GESTALT INTEREST POINTS WITH A NEURAL NETWORK FOR MAKEUP-ROBUST FACE RECOGNITION

Markus Hörhan *Horst Eidenberger*

Vienna University of Technology
Institute of Software Technology and Interactive System
Favoritenstrasse 9/1882, 1040 Vienna, Austria

ABSTRACT

In this paper, we propose a novel approach for the domain of makeup-robust face recognition. Most face recognition schemes usually fail to generalize well on these 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.

Index Terms— Face recognition, makeup-robust, person identification, GIP, CNN

1. INTRODUCTION

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. [1] 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 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 makeup robust face recognition method we utilize certain visual traits on higher semantic representation levels.

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 [2], [3]. Most high-accuracy FR systems today rely on deep-learning methods and they are already being deployed in commercial face-verification applications [4]. One problem of CNNs is their high computational complexity. In this work we present a prototype-based method which is faster than CNNs while delivering a very high categorization accuracy for the given application domain.

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 2. Section 3 explains the proposed approach in detail. In Section 4, the dataset and experimental results are presented.

2. RELATED WORK

To our knowledge, there is limited scientific literature on addressing the challenge of makeup-robust face recognition. Chen et al. [5] addressed this problem with a patch-based ensemble learning method. Song et al. [6] 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. [7] 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.

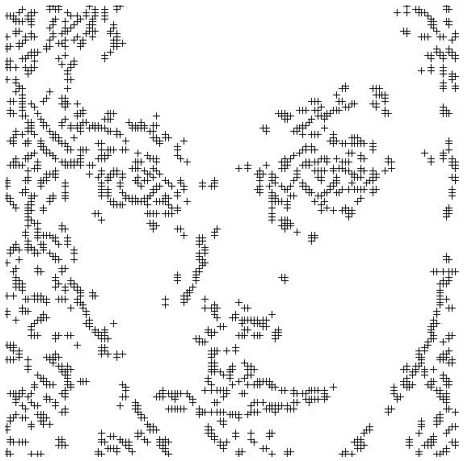

Fig. 1. Example output of the GIP algorithm. The GIP algorithm is fast and highly effective. Because it is inspired by cognition it extracts very little, but well-selected image information.

3. PROPOSED APPROACH

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 [8] described visual perception as a multi-stage 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 [9] 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 [10] 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) [11], 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 grey 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 [11] the algorithm and its continued development are explained in detail. Figure 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 feedforward network with a tangent sigmoid transfer function:

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

1) 2) 3) 4) 5) 6) 7) 8) 9) 10) 11) 12) 13) 14) 15) 16) 17) 18) 19) 20) 21) 22) 23) 24) 25) 26)

Fig. 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.

in the hidden layer, and a softmax transfer function

$$softmax(n) = \frac{e^n}{\sum(e^n)}, \tag{2}$$

in the output layer. For training the network its weight and bias values are updated according to the scaled conjugate gradient backpropagation method [12].

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.

4. EVALUATION

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

4.1. Dataset

In the works of Chen et al. (e.g. [13]) 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 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.

4.2. Baseline algorithms

We compare our method to several different state-of-the-art hand-crafted feature extraction algorithms, namely SIFT [14], SURF [15], BRISK [16] and FREAK [17]. After quantizing the extracted descriptors with the popular BoVW-algorithm [18] we categorize the resulting histograms with a neural network. Additionally, we compare our method to a CNN [19]. One drawback of CNNs is that they usually require a large amount of training data in order to avoid overfitting. Since our training data is relatively little to train a CNN, we used the pre-trained and well established AlexNet [20]. In a second stage we trained a multiclass 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.

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 5. We identified that 200 hidden layers maximize the mean accuracy for our application domain.

In Figure 3 some example ROC curves are shown. As expected the CNN-based baseline method is the strongest com-

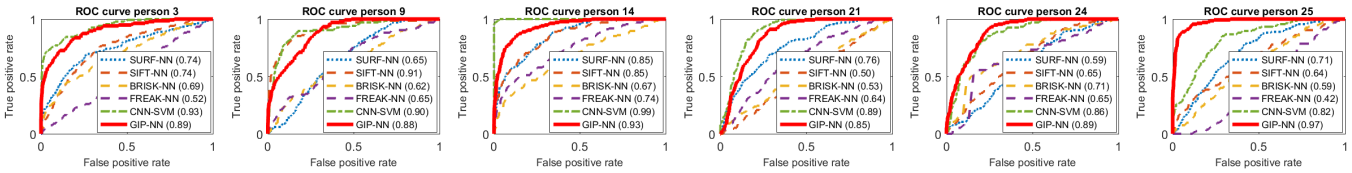

Fig. 3. ROC curves of our experiments for 6 of the 26 subjects. The numbers in parentheses in each legend indicate the Area Under the Curve (AUC) values.

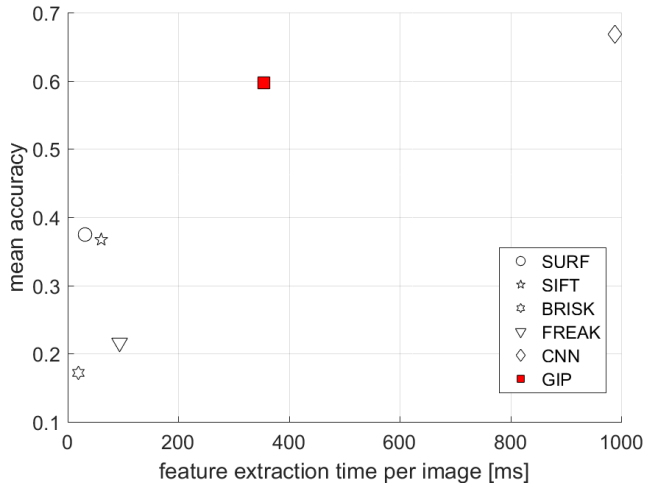

Fig. 4. The mean accuracies over feature extraction time of the different methods.

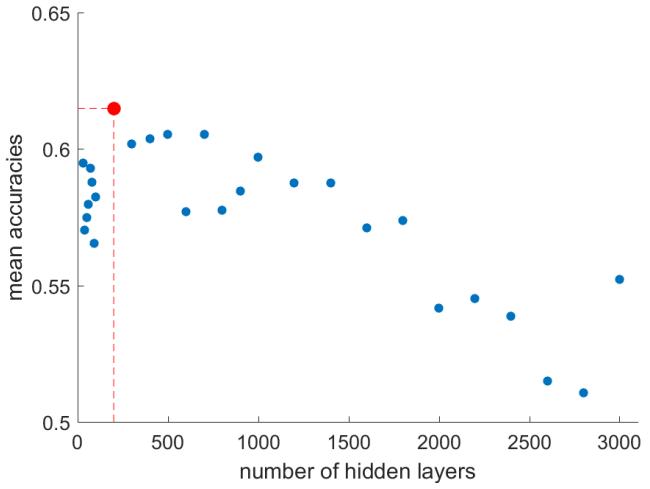

Fig. 5. The mean accuracies over different number of neural network's hidden layers. For our application domain a neural network with 200 hidden layers delivers the highest accuracy.

petitor. 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 2 shows that even for human beings it is difficult to identify these people because the make-up changes their faces drastically.

Figure 4 compares the mean accuracies over feature extraction time of the different methods. GIP-NN is cleary 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 [11]. That is, we need less disk space and processing power. This is very beneficial for a big data domain like face recognition.

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.

6. REFERENCES

[1] A. Dantcheva, C. Chen, and A. Ross, "Can facial cosmetics affect the matching accuracy of face recognition systems?," in *2012 IEEE Fifth International Confer-*

ence on Biometrics: Theory, Applications and Systems (BTAS), Sept 2012, pp. 391–398.

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

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

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

[5] Cunjian Chen, Antitza Dantcheva, and Arun Ross, "An ensemble of patch-based subspaces for makeup-robust face recognition," *Information Fusion*, vol. 32, pp. 80–92, 2015.

[6] Yi Li, Lingxiao Song, Xiang Wu, Ran He, and Tieniu Tan, "Anti-makeup: Learning A bi-level adversarial network for makeup-invariant face verification," *CoRR*, vol. abs/1709.03654, 2017.

[7] 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*, 2017, pp. 918–923.

[8] 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.

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

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

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

[12] Martin F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525–533, 1993.

[13] 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)*, Feb 2017, pp. 1–8.

[14] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[15] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

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

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

[18] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[20] 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*, June 2009, pp. 248–255.