Pattern Analysis and Applications
https://doi.org/10.1007/s10044-020-00904-6

THEORETICAL ADVANCES

Gestalt descriptions for deep image understanding

Markus Hörhan[1] · **Horst Eidenberger**[1]

Received: 10 August 2018 / Accepted: 11 June 2020

Abstract
In this work, we present a novel visual perception-inspired local description approach as a preprocessing step for deep learning. With the ongoing growth of 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, e.g., 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 makeup-robust face recognition.

Keywords Image analysis · Deep learning-based methods · Gestalt descriptors · Image classification · Face recognition · Person identification

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 [25]. 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 preprocessing step for deep learning. To the best of our knowledge there are no other scientific works about utilizing Gestalt laws to preprocess images for deep learning until now. The experiments in Sects. 3.2 and 3.3 were made to test the fundamental idea, different parameterization and some variations of our method. We came to the conclusion that it outperforms all of the baseline local description methods to which we compared it. However, the experiments of Sect. 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 overfitting for our application domain of makeup-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

✉ Markus Hörhan
markushoerhan@gmail.com

Horst Eidenberger
horst.eidenberger@tuwien.ac.at

[1] Institute of Visual Computing and Human-Centered Technology, Vienna University of Technology, Favoritenstrasse 11, 1040 Vienna, Austria

Published online: 01 July 2020

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 are 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 by backpropagation 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 [7]. A few years later the AlphaGo system [42] 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 preprocessing 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. [19]. 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 shortcuts to jump over some layers. The motivation for skipping over layers is to avoid the problem of vanishing gradients [17], by reusing information as residuals from a previous layer until the layer next to the current one has learned its weights.

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.

Marr [32] 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 [20] 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 drawing conclusions from incomplete data, based on previous experience.

Gestalt psychology [28] 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. In particular, 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 Fig. 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 Sects. 1.2 and 1.3, and then we provide a comprehensive

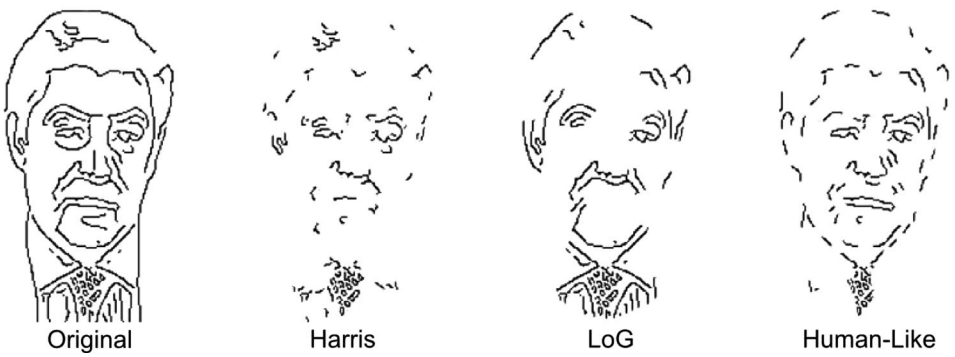


Fig. 1 The left edge map of a face is represented by points from a Harris corner detector and a Laplacian of Gaussian (LoG) operator. Many of the important face features (e.g., the eyes) are thereby lost. The rightmost image shows the GIP description. It preserves the perceptual features of the original stimulus well, but does not produce a longer description than the LoG operator [14]

overview of the Gestalt interest points (GIPs) algorithm in Sect. 2. The details of our Gestalt regions of interest (GROI) method are presented in Sect. 2.2.2. Experimental results are analyzed in Sect. 3, and conclusions are finally given in Sect. 4.

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. [47] 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 [12, 13]. In [41] 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 [4] 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 [26] proposed a context-based method for object recognition inspired by the Gestalt Laws of Proximity and Similarity. Qiu et al. [38] 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 [48] 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 [34] 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. [18] 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 [2] 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.

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 Sect. 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 [21–24] GIP feature which defines the fundamental basis of the GROIs. It can be used 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. (3) Additionally, we show our experimental results on various forensic application domains in Sects. 3.2 and 3.3, which can be also found in previously published material [22, 23].

2 Proposed approach

In this section we provide a detailed overview over the Gestalt interest of points (GIPs) algorithm [23]. 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.

2.1 Gestalt Interest Points Detection

The theories of visual perception mentioned in Sect. 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 Fig. 2. After the input image is converted to gray scale (Fig. 2a), the image gradient vectors are calculated (Fig. 2b). The gradient image is split into m by n (e.g., 16×16) macroblocks, 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.

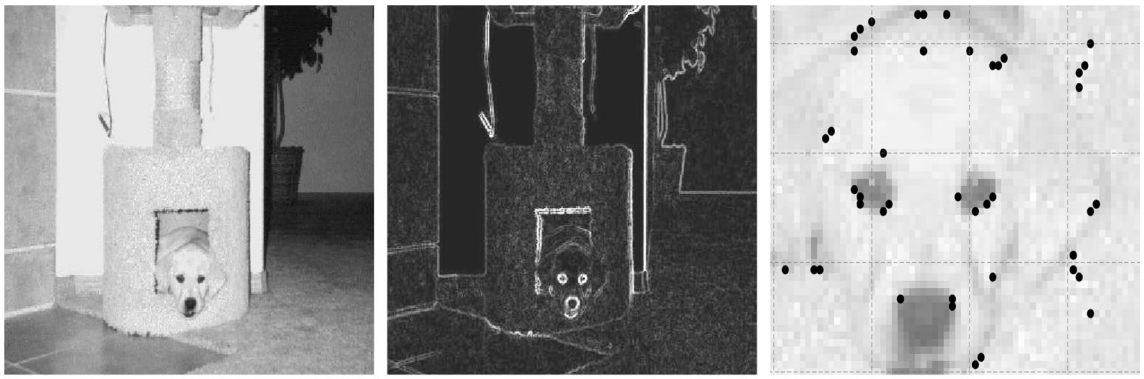


Fig. 2 GIP algorithm overview

According to this assumption, we assume that low-contrast image macroblocks 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 [35] 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 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 (Fig. 2c).

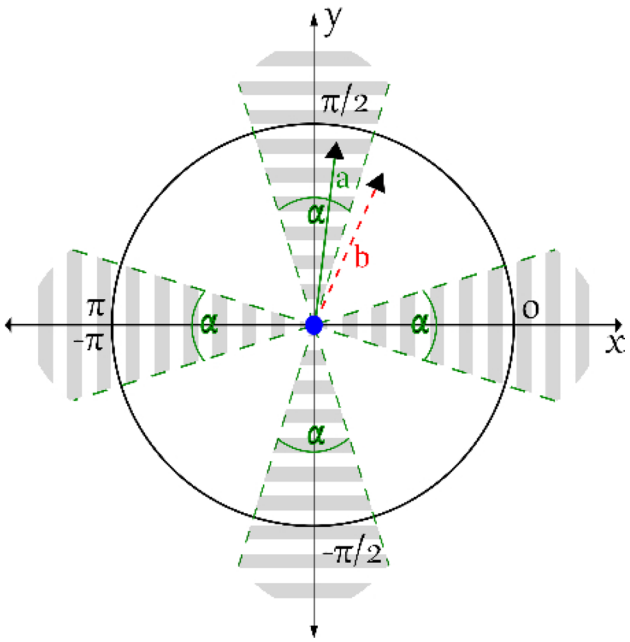


Fig. 3 The point in the origin indicates a GIP and vector **a** its gradient, which is within one of the four circle segments. In this case, the GIP will be accepted as an interest point. If vector **b** was the gradient of this GIP, the GIP would be discarded because its underlying edge has diagonal orientation

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:

$$F = \left(m_1\ m_2\ m_3\ p_1\ p_2\ p_3\ o_1\ o_2\ o_3\right) \tag{1}$$

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 macroblock. 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 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 pseudocode 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

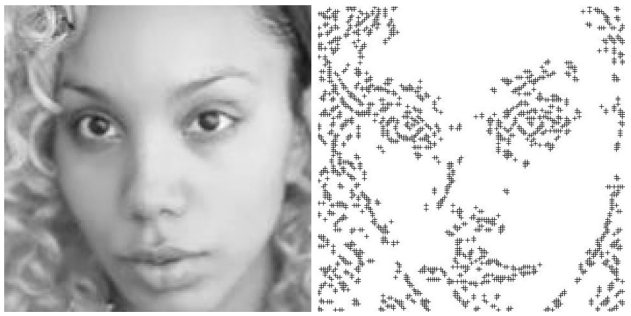


Fig. 4 A face image on the left and its GIP representation on the right. The GIP algorithm is fast and highly effective. Because it is inspired by cognition, it extracts very little but well-selected image information

2.2.1 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 5 visualizes this concept, and the result is one triangle within every image block.

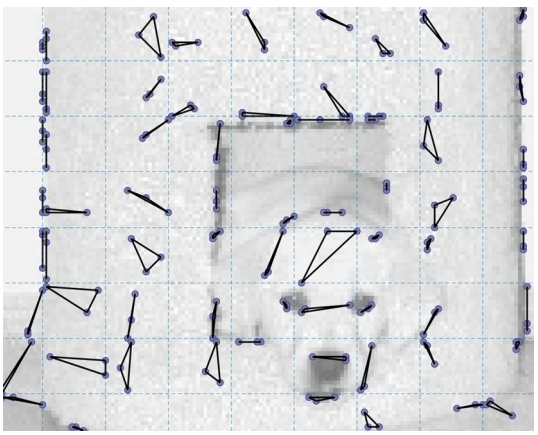


Fig. 5 Inter-GIP distances

Since a variety of different distance functions does 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 Eqs. (2)–(7).

$$D_{\text{Chebychev}} = \max(|x_2 - x_1|, |y_2 - y_1|) \tag{2}$$

$$D_{\text{Cityblock}} = |x_2 - x_1| + |y_2 - y_1| \tag{3}$$

$$D_{\text{Cosine}} = 1 - \frac{P'Q}{\sqrt{(P'P)(Q'Q)}} \tag{4}$$

$$D_{\text{Euclidean}} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

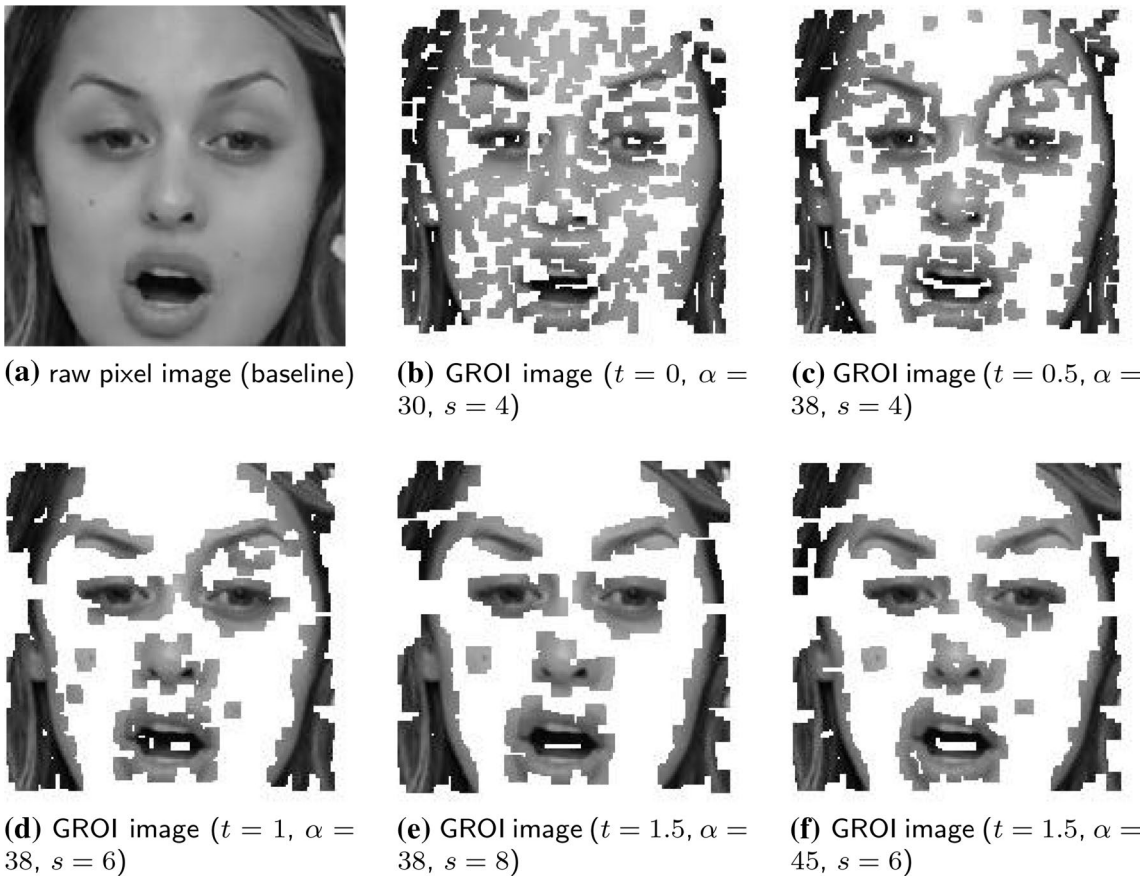
$$D_{\text{Minkowski}} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{6}$$

$$D_{\text{Jaccard}} = 1 - \frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} max(x_i, y_i)} \tag{7}$$

where $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ representing two points in the two-dimensional space. Minkowski distance (6) can be considered a generalization of three other distances, the Euclidean if $p = 2$, the Cityblock if $p = 1$ and the Chebyshev distance if $p = \infty$. For the experiment in Sect. 3.3 we defined $p = 3$. Please note that (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 [15].

2.2.2 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 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



(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 ($t = 1.5$, $\alpha = 45$, $s = 6$)

Fig. 6 An example face image and its Gestalt region of interest (GROI) image representations. Each GROI image was produced with different parameter combinations

follows: in the first step the GIPs are detected in the input image as described in Sect. 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-overfitting drastically. Figure 6 shows various GROI example images produced with different GIP parameters t and α (see Sect. 2.1 for detailed explanation) and GROI parameter s.

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.

3.1 Overview of experiments

In Sect. 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 [22]. As described in Sect. 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 Sect. 3.3 present the GIP-IGD algorithm applied on two different image classification tasks

[23]. 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 Sect. 2.2.1, 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 are presented in Sect. 3.4. As demonstrated in the experiments of Sect. 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.

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 [22].

We assumed that the ability of the GIP algorithm to select interest points within high-contrast image blocks and on non-diagonal edges (Sect. 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.

3.2.1 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 photographs for a group of fifteen persons who underwent significant weight change (at least 20 kg) in less than 1 year. The majority of the photographs were taken from a diet web forum [39]. Others were provided by acquaintances of the authors.

3.2.2 Experimental setup

Five local feature description methods were chosen for comparison with the GIP feature: SIFT [31], SURF [3], MSER [33], FREAK [36] and ORB [40]. After feature extraction with one of the above methods, we received multiple feature

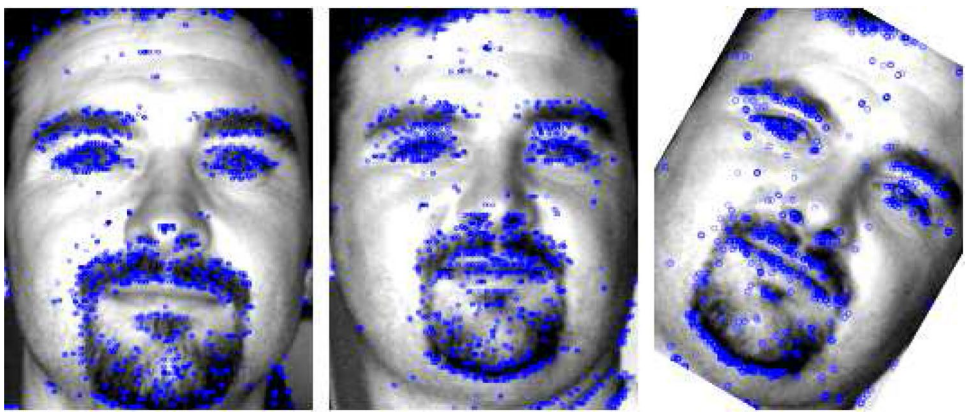


Fig. 7 The GIP algorithm was applied on pictures such as these examples. GIPs which are within low-contrast macroblocks, and many of the GIPs on diagonal edges were discarded. Left: the image shows a normal weight person and the detected GIP points, indicated as circles. Middle: shows the same person after 30 kg of weight gain and the detected GIP points. Right: shows the person image rotated by 30° and the detected GIP points

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 [9]. Each of our images was 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 photographs that show the higher weight. Figure 7 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. 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 photographs in the dataset are left in their original resolutions, ranging from 201×285 to 508×728 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.

3.2.3 Evaluation

The second column of Table 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%. However, to obtain this result an average

Table 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	8883
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

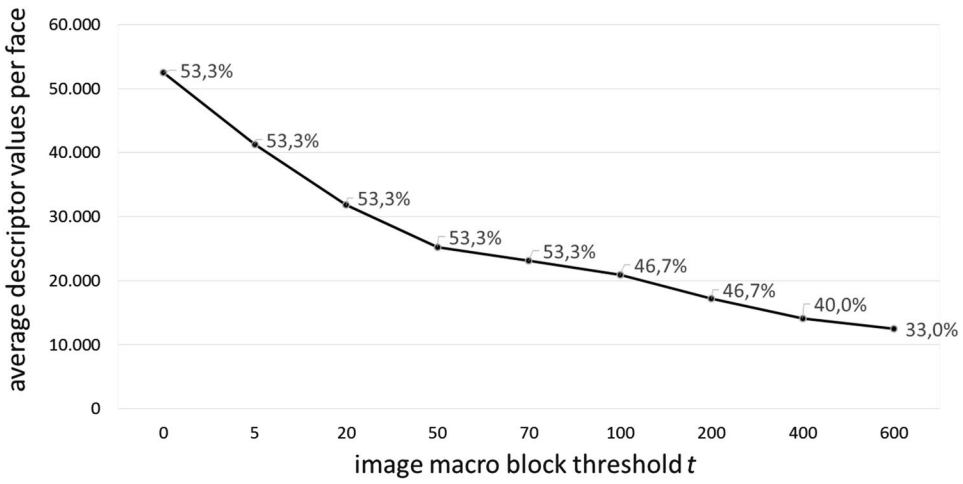


Fig. 8 The average number of description values per face and categorization accuracy in percent as a function of image macroblock variance threshold t. A value of $t = 0$ means that no image macroblocks are excluded from the recognition process

of 57,984 description values per face (last column) is 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% which is far below an acceptable rate for practical application.

Compared to the five baseline features above, the GIP description algorithm is by far more accurate in its original form with 53.3%. This performance is 20% ahead of the SURF algorithm that leads the baseline description methods. As Fig. 8 shows, discarding interest points that lie within low-contrast macroblocks 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 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%. 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 10 illustrates the behavior of the algorithm. An accuracy of 46.7% 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

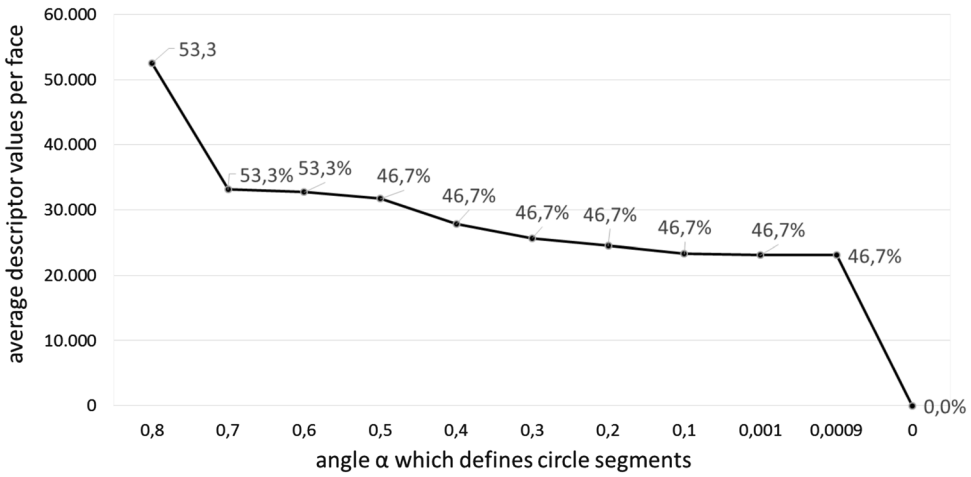


Fig. 9 The average description values per face and categorization accuracy in percent as a function of circle segment angle α. A value of $\alpha = 0.8$ radians means that no GIPs are discarded. If there are no perfect straight lines in the face image dataset, then for $\alpha = 0$ the accuracy will drop to zero

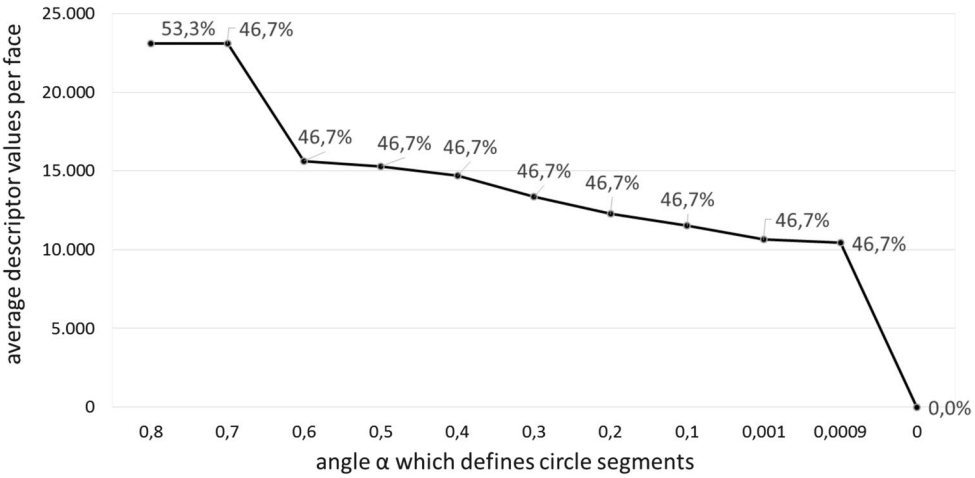


Fig. 10 The average number of description values per face and categorization accuracy in percent as a function of circle segment angle α with $t = 70$

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.

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°. The third column of Table 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 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 Sect. 3.4.

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 Sect. 2.2.1. 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 [23].

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 Sect. 3.4.

3.3.1 Datasets

In our first evaluation task, we tested the detection performance with the INRIA Horses dataset [16], 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 [44]. It consists of 2500 food images, which cover a wide variety of food items and 2500 randomly selected non-food images. Some food images also contain other objects or people. The Food-5K dataset with a total size of 5000 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.

3.3.2 Experimental setup

We compared our method to several different local feature description algorithms, namely SIFT [31], SURF [3], BRISK [29] and FREAK [36]. 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 [9] 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.

3.3.3 Evaluation

The experimental results of applying our algorithm on the INRIA Horses dataset are shown in Fig. 11, and the results for the Food-5K dataset are presented in Fig. 12. Figures 11a and 12a depict the F_1-scores over extraction time of our IGD

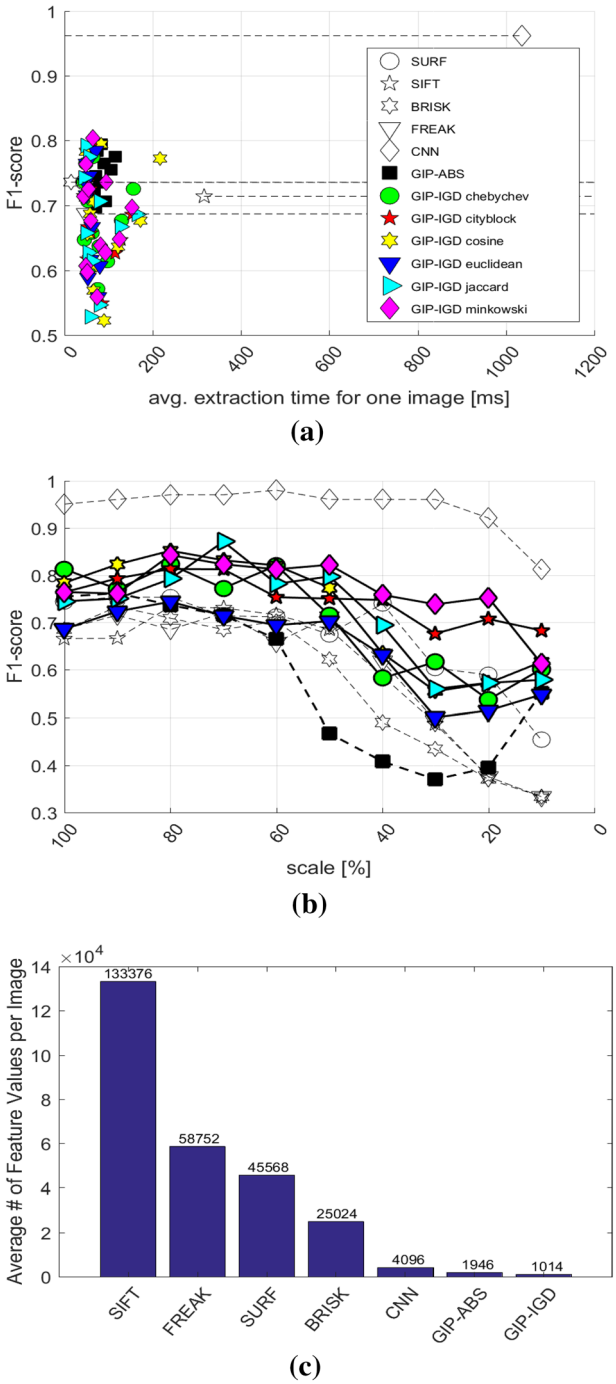


Fig. 11 The experimental results of applying our algorithm on the INRIA Horses dataset. Our algorithm is also compared to several different baseline methods. The different F_1-scores for each IGD distance measure in Fig. 11a arise through adjusting the two GIP parameters t and α, which are described in Sect. 2.1

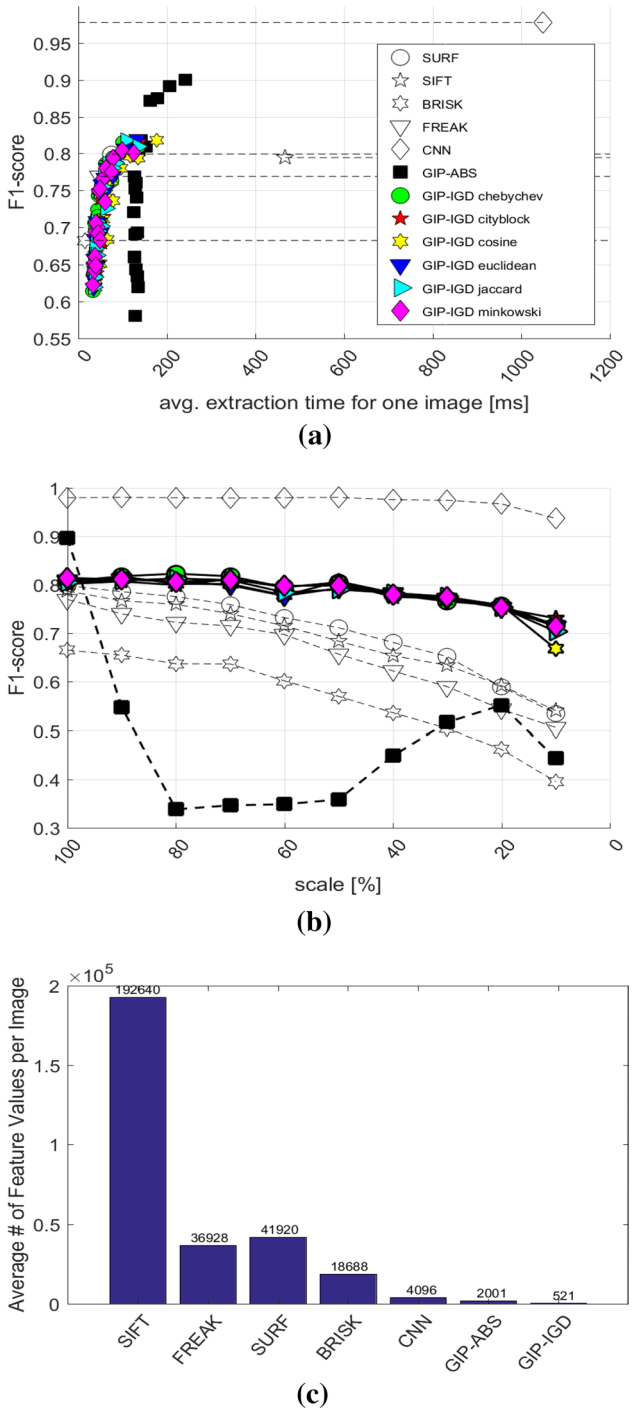


Fig. 12 The experimental results of applying our algorithm on the Food-5K dataset. Our algorithm is also compared to several different baseline methods. The different F_1-scores for each IGD distance measure in Fig. 12a arise through adjusting the two GIP parameters t and α, which are described in Sect. 2.1

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 11b and 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. In particular, 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 11c and 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 11 and 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 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.

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

Fig. 13 Before (top line) and after (bottom line) makeup examples of four subjects contained in our makeup dataset

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. [11] 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 Fig. 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. [10] addressed this problem with a patch-based ensemble learning method. Song et al. [30] 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. [49] 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 [37, 45, 46]. 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 6000 non-makeup face images of 6 subjects serve as input. Henceforth, the classification stage assigns each of the 1200 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.

3.4.1 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 [24]. This subset consists of 6 subjects with 1000 non-makeup face images per subject for training and 200 makeup images per subject for testing. Figure 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.

3.4.2 Experimental setup

We implemented a prototype for this experiment utilizing Python in combination with the machine learning framework Tensorflow [1] and the high-level neural networks API Keras [6]. The structure of the chosen CNN model is shown in Table 2. It is an adapted version of the VGG-like model from the Keras website. VGGNet [43] was invented by VGG (Visual Geometry Group) from University of Oxford. According to VGGNet we also use filters of size 3×3 because smaller filters generally provide better results. The number of layers was chosen to satisfy our requirements. On the one hand, we

Table 2 Structure of the adapted example CNN model from Keras website [5]

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)

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 [27] 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% of the training images. To avoid long training times and possible overfitting 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 to CNN training. For our experiments we decided to run them on Crestle [8]. The Crestle servers are equipped with NVIDIA Tesla K80 GPUs, and therefore, they have been considered adequate for our purposes.

3.4.3 Evaluation

Figure 14 shows the training accuracies for each epoch over the training period, and Fig. 15 the validation accuracies, respectively. See Sect. 2.2.2 for a detailed explanation of the parameters t, α and s. As mentioned above the validation set comprises 10% 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 Fig. 6. As can be seen in Figs. 14 and 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 makeup images from the test set. Each line marker in Fig. 16 denotes an accuracy produced using one of these stored models. After 30 training

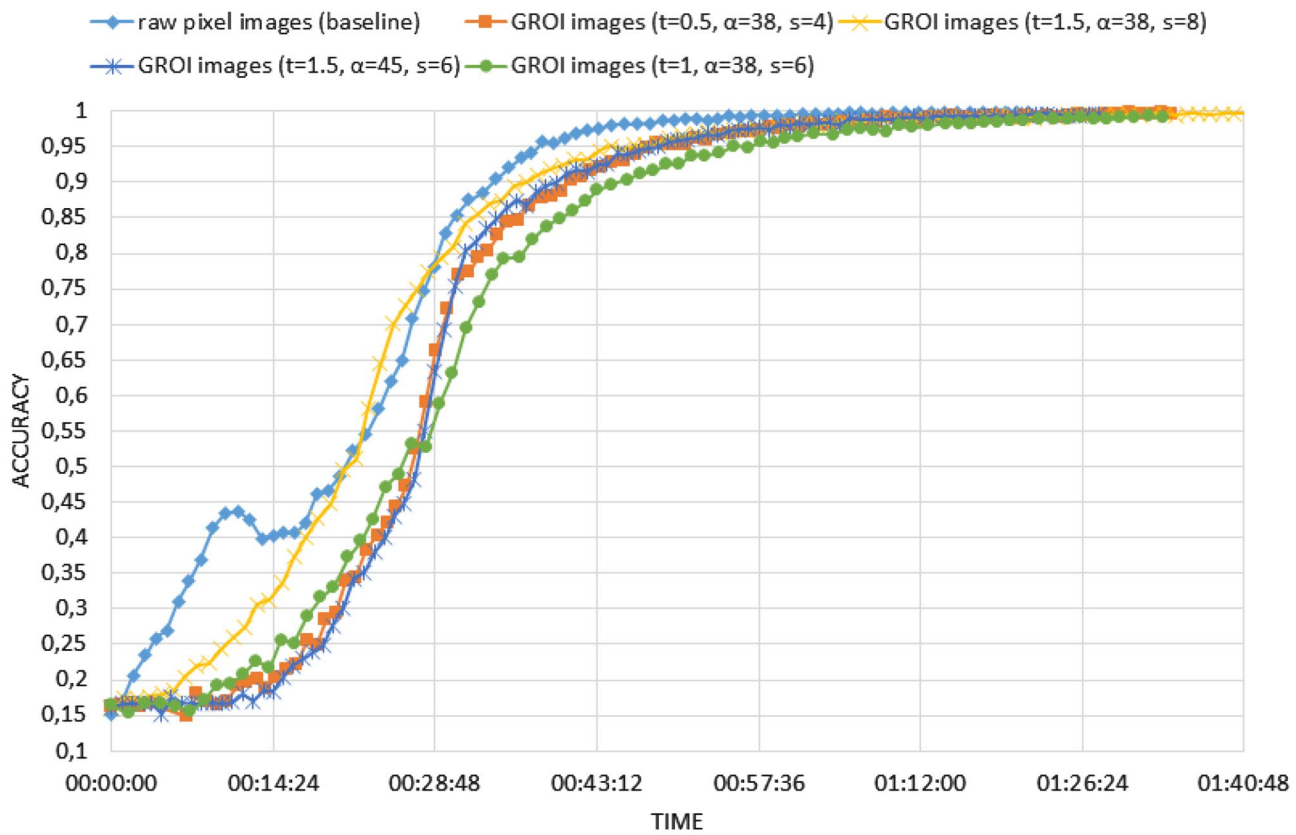


Fig. 14 Train accuracies for each epoch of the training process. Each line marker denotes one train epoch

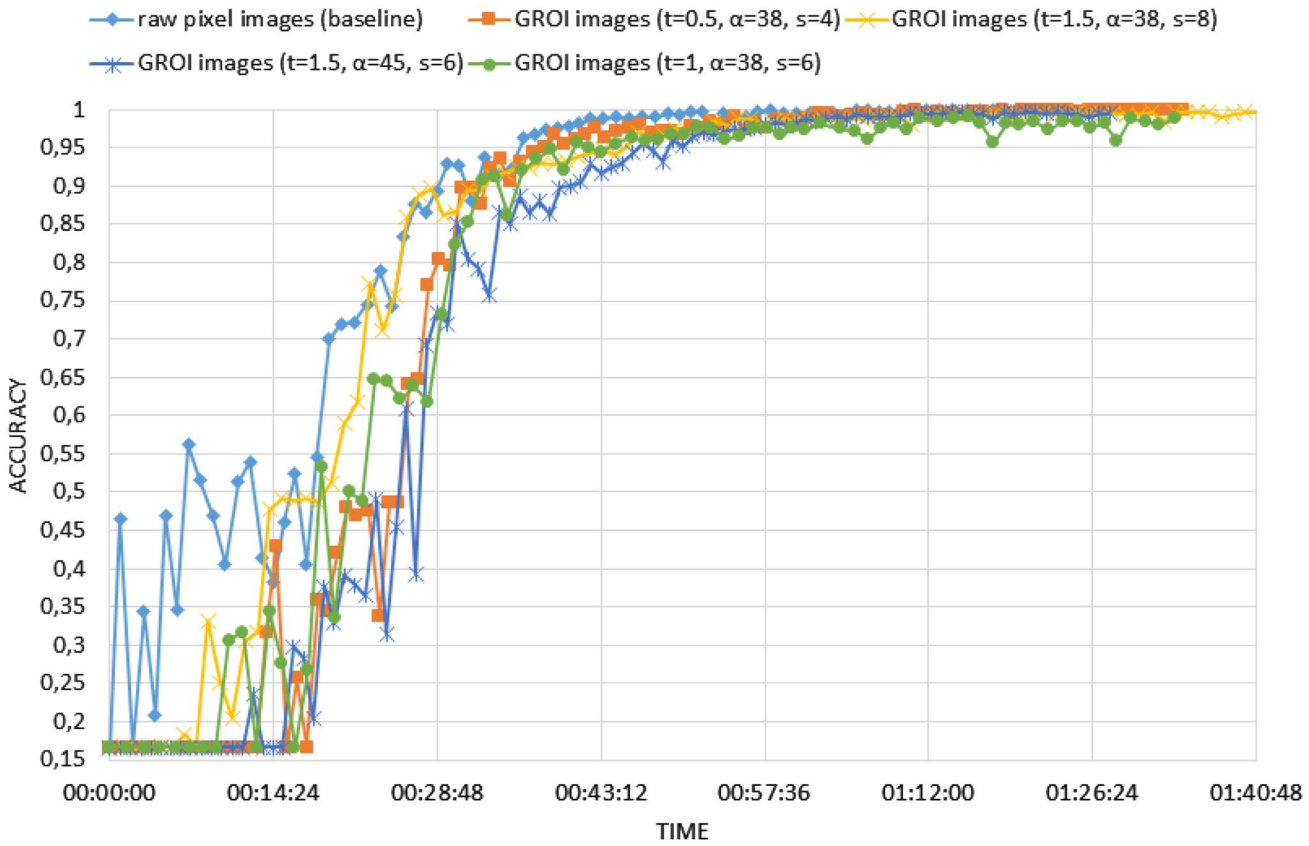


Fig. 15 Validation accuracies for each epoch of the training process. Each line marker denotes one train epoch

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

Figure 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 by 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° 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 by GROI images leads to slower training convergence in comparison with the CNN trained by raw pixel images. This fact in combination with the high test accuracies proves that our presented method is more robust against overfitting 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 background, thus requiring less disk space. This is a very important argument in big data domains such as face recognition.

4 Conclusion

In this work, we proposed a novel visual perception-inspired local description approach as a preprocessing 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 preprocessed 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 overfitting. Another advantage of the GROI approach is that it is

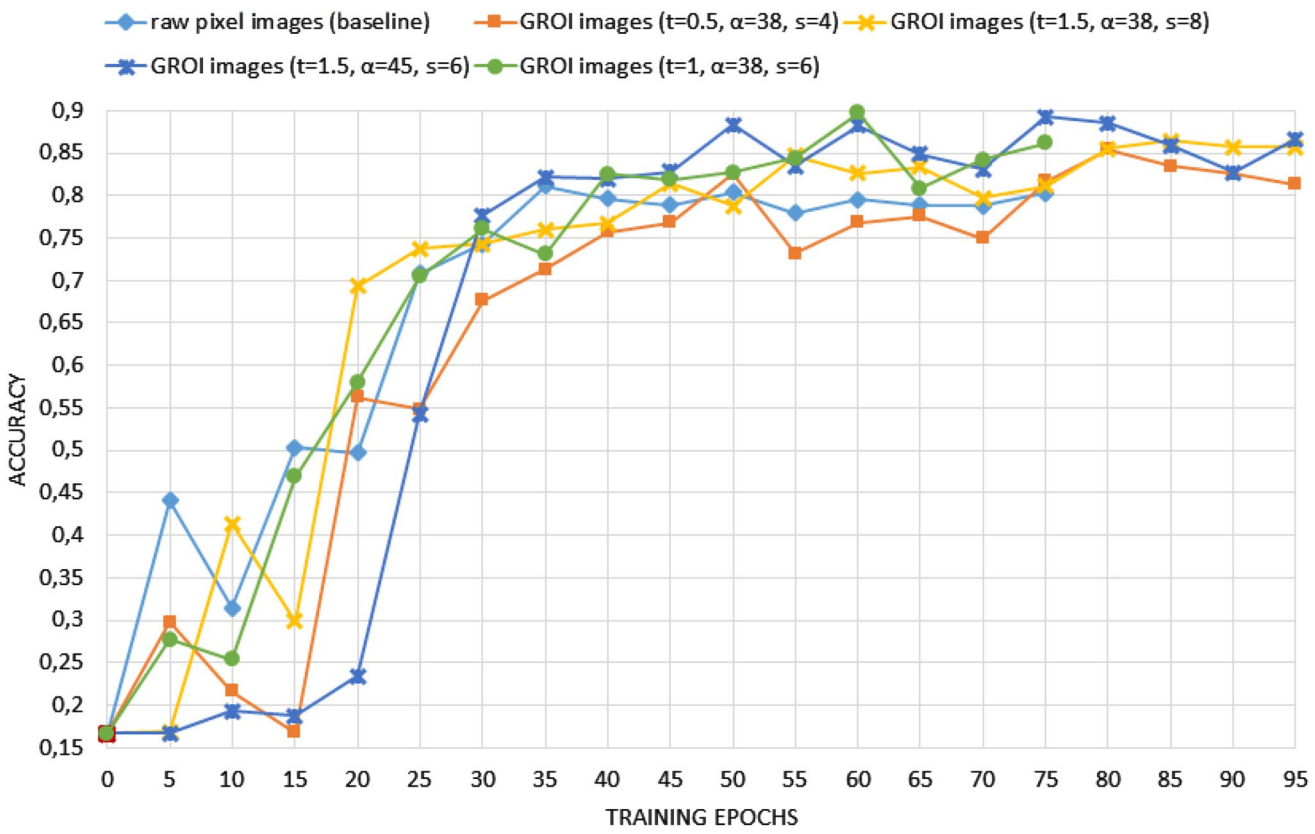


Fig. 16 Test accuracies for stored train models. For every fifth training epoch the resulting model was stored during the training process. These models were used to classify the makeup test images. Each line marker denotes an accuracy produced using one of these stored models

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. [10] reached a Rank-1 accuracy of 89.40% 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. [30] delivers up to 94.8% Rank-1 accuracy applied on three different datasets. Zheng et al. [49] proposed a new hierarchical feature learning framework and achieved an accuracy up to 81.11% with two different datasets. As we showed in our experiments, with our method an accuracy of 89.8% 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.

Acknowledgements Open access funding provided by TU Wien (TUW).

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016)Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on operating systems design and implementation, OSDI'16. USENIX Association, Berkeley, CA, USA, pp 265–283. http://dl.acm.org/citation.cfm?id=3026877.3026899
2. Al-Shabi M, Cheah WP, Tee C (2017) Facial expression recognition using a hybrid cnn-sift aggregator. In: MIWAI
3. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (surf). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014
4. Bileschi S, Wolf L (2007) Image representations beyond histograms of gradients: the role of gestalt descriptors. In: 2007 IEEE conference on computer vision and pattern recognition, pp 1–8. https://doi.org/10.1109/CVPR.2007.383122
5. Chollet F (2018) Keras model examples (last visited on August 5th 2018). https://keras.io/getting-started/sequential-model-guide/
6. Chollet F et al (2015) Keras. https://keras.io
7. Cireşan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the twenty-second International Joint Conference on Artificial Intelligence, IJCAI'11, vol 2. AAAI Press, pp 1237–1242. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-210
8. Crestle: Crestle effortless infrastructure for deep learning (last visited on June 21th 2018). https://www.crestle.com/
9. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, pp 1–22
10. Cunjian C, Dantcheva A, Ross A (2015) An ensemble of patch-based subspaces for makeup-robust face recognition. Inf Fusion. https://doi.org/10.1016/j.inffus.2015.09.005
11. Dantcheva A, Chen C, Ross A (2012) Can facial cosmetics affect the matching accuracy of face recognition systems? In: 2012 IEEE fifth international conference on biometrics: theory, applications and systems (BTAS), pp 391–398. https://doi.org/10.1109/BTAS.2012.6374605
12. Desolneux A, Moisan L, Morel JM (2004) Gestalt theory and computer vision. In: Seeing, thinking and knowing. Springer, pp 71–101
13. Dickinson S, Pizlo Z (2015) Shape perception in human and computer vision. Springer, Berlin
14. Eidenberger H (2011) Fundamental media understanding. atpress, Vienna
15. Eidenberger H (2012) Handbook of multimedia information retrieval. atpress, Vienna
16. Ferrari V, Jurie F, Schmid C (2010) From images to shape models for object detection. Int J Comput Vis 87(3):284–303. https://doi.org/10.1007/s11263-009-0270-9
17. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Teh YW, Titterington M (eds) Proceedings of the thirteenth international conference on artificial intelligence and statistics, Proceedings of Machine Learning Research, vol 9. PMLR, pp 249–256 http://proceedings.mlr.press/v9/glorot10a.html
18. Guo Q, Xiao J, Hu X (2018) New keypoint matching method using local convolutional features for power transmission line icing monitoring. Sensors 18:698
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR. IEEE Computer Society, pp 770–778
20. Helmholtz H (1925) Handbuch der physiologischen Optik. Leopold Voss, Leipzig
21. Hörhan M, Eidenberger H (2013) 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
22. Hörhan M, Eidenberger H (2014) Gestalt interest points for image description in weight-invariant face recognition. In: SPIE Visual Communications Proceedings. SPIE
23. Hörhan M, Eidenberger H (2017) 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
24. Hörhan M, Eidenberger H (2018) Gestalt interest points with a neural network for makeup-robust face recognition. In: 2018 25th IEEE international conference on image processing (ICIP), pp 2391–2395. https://doi.org/10.1109/ICIP.2018.8451075
25. Kandel E (2013) Principles of neural science, 5 edn. Principles of neural science. McGraw-Hill Education. https://books.google.at/books?id=s64z-LdAIsEC
26. Kim S, Yoon KJ, Kweon IS (2006) 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), pp 193–193. https://doi.org/10.1109/CVPRW.2006.146
27. Kingma D, Ba J (2014) Adam: a method for stochastic optimization
28. Koffka K (1935) Principles of gestalt psychology. Lund Humphries, London
29. Leutenegger S, Chli M, Siegwart RY (2011) Brisk: binary robust invariant scalable keypoints. In: Proceedings of the 2011 International Conference on Computer Vision, ICCV '11. IEEE Computer Society, Washington, DC, USA, pp 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542
30. Li Y, Song L, Wu X, He R, Tan T (2018) Anti-makeup: learning a bi-level adversarial network for makeup-invariant face verification. In: AAAI
31. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
32. Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. Henry Holt and Co. Inc, New York
33. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British Machine Vision Conference. BMVA Press, pp 36.1–36.10. https://doi.org/10.5244/C.16.36
34. Nguyen TD, Pham TD, Baek NR, Park KR (2018) Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors. In: Sensors
35. Olson RK, Attneave F (1970) What variables produce similarity-grouping. Am J Psychol 83:1–21
36. Ortiz R (2012) Freak: fast retina keypoint. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition (CVPR), CVPR '12. IEEE Computer Society, Washington,

DC, USA, pp 510–517. http://dl.acm.org/citation.cfm?id=2354409.2354903
37. Parkhi OM, Vedaldi A, Zisserman A et al (2015) Deep face recognition. In: bmvc 1:6
38. Qiu S, Wen D, Cui Y, Feng J (2016) Lung nodules detection in ct images using gestalt-based algorithm. Chin J Electron 25(7):711–718
39. Reddit: Diet progress pictures (last visited on May 28th 2014). http://www.reddit.com/r/progresspics/
40. Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: an efficient alternative to sift or surf. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE, pp 2564–2571
41. Shen IC, Cheng WH (2015) Gestalt rule feature points. IEEE Trans Multimed 17(4):526–537. https://doi.org/10.1109/TMM.2015.2405350
42. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489. https://doi.org/10.1038/nature16961
43. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
44. Singla A, Yuan L, Ebrahimi T (2016) 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. ACM, New York, NY, USA, pp 3–11. https://doi.org/10.1145/2986035.2986039
45. Sun Y, Liang D, Wang X, Tang X (2015) Deepid3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873
46. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, pp 499–515
47. Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt. II. Psychologische Forschung 4(1):301–350. https://doi.org/10.1007/BF00410640
48. Yan Y, Ren J, Sun G, Zhao H, Han J, Li X, Marshall S, Zhan J (2018) Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. Pattern Recognit 79:65–78. https://doi.org/10.1016/j.patcog.2018.02.004
49. Zheng Z, Kambhamettu C (2017) Multi-level feature learning for face recognition under makeup changes. In: Proceedings—12th IEEE international conference on automatic face and gesture recognition, pp 918–923. https://doi.org/10.1109/FG.2017.131