two sub problems, namely encoding the input stimuli in some kind of description and similarity measurement, which is typically solved by convolution. The convolution output will be highest for

[1] Published in: 50th Asilomar Conference on Signals, Systems and Computers, 2016. [HE16]

areas which most closely match the template. In contrast to one-dimensional template matching the input images are usually not smoothed in visual template matching. In the past, template matching was often only used in dedicated hardware solutions because of the computational complexity. The presented method reduces computation complexity drastically.

The main contribution of this paper is proposing a fast DCT-based PC method for detecting and precisely localizing objects in images, which requires little training effort. A combination of DCT and PC can be found in [PP15] but for the purpose of image mosaicing. In many applications like weapons targeting or videogrammetry, it is important to precisely localize an object of interest instead of simply recognizing that there is an object somewhere in the image. Because of the relatively short time until the algorithm is trained, it is especially suitable for ad hoc queries where a long training time is not acceptable or not economical, e.g. in manufacturing as a part of quality control. Additionally, a big advantage of DCT is that a lot of digital media is stored as DCT transformed data, e.g. the widely used MPEG and JPEG compressed files. For these types of media, the first step of our algorithm can be omitted and makes it more efficient again.

Below, we describe our DCT-PC algorithm, the test dataset, the evaluation process and the baseline method to which our approach is compared. Finally, we present the results of our experiments. It turns out that on the given domain the DCT-PC algorithm outperforms the state-of-the-art viola jones method. It dominates it both in terms of detection accuracy and training effort. In summary, the method described in this paper requires considerably less training time to detect objects much more accurately.

5.2 Proposed Approach

We developed a system which, given an image as input, returns a list of locations at which instances of the searched object class are detected in the image. Note that this problem is distinct from and more challenging than the problem of simply deciding an input image contains an instance of the searched object class or not. Evaluation criteria for the detection problem are discussed later in the paper.

Below, we explain the theory behind our algorithm and after that the algorithm itself. Our approach for learning to detect objects consists broadly of two stages, which are outlined briefly below:

1. ***DCT***
   DCT is a variant of Fourier transform suitable for many image processing applications but using only real numbers. It represents a signal as a sum of cosine functions of different frequencies. There are several slightly different variants of DCT. The DCT-II is probably the most commonly used form and for the one-dimensional case it is defined as follows:

$$X_k = \sum_{n=0}^{N-1} x_n \cos[\frac{\Pi}{N}(n+\frac{1}{2})k] \qquad k = 0, ..., N-1. \tag{5.1}$$

A two-dimensional DCT is simply a separable product of 1D DCTs along each dimension. Typically, an image is transformed applying 2D DCT on 8 by 8 image blocks. The dimensionality of DCT coefficients matrix is equal to the size of the input image. To save computing time and storage requirements the number of DCT coefficients can be reduced (similarly as in JPEG compression). Most of the image information is concentrated in a few DCT coefficients, which are located in the upper left corner of the DCT coefficients matrix. With the inverse cosine transform (IDCT) it is possible to transform an image back into spatial domain, which can be done lossless, if all the DCT coefficients are involved.

We wanted to find a quick way to compute the DCT and compared 3 different implementations, namely the Matlab built-in implementation, Narasimha's implementation [NMP78] and the DCT from the Medical Image Registration Toolbox (MIRT) [Myr]. It turned out that the MIRT-DCT implementation is approximately one third faster than the Matlab DCT and therefore the fastest among the three considered followed by Narasimha's implementation. Due to these findings we used the MIRT-DCT implementation for our experiments, although probably even faster methods exist.

2. ***Phase Correlation***
If an image and a translated version of this image is given, the PC can be used to find the displacement between these two images. The shifting property of the Fourier transform states that the coordinate displacement of two functions in spatial domain is transformed as difference of phases in Fourier domain. A difference of phase means that for example two sine waves with equal period duration have different zero crossings. PC is based on this property. The main part of PC is calculating the so called cross-power spectrum R:

$$R = \frac{G_a \circ G_b^*}{|G_a \circ G_b^*|} \tag{5.2}$$

Ga and Gb are the Fourier transformed input images. Ga is element-wise multiplied with the complex conjugate of Gb and this product is then element-wise normalized. Finally, R is transformed back into spatial domain using the inverse Fourier transform and the peak in the resulting function denotes the translation of the two input images. The surface of PC is characterized by a sharp symmetric peak at the location of a found object instance, which can be interpreted as a belief score plus very low amplitude peaks at other locations. PC is used for image registration [FZB02], motion estimation [KABN12] and as it is presented in this work also for object detection.

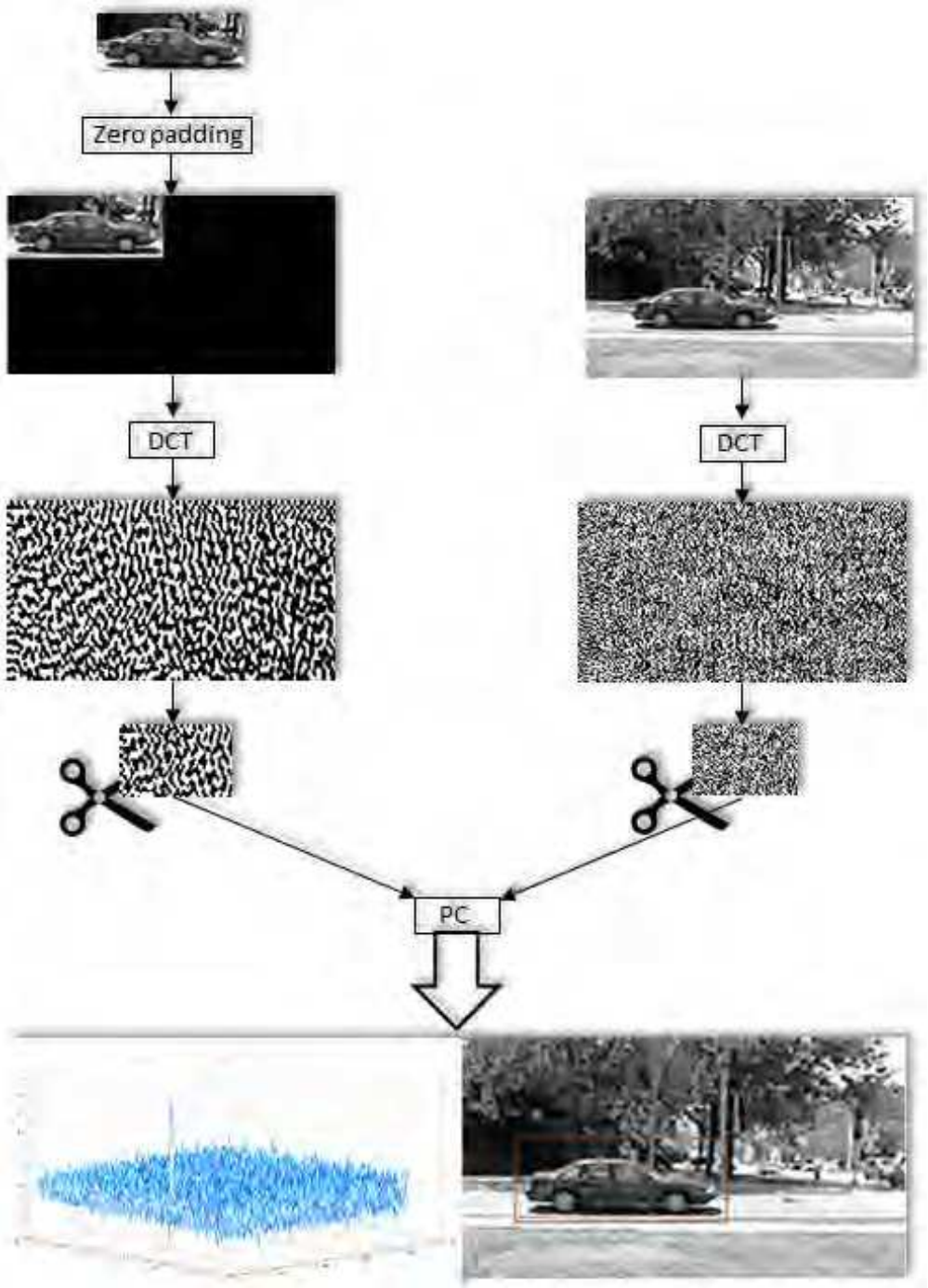


Figure 5.1: Overview of our DCT-PC algorithm. An object template (top left) and a search image (top right) serve as input. After zero padding the template both input images are transformed into frequency domain and a reduced set of the resulting DCT coefficients is then correlated with PC. The algorithm provides the template location in the search image as output.

According to [PP15] instead using PC someone could also use cross correlation (CC). But the authors demonstrated that this method is not that suitable for their application of image mosaicing in comparison to PC. From these findings and the already mentioned advantages of DCT combined with performing correlation in frequency domain, we can clearly conclude, that PC is the preferable method for the proposed approach. CC is not invariant to lighting conditions and the main problem is that the CC function is characterized by a lower broader peak centered at the registration point, which is only a local maximum, accompanied by other peaks. The sharp and global PC's peak instead indicates the percentage of the overlapping area.

The FFT-based PC technique is efficient for many applications but as already mentioned, replacing FFT with DCT makes it possible to compress images by using only a very limited number of coefficients where all the energy is concentrated. Therefore, it is computationally less expensive and reduces storage requirements. The proposed DCT-PC method is represented schematically in Fig. 5.1 and the steps are described as follows:

1: **function** DETECTOBJECTS(I, T, C)
Input: Gray-scale search image I and object templates T
Output: Array C representing search image locations at which instances of the object class are detected
2: Read search image I and convert it from RGB to gray-scale
3: **for all** $t \in T$ **do**
4: Enlarge t to the same size as I using zero padding
5: Compute DCT of I and take the complex conjugate of t
6: Filter out less important DCT coefficients
7: Compute the normalized cross power spectrum
8: Apply IDCT to get the coefficients matrix
9: Save matrix indices C of coefficients $>$ threshold T_0
10: **end for**
11: **return** C
12: **end function**

Fig. 5.2 depicts the impact on a car template when it will be transformed into frequency domain by DCT and then transformed back into spatial domain by applying inverse DCT on a reduced set of the coefficients. With 50 percent coefficients the car is clearly identifiable but with only 4 percent it is hardly possible to recognize it for human beings. How many coefficients are necessary for our DCT-PC method to identify cars will be presented later in this work.

5.3 Results

In this Section the used dataset and the experimental results are presented and discussed.

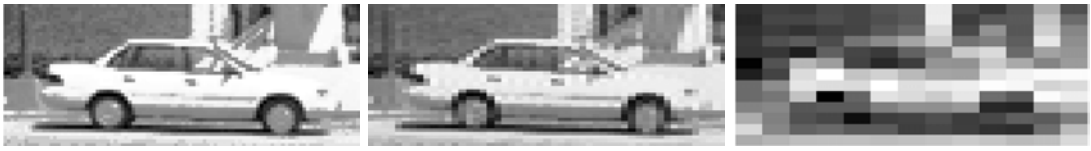


Figure 5.2: *Left:* Template of a car. *Middle:* The template omitting 50 percent DCT coefficients. *Right:* The template omitting 96 percent DCT coefficients.

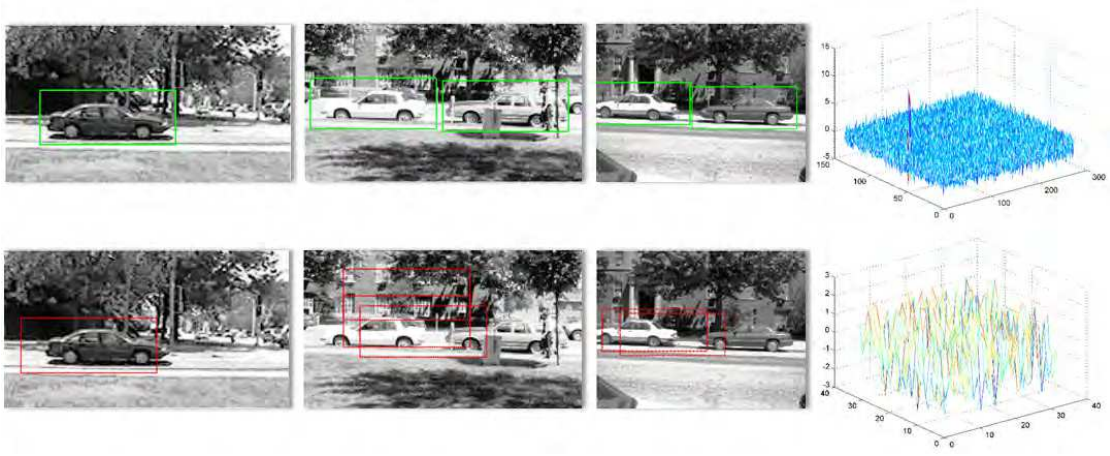


Figure 5.3: Examples of detection on the UIUC dataset. The top row shows correct detection results with 100 percent DCT coefficients and a typical PC output at far right. The bottom row shows some false positive examples with only four percent DCT coefficients and a typical PC output at far right.

5.3.1 Dataset

To test detection performance, we used the UIUC car dataset [AAR04]. It is divided into two parts: The first part includes 170 gray-scale images, containing a total of 200 side views of cars, with some images containing multiple cars. All the car images in this test set are of size 100 x 40. The second part consists of 108 gray-scale images containing 139 cars at various sizes with a ratio between the largest and smallest cars of about 2.5. In the next Section we will study the properties of this approach for solving the task of car detection.

5.3.2 Evaluation

Two different baseline algorithms, namely FFT-based phase correlation and the detection framework by viola and jones [VJ01] have been considered for qualitative as well as quantitative comparison. Viola and Jones combine Haar-like features with the adaptive boosting (AdaBoost) [FS97] algorithm, which can be used to improve the performance of a classifier. Although the framework is fast, it is prone to over-fitting in some applications and training can be very time consuming. Like many others in the object detection

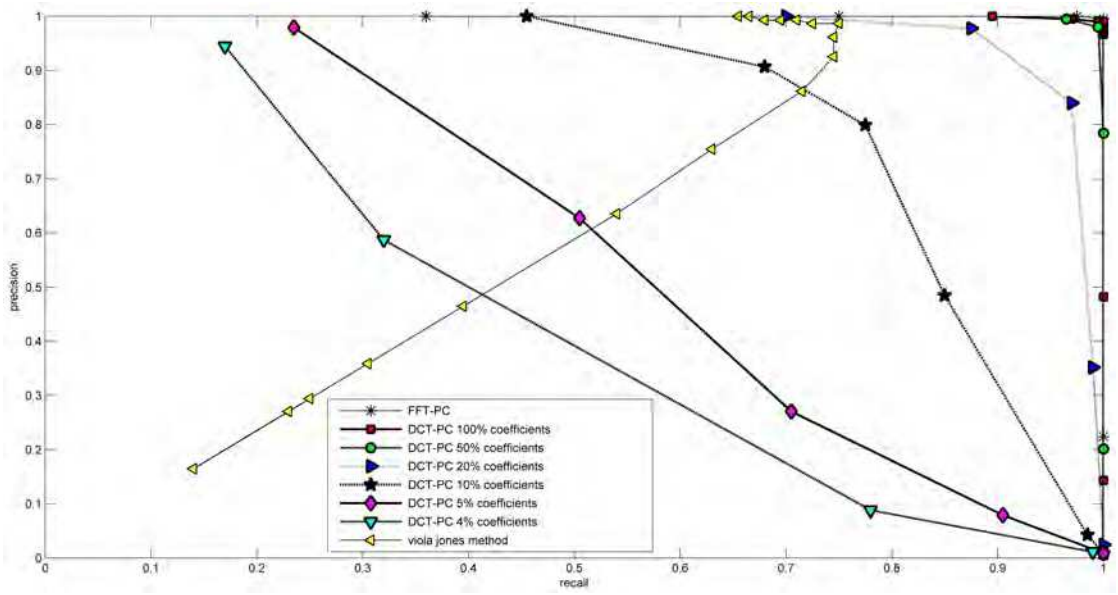


Figure 5.4: PR curves for the UIUC single-scale dataset. Note that the performance is getting worse as the number of DCT coefficients decreases. But with only 20 percent of the DCT coefficients the performance of our method is still better than the viola jones method.

domain, we present our results utilizing Precision-Recall (PR) curves. A PR curve illustrates the trade-off between recall and precision over a range of threshold parameter values. Such a threshold is included in many object detection algorithms and in our work, it is called T_0.

Similar to the viola jones method we have to train our DCT-PC algorithm with templates of the relevant objects. The goal of the training stage is finding the threshold value T_0 which achieves the highest F-measure. This value represents the best trade-off between recall and precision. After determining T_0 the algorithm can be employed to find object instances which are not included in the train set.

The top row of Fig. 5.3 shows some detection results with 100 percent DCT coefficients classifier. All the cars are correctly detected and the corresponding PC function (top right) is characterized by a sharp peak at the image location where a car was found. More precisely this peak denotes the upper left corner of the bounding box in the first search image (top left). The bottom row shows some false positive examples. In this case only four percent of the DCT coefficients were used for detection and the results are not as good as with 100 percent coefficients. Note that there is no sharp peak in the corresponding PC function (bottom right). Therefore, no certain statement can be made about the location of a potential car in the test image.

Fig. 5.4 depicts the PR curves of our experiments for the single-scale car dataset

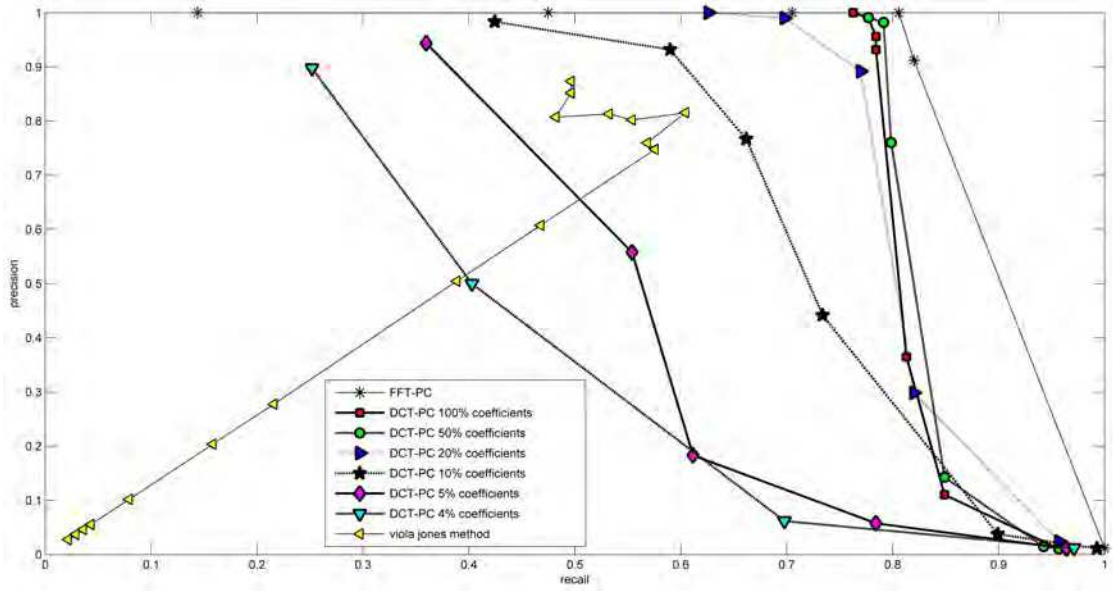


Figure 5.5: PR curves for the UIUC multi-scale dataset. With only 10 percent of the DCT coefficients the performance of our method is still better than the viola jones method.

with different numbers of DCT coefficients and Table 5.1 compares the corresponding computation times. Each point of a curve represents a PR value pair for a certain value of T0. The PR curve of the viola jones method is made upon PR values of different classifier stages. With 100 percent DCT coefficients the best classification results are achieved. The results are very similar to the FFT-PC results. However, according to Table 5.1 FFT-PC is with 742 ms per image the slowest method. DCT-PC with 100 percent coefficients is per image about 100 ms faster than FFT-PC and with a recall of 100 percent and a precision of 98 percent it clearly outperforms the viola jones method for our domain. Omitting 50 percent of the DCT coefficients has hardly any effect on the classification results and the computation time. With only 20 percent of the coefficients the algorithm still achieves much better results than the viola jones method and with 442 ms per image it is about 300 ms faster than the DCT-PC with 100 percent coefficients. This proves that the algorithm becomes faster as the number of DCT coefficients decreases. The DCT-PC results fall below viola jones results only after the number of coefficients comes down to 10 percent. These experiments demonstrated that omitting up to 80 percent of DCT coefficients, which corresponds to a strong image compression, still delivers comparable high classification results and is about twice as fast as classification with 100 percent coefficients.

Fig. 5.5 shows the PR curves of our experiments for the multi-scale car dataset with different numbers of DCT coefficients and Table 5.2 compares the corresponding computation times. From the results it can quite clearly be seen that classifying this multi scaled cars is much more challenging than classifying the single-scale car dataset. With at least 50

Method	Training time (hours)	Average time per image (ms)
FFT-PC	0.5	742
DCT-PC with 100% DCT coefficients	0.5	638
DCT-PC with 50% DCT coefficients	0.5	640
DCT-PC with 20% DCT coefficients	0.5	442
DCT-PC with 10% DCT coefficients	0.5	378
DCT-PC with 5% DCT coefficients	0.5	345
DCT-PC with 4% DCT coefficients	0.5	335
viola jones	48	1.4

Table 5.1: Comparison of computation times using UIUC single-scale dataset.

Method	Training time (hours)	Average time per image (ms)
FFT-PC	0.5	853
DCT-PC with 100% DCT coefficients	0.5	654
DCT-PC with 50% DCT coefficients	0.5	736
DCT-PC with 20% DCT coefficients	0.5	503
DCT-PC with 10% DCT coefficients	0.5	399
DCT-PC with 5% DCT coefficients	0.5	352
DCT-PC with 4% DCT coefficients	0.5	345
viola jones	48	3.7

Table 5.2: Comparison of computation times using UIUC multi-scale dataset.

percent of the DCT coefficients the recall is about 80 percent and the precision is close to 100 percent. Although the FFT-PC method delivers slightly higher classification results it is about 200 ms per image slower. With only 10 percent coefficients the DCT-PC method delivers a recall equal to 66 percent and a precision equal to 78 percent. This is much better than the viola jones method. Note that for the single-scale dataset at least 20 percent coefficients were necessary to outperform the viola jones method. When reducing the set of coefficients to 5 percent the algorithm performs not as good as the viola jones method anymore. So, omitting up to 90 percent of DCT coefficients, which corresponds to an even stronger image compression than classifying the single-scale car dataset, the algorithm still delivers comparable high classification results.

Table 5.1 indicates that the viola jones method takes in average 1.4 ms to detect a car in an image. Although this is faster than our DCT-PC algorithm the 48 hours which are necessary for training is a lot longer in comparison to 30 minutes for our algorithm. Furthermore, the recall and the precision of the viola jones method is relatively low. In other words, the training of our algorithm is one hundred times faster while delivering a much higher classification performance for our application domain.

5.4 Conclusions and Future Work

In this work, we have presented a novel approach for object detection in images. The use of DCT reduces the computational complexity significantly, thereby making the method suitable for applications involving compression and transmission of images as well as videos. Experimental results on a publicly available vehicle dataset have shown the effectiveness of the proposed method compared to the state-of-the-art viola jones method. The experiments demonstrated that omitting up to 90 percent of DCT coefficients, which corresponds to strong image compression still delivers comparable high classification results and is about twice as fast as classification with all coefficients. For the UIUC car dataset the proposed algorithm classifies cars much more accurately and the training is one hundred times faster than the baseline method, which makes it a good choice for ad hoc queries.

For the future, we plan to speed up the matching process. We are positive that this can be accomplished through integrating an image pyramid into the algorithm. The lower resolution images can then be searched for the template, in order to yield possible start positions for searching at the larger scales. The larger images can then be searched in a small window around the start position to find the best template location.

CHAPTER 6

The Gestalt Interest Points Distance Feature for Compact and Accurate Image Description[1]

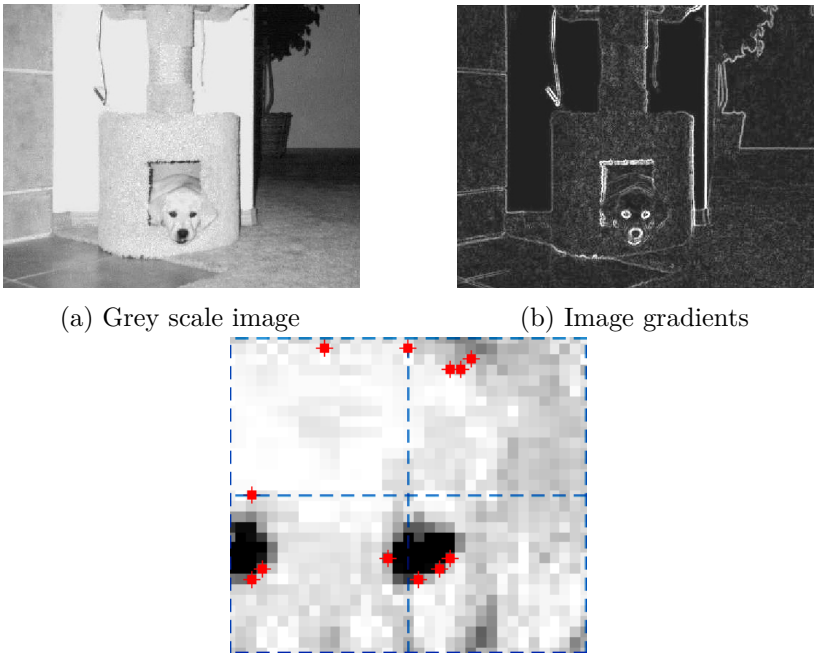
6.1 Introduction

The main contribution of this paper was the modification of a previously defined method for local description of visual media based on the Gestalt Laws. We employed established distance measures such as the Jaccard coefficient in the feature extraction process, aiming at making the descriptions more expressive and robust against image transformations and noise. The result is a competitive and compact local visual descriptor that can be used in combination with various machine learning methods, including deep networks.

Deep learning is the predominant method in visual information retrieval today. Though frequently applied on the pixel level, there are good reasons to combine deep networks with signal processing-based feature extraction methods in order to create a powerful visual media analysis scheme. For once, there appears to be sufficient evidence that a similar approach is also taken in the human brain [Kan13]. Then, decades of fruitful scientific research have yielded a multitude of sophisticated visual description methods. Eventually, in particular the local point-based description methods are able to provide strong descriptions of visual cues that are in-line with the findings about the processing of information in the visual cortex.

The remainder of the paper is structured as follows. Section 2 introduces Gestalt-based Interest Points and motivates our approach for refinement, which is described in technical

[1] Published in: International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2017. [HE17]



(a) Grey scale image (b) Image gradients

(c) Three GIP per block

Figure 6.1: GIP algorithm overview.

detail in Section 3. Experiments and results are given in Section 4 with the conclusions and an outlook on future work in the last Section.

6.2 Background and Motivation

Below, we briefly explain the theory behind the Gestalt Interest Points (GIP) algorithm.

6.2.1 Gestalt Interest Points

The GIP algorithm was introduced in [HE13] and is based on the Gestalt Laws of closure and continuity, i.e. the idea that, unlike in other local image description methods, certain weaker candidates are – in addition to the local extrema – also useful as interest points. The algorithm works as depicted in Figure 6.1. After the input image is converted to gray scale (Figure 6.1a) the image gradient vectors are calculated (Figure 6.1b). The gradient image is split into m by n (e.g. 16x16) macro blocks. For each block, the three largest gradient magnitudes are identified. The pixel positions which correspond to these magnitudes are the so-called GIP (Figure 6.1c).

After detecting these points, feature vectors are computed to describe the image. Each feature vector describes one image block and is defined by:

$$F = (m_1, m_2, m_3, p_1, p_2, p_3, o_1, o_2, o_3) \tag{6.1}$$

where m_1, m_2, m_3 are the three gradient magnitude values, p_1, p_2, p_3 are the three absolute positions and o_1, o_2, o_3 are the three orientations of the interest points, which were chosen within one macro block. Experiments have shown that this simple recipe results in very compact descriptions that satisfy the major Gestalt Laws [HE13]. Note that, in the original version of the algorithm, the absolute pixel positions were chosen as a feature. This can have a negative impact on the classification accuracy in certain cases and will be discussed extensively later in this work.

Two major algorithmic improvements were made in [HE14]. According to these improvements, interest points in low-contrast macro blocks (below threshold t) and interest points on diagonal edges (not within inclination angle α) were discarded. It could be shown that they are not yet sufficiently discriminative for the recognition process.

6.3 Proposed Approach

This Section describes general properties of image features and after that we propose a new feature as part of the GIP descriptor.

6.3.1 Features

The goal of a feature descriptor is to provide a unique and robust description of an image feature, e.g., by describing the intensity distribution of the pixels within the neighborhood of the point of interest. Most descriptors are thus computed in a local manner; hence a description is obtained for every point of interest identified previously. The dimensionality of the descriptor has direct impact on both its computational complexity and matching accuracy. A short descriptor may be more robust against appearance variations like different scales, translations or rotations, but may not offer sufficient discrimination and thus give too many false positives.

As already mentioned, the GIP descriptor contains the three absolute positions of the three interest points detected within one image block. The problem with the absolute positions is that they are not sufficiently robust against image transformation, e.g. scaling. To improve the GIP descriptor, we propose a new feature, the so-called *Inter-GIP Distances (IGD)*. They are intended to replace the interest point's absolute positions in the GIP descriptor.

6.3.2 Inter-GIP Distances (IGD)

As described previously, the GIP algorithm detects three interest points inside each image block. These three points could also be interpreted as the corner points of a triangle. The distances between these points could therefore be seen as the triangle's side lengths

and can serve as features. Figure 6.2 visualizes this concept and the result is one triangle within every image block.

The term *distance* can have different meanings. For instance, the so-called Cityblock distance (see equation 6.3) between two points is calculated as the distance in x plus the distance in y, which is similar to the way we move in a city. The Euclidean distance (6.5) is instead calculated as the length of the line segment connecting two points. The Chebychev distance (6.2) is also known as chessboard distance, since in the game of chess the minimum number of moves needed by a king to go from one square on a chessboard to another equals the Chebychev distance between the squares. The Minkowski distance (6.6) can be considered as a generalization of three other distances, the Euclidean if $p = 2$, the Cityblock if $p = 1$ and the Chebychev distance if $p = \infty$. For our experiments we defined $p = 3$. The Jaccard distance (6.7) measures dissimilarity between sample sets.

Since a variety of different distance functions do exist, the question arose, which one would be the best for our purpose. We also wanted to measure, how the choice of a certain distance function affects the classification accuracy/speed. Therefore, one goal of this work was to find the most suitable distance measure to compute the IGD. We decided to test our algorithm with several known distance functions, which are listed for the two-dimensional case in equations (6.2)-(6.7).

$$D_{Chebychev} = max(|x_2 - x_1|, |y_2 - y_1|) \tag{6.2}$$

$$D_{Cityblock} = |x_2 - x_1| + |y_2 - y_1| \tag{6.3}$$

$$D_{Cosine} = 1 - \frac{PQ'}{\sqrt{(PP')(QQ')}} \tag{6.4}$$

$$D_{Euclidean} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{6.5}$$

$$D_{Minkowski} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{6.6}$$

$$D_{Jaccard} = 1 - \frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)} \tag{6.7}$$

where $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ representing two points in the two-dimensional space. Please note that (6.4) is actually a similarity measure, hence inverse to the others. That, however, has no effect on the discriminative value of the descriptor. In future work we will analyze the performance of our algorithm with other metrics. Further information about distance functions can be found in [Eid12b]. We present our experimental results in the next Section.

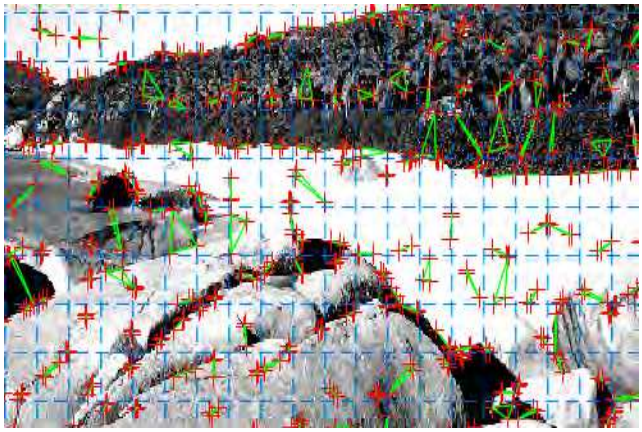


Figure 6.2: Inter-GIP Distances

6.4 Experimental Results

In this Section, the datasets on which we applied our algorithm and an extensive evaluation are presented and discussed. One goal of our experiments was to find the IGD distance measure which maximizes the categorization accuracy while keeping the computational complexity as low as possible. Moreover, we wanted to find out how robust our modified GIP algorithm is against image scaling.

We compare our method to several different state-of-the-art algorithms, namely CNN [LBBH98], SIFT [Low04], SURF [BETVG08], BRISK [LCS11] and FREAK [Ort12]. Additionally, we compare GIP-IGD to the original GIP method (GIP-ABS) [HE13][HE14], which produces feature vectors containing absolute pixel positions. Recently, the Convolutional Neural Network (CNN) offers a very accurate state-of-the-art technique for many general image classification problems. The SIFT and SURF descriptors are both vectors containing floating point values. More recent binary descriptor methods like BRISK and FREAK are less computationally expensive but on the other hand their accuracy is lower.

After quantizing the extracted descriptors with the popular BoVW-algorithm [CDF+04] we fed the resulting histograms into Matlab's Classification Learner App. The app compares several different Classifiers, e.g., different variations of Trees, Support Vector Machines (SVM), Nearest Neighbor Classifiers, Ensemble Classifiers and so forth. It turns out that the Medium Gaussian SVM is best suited for our categorization problems. The F1 measure is used for evaluation. It is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. The performance metric was evaluated with 5-fold cross validation.

6.4.1 INRIA Horses Categorization

In our first evaluation task, we tested the detection performance with the INRIA Horses dataset [FJS10], consisting of 170 images containing horses, and 170 without horses. The

goal of the evaluation task was to categorize the images into images containing horses and images without horses. Because the horses appear at different scales, and against cluttered backgrounds, the categorization is considered to be a sophisticated problem.

6.4.2 Food-5K Categorization

The second dataset which we used to test our algorithm is the Food-5K dataset [SYE16]. It consists of 2,500 food images, which cover a wide variety of food items and 2,500 randomly selected non-food images. Some food images also contain other objects or people. The Food-5K dataset is with a total size of 5,000 images significantly bigger than the INRIA horses dataset. Figure 6.3 shows some examples of food and non-food images in Food-5K.

Food image classification plays a very important role for dietary assessment, which became a health issue of great interest in recent years. Many food items look similar and even for human beings they are sometimes hard to distinguish. Therefore, high-accuracy food classification is a hard problem to solve. The goal of this evaluation task is to categorize the images into food and non-food images.

6.4.3 Discussion

The experimental results of applying our algorithm on the INRIA Horses dataset are shown in the left column and the results for the Food-5K dataset are presented in the right column of Figure 6.4. Figures 6.4a and 6.4b depict the F1-scores over extraction time of our IGD experiments with different distance measures. Adjusting the values of the two GIP parameters t and α causes the algorithm to extract more or less image feature vectors. Therefore, these parameters indirectly affect the extraction time per image and the categorization accuracy because they determine the number of extracted feature vectors. With higher t and lower α, the number of extracted feature vectors per image decreases. It is assumed that the remaining descriptors carry a considerable amount of information and descriptors with less information are omitted. A small set of descriptors for the categorization task reduces the computational complexity significantly.

As already mentioned, the binary descriptor methods BRISK and FREAK are very fast and therefore strong competitors when it comes to computational complexity. Although they are fast, their F1-scores are relatively low. The F1-scores of SURF and SIFT are higher but not as high as the F1-scores of GIP. SIFT is with about 400 ms extraction time per image, comparatively slow. The clear winner in terms of F1-score is the CNN but the serious drawback is the high computational complexity. The CNN needs more than a second to extract the features from one image and therefore it is by far the slowest method. A CNN can achieve extremely high accuracies, but this advantage does not come without a price. CNNs in general are computationally expensive and relatively slow, even with graphical processing units. Additionally, a huge set of training data is needed, which can be difficult to provide and the training process itself can be very time consuming.

Figure 6.3: Example images of Food-5K dataset. The top row shows food images and the bottom row non-food images.

Figures 6.4c and 6.4d show the F1-scores over scaled versions of the test images. As mentioned earlier, GIP-ABS does not work well when it comes to categorizing scaled images. In contrast, GIP-IGD is to a certain degree scale-invariant. Especially, GIP-IGD in combination with the Minkowski distance measure delivers outstanding results in the case of horse categorization, and in the case of food categorization GIP-IGD delivers good results in general, no matter which distance measure is used. For our application domains GIP-IGD is more robust against scaling than SURF, SIFT, BRISK and FREAK. The CNN has the highest accuracy but as mentioned above, it is very slow.

Figures 6.4e and 6.4f depict the average number of feature values extracted from one image. The description vectors of SURF, BRISK and FREAK are 64-dimensional and the SIFT vector has 128 elements. Hence, they are more memory-consuming than the 9-dimensional GIP-IGD feature vectors. For example, in case of using FREAK for horse categorization a total number of $918 * 64 = 58,752$ feature values per image are necessary to get a comparatively poor F1-score of 69%. In other words, Figure 6.4 demonstrates that GIP outperforms SIFT's, SURF's, BRISK's and FREAK's accuracy while reducing the descriptor values per image to only a few percent.

6.5 Conclusion

In our experiments we demonstrated that it is possible to classify images fast and with high accuracy using only a few and very compact GIP image descriptors. The GIP feature vector is much more compact than all the feature vectors of the evaluated baseline competitor methods. Furthermore, viewer descriptor values have a positive impact on classifier training time and storage requirements. For example, this is an important advantage for low-power devices, mobile devices and in the big data domain. We also demonstrated that the replacement of absolute pixel positions through IGD feature makes the GIP feature vector more robust against image scaling. In future work we will investigate, how efficient and accurate GIP is in combination with Deep Learning methods.

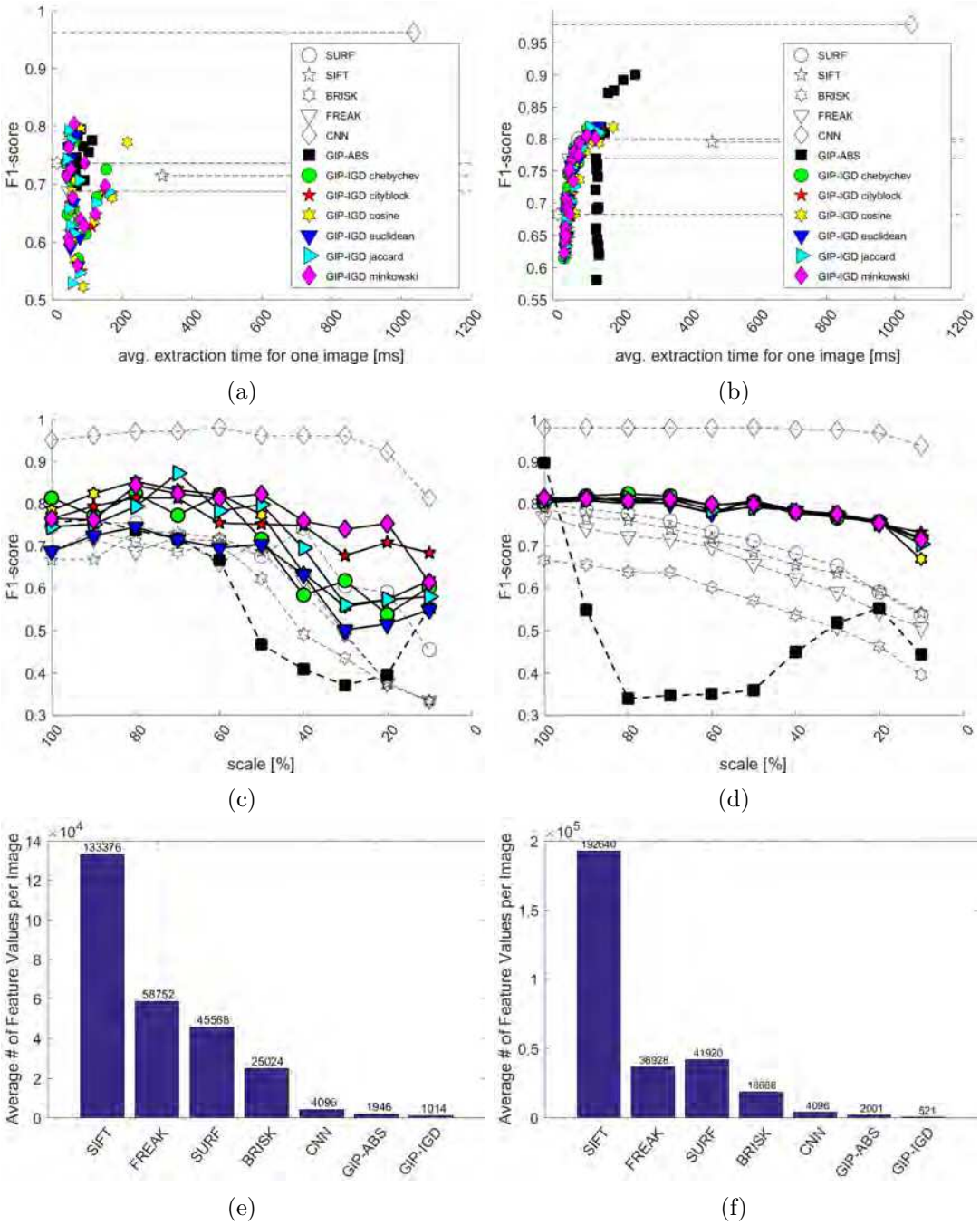


Figure 6.4: The experimental results of applying our algorithm on the INRIA Horses dataset are shown in the left column and the results for the Food-5K dataset are presented in the right column. Our algorithm is also compared to several different baseline methods. The different F1-scores for each IGD distance measures in figures 6.4a and 6.4b arise through adjusting the two GIP parameters t and α, which are described in Section 6.2.1.

CHAPTER 7

Gestalt Interest Points with a Neural Network for Makeup-Robust Face Recognition[1]

7.1 Introduction

In this work we presented a prototype-based method which is faster than CNNs while delivering a very high categorization accuracy for makeup-robust face recognition. Face recognition has been an active topic of scientific research for decades now. The rapid evolution of face recognition systems into real-time applications has raised new concerns about their ability to resist presentation attacks, particularly in unattended application scenarios such as automated border control.

Dantcheva et al. [DCR12] claimed in their study that the application of facial cosmetics significantly decreases the performance of both academic face verification approaches and commercial approaches. As shown in Figure 7.2, significant appearance changes can be observed for individuals with and without makeup. Obviously, the faces with makeup have smoother skin, longer eyelashes, etc. Thus, there might be a large gap between non-makeup and makeup domains. However, if we look further through higher semantic representation levels, the gap becomes smaller. The intuition behind this is that some visual traits remain unchanged regardless of makeup. Eventually, as the representation goes up to semantic levels, the two images are both described as faces, and hence the gap diminishes. In our proposed make-up robust face recognition method, we utilize certain visual traits on higher semantic representation levels.

[1] Published in: 25th International Conference on Image Processing (ICIP). IEEE, 2018. [HE18]

Convolutional neural networks (CNNs) have become very popular in recent years, due to their near-perfect recognition accuracy on unconstrained datasets. CNNs have been shown to be extremely accurate in face recognition (FR) tasks [PVZ15a], [SLWT15a]. Most high-accuracy FR systems today rely on deep-learning methods and they are already being deployed in commercial face-verification applications [SKP15]. One problem of CNNs is their high computational complexity.

The main contribution of this paper is a fast and accurate face recognition method that is inspired by cognitive science and robust to cosmetic changes. Additionally, the dataset for the experiments reported in the paper will be made publicly available and can be obtained by sending an e-mail request. It turns out that on the given domain the proposed approach outperforms state-of-the-art description methods such as SIFT, SURF and others. It dominates them both in terms of recognition accuracy and in description compactness.

The remainder of this paper is organized as follows. Related research work is reviewed in Section 7.2. Section 7.3 explains the proposed approach in detail. In Section 7.4, the dataset and experimental results are presented.

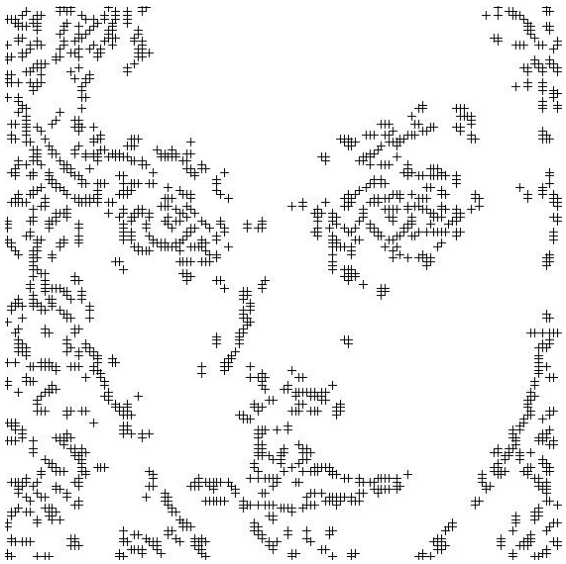


Figure 7.1: Example output of the GIP algorithm. The GIP algorithm is fast and highly effective. Because it is inspired by cognition it extracts very little, but well-selected image information.

7.2 Related Work

To our knowledge, there is limited scientific literature on addressing the challenge of makeup-robust face recognition. Chen et al. [CDR15] addressed this problem with a patch-based ensemble learning method. Song et al. [LSW+18] synthesize a non-makeup

image from a face image with makeup via a generative network. After that, deep features are extracted from the synthesized image to further accomplish the makeup-robust face recognition. Zheng et al. [ZK17] proposed a hierarchical feature learning framework for face recognition under makeup changes. Their method seeks transformations of multilevel features because these features tend to be more invariant on higher semantic levels, and less invariant on the lower levels.

7.3 Proposed Approach

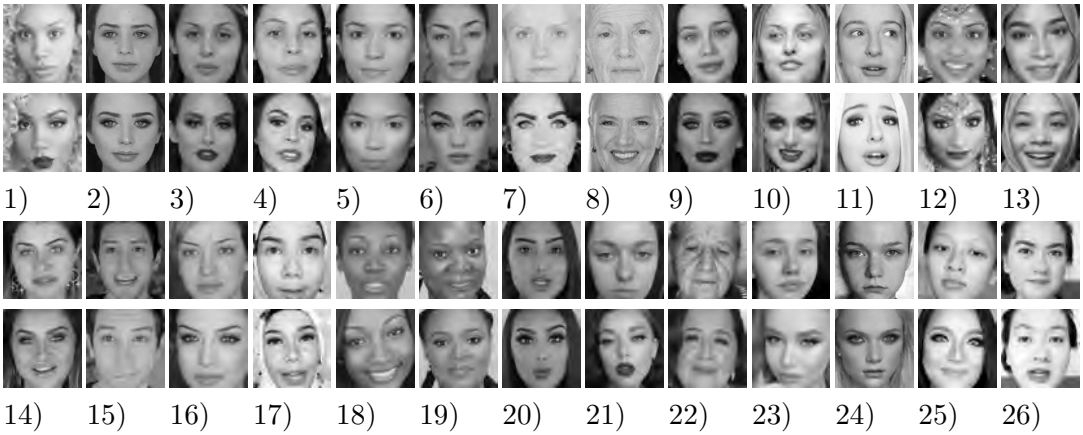


Figure 7.2: Some example images of the 26 subjects contained in our self-compiled dataset. The images are collected from YouTube makeup tutorials. The top row shows images of people without makeup and the bottom row shows images of the same individuals with makeup. Note the variations in pose, illumination and expression and the significant dissimilarities of the same identities.

Cognitive computing methods often make use of a variety of cognitive concepts. Our proposed method is, on the one hand, inspired by visual perception and by biological neural networks, on the other hand. The psychological theories behind our proposed method are described below.

David Marr [Mar82] described visual perception as a multistage process. In the first stage a 2D sketch of the retina image is generated, based on feature extraction of fundamental components of the scene, including edges, regions and so forth. The second stage extracts depth information by detecting textures. Finally, a 3D model is generated out of the previously gathered information.

Hermann von Helmholtz examined in his work [Hel25] about visual perception that the information gathered via the human eye is a very simplified version of the real world. He therefore concluded that most of the visual perception processes take place in the brain. In his theory vision could only be the result of making assumptions and conclusions from incomplete data, based on previous experience.

Gestalt psychology [Kof35] is an attempt to understand the laws behind the ability to acquire and maintain meaningful perceptions in an apparently chaotic world. According to this theory, there are eight so-called Gestalt Laws that determine how the visual system automatically groups elements into patterns: Proximity, Similarity, Closure, Symmetry, Common Fate, Continuity as well as Good Gestalt and Past Experience.

The psychological theories mentioned above build the foundation of the Gestalt Interest Points algorithm (GIP) [HE17], which serves as the artificial visual perception building block of our proposed method. Firstly, as inspired by David Marr the GIP algorithm extracts certain edge and texture information. Secondly, inspired by the way Helmholtz described visual perception, the information gathered by the GIP algorithm greatly simplifies the input image. Therefore, the algorithm is fast and highly effective because it extracts very little but well-selected image information. Thirdly, the GIP algorithm is based on the Gestalt Laws of Closure and Continuity, i.e. the idea that, unlike in other local image description methods, certain weaker candidates are – in addition to the local extrema – also useful as interest points. The GIP algorithm works as follows. After the input image is converted to gray scale the image gradient vectors are calculated. The gradient image is split into m by n (e.g. 16x16) macro blocks. For each block, the three largest gradient magnitudes are identified. The pixel positions which correspond to these magnitudes are the so-called GIP. Interest points in low-contrast macro blocks (below threshold t) and interest points on diagonal edges (not within inclination angle α) are discarded. It could be shown that they are not yet sufficiently discriminative for the recognition process. After interest point detection, feature vectors are computed to describe the image. Each feature vector describes one image block and is defined by:

$$\vec{F} = (m_1, m_2, m_3, p_1, p_2, p_3, o_1, o_2, o_3),$$

where m_1, m_2, m_3 are the three gradient magnitude values, p_1, p_2, p_3 are the three absolute positions and o_1, o_2, o_3 are the three orientations of the interest points, which were chosen within one macro block. Experiments have shown that this simple recipe results in very compact descriptions that satisfy the major Gestalt Laws. In [HE17] the algorithm and its continued development are explained in detail. Figure 7.1 depicts an example output of the GIP algorithm.

The second building block of our proposed method is an artificial neural network (ANN). ANNs are computing systems inspired by the biological neural networks that constitute the brains of humans and animals. Such systems learn (progressively improve performance on) tasks by considering examples. The idea behind ANNs is not new, but it has been popularized more recently because we now have lots of data and GPU-based processors that can achieve successful results on hard problems. There are many kinds of ANNs, but in general they consist of systems of nodes with weighted interconnections among them. Typically, neural networks learn by updating the weights of their interconnections. Nodes are arranged in multiple layers, including an input layer where the data is fed into the system; an output layer where the answer is given; and one or more hidden layers, for the learning of example patterns. Our applied ANN is a feed-forward network with a

tangent sigmoid transfer function:

$$tansig(n) = \frac{2}{(1 + e^{-2*n}) - 1} \tag{7.1}$$

in the hidden layer, and a softmax transfer function

$$softmax(n) = \frac{e^n}{\sum(e^n)}, \tag{7.2}$$

in the output layer. For training the network its weight and bias values are updated according to the scaled conjugate gradient backpropagation method [Mol93].

Our proposed approach is a combination of GIP feature extraction with ANN classification (GIP-NN). GIP and ANNs are both inspired by cognition. Therefore, the logical consequence for us was to combine both concepts into a powerful recognition system. The detected Gestalt Interest Points serve as input for the ANN. One advantage of our approach is that we do not need color information, which is often not available, e.g. frames of surveillance cameras. Actually, it is very likely that a color-based recognition approach would perform worse, because makeup changes the skin color and therefore the recognition process may leads to false positives.

7.4 Evaluation

In this Section, the datasets on which we applied our algorithm and an extensive evaluation are presented and discussed.

7.4.1 Dataset

In the works of Chen et al. (e.g. [CDSR17]) they made certain datasets publicly available. However, these datasets are not appropriate for our application domain because they consist of only a few images per subject. Since one component of our recognition system is a neural network, we need a large set of face images to train it. Therefore, we decided to compile a dataset by ourselves, consisting of 26 subjects from YouTube makeup tutorials. Figure 7.2 shows some example images. In total 23,145 video frames of the subjects before and after the application of makeup were captured. We used Matlab's cascade face detection strategy to crop out faces from the frames. The makeup in the resulting face images varies from subtle to heavy. The cosmetic alteration affects the quality of the skin due to the application of foundation and change in lip color and the accentuation of the eyes by diverse eye makeup products. This dataset includes some variations in expression and pose. The illumination condition is reasonably constant over multiple shots of the same subject. In a few cases, the hair style before and after makeup changes drastically.

7.4.2 Baseline algorithms

We compare our method to several different state-of-the-art hand-crafted feature extraction algorithms, namely SIFT [Low04], SURF [BETVG08], BRISK [LCS11] and FREAK [Ort12]. After quantizing the extracted descriptors with the popular BoVW-algorithm [CDF+04] we categorize the resulting histograms with a neural network. Additionally, we compare our method to a CNN [LBBH98]. One drawback of CNNs is that they usually require a large amount of training data in order to avoid over-fitting. Since our training data is relatively little to train a CNN, we used the pre-trained and well established AlexNet [DDS+09]. In a second stage we trained a multi-class linear SVM with CNN features extracted from our own training data. This is a common technique when it comes to applying CNNs to problems with small training sets and furthermore, it saves a significant amount of training time.

7.4.3 Experimental results

The following experiment was designed for exploring the effectiveness of the GIP-NN method in matching after-makeup against before-makeup face samples and for comparing our approach to the different baseline methods. Note that there is no overlap between training images and test images of the subjects and therefore this experiment is a very sophisticated recognition task. For the training stage 19,635 non-makeup face images of 26 subjects serve as input. Henceforth, the classification stage assigns each of the 3,510 makeup test images to one of the 26 subjects. Experiments were conducted using Matlab R2016b on a 64-bit Windows operating system with Intel Core i7-3632QM 2.20 GHz CPU and 8 GB RAM.

The number of hidden layers of a neural network has great impact on its performance. To find the optimal number of hidden layers we applied our algorithm on a small subset of our dataset with different numbers of hidden layers. The result is depicted in Figure 7.5. We identified that 200 hidden layers maximize the mean accuracy for our application domain.

In Figure 7.3 some example ROC curves are shown. As expected, the CNN-based baseline method is the strongest competitor. For subject 14 the CNN-based baseline method is the most accurate but for subject 25 our method clearly outperforms all the baseline methods. The subjects 3, 9, 21 and 24 are a big challenge for all methods. Figure 7.2 shows that even for human beings it is difficult to identify these people because the make-up changes their faces drastically.

Figure 7.4 compares the mean accuracies over feature extraction time of the different methods. GIP-NN is clearly more accurate than all the hand-crafted feature extraction algorithms, yet with 59.5 percent less accurate than the CNN-based method with 68 percent. But this advantage of the CNN-based method does not come without a price because it is significantly slower than the proposed approach. Our method extracts the features of one image in 350 ms on average. In contrast the CNN-based baseline method needs on average 980 ms for feature extraction. Furthermore, with our GIP-NN

approach the whole makeup-robust face recognition experiment took only about two hours to complete. In contrast the same experiment lasted more than 5 hours with the CNN-based recognition method and the same hardware. Another advantage of the GIP algorithm is that it describes images more compactly than all the other feature extraction baseline methods [HE17]. That is, we need less disk space and processing power. This is very beneficial for a big data domain like face recognition.

7.5 Conclusion

In this work we introduce a novel approach for makeup-robust face recognition based on the GIP algorithm and an artificial neural network. The approach is, on the one hand, inspired by visual perception and by biological neural networks, on the other hand. We evaluated our method empirically with a self-compiled dataset composed by YouTube makeup tutorials of 26 subjects. Our experiments showed that GIP-NN is very accurate and almost three times faster than the CNN-based baseline method. Especially for surveillance fast and accurate face recognition is essential. We demonstrated that our method is highly effective for the domain of makeup-robust face recognition.

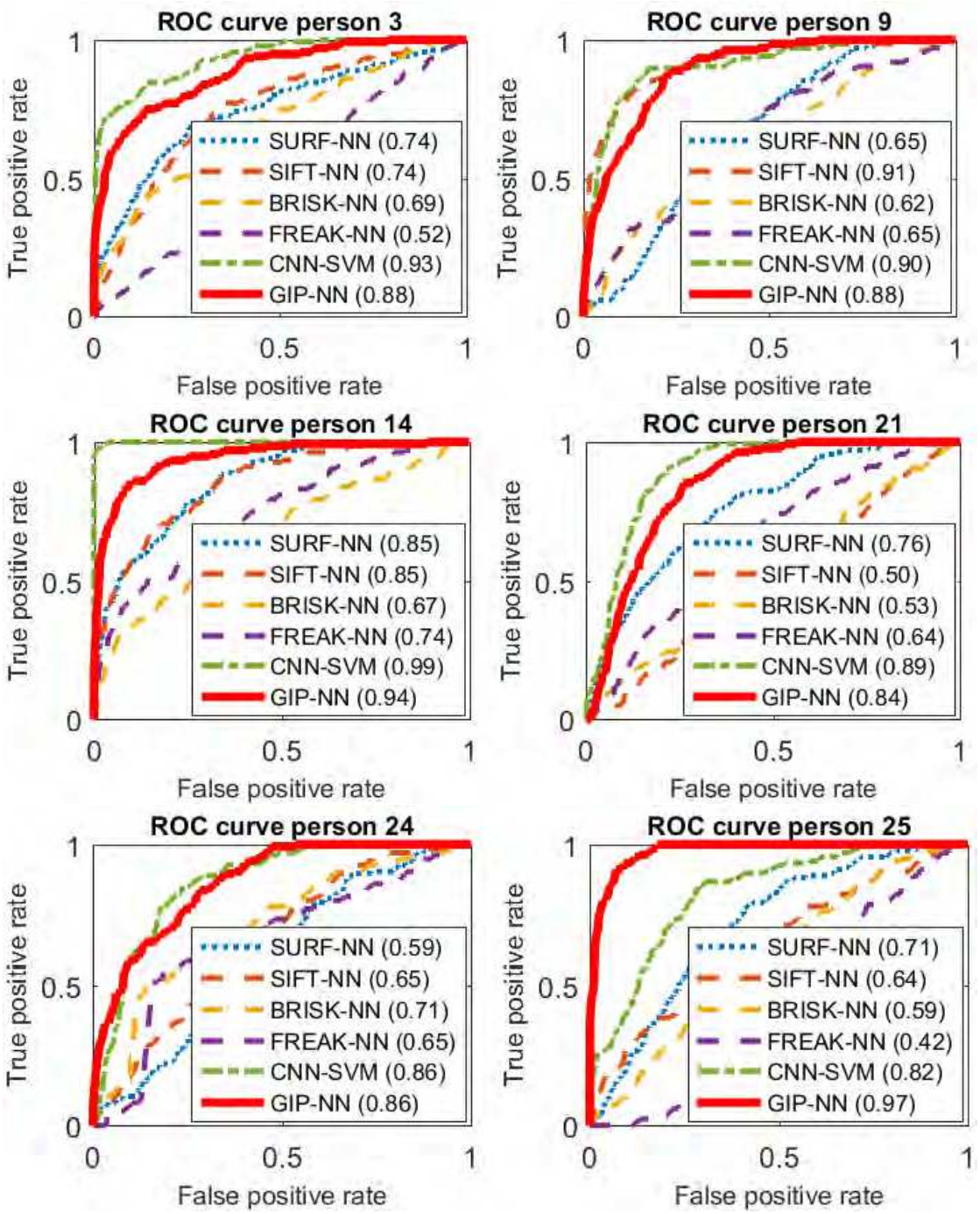


Figure 7.3: ROC curves of our experiments for 6 of the 26 subjects. The numbers in parentheses in each legend indicate the Area Under the Curve (AUC) values.

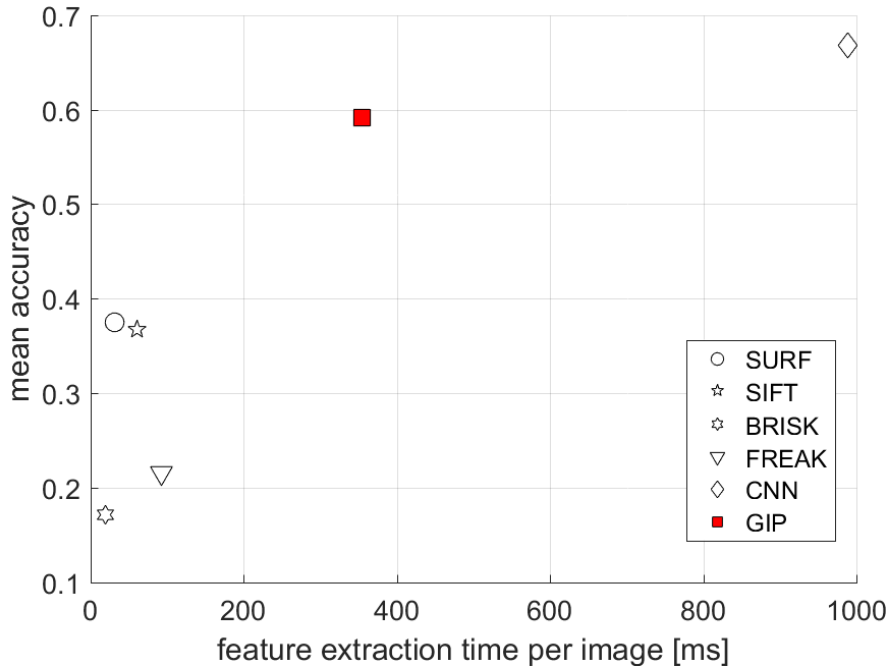


Figure 7.4: The mean accuracies over feature extraction time of the different methods.

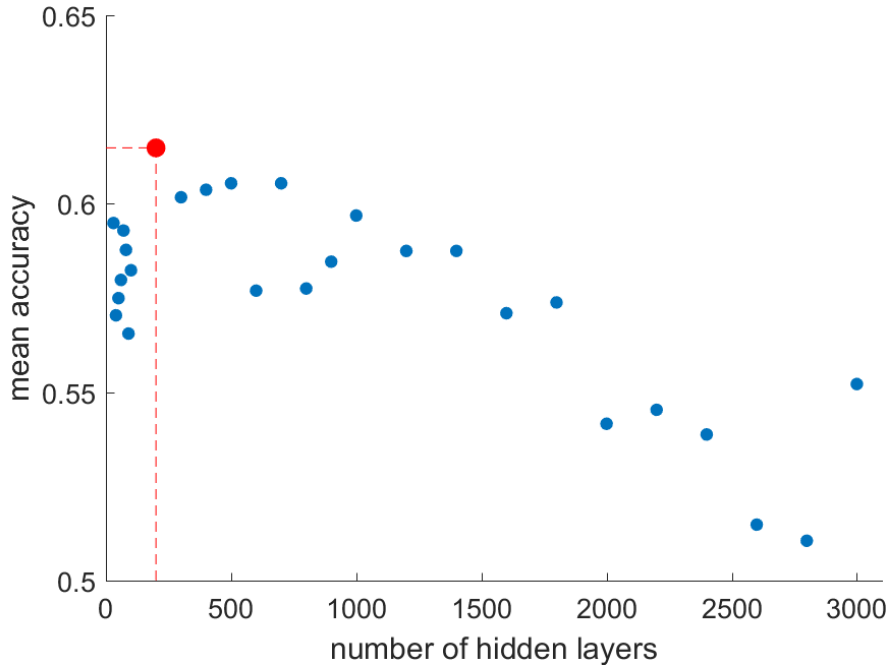


Figure 7.5: The mean accuracies over different number of neural network's hidden layers. For our application domain a neural network with 200 hidden layers delivers the highest accuracy.

CHAPTER 8

Gestalt Descriptions for Deep Image Understanding[1]

8.1 Introduction

Deep learning is a predominant method in visual information retrieval today. Though typically applied on the pixel level, there are good reasons to combine deep learning methods with signal processing-based feature extraction methods in order to create a powerful visual media analysis scheme. For once, there appears to be sufficient evidence that a similar approach is also taken in the human brain [Kan13]. Then, decades of fruitful scientific research have yielded a multitude of sophisticated visual description methods. Eventually, the local point-based description methods in particular are able to provide strong descriptions of visual cues that are in-line with the findings about the processing of information in the visual cortex.

In this work, we present a novel local description approach inspired by the Gestalt Laws as a pre-processing step for deep learning. To the best of our knowledge there are no other scientific works about utilizing Gestalt Laws to pre-process images for deep learning until now. The experiments in Section 8.3.2 and 8.3.3 were made to test the fundamental idea, different parametrization and some variations of our method. We concluded that it outperforms all of the baseline local description methods to which we compared it. However, the experiments of Section 8.3.3 revealed that a general-purpose CNN is often more accurate, despite much slower, than our approach. Based on our findings, we decided to fuse our method with the CNN approach to build an even more powerful image recognition system. It turns out that feeding the output of our method into a CNN makes the image recognition process more accurate and robust against over-fitting

[1] Under review for: Pattern Analysis and Applications Journal. Springer Press.

for our application domain of make-up-robust face recognition. This is due to the heavily compressed and content-rich image description produced by our approach.

In machine learning, a CNN is a class of Deep Neural Networks (DNNs), most commonly applied to describing visual imagery. CNNs are computing systems inspired by the biological neural networks that constitute the brains of humans and animals. Such systems learn tasks by considering examples utilizing a sophisticated learning algorithm. Typically, CNNs learn by updating the weights of their interconnections. CNNs are arranged in multiple layers, including an input layer where the data is fed into the system; an output layer where the answer is given; and several hidden layers, for the learning of example patterns. Although CNNs trained with back-propagation had been around for decades, and GPU implementations of Neural Networks for years, including CNNs, fast implementations of CNNs with max-pooling on GPUs in the style of Ciresan and colleagues helped to make progress on computer vision. For the first time, in 2011 this approach achieved superhuman performance in a visual pattern recognition contest [CMM+11]. A few years later the AlphaGo system [SHM+16] was very important to generate wide public awareness of DNNs and thus also for CNNs.

As already mentioned, one part of this work demonstrates the effectiveness of our method as a pre-processing step for a CNN. However, a CNN is only one type of Deep Network, and our method could also be combined with other types, e.g. Deep Residual Networks (ResNets) proposed by Kaiming et al. [HZRS16]. One could assume that building more accurate deep learning models could be performed by simply stacking more and more layers. Kaiming et al. demonstrated the depth problem, i.e. to some point, accuracy would improve, but beyond about 25+ layers, accuracy tends to drop. As a solution for this problem, Kaiming et al. presented the ResNets which have since allowed the training of over 2000 layers with increasing accuracy. A ResNet builds on constructs known from pyramidal cells in the cerebral cortex. ResNets do this by utilizing skip connections or short-cuts to jump over some layers. The motivation for skipping over layers is to avoid the problem of vanishing gradients [GB10], by reusing information as residuals from a previous layer until the layer next to the current one has learned its weights.

8.1.1 Theories of Visual Perception

Cognitive computing methods often make use of a variety of cognitive concepts. Our proposed method is inspired by visual perception. The psychological theories behind our proposed method are described below.

David Marr [Mar82] described visual perception as a multistage process. In the first stage a 2D sketch of the retina image is generated, based on feature extraction of fundamental components of the scene, including edges, regions and so forth. The second stage extracts depth information by detecting textures. Finally, a 3D model is generated out of the previously gathered information.

Hermann von Helmholtz examined in his work [Hel25] about visual perception that the information gathered via the human eye is a very simplified version of the real world.

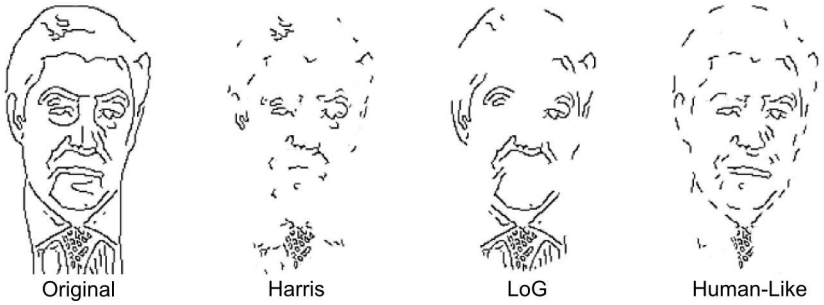


Figure 8.1: The left edge map of a face is represented by points from a Harris corner detector and a Laplacian of Gaussian (LoG) operator. Many of the important face features (e.g. the eyes) are thereby lost. The rightmost image shows the GIP description. It preserves the perceptual features of the original stimulus well but does not produce a longer description than the LoG operator [Eid11].

He therefore concluded that most of the visual perception processes take place in the brain. In his theory, vision could only be the result of making assumptions and drawing conclusions from incomplete data, based on previous experience.

Gestalt psychology [Kof35] is an attempt to understand the laws behind the ability to acquire and maintain meaningful perceptions in an apparently chaotic world. According to this theory, there are eight so-called Gestalt Laws that determine how the visual system automatically groups elements into patterns: Proximity, Similarity, Closure, Symmetry, Common Fate, Continuity as well as Good Gestalt and Past Experience. Especially the Gestalt Law of Closure was of great interest for our work. The Gestalt Law of closure states that the perception of individuals fills in visual gaps in incomplete shapes. For example, humans are able to recognize a whole circle, even if there are gaps in its contour. For our approach this means that due to the Gestalt Law of closure it is still possible to recognize what an image depicts, only by considering its local representation. This effect is shown in Figure 8.1. Obviously, such interest point sets are more useful for media understanding than points from which humans cannot identify the semantic content of an image. If the user cannot reconstruct the object from the interest points, how should the machine?

The remainder of the paper is arranged as follows: we first discuss the related work and contributions of this paper in Sections 8.1.2 and 8.1.3, and then we provide a comprehensive overview of the Gestalt Interest Points (GIPs) algorithm in Section 8.2. The details of our Gestalt Regions of Interest (GROI) method are presented in Section 8.2.2. Experimental results are analyzed in Section 8.3, and conclusions are finally given in Section 8.4.

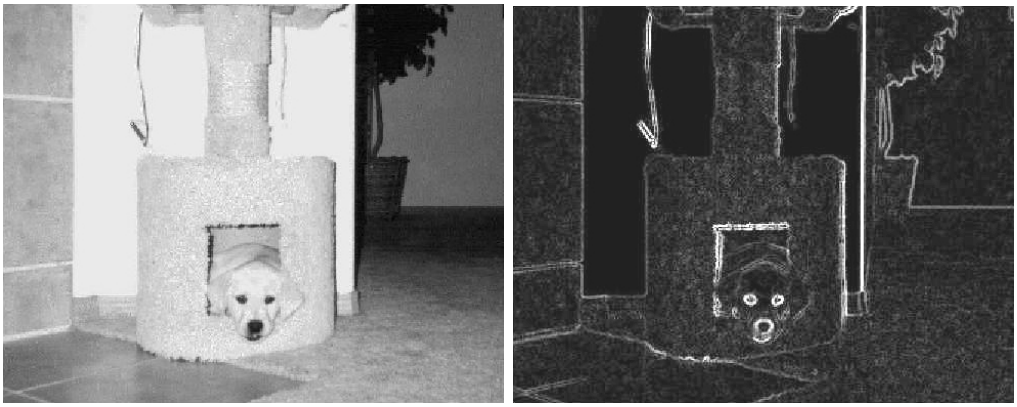
8.1.2 Related Work

A fundamental aspect of our work is the deployment of Gestalt Laws to describe images in a meaningful and efficient way. The basic Gestalt rules were first proposed by Wertheimer et al. [Wer23] for specifying the perceptual relationship between the human vision system and the perceived visual world. Some important problems in computer vision are modeled by utilizing the Gestalt principles [DMM04, DP15]. In [SC15] the authors proposed a novel method for establishing visual correspondences between images based on Gestalt theory. Their method detects visual features from images, with a particular focus on improving the repeatability of the local features in those images containing the same semantic contents. In [BW07] four new image features are presented, inspired by the Gestalt Laws of Continuity, Symmetry, Closure and Repetition. The resulting image representations are used jointly with existing state-of-the-art features to improve the accuracy of object detection systems. The authors of [KYK06] proposed a context-based method for object recognition inspired by the Gestalt Laws of Proximity and Similarity. Qiu et al. [QWCF16] presented a novel lung nodule detection scheme based on the Gestalt visual cognition theory. The proposed scheme involves two parts which simulate human eye cognition features such as simplicity, integrity and classification. In [YRS+18] the authors presented a method for image salient object detection with Gestalt Laws guided optimization.

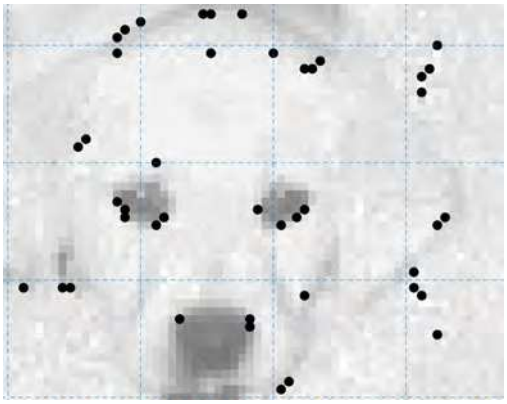
The second research direction that is related to our work is the development of methods which combine deep and handcrafted image features. For instance, in [NPBP18] the authors combined deep and handcrafted image features for Presentation Attack Detection in face recognition systems. Their method uses a CNN to extract deep image features and the multi-level local binary pattern (MLBP) method to extract skin detail features from face images. Qiangliang et al. [GXH18] detect keypoints with a method utilizing the Difference of Gaussian (DOG) operator. Then, they describe the keypoints by the proposed local convolutional features which are inspired by a CNN. In their work they showed results of applying the proposed method on the domain of power transmission line icing monitoring. In [ASCT17] they merged SIFT with CNN features for facial expression recognition. Because local methods like SIFT do not require extensive training data to generate useful features, the authors achieved comparatively high performance on small data.

8.1.3 Contributions

We list the main contributions of this work as follows: (1) We present the combination of the novel Gestalt Region of Interest (GROI) method with a CNN in Section 8.3.4. We applied it on the problem of makeup-robust face recognition and our experimental results show that it outperforms a conventional CNN for the given task. The presented GROI method and the results of the makeup-robust face recognition experiments are completely new and have not yet been made publicly available by us in previous works. (2) We provide a detailed overview of our previously presented [HE13, HE14, HE17, HE18] GIP feature which defines the fundamental basis of the GROIs. It can be used as a feature



(a) Grey scale image (b) Image gradients



(c) Gestalt Interest Points

Figure 8.2: GIP algorithm overview.

in itself without a CNN for image understanding tasks where a long training time is unacceptable and / or a huge amount of training data is unavailable. (3) Additionally, we show our experimental results on various forensic application domains in sections 8.3.2 and 8.3.3, which can be also found in previously published material [HE14, HE17].

8.2 Proposed Approach

In this Section we provide a detailed overview over the Gestalt Interest Points (GIPs) algorithm [HE17]. Below, we illustrate how the GIPs are detected and described by feature vectors. Furthermore, it is shown how to interconnect the GIP method via GROIs with a CNN to exploit the strengths of a highly effective local description method and deep learning.

8.2.1 Gestalt-Interest-Points Detection

The theories of visual perception mentioned in Section 8.1.1 build the foundation of the GIP algorithm. Firstly, as inspired by David Marr the GIP algorithm extracts edge and texture information. Secondly, inspired by the way Helmholtz described visual perception, the information gathered by the GIP algorithm greatly simplifies the input image. Therefore, the algorithm is fast and highly effective because it extracts very little but well-selected image information. Thirdly, the GIP algorithm is based on the Gestalt Laws of Closure and Continuity, i.e. the idea that, unlike in other local image description methods, certain weaker candidates may – in addition to the local extrema – also be useful as interest points.

The algorithm works as depicted in Figure 8.2. After the input image is converted to gray scale (Figure 8.2a) the image gradient vectors are calculated (Figure 8.2b). The gradient image is split into m by n (e.g. 16x16) macro blocks but not every block is interesting for further processing. For human perception edges appear to carry far more of the important image semantics than areas with low contrast. According to this assumption, we assume that low-contrast image macro blocks may sometimes be omitted for the benefit of better edge description elsewhere. For each block, we calculate the variance of gray values. If the variance of a block is below a certain threshold t, then the block is excluded from subsequent processing steps. During our experiments which are presented later in this work, we investigated the influence of t on the recognition accuracy. For each remaining image block, the three points with the largest gradient magnitudes are identified. This point set is called P and a subset of points $Q \subseteq P$ is selected according to the strategy described in the following paragraph.

The similarity grouping experiments of Olson and Attneave [OA70] showed that human beings are significantly faster in grouping horizontal or vertical lines than of diagonals or other patterns. As an explanation for this observation they assumed that significantly larger parts of the receptive field are oriented horizontally and vertically rather than diagonally. This concept inspired us to experiment with discarding interest points that are not on horizontal or vertical edges, as these might be less expressive for the description and recognition process. Figure 8.3 depicts the basic idea of the implementation. The adjustable inclination angle α defines circle segments. We apply the inverse tangent function on the gradient vectors of each image point from P to get the gradient directions. All image points with gradient vectors pointing in a direction within one of the circle segments are added to Q. If one gradient vector does not point in a direction within one of the circle segments, the underlying edge is considered to be diagonal and therefore we suppose that its interest points – so the hypothesis – are of insufficient use for the recognition process. Describing an image with less information should have a positive effect on resource usage and the performance of the recognition process. The remaining image points contained in Q are the so-called GIP (Figure 8.2c).

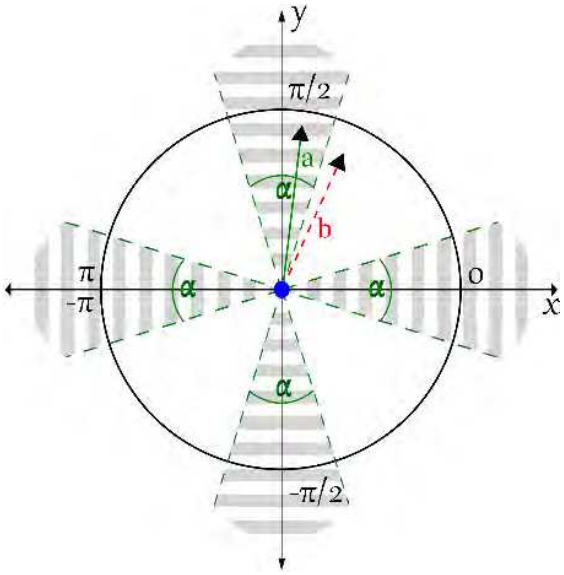


Figure 8.3: The point in the origin indicates a GIP and vector **a** its gradient, which is within one of the four circle segments. In this case, the GIP will be accepted as an interest point. If vector **b** was the gradient of this GIP, the GIP would be discarded because its underlying edge has diagonal orientation.

8.2.2 Gestalt-Interest-Points Description

After detecting the GIPs, feature vectors are computed to describe the image. Each feature vector describes one image block and is defined by:

$$F = \begin{pmatrix} m_1 & m_2 & m_3 & p_1 & p_2 & p_3 & o_1 & o_2 & o_3 \end{pmatrix} \tag{8.1}$$

where m_1, m_2, m_3 are the three gradient magnitude values, p_1, p_2, p_3 are the three absolute positions and o_1, o_2, o_3 are the three orientations of the interest point's gradients, which were chosen within one macro block. Since this is the basic version of the GIP feature vector that employs absolute pixel position values, we denote the GIP algorithm utilizing the feature vector described in this Section as GIP-ABS.

Experiments have shown that this simple recipe results in very compact descriptions that satisfy the major Gestalt Laws. Figure 8.4 depicts an example output of the GIP algorithm. Among the advantages of this straightforward scale-less implementation are the guarantee that the visual object shape is preserved in the description, and that clusters of high-curvature interest points in close proximity are avoided: compared to SIFT, SURF and related methods the local description is more evenly distributed over the entire input signal without ending up in a global description. The GIP-ABS pseudo code is presented in Algorithm 1.

Algorithm 1 The Gestalt Interest Points detection algorithm

function detectGIPs(im, t, α, Q, F)
Input: Input image im, variance threshold t, inclination angle α
Output: Gestalt Interest Point set Q and Gestalt Interest Point descriptors F
  $imgrey \leftarrow convert(im)$
  $[FX, FY] \leftarrow gradients(imgrey)$ ▷ gradient velocity components FX and FY
  $M \leftarrow \sqrt{FX. * FX + FY. * FY}$ ▷ gradient magnitudes
  $imgCube \leftarrow [FX, FY, M]$ ▷ 3 layers, each layer size == size(im)
  $cubes \leftarrow divide(imgCube, 16)$ ▷ divide into [16x16x3] cubes
  for all $C \in cubes$ **do**
    $MAGS \leftarrow C_M$ ▷ gradient magnitudes, [16x16] matrix
    $VX \leftarrow C_{FX}$ ▷ gradient velocity components x-direction, [16x16] matrix
    $VY \leftarrow C_{FY}$ ▷ gradient velocity components y-direction, [16x16] matrix
    if $Var(MAGS) < t$ **then**
      $continue$ ▷ discarding low contrast image blocks
    end if
    $M_{max} \leftarrow find3GreatestMagnitudes(MAGS)$
    $indices \leftarrow find(MAGS == M_{max})$ ▷ find matrix indices of magnitudes
    $ORIENTATIONS \leftarrow abs(atan2(VY[indices], VX[indices]))$
    if $diagonal(ORIENTATIONS, \alpha)$ **then**
      $continue$ ▷ discarding interest points on diagonal edges
    end if
    $absIndices \leftarrow calcAbsoluteImgIndices(indices)$
    $Q.add(absIndices)$
    $F.add([M_{max}, absIndices, ORIENTATIONS])$
  end for
  return Q, F
end function

function diagonal(O, α, $diagonal$)
Input: gradient orientations O, inclination angle α
Output: boolean $diagonal$
  $a \leftarrow 90 - \alpha$
  $b \leftarrow 90 + \alpha$
  $c \leftarrow 180 - \alpha$
  for all $o \in O$ **do**
    if not[$o <= \alpha$ OR ($o > a$ AND $o < b$) OR $o >= c$] **then**
      return $diagonal \leftarrow true$
    end if
  end for
  return $diagonal \leftarrow false$
end function

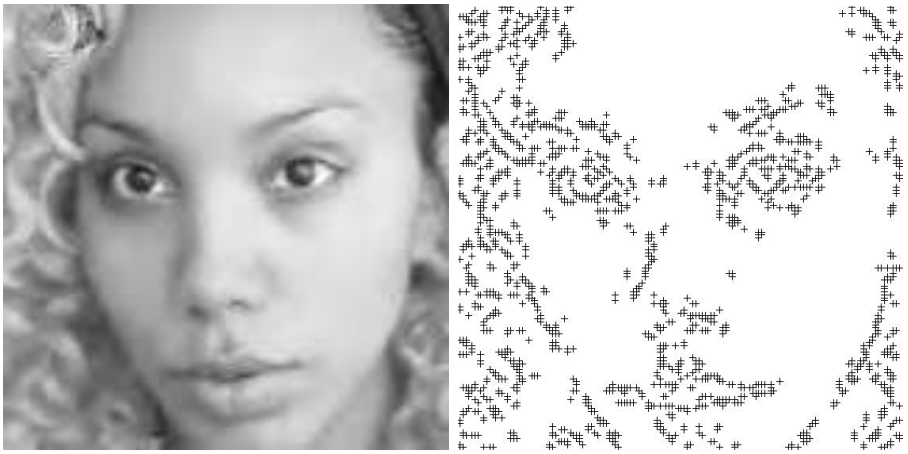


Figure 8.4: A face image on the left and its GIP representation on the right. The GIP algorithm is fast and highly effective. Because it is inspired by cognition, it extracts very little but well-selected image information.

Inter-GIP Distances (IGD)

As described above, the GIP-ABS descriptor contains the three absolute positions of the three interest points detected within one image block. The absolute positions are causing the GIP-ABS descriptor to be neither translation-invariant nor scale-invariant and are therefore not appropriate for some application domains where translation- and scale-invariance are desired. To address this issue, we developed a GIP descriptor which contains the so-called *inter-GIP distances (IGD)*. They are intended to replace the interest points absolute positions in the GIP descriptor when needed. During our experiments which are presented later in this work, we tried both variants of the GIP descriptor and investigated their influence on the recognition process. The idea of GIP-IGD is as follows. As described previously, the GIP algorithm detects three interest points inside each image block of an image. These three points are interpreted as the corner points of a triangle. The distances between these points could therefore be seen as the triangle's side lengths and can serve as features. Figure 8.5 visualizes this concept and the result is one triangle within every image block.

Since a variety of different distance functions do exist, the question arose, which one would be the best for the GIP-IGD operator. We also wanted to measure, how the choice of a certain distance function affects the classification accuracy and speed. Therefore, one goal of this work was to identify the most suitable distance measure to compute the IGD. We decided to test our algorithm with several known distance functions, which are listed for the two-dimensional case in equations (8.2)-(8.7).

$$D_{Chebychev} = max(|x_2 - x_1|, |y_2 - y_1|) \tag{8.2}$$

$$D_{Cityblock} = |x_2 - x_1| + |y_2 - y_1| \tag{8.3}$$

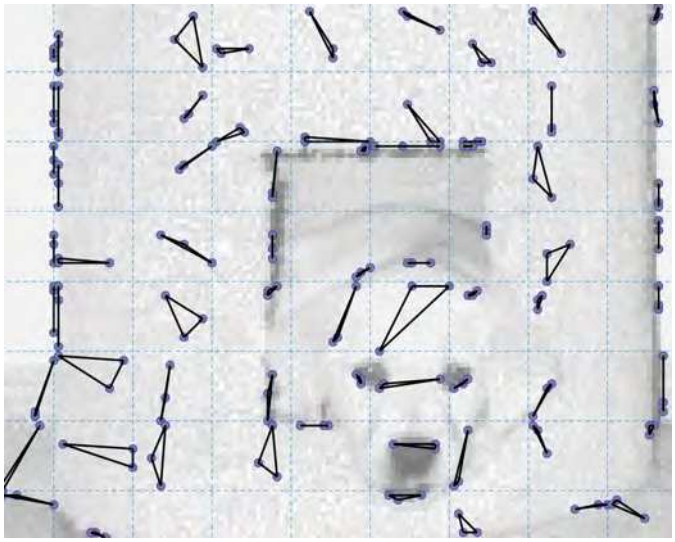


Figure 8.5: Inter-GIP Distances.

$$D_{Cosine} = 1 - \frac{P'Q}{\sqrt{(P'P)(Q'Q)}} \tag{8.4}$$

$$D_{Euclidean} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{8.5}$$

$$D_{Minkowski} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{8.6}$$

$$D_{Jaccard} = 1 - \frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)} \tag{8.7}$$

where $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ representing two points in the two-dimensional space. The Minkowski distance (8.6) can be considered a generalization of three other distances, the Euclidean if $p = 2$, the Cityblock if $p = 1$ and the Chebychev distance if $p = \infty$. For the experiment in Section 8.3.3 we defined $p = 3$. Please note that (8.4) is actually a similarity measure, hence inverse to the others. That, however, has no effect on the discriminative value of the descriptor. Further information about distance functions can be found in [Eid12b].

Gestalt Regions of Interest (GROI)

Later in this paper we present our experimental results of combining GIP with a CNN. For feeding the output of the GIP algorithm into a CNN we enhanced the GIP algorithm to produce so-called Gestalt Regions of Interest (GROI) images. Since GIPs are the basis for GROIs, the GROI images are also based on the Gestalt principles. They are

(a) raw pixel image (baseline) (b) GROI image ($t = 0$, $\alpha = 30$, $s = 4$) (c) GROI image ($t = 0.5$, $\alpha = 38$, $s = 4$)

(d) GROI image ($t = 1$, $\alpha = 38$, $s = 6$) (e) GROI image ($t = 1.5$, $\alpha = 38$, $s = 8$) (f) GROI image ($t = 1.5$, $\alpha = 45$, $s = 6$)

Figure 8.6: An example face image and its Gestalt Region Of Interest (GROI) image representations. Each GROI image was produced with different parameter combinations.

intended to be produced in a preprocessing step of a CNN to feed the CNN only with the most interesting image regions. Converting images into GROI images works as follows: In the first step the GIPs are detected in the input image as described in Section 8.2.1. These GIPs are serving as center pixels for quadratic regions of interest. The size of these squares is controlled through parameter s. The remaining pixels of the image, which are not within the squares are set to white. They are considered as not useful enough for the recognition process. Furthermore, we claim that preserving only the GROIs for training a CNN, instead of using the whole images, reduces the risk of data-over-fitting drastically. Figure 8.6 shows various GROI example images produced with different GIP parameters t and α (see Section 8.2.1 for detailed explanation) and GROI parameter s.

8.3 Experiments and Results

In this Section we present the experimental results of applying the GIP algorithm on various application domains – all related to concrete forensic applications. We used two different evaluation measures for our experiments, namely accuracy and F_1-score. In

terms of statistical significance, accuracy is the better choice when a huge number of test samples is unavailable. Based upon the size of the test set, we decided to use the accuracy for some experiments, and for others we utilized the F_1-score.

8.3.1 Overview of Experiments

In Section 8.3.2 we show the results of applying the GIP-ABS algorithm for the recognition of faces of people that have undergone significant body weight change [HE14]. As described in Section 8.2.1 the GIP detection algorithm is based on the assumption that some interest points contribute more to the description of images than others. This experiment was designed to find out which GIPs can be eliminated to make the whole method more efficient while retaining our classification results. The remaining GIPs are the fundamental basis for our final GROI-CNN experiments. Furthermore, we investigated the robustness of GIP against image rotation.

The experimental results of Section 8.3.3 present the GIP-IGD algorithm applied on two different image classification tasks [HE17]. This experiment shows that only a few of the very compact GIP-IGD image descriptors are necessary to quickly classify the images from the datasets with high accuracy. Furthermore, we compared our results to several local point-based description methods and to a CNN. As mentioned in Section 8.2.2, GIP-ABS does not work well when it comes to categorizing scaled images. In contrast this experiment shows that GIP-IGD is resilient to some scale changes.

The final experiment and the ultimate goal of this work is presented in Section 8.3.4. As demonstrated in the experiments of Section 8.3.3 our method outperforms the other applied local description methods. Nevertheless, the CNN dominates our method and all the other applied local description methods in terms of accuracy for the given application domain, but it is significantly slower. Therefore, we decided to merge the GIP method and CNNs to create an even more powerful recognition system. This experiment shows that a special variant of the GIP algorithm as a preliminary stage for a CNN outperforms a conventional CNN for the given application domain.

8.3.2 GIP for Weight-Invariant Face Recognition

This Section describes the application of GIP-ABS on the description of face images in a way that outperforms the baseline methods for the domain of significant change of person weight. In addition, the experimental results of investigating the influence of the GIP parameters t and α on the recognition accuracy are presented. Adjusting t and α causes the algorithm to extract more or fewer GIPs from the image. This experiment was designed to find out which GIPs can be omitted to make the whole method more efficient while retaining our classification results. The remaining GIPs are the fundamental basis for our final GROI-CNN experiments. Eventually, we show our evaluation of GIP with respect to sensitivity against rotation. The results are an extension of a previously published work [HE14].

We assumed that the ability of the GIP algorithm to select interest points within high-contrast image blocks and on non-diagonal edges (Section 8.2.1) should increase the face recognition performance. Both are targeted at typical properties of face images: On the one hand, face features are often distinguished by high contrast which is to a certain degree due to the morphology of the human skull. On the other hand, face features tend to have a clear orientation. Both aspects are influenced by weight change: Weight gain reduces the availability and contrast of face features which also influences their orientation. The investigation of the reasonability of these assumptions and their implementation are – next to the identification of the best-performing GIP parameters – a second target of our research.

Dataset

To our knowledge, a standardized dataset for the recognition of faces of overweight people is currently not available. The commonly used databases (UMIT, FERET, etc.) do not include such material. This is unfortunate as the problem is of high practical relevance, in particular in the forensic application of face recognition. As a consequence, we had to compile a dataset for our experiments. It turned out that pairs of face images with significant weight gain/loss in-between are hard to find. Eventually, we succeeded in assembling a dataset of face photos for a group of fifteen persons who underwent significant weight change (at least twenty kilograms) in less than one year. The majority of the photos were taken from a diet web forum [Red]. Others were provided by acquaintances of the authors.

Experimental Setup

Five local feature description methods were chosen for comparison with the GIP feature: SIFT [Low04], SURF [BETVG08], MSER [MCUP02], FREAK [Ort12] and ORB [RRKB11]. After feature extraction with one of the above methods, we received multiple feature vectors for each image. Then we generated a vocabulary composed of 300 visual words via the k-means clustering algorithm and quantized all the feature vectors with the popular BoVW-algorithm [CDF+04]. Each of our images were now represented by a single histogram. For classification of the features, we employed the Euclidean distance. Hence, all standard descriptors as well as our approach are employed in exactly the same way. This is a mandatory requirement for comparing the description performance for the recognition problem at hand. During our experiments it turned out that on the given application domain the GIP algorithm outperforms the above-mentioned state-of-the-art description methods. It dominates them both in terms of recognition accuracy and of description compactness. In summary, the GIP algorithm produces shorter description that contains more weight-invariant face information.

For the experiments, without loss of generality we employ the face images with lower weight as the training set. The test set consists of the face photos that show the higher weight. Figure 8.7 shows three example images and descriptions extracted by the GIP algorithm. The evaluation task is to associate each test image with the corresponding

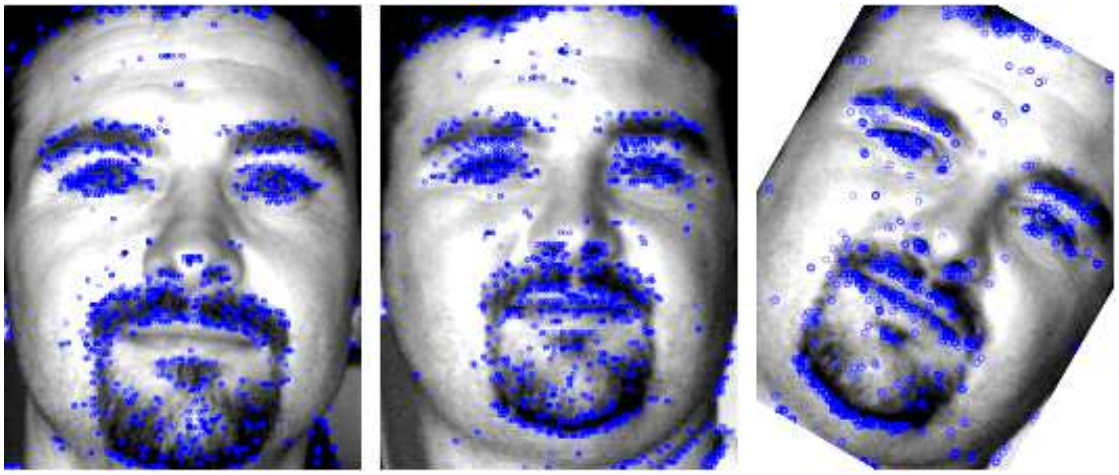


Figure 8.7: The GIP-algorithm was applied on pictures such as these examples. GIPs which are within low-contrast macro blocks and many of the GIPs on diagonal edges were discarded. *Left*: The image shows a normal weight person and the detected GIP points, indicated as dark blue circles. *Middle*: Shows the same person after 30 kilograms of weight gain and the detected GIP points. *Right*: Shows the person image rotated by 30 degrees and the detected GIP points.

training image. Due to the small number of samples success is measured as accuracy, i.e. here the number of true positives. The ground truth is provided by the authors.

Remark: In practical forensic application, pictures of suspects (e.g. taken by a surveillance camera) are typically of very low quality. To evaluate how well ours and the state-of-the-art local description algorithms can deal with this aspect, the photos in the dataset are left in their original resolutions, ranging from 201x285 to 508x728 pixels. However, the contrast of the test images was adapted to the contrast of the training images using histogram equalization because this step improves the overall classification performance without limiting the generality of the experiment.

Evaluation

The second column of Table 8.1 shows that using the five baseline interest point features SIFT, SURF, MSER, FREAK and ORB to identify the faces of people who experienced significant weight change delivers only moderate classification accuracies. Of all five features, SURF provides the best results with 33 percent. However, to obtain this result an average of 57,984 description values per face (last column) are necessary. This number is calculated as the product of the average number of description vectors per face (453) times the size of one description vector (128). Using the MSER feature, only 132 description values per face are required on average. In return, the classification accuracy is only 6.7 percent which is far below an acceptable rate for practical application.

Compared to the five baseline features above, the GIP description algorithm is by far

Table 8.1: A comparison of classification accuracies and the average number of description values per face for identifying faces of people who experienced significant weight change.

Method	Acc.	Accuracy 30° rotated	Average Number of Description Values per Face
BoVW+SIFT	20%	13.3%	25,309
BoVW+SURF	33%	13.3%	57,984
BoVW+MSER	6.7%	6.7%	132
BoVW+FREAK	20%	20%	59,473
BoVW+ORB	13.3%	13.3%	8,883
BoVW+GIP	53.3%	53.3%	52,536
BoVW+GIP $t = 70$	53.3%	53.3%	23,086
BoVW+GIP $\alpha = 0.0009$	46.7%	46.7%	23,101
BoVW+GIP $t = 70$ $\alpha = 0.0009$	46.7%	46.7%	10,425

more accurate in its original form with 53.3 percent. This performance is 20 percent ahead of the SURF algorithm that leads the baseline description methods. As Figure 8.8 shows, discarding interest points that lie within low-contrast macro blocks does not affect the accuracy until the threshold reaches $t = 70$. At the same time, the average number of description values per face goes down from 52,536 to 23,086. That is, by discarding low-contrast blocks, we maintain the original accuracy of the GIP algorithm but reduce the amount of data to just 44%. The required information for this result is even less than the information SIFT, SURF and FREAK need to achieve their lower performance. Hence, we consider it justified to say that for the given domain, the GIP approach clearly dominates the baseline local description methods.

Figure 8.9 shows that the elimination of diagonal edges in the GIP algorithm does not affect the accuracy until an $\alpha = 0.64$ is reached, but this modification reduces the average number of description values per face drastically. With $\alpha = 0.0009$ and 23,101 description values, the algorithm still reaches an accuracy of 46.7 percent. A classification accuracy of 46.7 percent is still significantly higher than the results reached by the commonly used local feature transformations. We find these results encouraging to employ these modifications also in other application domains.

Selecting interest points within high-contrast image blocks and on non-diagonal edges leads to a significant reduction of the average number of description values required to 10,425 values per face. Figure 8.10 illustrates the behavior of the algorithm. An accuracy of 46.7 percent is still significantly higher than the results of the baseline features. This result supports our hypothesis that interest points on almost horizontal or vertical edges are more useful for face description than other points. Furthermore, it indicates that our hypothesis (certain interest points in low contrast areas can be neglected) has empirical substance. There appears to exist a trade-off between Gestalt perception and focusing on salient points.

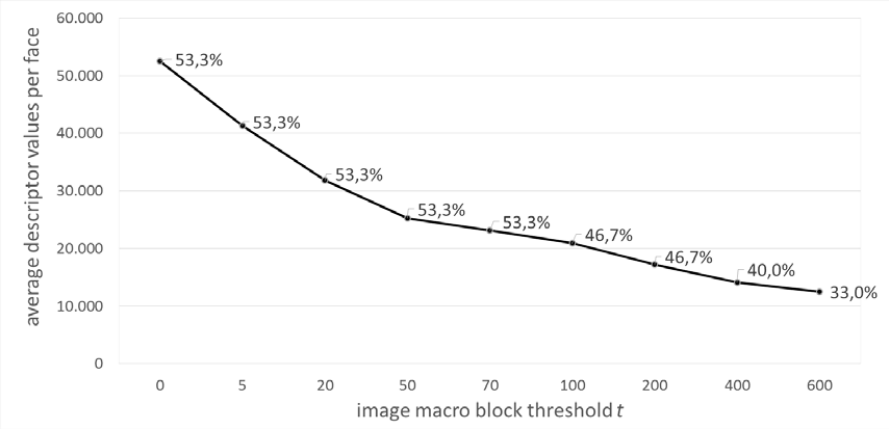


Figure 8.8: The average number of description values per face and categorization accuracy in percent as a function of image macro block variance threshold t. A value of $t = 0$ means that no image macro blocks are excluded from the recognition process.

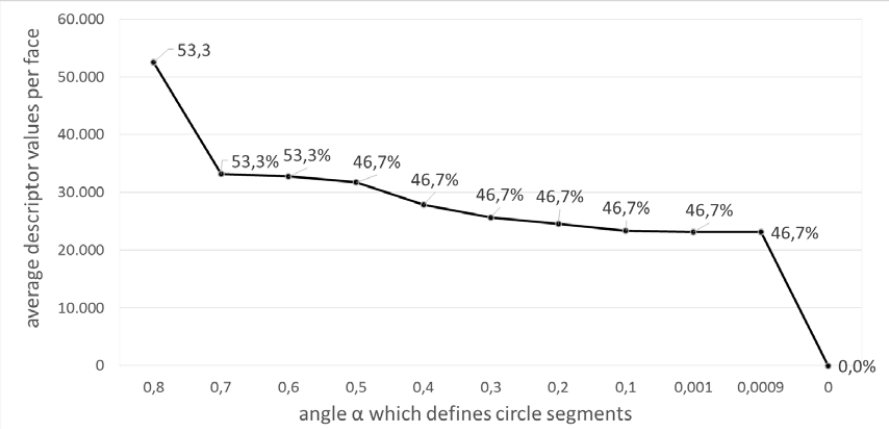


Figure 8.9: The average description values per face and categorization accuracy in percent as a function of circle segment angle α. A value of $\alpha = 0.8$ radians means that no GIPs are discarded. If there are no perfect straight lines in the face image dataset then for $\alpha = 0$ the accuracy will drop to zero.

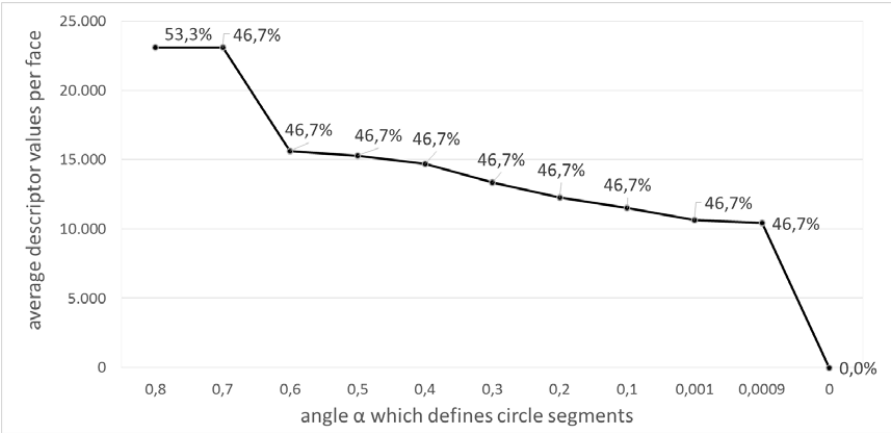


Figure 8.10: The average number of description values per face and categorization accuracy in percent as a function of circle segment angle α with $t = 70$.

Eventually, we evaluated the sensitivity of our approach against rotation. All baseline feature extraction methods are to a certain degree rotation-invariant. To find out how robust the GIP approach is against rotation, we conducted an experiment with all test images rotated by 30 degrees. The third column of Table 8.1 depicts the outcome. SIFT and SURF are known from literature as scale and rotation-invariant features. In many works they have been very successfully applied in numerous different application domains. However, for our specific application domain, Table 3.1 shows that SIFT and SURF deliver lower accuracy for rotated images. The accuracy of MSER, FREAK and ORB remains constant. Likewise, GIP is not affected by rotation: The performance remains constant. Hence, we consider it fair to conclude that the GIP approach is a highly competitive local description approach for the problem under consideration.

In summary, it appears that the GIP approach describes faces in a weight-invariant way to a sufficiently higher degree than the baseline methods do. Its accuracy is at least 20% better than the first competitor (SURF). As assumed, the relative completeness of Gestalt interest points makes a clear difference in recognition performance. GIP descriptions are more compact than most other descriptions and they are rotation-invariant. That is, we need less disk space and processing power for description storage and evaluation. This is an important advantage in a big data domain such as face recognition. Rotation invariance is a simple requirement satisfied by most – yet not all – algorithms.

The GIP algorithm is based on the assumption that some interest points contribute more to the description of images than others. The experiment demonstrated that certain well-selected GIPs can be omitted in order to make the whole method more efficient while retaining our classification results. The remaining GIPs are the fundamental basis for our final GROI-CNN experiments presented in Section 8.3.4.

8.3.3 GIP-IGD for Image Categorization

In this Section, we present an extensive evaluation of applying the GIP algorithm in combination with the IGD feature vector (GIP-IGD) on image categorization. GIP-IGD is described in Section 8.2.2. One goal of the following experiments was to find the IGD distance measure which maximizes the categorization accuracy while keeping the computational complexity as low as possible. Moreover, we wanted to test how robust our GIP-IGD algorithm is against image scaling. The presented results are an extension of a previously published work [HE17].

As demonstrated in the experiments of this Section our method outperforms all of the other applied local description methods. Nevertheless, the CNN dominates our method and all the other applied local description methods in terms of accuracy for the given application domain, though it is much slower. Therefore, we decided to build a bridge between the GIP method and CNNs to create an even more powerful recognition system. The experimental results addressing this issue are presented in Section 8.3.4.

Datasets

In our first evaluation task, we tested the detection performance with the INRIA Horses dataset [FJS10], consisting of 170 images containing horses, and 170 without horses. The goal of the evaluation task was to categorize the images into images containing horses and images without horses.

The second dataset which we used to test our algorithm is the Food-5K dataset [SYE16]. It consists of 2,500 food images, which cover a wide variety of food items and 2,500 randomly selected non-food images. Some food images also contain other objects or people. The Food-5K dataset with a total size of 5,000 images is significantly bigger than the INRIA horses dataset. The goal of this evaluation task was to categorize the images into food and non-food images.

Experimental Setup

We compared our method to several different local feature description algorithms, namely SIFT [Low04], SURF [BETVG08], BRISK [LCS11] and FREAK [Ort12]. Additionally, we compared GIP-IGD to GIP-ABS and to a CNN. Recently, the CNN offers a very accurate state-of-the-art technique for many general image classification and object recognition problems. The SIFT and SURF descriptors are both vectors containing floating point values. More recent binary descriptor methods like BRISK and FREAK are less computationally expensive, but their accuracy is lower.

After quantizing the extracted local descriptors with the BoVW-algorithm [CDF+04] we fed the resulting histograms into Matlab's Classification Learner App. The app compares several different Classifiers, e.g., different variations of Trees, Support Vector Machines (SVM), Nearest Neighbor Classifiers, Ensemble Classifiers and so forth. It turned out that the Medium Gaussian SVM appeared to be best suited for our categorization problems.

Evaluation

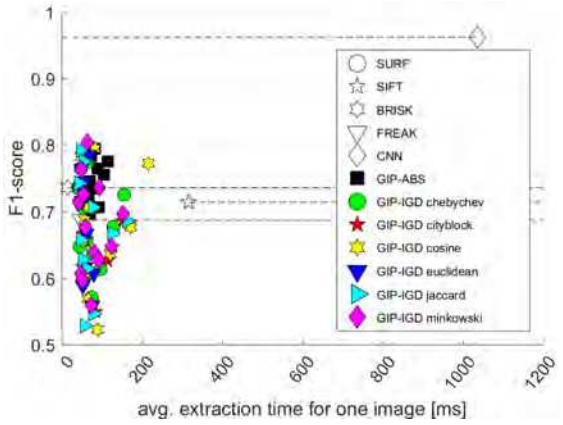
The experimental results of applying our algorithm on the INRIA Horses dataset are shown in Figure 8.11 and the results for the Food-5K dataset are presented in Figure 8.12. Figures 8.11a and 8.12a depict the F_1-scores over extraction time of our IGD experiments with different distance measures. Adjusting the values of the two GIP parameters t and α causes the algorithm to extract more or fewer image feature vectors. Therefore, these parameters indirectly affect the extraction time per image and the categorization accuracy because they determine the number of extracted feature vectors. With higher t and lower α, the number of extracted feature vectors per image decreases. It is assumed that the remaining descriptors carry a considerable amount of information and descriptors with less information are omitted. A small set of descriptors for the categorization task reduces the computational complexity significantly.

As already mentioned, the binary descriptor methods BRISK and FREAK are very fast and therefore strong competitors when it comes to computational complexity. Yet, their F_1-scores are relatively low. The F_1-scores of SURF and SIFT are higher but not as high as the F_1-score of GIP. SIFT is with about 400 ms extraction time per image, comparatively slow. The clear winner in terms of F_1-score is the CNN but the serious drawback is the high computational complexity. The CNN needs more than one second to extract the features from one image and therefore it is by far the slowest method.

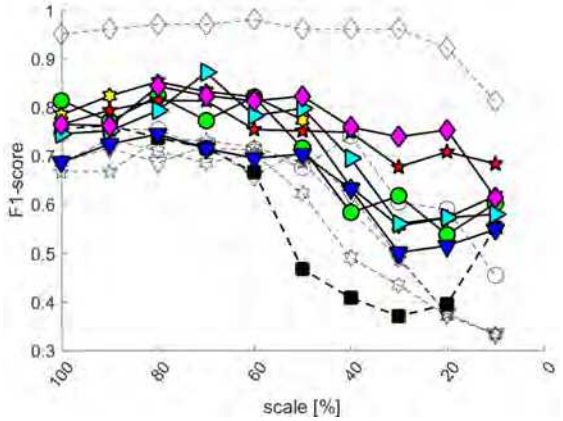
Figures 8.11b and 8.12b show the F_1-scores over scaled versions of the test images. As mentioned earlier, GIP-ABS does not perform well when it comes to categorizing scaled images. In contrast, GIP-IGD is resilient to some scale changes. Especially, GIP-IGD in combination with the Minkowski distance measure delivers outstanding results in the case of horse categorization, and in the case of food categorization GIP-IGD delivers good results in general, no matter which distance measure is used. For our application domains GIP-IGD is more robust against scaling than SURF, SIFT, BRISK and FREAK. The CNN has the highest accuracy but as mentioned above, it is significantly slower.

Figures 8.11c and 8.12c depict the average numbers of feature values extracted from one image. The description vectors of SURF, BRISK and FREAK are 64-dimensional and the SIFT vector has 128 elements. Hence, they are more memory-consuming than the 9-dimensional GIP-IGD feature vectors. For example, in case of using FREAK for horse categorization a total number of $918 * 64 = 58,752$ feature values per image are necessary to get a comparatively poor F_1-score of 69%. In other words, Figures 8.11 and 8.12 demonstrate that GIP outperforms SIFT's, SURF's, BRISK's and FREAK's accuracy while reducing the descriptor length per image to only a few percent.

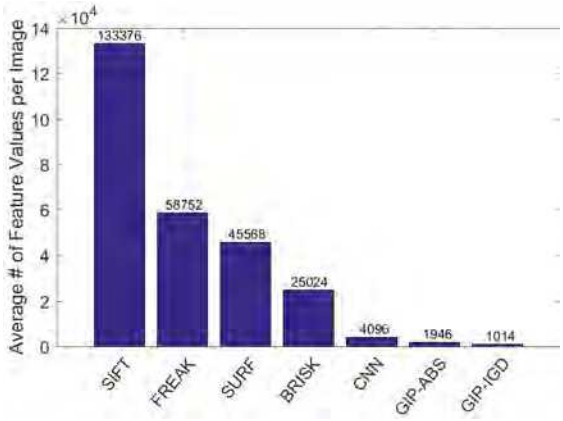
A CNN can achieve extremely high accuracies, but this advantage does not come without a price. CNNs in general are computationally expensive and slow compared to interest point features, even with graphical processing units. Additionally, a huge set of training data is needed, which can be difficult to provide and the training process itself can be very time consuming. We showed above that it is possible to use the GIP feature for image understanding tasks where a long training time is unacceptable and / or a huge



(a) Grey scale image

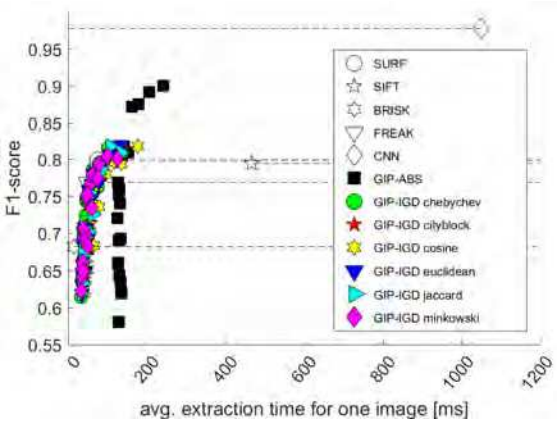


(b) Image gradients

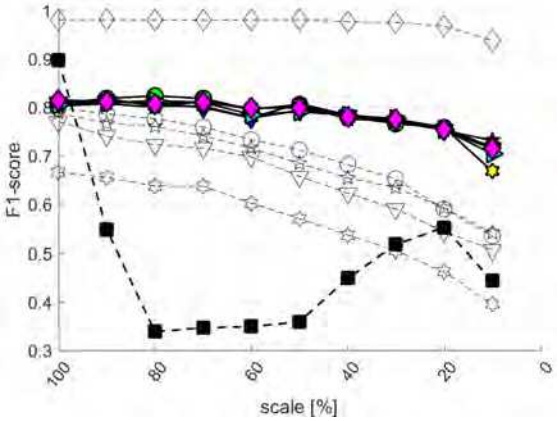


(c) Gestalt Interest Points

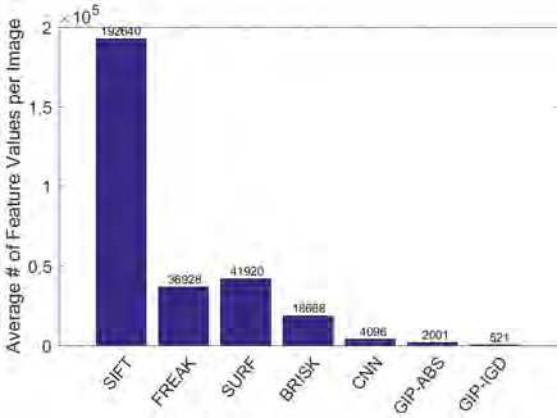
Figure 8.11: The experimental results of applying our algorithm on the INRIA Horses dataset. Our algorithm is also compared to several different baseline methods. The different F_1-scores for each IGD distance measure in Figure 8.11a arise through adjusting the two GIP parameters t and α, which are described in Section 8.2.1.



(a) Grey scale image



(b) Image gradients



(c) Gestalt Interest Points

Figure 8.12: The experimental results of applying our algorithm on the Food-5K dataset. Our algorithm is also compared to several different baseline methods. The different F_1-scores for each IGD distance measure in Figure 8.12a arise through adjusting the two GIP parameters t and α, which are described in Section 8.2.1.

Figure 8.13: Before (top line) and after (bottom line) makeup examples of four subjects contained in our makeup dataset.

amount of training data is unavailable. As demonstrated our method outperforms all the other applied local description methods. Nevertheless, the CNN dominates these methods including ours in terms of accuracy for the given application domain. Therefore, we decided to build a bridge between the GIP method and CNNs to create an even more powerful recognition system. The next Section shows that the GROI variant of the GIP algorithm merged with a CNN outperforms a conventional CNN for the given application domain.

8.3.4 Deep Gestalt Regions of Interest for Makeup-Robust Face Recognition

In our last experiment we present the results of training a CNN with the novel GROI images for the domain of makeup-robust face recognition. The rapid evolution of face recognition systems into real-time applications has raised new concerns about their ability to resist presentation attacks, particularly in unattended application scenarios such as automated border control. Research about makeup-robust face recognition is still very limited and we think that our work could be beneficial in solving this problem. Dantcheva et al. [DCR12] claimed in their study that the application of facial cosmetics significantly decreases the performance of both academic face verification approaches and commercial approaches. As shown in Figure 8.13, significant appearance changes can be observed for individuals with and without makeup.

To our knowledge, there is only limited scientific literature that addresses the challenge of makeup-robust face recognition. Chen et al. [CDR15] addressed this problem with a patch-based ensemble learning method. Song et al. [LSW+18] synthesize a non-makeup image from a face image with makeup via a generative network. After that, deep features are extracted from the synthesized image to further accomplish the makeup-robust face

recognition. Zheng et al. [ZK17] proposed a hierarchical feature learning framework for face recognition under makeup changes. Their method seeks transformations of multilevel features because these features tend to be more invariant on higher semantic levels, and less invariant on the lower levels.

Many recent works on face recognition have proposed numerous variants of CNN architectures [PVZ+15b, WZLQ16, SLWT15b]. GROI images and CNNs are both inspired by cognition. Therefore, it appears reasonable to merge both concepts into one powerful face recognition system. In this experiment after-makeup against before-makeup face samples were matched and it was designed for exploring the effectiveness of feeding GROI images into a CNN. To obtain baseline results to which we can compare our method, we decided to feed the unmodified raw pixel images into the same CNN which we fed with the GROI images. Note that there is no overlap between training images and test images of the subjects and therefore this experiment is a very sophisticated recognition task. For the training stage 6,000 non-makeup face images of 6 subjects serve as input. Henceforth, the classification stage assigns each of the 1,200 makeup test images to one of the 6 subjects. One advantage of our approach is that we do not need color information, which is often not available, e.g. frames of surveillance cameras. Actually, it is very likely that a color-based recognition approach would perform worse, because makeup changes the skin color and therefore the recognition process may lead to false positives.

Dataset

Since we wanted to keep CNN training times as low as possible, we decided to utilize a subset of the self-compiled YouTube makeup dataset, which we presented in an earlier work [HE18]. This subset consists of 6 subjects with 1,000 non-makeup face images per subject for training and 200 makeup images per subject for testing. Figure 8.13 shows some example images. On the one hand, the dataset is small and therefore it saves training time but, on the other hand, it is big enough to deliver reasonable experimental results. However, we plan to employ the GROI method on bigger datasets in future work. The makeup in the test face images varies from subtle to heavy. The cosmetic alteration affects the quality of the skin due to the application of foundation and change in lip color and the accentuation of the eyes by diverse eye makeup products. This dataset includes some variations in expression and pose. The illumination condition is reasonably constant over multiple shots of the same subject. In a few cases, the hair style before and after makeup changes drastically.

Experimental Setup

We implemented a prototype for this experiment utilizing Python in combination with the machine learning framework Tensorflow [ABC+16] and the high-level neural networks API Keras [C+15]. The structure of the chosen CNN model is shown in Table 8.2. It is an adapted version of the VGG-like model from the Keras website. VGGNet [SZ14] was invented by VGG (Visual Geometry Group) from University of Oxford. According to VGGNet we also use filters of size 3x3 because smaller filters generally provide better

Table 8.2: Structure of the adapted example CNN model from Keras website [Cho]

Layer Name (type)	Output Shape
conv2d_1 (Conv2D)	(158, 158, 32)
conv2d_2 (Conv2D)	(156, 156, 32)
max_pooling2d_1 (MaxPooling2)	(78, 78, 32)
dropout_1 (Dropout)	(78, 78, 32)
conv2d_3 (Conv2D)	(76, 76, 64)
conv2d_4 (Conv2D)	(74, 74, 64)
max_pooling2d_2 (MaxPooling2)	(37, 37, 64)
dropout_2 (Dropout)	(37, 37, 64)
conv2d_5 (Conv2D)	(35, 35, 64)
max_pooling2d_3 (MaxPooling2)	(35, 11, 64)
dropout_3 (Dropout)	(35, 11, 64)
conv2d_6 (Conv2D)	(35, 10, 64)
max_pooling2d_4 (MaxPooling2)	(17, 10, 64)
dropout_4 (Dropout)	(17, 10, 64)
flatten_1 (Flatten)	(10880)
dense_1 (Dense)	(256)
dropout_5 (Dropout)	(256)
dense_2 (Dense)	(6)

results. The number of layers was chosen to satisfy our requirements. On the one hand, we wanted a CNN with enough layers to ensure high accuracies, and on the other hand, limiting the number of layers for shorter training times was a second important requirement.

During training of a CNN its network weights are updated iteratively by an optimization algorithm. The choice of this optimization algorithm is crucial for the performance of a CNN. We empirically identified that the Adam optimization algorithm [KB14] with an initial learning rate $lr = 0.00001$ and the categorical cross entropy loss function leads to fast training accuracy convergence for our dataset. Each epoch the training progress was validated using 10 percent of the training images. To avoid long training times and possible over-fitting we decided to use an early stop strategy. A patience value of 15 was set, i.e. the number of epochs to wait before early stop, if the validation accuracy stagnates.

A powerful hardware infrastructure is necessary when it comes CNN training. For our experiments we decided to run them on Crestle [Cre]. The Crestle servers are equipped with NVIDIA Tesla K80 GPUs and therefore they have been considered adequate for our purposes.

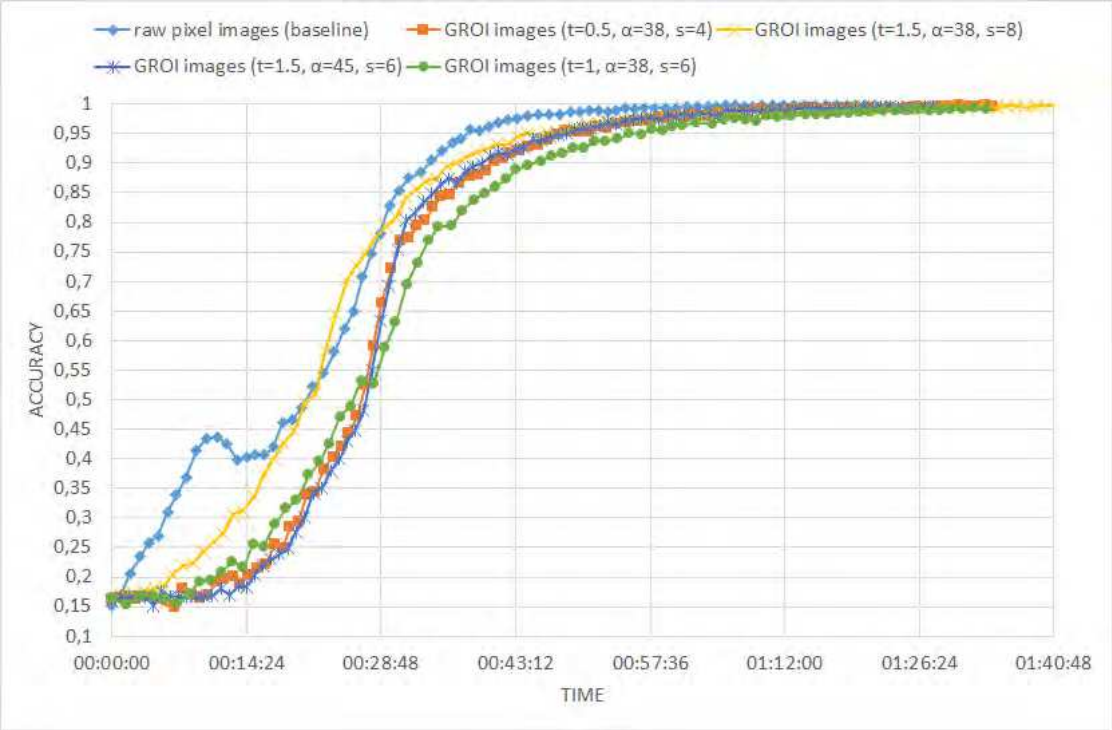


Figure 8.14: Train accuracies for each epoch of the training process. Each line marker denotes one train epoch.

Evaluation

Figure 8.14 shows the training accuracies for each epoch over the training period and Figure 8.15 the validation accuracies, respectively. See Section 8.2.2 for a detailed explanation of the parameters t, α and s. As mentioned above the validation set comprises 10 percent of the train images. We trained six types of CNNs, one with the raw pixel images and 5 with different versions of GROI images. For a visual overview of the different input image types see Figure 8.6. As can be seen in Figure 8.14 and Figure 8.15 the CNN fed by the raw pixel images leads to the fastest convergence, closely followed by the CNN fed by GROI images with parameters $t = 1.5$, $\alpha = 38$ and $s = 8$. A greater value for s causes the algorithm to produce bigger GROIs. We assume that this is the reason why the training employing GROI images produced with $s = 8$ leads to similar convergence as with raw pixel images. The GROI images with $s < 8$ leading to slower training and validation accuracy convergence.

For test purposes the resulting model was stored after every fifth training epoch during the training process. These models were used to classify the make up images from the test set. Each line marker in Figure 8.16 denotes an accuracy produced using one of these stored models. After 30 training epochs the CNN model trained with the GROI images ($t = 1.5$, $\alpha = 45$, $s = 6$) starts to outperform the baseline CNN trained with the

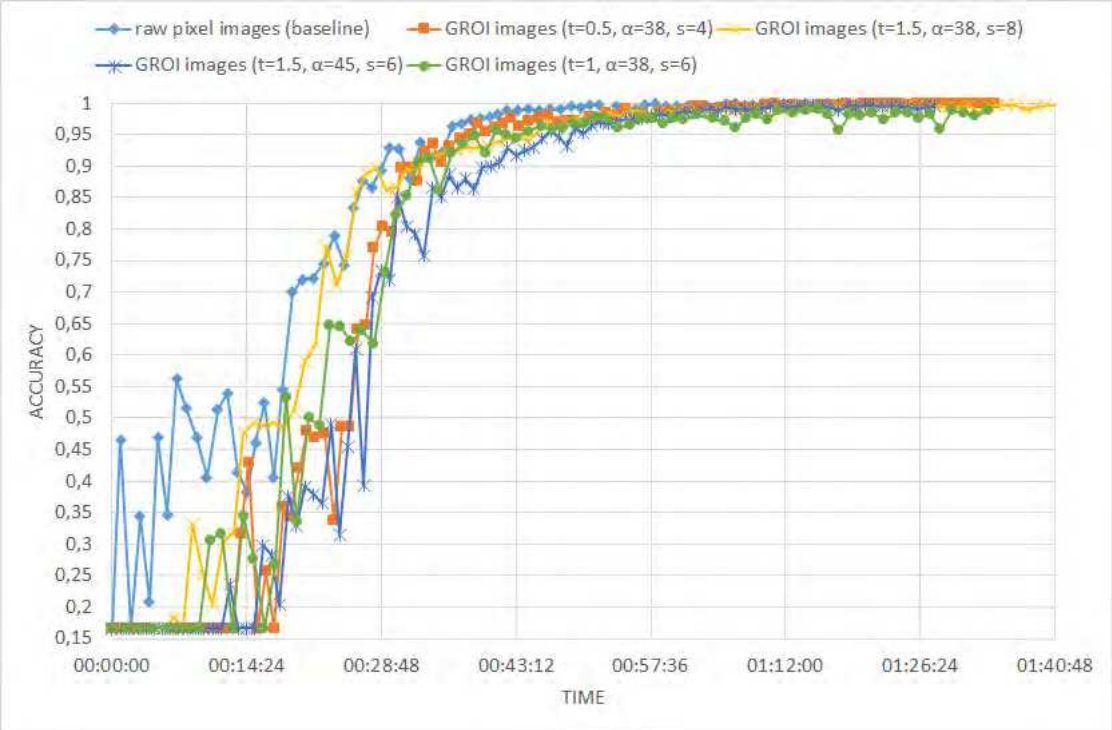


Figure 8.15: Validation accuracies for each epoch of the training process. Each line marker denotes one train epoch.

unmodified images. With the model trained for 50 epochs by the GROI images ($t = 1.5$, $\alpha = 45$, $s = 6$) 88.3 percent of the test images are classified correctly. The baseline model in comparison delivered only 80 percent accuracy after 50 epochs of training. The peak of 89.8 percent was produced after 60 epochs with the model trained with the GROI images ($t = 1$, $\alpha = 38$, $s = 6$).

Figure 8.16 demonstrates that training a CNN by GROI images clearly outperforms a CNN trained from raw pixel images for the domain of makeup-robust face recognition. The model trained with GROI images ($t = 1.5$, $\alpha = 45$, $s = 6$) produces the highest accuracies among all models. With a greater parameter t more low-contrast GROIs are omitted. A value of 45 degrees is the maximum for α and this means that the parameter does not have any effect on producing GROI images.

As described above the CNN trained with GROI images leads to slower training convergence in comparison to the CNN trained with raw pixel images. This fact in combination with the high test accuracies proves that our presented method is more robust against over-fitting than the conventional method, training a CNN by raw pixel images. Another advantage of the GROIs is that it is possible to describe the semantic content of images more compactly than with whole images. For example, it would be possible to store only the GROIs and their center point coordinates instead of storing GROIs on white

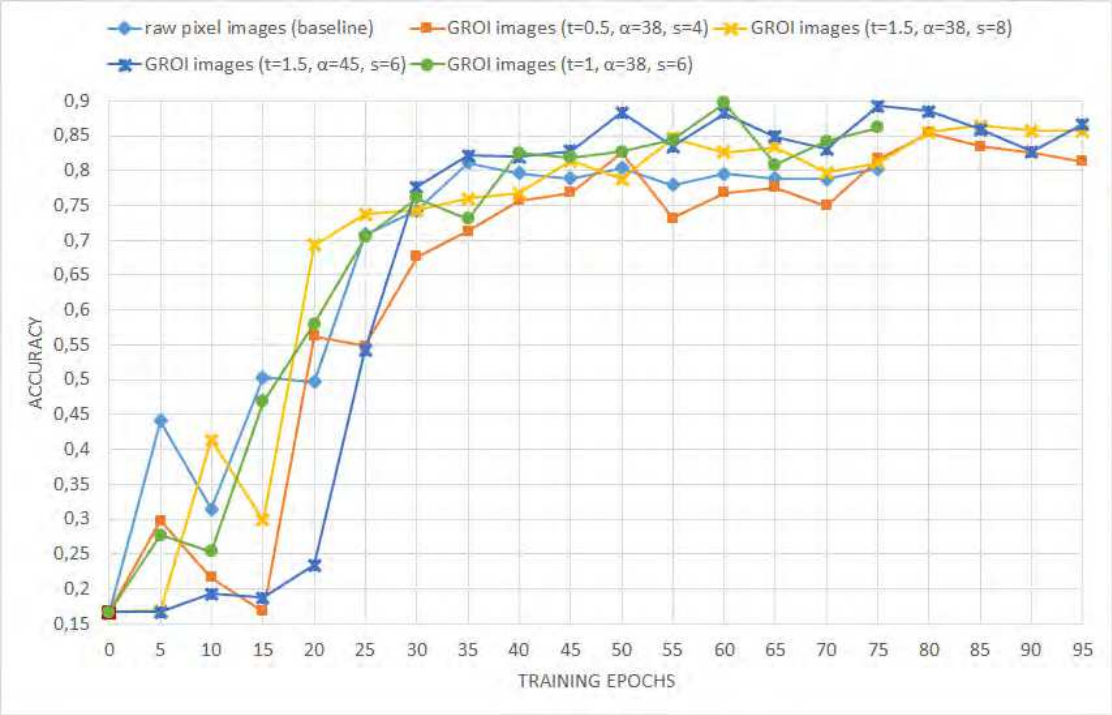


Figure 8.16: Test accuracies for stored train models. For every fifth training epoch the resulting model was stored during the training process. These models were used to classify the make up test images. Each line marker denotes an accuracy produced using one of these stored models.

background, thus requiring less disk space. This is a very important argument in big data domains such as face recognition.

8.4 Conclusion

In this work, we proposed a novel visual perception inspired local description approach as a pre-processing step for deep learning. To show the effectiveness of our GROI method we fed its output into a state-of-the-art convolutional neural network. Our experimental results revealed that it outperforms a CNN that is trained on images which are not pre-processed by our method in the domain of makeup-robust face recognition. The problem of makeup-robust face recognition is of high relevance for practical life and our method could be helpful in solving this problem. The proposed GROI method interconnected with a CNN dominates a conventional CNN in terms of accuracy and robustness against over-fitting. Another advantage of the GROI approach is that it is possible to describe the semantic content of images more compactly than with whole images.

In our opinion, a serious comparison between the results of this work and results of

other works in a scientifically substantiated way is not possible based on the facts (i) we could not find many works about makeup-robust face recognition and, (ii) we had to assembly our own dataset to fit our needs. Nevertheless, we want to list the results of some other works. Chen et al. [CDR15] reached a Rank-1 accuracy of 89.40 percent applying their patch-based ensemble learning method in combination with Commercial Off-The-Shelf (COTS) Systems on the YMU-dataset. The bi-level adversarial network (BLAN) proposed by Song et al. [LSW+18] delivers up to 94.8 percent Rank-1 accuracy applied on three different datasets. Zheng et al. [ZK17] proposed a new hierarchical feature learning framework and achieved an accuracy up to 81.11 percent with two different datasets. As we showed in our experiments, with our method an accuracy of 89.8 percent was reached through applying the GROI method on our self-compiled makeup faces dataset. These results could be a baseline for future work.

The GROI feature is based on the earlier presented GIP feature. We showed that it is possible to use the GIP feature as a feature in itself without a CNN for image understanding tasks where a long training time is unacceptable and / or a huge amount of training data is unavailable. Experiments have demonstrated that the GIP algorithm results in very compact descriptions that satisfy the major Gestalt Laws.

However, a CNN is only one – but successful – example of a deep learning method and our approach could also be combined with other methods, e.g. ResNets. As is evident from our experiments, the output of our algorithm consists of heavily compressed content-rich information. We assume that adding this information as residuals to the output of ResNet convolution operations could improve the ResNet in a similar way as the CNN was improved during our experiments. Furthermore, with higher accuracy it would be possible to use fewer network layers and thus shorten the training time of the network. Experiments addressing this topic are planned for future work.

List of Figures

List of Tables

List of Algorithms

Bibliography

[AAR04] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, Nov 2004.

[ABC+16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[Ahm19] Mohiuddin Ahmed. Data summarization: A survey. *Knowl. Inf. Syst.*, 58(2):249–273, February 2019.

[ASCT17] Mundher Al-Shabi, Wooi Ping Cheah, and Connie Tee. Facial expression recognition using a hybrid cnn-sift aggregator. In *MIWAI*, 2017.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[BFL+18] Qirong Bo, Jun Feng, Pan Li, Zhaohui Lv, and Jing Zhang. Towards better soft-tissue segmentation based on gestalt psychology. *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 262–267, 2018.

[BM10] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 282–295, Berlin, Heidelberg, 2010. Springer-Verlag.

[Bra00] G. Bradski. Opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

[BSCL14] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 584–599, Cham, 2014. Springer International Publishing.

[BW07] S. Bileschi and L. Wolf. Image representations beyond histograms of gradients: The role of gestalt descriptors. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[C+15] François Chollet et al. Keras. `https://keras.io`, 2015.

[CDF+04] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CDR15] Chen Cunjian, Antitza Dantcheva, and Arun Ross. An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, October 2015.

[CDSR17] C. Chen, A. Dantcheva, T. Swearingen, and A. Ross. Spoofing faces using makeup: An investigative study. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–8, Feb 2017.

[Cho] François Chollet. Keras model examples. `https://keras.io/getting-started/sequential-model-guide/`. last visited on August 5th 2018.

[CHWS11] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 119 –124, aug. 2011.

[CMM+11] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, volume 2 of *IJCAI'11*, pages 1237–1242. AAAI Press, 2011.

[Cre] Crestle. Crestle effortless infrastructure for deep learning. `https://www.crestle.com/`. last visited on June 21th 2018.

[CSWL09] L.H. Chen, C.W. Su, C.F. Weng, and H.Y.M. Liao. Action scene detection with support vector machines. *Journal of Multimedia*, 4(4):248–253, 2009.

[CTYH12] Yang Cai, Wei Tong, Linjun Yang, and Alexander G. Hauptmann. Constrained keypoint quantization: Towards better bag-of-words model for

large-scale multimedia retrieval. In *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 16:1–16:8, New York, NY, USA, 2012. ACM.

[DCR12] A. Dantcheva, C. Chen, and A. Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 391–398, Sept 2012.

[DD95] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[DDS+09] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[DISB13] S. Du, M. Ibrahim, M. Shehata, and W. Badawy. Automatic license plate recognition (alpr): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):311–325, Feb 2013.

[DMM04] Agné Desolneux, Lionel Moisan, and Jean-Michel Morel. Gestalt theory and computer vision. In *Seeing, Thinking and Knowing*, pages 71–101. Springer, 2004.

[DP15] Sven Dickinson and Zygmunt Pizlo. *Shape perception in human and computer vision*. Springer, 2015.

[EG18] E. Etemad and Q. Gao. Image representation using bag of perceptual curve features. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Dec 2018.

[Eid11] Horst Eidenberger. *Fundamental Media Understanding*. atpress, Vienna, 2011.

[Eid12a] Horst Eidenberger. *Frontiers of Media Understanding*. atpress, Vienna, 2012.

[Eid12b] Horst Eidenberger. *Handbook of Multimedia Information Retrieval*. atpress, Vienna, 2012.

[FJS10] Vittorio Ferrari, Frédéric Jurie, and Cordelia Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010.

[FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997.

[FZB02] H. Foroosh, J. B. Zerubia, and M. Berthod. Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, 11(3):188–200, March 2002.

[FZS12] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, 2012.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. PMLR, 13–15 May 2010.

[GK13] L. Gómez and D. Karatzas. Multi-script text extraction from natural scenes. In *2013 12th International Conference on Document Analysis and Recognition*, pages 467–471, Aug 2013.

[GMK+10] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In S. Konstantopoulos et al., editor, *Artificial Intelligence: Theories, Models and Applications*, volume 6040 of *Lecture Notes in Computer Science*, pages 91–100. Springer Berlin / Heidelberg, 2010.

[GPK18] S. S. Gangonda, P. P. Patavardhan, and K. J. Karande. An extensive survey of prominent researches in face recognition under different conditions. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6, Aug 2018.

[GWM14] RU Gobithaasan, Yip Siew Wei, and Kenjiro T Miura. Log-aesthetic curves for shape completion problem. *Journal of Applied Mathematics*, 2014, 2014.

[GXH18] Qiangliang Guo, Jin Xiao, and Xiaoguang Hu. New keypoint matching method using local convolutional features for power transmission line icing monitoring. *Sensors*, 18:698, 02 2018.

[GXYF06] Y.L. Geng, D. Xu, J.Z. Yuan, and S.H. Feng. Two important action scenes detection based on probability neural networks. *Advances in Neural Networks-ISNN 2006*, pages 448–453, 2006.

[HE13] Markus Hörhan and Horst Eidenberger. New content-based features for the distinction of violent videos and martial arts. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE Press, 2013.

[HE14] Markus Hörhan and Horst Eidenberger. Gestalt interest points for image description in weight-invariant face recognition. In *SPIE Visual Communications Proceedings*. SPIE, 2014.

[HE16] M. Hörhan and H. Eidenberger. An efficient dct template-based object detection method using phase correlation. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 444–448, Nov 2016.

[HE17] Markus Hörhan and Horst Eidenberger. The gestalt interest points distance feature for compact and accurate image description. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Bilbao, Spain, December 2017.

[HE18] Markus Hörhan and Horst Eidenberger. Gestalt interest points with a neural network for makeup-robust face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2391–2395, Oct 2018.

[Hel25] Hermann Helmholtz. *Handbuch der physiologischen Optik*. Leopold Voss, Leipzig, 1925.

[HKP07] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[HOS15a] Takahiro Hayashi, Tatsuya Ooi, and Motoki Sasaki. Contour completion of partly occluded objects based on figural goodness. *International Journal of Networked and Distributed Computing*, 3(3):185–192, 2015.

[HOS15b] Takahiro Hayashi, Tatsuya Ooi, and Motoki Sasaki. Decomposition of partly occluded objects based on evaluation of figural goodness. In *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*, pages 307–312. IEEE, 2015.

[HSH07] Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu. Movie classification using visual effect features. In *Signal Processing Systems, 2007 IEEE Workshop on*, pages 295 –300, oct. 2007.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[JC12] Hervé Jégou and Ondřej Chum. Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *ECCV*, 2012.

[JDS08] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[JDSP10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, June 2010.

[KABN12] Sanjeev Kumar, Haleh Azartash, Mainak Biswas, and Truong Nguyen. Real-time affine global motion estimation using phase correlation and its application for digital image stabilization. *Image Processing, IEEE Transactions on*, 20:3406 – 3418, 01 2012.

[Kan13] E. Kandel. *Principles of Neural Science, Fifth Edition*. Principles of Neural Science. McGraw-Hill Education, 2013.

[KB14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[Kof35] K. Koffka. *Principles of Gestalt Psychology*. Lund Humphries / London, 1935.

[Kor] Vlad Korolev. Vlad's blog. `http://dovgalecs.com/blog/freak-descriptor-in-matlab/`. last visited on June 12th 2014.

[Kos94] Stephen M. Kosslyn. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA, USA, 1994.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[KYK06] Sungho Kim, Kuk-Jin Yoon, and In So Kweon. Object recognition using a generalized robust invariant feature and gestalt's law of proximity and similarity. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 193–193, June 2006.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[LCS11] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2548–2555, Washington, DC, USA, 2011. IEEE Computer Society.

[LKZT16] Ying Li, Xiangwei Kong, Liang Zheng, and Qi Tian. Exploiting hierarchical activations of neural network for image retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 132–136, New York, NY, USA, 2016. ACM.

[LLH16] Li Jun Tao, Liu Yin Hong, and Hao Yan. The improvement and application of a k-means clustering algorithm. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 93–96, July 2016.

[LLH18] Mingchen Li, Yiyang Li, and Yifan He. Makeup removal system with deep learning. *CS230*, 2018.

[LLZ+08] A. Liu, J. Li, Y. Zhang, S. Tang, Y. Song, and Z. Yang. An innovative model of tempo and its application in action scene detection for movie analysis. In *Applications of Computer Vision, 2008. IEEE Workshop on*, pages 1–6. IEEE, 2008.

[LOM04] B. Lehane, N.E. O'connor, and N. Murphy. Action sequence detection in motion pictures. In *The international Workshop on Multidisciplinary Image, Video, and Audio Retrieval and Mining*, 2004.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[LOW+18] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey, 2018.

[LSD+17] B. Liu, J. Sheng, J. Dun, S. Zhang, Z. Hong, and X. Ye. Locating various ship license numbers in the wild: An effective approach. *IEEE Intelligent Transportation Systems Magazine*, 9(4):102–117, winter 2017.

[LSW+18] Yi Li, Lingxiao Song, Xiang Wu, Ran He, and Tieniu Tan. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. In *AAAI*, 2018.

[LW09] J. Lin and W. Wang. Weakly-supervised violence detection in movies with audio and video based co-training. *Advances in Multimedia Information Processing-PCM 2009*, pages 930–935, 2009.

[LZL+19] C. Leng, H. Zhang, B. Li, G. Cai, Z. Pei, and L. He. Local feature descriptor for image matching: A survey. *IEEE Access*, 7:6424–6434, 2019.

[Mar82] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[Mat] Johan Mathe. Shotdetect. `http://shotdetect.nonutc.fr/`. last visited on November 8th 2012.

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 36.1–36.10. BMVA Press, 2002. doi:10.5244/C.16.36.

[ML14] M. Muja and D. G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2227–2240, Nov 2014.

[MM19] Eckart Michaelsen and Jochen Meidow. Hierarchical perceptual grouping for object recognition. In *Advances in Computer Vision and Pattern Recognition*, pages 1–5, 2019.

[Mol93] Martin F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *NEURAL NETWORKS*, 6(4):525–533, 1993.

[Myr] Andriy Myronenko. Medical image registration toolbox (mirt). `https://sites.google.com/site/myronenko/software`. last visited on April 22th 2016.

[NA19] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019.

[New92] Fiona N. Newell. Perceptual recognition of familiar objects in different orientations. *Doctoral thesis*, 1992.

[NMP78] Madihally Narasimha and Allen M. Peterson. On the computation of the discrete cosine transform. *Communications, IEEE Transactions on*, 26:934 – 936, 07 1978.

[NPBP18] Tien Dat Nguyen, Tuyen Danh Pham, Na Rae Baek, and Kang Ryoung Park. Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors. In *Sensors*, 2018.

[NS06] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168, June 2006.

[NYD15] Joe Yue-Hei Ng, Fan Yang, and Larry S. Davis. Exploiting local features from deep networks for image retrieval, 2015.

[OA70] R. K. Olson and F. Attneave. What variables produce similarity-grouping. *American Journal of Psychology*, 83:1–21, 1970.

[OLH+12] Toshiyuki Okada, Marius George Linguraru, Masatoshi Hori, Yuki Suzuki, Ronald M Summers, Noriyuki Tomiyama, and Yoshinobu Sato. Multi-organ segmentation in abdominal ct images. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3986–3989. IEEE, 2012.

[OMB14] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187 – 1200, Jun 2014. Preprint.

[Ort12] Raphael Ortiz. Freak: Fast retina keypoint. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 510–517, Washington, DC, USA, 2012. IEEE Computer Society.

[OT06] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[PCI+07] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[PHK+18] K. Pitstick, J. Hansen, M. Klein, E. Morris, and J. Vazquez-Trejo. Applying video summarization to aerial surveillance. In Michael A. Kolodny, Dietrich M. Wiegmann, and Tien Pham, editors, *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, volume 10635, pages 55 – 64. International Society for Optics and Photonics, SPIE, 2018.

[PP15] A. Pandey and U. C. Pati. An improved dct-based phase correlation method for image mosaicing. In *2015 Third International Conference on Image Information Processing (ICIIP)*, pages 265–270, Dec 2015.

[PSM10] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[PVZ15a] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[PVZ+15b] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[Pyl02] Zenon W. Pylyshyn. Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25(2):157–182, 2002.

[QWCF16] Shi Qiu, Desheng Wen, Ying Cui, and Jun Feng. Lung nodules detection in ct images using gestalt-based algorithm. *Chinese Journal of Electronics*, 25:711–718(7), July 2016.

[Red] Reddit. Diet progress pictures. `http://www.reddit.com/r/progresspics/`. last visited on May 28th 2014.

[RRKB11] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[SC15] I. C. Shen and W. H. Cheng. Gestalt rule feature points. *IEEE Transactions on Multimedia*, 17(4):526–537, April 2015.

[SCVJA10] Fillipe D. M. de Souza, Guillermo C. Chavez, Eduardo A. do Valle Jr., and Arnaldo de A. Araujo. Violence detection in video using spatio-temporal features. In *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, SIBGRAPI '10, pages 224–230, Washington, DC, USA, 2010. IEEE Computer Society.

[Sej18a] Terrence J. Sejnowski. *The Deep Learning Revolution*, pages 28–30. MIT Press, Cambridge, MA, 2018.

[Sej18b] Terrence J. Sejnowski. *The Deep Learning Revolution*. MIT Press, Cambridge, MA, 2018.

[Sej18c] Terrence J. Sejnowski. *The Deep Learning Revolution*, page 32f. MIT Press, Cambridge, MA, 2018.

[SH11] R. Sorschag and M. Hörhan. Action scene detection from motion and events. In *2011 18th IEEE International Conference on Image Processing*, pages 3641–3644, Sep. 2011.

[SHM+16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016.

[SKA+15] V. Sharma, A. Kumar, N. Agrawal, P. Singh, and R. Kulshreshtha. Image summarization using topic modelling. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 226–231, Oct 2015.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[SLWT15a] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015.

[SLWT15b] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[SOPP18] Luiz Souza, Luciano Oliveira, Mauricio Pamplona, and Joao Papa. How far did we get in face spoofing detection? *Engineering Applications of Artificial Intelligence*, 72:368 – 381, 2018.

[Sor] Robert Sorschag. Action detection groundtruth. `www.ims.tuwien.ac.at/sor/ActionDetectionGroundTruth.zip`. last visited on May 28th 2011.

[Sor11] R. Sorschag. Cori: A configurable object recognition infrastructure. In *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, pages 138 –143, nov. 2011.

[SPP16] Manasa Srinivas, M.M. Manohara Pai, and Radhika M. Pai. An improved algorithm for video summarization – a rank based approach. *Procedia Computer Science*, 89:812 – 819, 2016.

[SQX+09] Suhuai Luo, Qingmao Hu, Xiangjian He, Jiaming Li, J. S. Jin, and M. Park. Automatic liver parenchyma segmentation from abdominal ct images using support vector machines. In *2009 ICME International Conference on Complex Medical Engineering*, pages 1–5, April 2009.

[SRM+19] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019.

[SS02] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press / Cambridge, MA, 2002.

[SS20] Anurag Singh and Deepak Kumar Sharma. *Image Collection Summarization: Past, Present and Future*, pages 49–78. Springer International Publishing, Cham, 2020.

[SSKG20] Himanshu Shekhar, Sujoy Seal, Saket Kedia, and Amartya Guha. *Survey on Applications of Machine Learning in the Field of Computer Vision*, pages 667–678. Springer, Singapore, 01 2020.

[SYE16] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2Nd International Workshop on Multimedia Assisted*

Dietary Management, MADiMa '16, pages 3–11, New York, NY, USA, 2016. ACM.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sze19] R Szeliski. Computer vision: Algorithms and applications. *Instructor*, 201901, 2019.

[TAJ13] G. Tolias, Y. Avrithis, and H. Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *2013 IEEE International Conference on Computer Vision*, pages 1401–1408, Dec 2013.

[TSJ15] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[TSOP15] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, Nov 2015.

[VF08] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, Dec 2001.

[WCM+13] Robin Wolz, Chengwen Chu, Kazunari Misawa, Michitaka Fujiwara, Kensaku Mori, and Daniel Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE transactions on medical imaging*, 32(9):1723–1730, 2013.

[Wer23] Max Wertheimer. Untersuchungen zur Lehre von der Gestalt. ii. *Psychologische Forschung*, 4(1):301–350, Jan 1923.

[WZLQ16] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[YLYH07] Linjun Yang, Jiemin Liu, Xiaokang Yang, and Xian-Sheng Hua. Multimodality web video categorization. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07, pages 265–274, New York, NY, USA, 2007. ACM.

[YRS+18] Yijun Yan, Jinchang Ren, Genyun Sun, Huimin Zhao, Junwei Han, Xuelong Li, Stephen Marshall, and Jin Zhan. Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognition*, 79:65 – 78, 2018.

[YUG+13] Victoria Yanulevskaya, Jasper Uijlings, Jan-Mark Geusebroek, Nicu Sebe, and Arnold Smeulders. A proto-object-based computational model for visual saliency. *Journal of vision*, 13(13):27–27, 2013.

[YWX17] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336, 2017.

[YXGS16] J. Yu, G. Xia, C. Gao, and A. Samal. A computational model for object-based visual saliency: Spreading attention along gestalt cues. *IEEE Transactions on Multimedia*, 18(2):273–286, Feb 2016.

[Zen17] Y. Zeng. Decomposition and construction of object based on law of closure in gestalt psychology. In *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pages 1–4, Oct 2017.

[ZK17] Zhenzhu Zheng and Chandra Kambhamettu. Multi-level Feature Learning for Face Recognition under Makeup Changes. In *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 918–923, 2017.

[ZYT18] L. Zheng, Y. Yang, and Q. Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, May 2018.

[ZZH16] W. Zhao, Z. Zhang, and K. Huang. Joint crowd detection and semantic scene modeling using a gestalt laws-based similarity. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1220–1224, Sep. 2016.

[ZZW+16] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in CNN feature transfer. *CoRR*, abs/1604.00133, 2016.