Shallow CNNs for the Reliable Detection of Facial Marks

Chris Zeinstra
University of Twente
Faculty of EEMCS, DMB Group
Enschede, The Netherlands
c.g.zeinstra@utwente.nl

Erwin Haasnoot
University of Twente
Faculty of EEMCS, DMB Group
Enschede, The Netherlands
e.haasnoot@utwente.nl

***Abstract*—Facial marks are local irregularities of skin texture. Their type and/or spatial pattern can be used as a (soft) biometric modality in several applications. A key requirement for a biometric system that utilises facial marks is their reliable detection. Detection methods typically use a blob detector followed by heuristic post processing steps to reduce the number of false positives. In this paper, we consider shallow Convolutional Neural Networks (CNNs) for facial mark detection. The choice of this network type seems natural as it learns multiple (non) blob detectors; shallow refers to the fact that we only consider CNNs up to three layers. We show that (a) these CNNs successfully address the false positive problem, (b) remove the need for post processing steps, and (c) outperform a classic blob detector, approaches taken in previous studies and some other non CNN type classifiers in terms of EER and FMR at TMR=0.95.**

***Index Terms*—Facial Marks, Image Processing, Forensics, CNN.**

Facial marks are local irregularities of skin texture and include moles, pockmarks, raised skin, and scars [1], [2]. Facial marks have a potential to be highly discriminating, an observation that is reflected by their inclusion in the Bertillonage system [3], [4] that was used to describe persons for law enforcement purposes more than a century ago in France. Figure 1 shows two examples of facial marks and an annotation within the Bertillonage system.

Recent studies have considered facial marks and the spatial patterns they form, either to augment face recognition systems [5]–[7] and/or as a single biometric modality [1], [2], [5], [8]. Applications include identification (querying mugshot databases for matches to a facial mark spatial pattern of a perpetrator [9]) and verification (using facial marks to distinguish between mono zygotic twin [10], [11] and using their spatial pattern to calculate strength of evidence to be used in a court of law [2]). The latter study also clearly showed that in a particular setting up to 30% of the considered subjects could perfectly be identified solely based on their facial mark pattern. Finally, a series of related studies [12]–[14] consider the detection and performance of Relatively Permanent Pigmented or Vascular Skin Marks (RPPVSM) found on the back torso.

A key requirement for a biometric system that utilises facial marks is their reliable detection. Several studies mentioned here above indicate that the reduction of the number of false positive matches remains a challenging task. A common approach in these studies is to

1) apply geometric (for example affine transformation and cropping) and photo metric (for example gray scale conversion and illumination compensation) transformations,
2) detect blobs (for example Laplacian of Gaussian or Fast Radial Symmetry Transform [15]), and
3) heuristically reduce the number of false positive matches (for example skin detection and masks to exclude facial parts).

In this work, we present a different approach to the second step in the detection procedure. It successfully addresses the false positive problem and removes the need for the third step. Instead of using a single blob detector, we explore Convolutional Neural Networks (CNNs) that essentially train a collection of blob detectors. CNNs are layered computational structures that have shown impressive results over more traditional approaches in many domains, see for example [16] for selected applications in the biometric domain. The aim of this paper is not to present the best possible performance, but rather to explore the potential of shallow CNNs to detect facial marks. Shallow refers to a limited number of layers; blob detection is a straightforward problem that might be best served by a simple, straightforward approach.

We select skin patches that either contain a facial mark or not. We study to which extent the skin patch size and the inclusion of a photo metric pre processing step influences the detection performance. Furthermore, we compare the CNN performances not only to the traditional Laplacian of Gaussian approach and results of previous studies but also to other

a) b)

Fig. 1. From left to right: a) Two examples of facial marks, b) annotation within the Bertillonage system.

a) b) c) d)

Fig. 2. From left to right: a) Orginal colour image, b) ORG image, c) CLAHE image, d) ORG image with facial mark patches (white squares) and random non facial mark patches (black squares).

non CNN classifiers to create a broader overview of possible performance.

The remainder of this paper is structured as follows. We describe the experimental setup in Section I. In Section II we present and discuss results and compare them to some results found in literature. Section III contains the conclusion and presents possible extensions to this work.

I. Experimental Setup

A. Datasets

We use manually annotated facial mark locations to generate facial mark and non facial mark patches. These locations were collected by [2] and belong to 12306 images of 568 subjects of a subset of FRGCv2. This subset contains images of subjects with a neutral expression taken under a controlled studio condition. The reasons that we confine this study to this set are (a) to our knowledge it is the only large set with this type of annotation[1] and (b) creation of an additional comparable dataset would require a significant time investment.

B. Image Pre processing

Related studies apply geometric and photo metric transformations prior to the facial mark detection. We determine the affine transformation that maps the pupil coordinates to fixed locations (200, 250) and (400, 250), apply a bicubic texture interpolation and then crop the image to (800, 600). This transformation ensures that the interpupillary distance is constant and is chosen as a trade-off between a reasonable image size and facial mark visibility. The photo metric transformation consists of a gray scale transformation and CLAHE [17], a specific adaptive histogram equalisation method. Figures 2a,b,c show an original colour image, its geometrically and gray scale transformed image (henceforth referred to as ORG) and the effect of CLAHE respectively.

C. Patch Generation

We sort the subjects of FRGC dataset on their subject identification number in ascending order and use the first 394 subjects to create the training set and the remaining subjects to create the test set.

[1]For example, we consider for example the CFM dataset with 150 images described in [5] small.

For both sets, we use the same procedure to generate facial mark and non facial mark patches. To obtain a single facial mark patch, we randomly select a subject, then randomly select an image and finally randomly select a facial mark location within that image to generate the facial mark patch around that location. Prior to obtaining non facial mark patches, for each image we create a number of random non facial mark patches that (a) do not overlap with each other and (b) do not overlap with facial mark patches. To obtain a single non facial mark patch, we randomly select a subject, then randomly select an image and finally randomly select a non facial mark patch from the set described here above. Figure 2d highlights some facial mark and non facial mark patches.

We use this procedure to generate 50000 and 10000 facial mark patches and 50000 and 10000 non facial mark patches for the train and test set respectively. In order to investigate the effect of patch size, we take a patch size of 15×15, 19×19, and 23×23 around every centre chosen by our procedure. Moreover, to study the effect of photo metric pre processing, we extract exactly the same patches from the images that have been pre processed with CLAHE.

D. Experiments

We define two experiments that compare classifiers in terms of EER and ROC curves. In both experiments the six combinations of patch size and ORG/CLAHE are considered.

Experiment 1 determines a baseline performance. We apply the Laplacian of Gaussian (LoG) with kernel size $3 <= k <= 13$ to a patch, followed by a linear min-max mapping to $[0, 1]$ yielding a matrix $L = (L_{ij})$. To this patch, we assign a score $s = -\sum_{ij} L_{ij}$. We expect that facial mark patches yield small negative score values, whereas non facial mark patches in general yield larger negative scores. Figures 3a (facial mark patch/ORG), 3b its matrix L, 3e (non facial mark patch/ORG), and 3f its matrix L illustrate this. The other patches contained in Figure 3 are the corresponding CLAHE patches and their matrices L.

If we interpret every patch as a vector, we can apply other non CNN approaches as well. We consider $K = 3$ Nearest Neighbors, Gaussian Kernel Density Estimation with bandwidth 0.15, Linear ($C = 1$) and RBF ($C = 1000$) Support Vector Machines, Random Forest (ensemble of 300 decision trees with depth 200) and a classic fully connected neural network consisting of three densely connected layers in which the middle layer is wide (10 times the input size).

Experiment 2 compares the performance of six similar shallow CNNs to the classifiers studied in Experiment 1. Their architectures are shown in Figure 4. These variations are chosen to address the impact of small architectural differences such as the number of layers and the choice of layer type. All CNNs accept patches and output the facial mark probability. We use CNN A to determine an odd kernel size $3 <= k <= 13$ that yields the best average EER; this kernel size is subsequently used in the other CNNs.

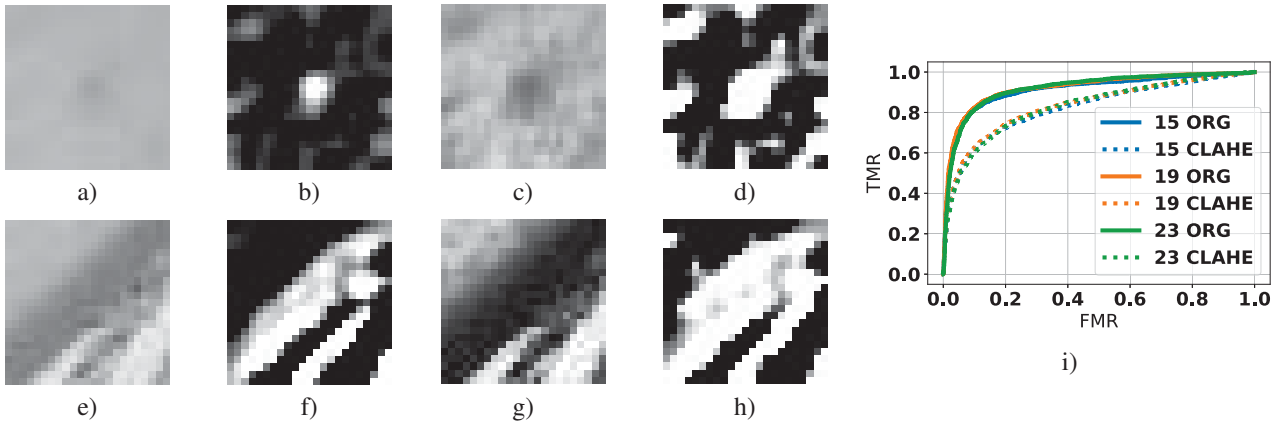


Fig. 3. Left part contains patches and the effect of the Laplacian of Gaussian with kernel size $k = 5$.Top row contains a facial mark patch. From left to right: a) ORG, b) matrix L, c) CLAHE , d) matrix L. Bottom row contains a non facial mark. From left to right: e) ORG, f) matrix L, g) CLAHE , h) matrix L. Right part i): ROC curves of Laplacian of Gaussian.

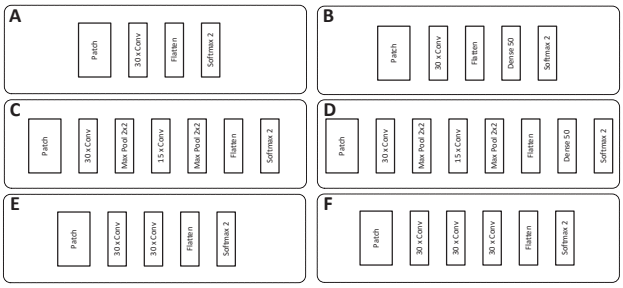


Fig. 4. Architecture of shallow CNNs A to F.

Classifier	15 ORG	15 CLAHE	19 ORG	19 CLAHE	23 ORG	23 CLAHE
LoG	0.146	0.245	0.137	0.233	0.141	0.238
KNN	0.114	0.154	0.146	0.179	0.184	0.211
KDE	0.126	0.264	0.157	0.342	0.188	0.384
Linear SVM	0.143	0.176	0.136	0.183	0.130	0.194
RBF SVM	0.108	0.0900	0.102	0.0915	0.103	0.102
Random Forest	0.0754	0.0852	0.0959	0.108	0.127	0.132
Classic NN	0.0651	0.0772	0.0770	0.0753	0.0768	0.0760
CNN A	0.0541	0.0554	0.0459	0.0555	0.0455	0.0543
CNN B	0.0452	0.0628	0.0460	0.0555	0.0434	0.0571
CNN C	0.0436	0.0589	0.0357	0.0433	0.0279	0.0408
CNN D	0.0437	0.0588	0.0329	0.0458	0.0293	0.0383
CNN E	0.0340	0.0542	0.0293	0.0425	0.0292	0.0428
CNN F	0.0397	0.0553	0.0318	0.0462	0.0281	0.0422

TABLE I
EERS OF ALL CLASSIFIERS. BEST PERFORMANCE FOR EACH COMBINATION UNDERLINED.

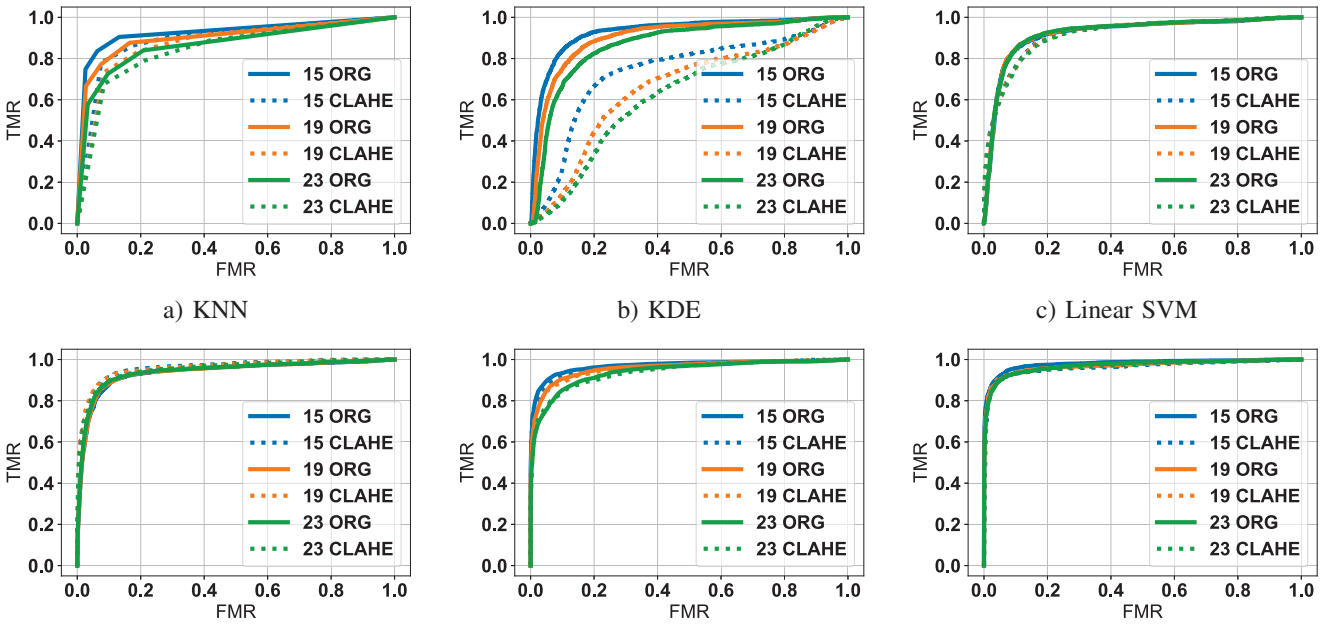
II. RESULTS AND DISCUSSION

A. Experiment 1 - Baseline

In the application of the LoG we found that kernel size $k = 5$ yields both visually a clear distinction between facial and non facial patches as well in terms of EER. Figure 3i shows their ROC curves for the six combinations. They clearly demonstrate the false positive problem: setting a reasonable value for the TMR (for example 0.95) yields a high value for the FMR (larger than 0.42 in all cases). Figures 5 present ROC curves of several non CNN classifiers. Some classifiers still possess the false positive problem (for example KNN and KDE), but increasingly exhibit better performance (for example the classic neural network). The latter classifier is the only non CNN classifier that consistently has an EER< 0.08, see Table I. For all classifiers in Experiment 1 we find that (a) almost always the patch size has a negative influence of the EER and (b) the application of CLAHE has a variable negative influence on the EER. Especially the case of the LoG, the severe deterioration of the performance after the application of CLAHE can be observed. An explanation can be found in Figures 3b,d,f and h in which the effect of CLAHE on matrices L of both facial and non facial patches is shown, resulting in a larger overlap between the score distributions of facial and non facial patches. Differences between patch sizes and ORG/CLAHE are mostly present in KNN, KDE and the Random Forest. We believe that the deteriorating performance of KDE can be attributed to the fixed bandwidth while the dimensionality increases significantly. Finally, if we fix TMR=0.95, it can be shown that FMR almost always exceeds 0.15.

B. Experiment 2 - Shallow CNN

Figure 6 shows the trained filters of CNN A for a 13×13 convolution kernel size. Although filters of this size clearly reflect the two input classes, we found that setting $k = 5$ in CNN A yields the lowest overall EER taken over the six possible combinations. Therefore we use this convolution kernel size for all CNNs. CNN A consist only of one convolutional layer, whereas CNNs typically contain more and different type layers. For example, CNN C contains convolutional layers followed by max pooling layers. Table I shows that this architecture improves the EER, the lowest EER is 0.0279 in the 23×23/ORG case. Another layer type is the dense layer used in classic neural networks. If we add one such layer to CNNs A and C prior to the last softmax layer, we obtain CNNs B and D, respectively. In both cases, A vs. B and C vs. D, we do not see an improvement. Although CNN C has the lowest



a) KNN b) KDE c) Linear SVM

d) RBF SVM e) Random Forest e) Classic Neural Network

Fig. 5. ROC curves of non CNN classifiers.

overall EER, we are also interested in the effect of adding more convolutional layers to CNN A.

CNN E and F contain two and three layers, respectively. CNN A improves by adding layers, however having two layers is seems somewhat better than three layers. Since we do not train and evaluate any of our classifiers on different partitions of the dataset, we will and cannot claim any significant difference(s) between the classifiers. We notice that (a) in general the EER improves if the patch size is larger and (b) the application of CLAHE has a variable negative influence on the EER. All CNN ROC curves are plotted in Figure 7.

Finally, if we fix TMR=0.95, it can be shown that FMR almost always is below 0.08 and in some cases as low as 0.01.

C. Discussion

The results of Experiments 1 and 2 seem to indicate that shallow CNNs have addressed the false positive problem.

Only a few studies mentioned in the Introduction do report solely on performance of the detection of facial marks. For example, [8] reports 4.2% FNMR and 4.2% FMR on a set of 120 subjects and 7.1% FNMR and 2.4% FMR on a set of 85 subjects. The study [1] shows that their automatic method has an EER of 0.155. The semiautomatic approach yielded an EER of 0.122. Finally, [5] reports 73.1% precision and 57.0% recall on a dataset of 265 subjects and 13.6% precision and 16.2% recall on another dataset of 30 celebrities.

Although different datasets were used for above results and our own, we conclude that there is a strong indication that our approach is state-of-the-art.

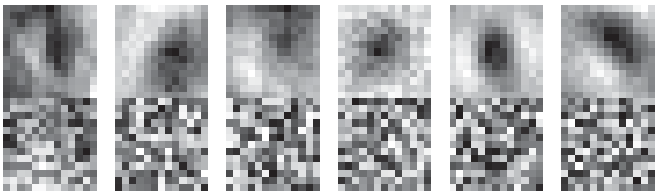
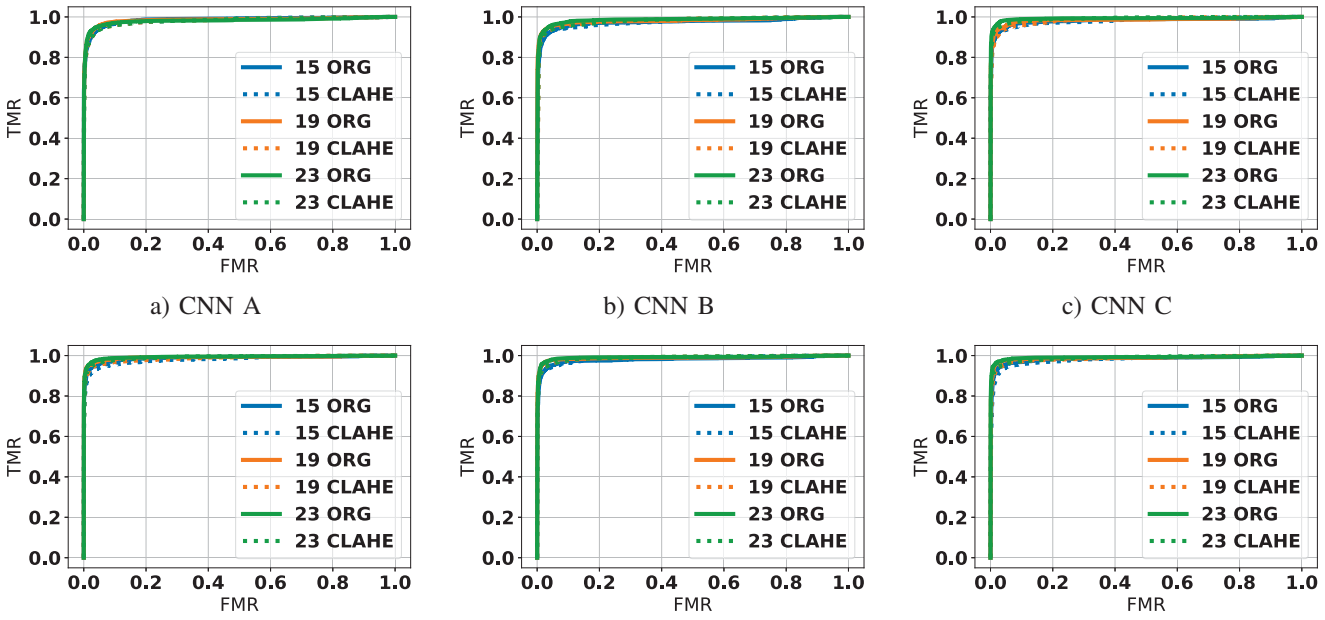


Fig. 6. Selection of filters of CNN A, trained on patch size 15×15/ORG using convolution kernel size $k = 13$.

III. Conclusion and Future Work

We were able to reproduce the false positive problem of LoG. We employed various traditional classifiers and found that their EER and ability to reduce the false positive problem varied. These results and some other results found in literature are surpassed by several shallow CNN architectures. In particular an architecture that consists of two convolutional and max pooling layers yielded an EER=0.0279 and FMR=0.01 at TMR=0.95, which in our opinion shows that the false positive problem has been addressed in a satisfactory manner. Although the primary motivation for using CNN is to consider it as a collection of blob detectors, we actually found that using as small kernel size had comparable performance to a larger size, emphasising the general expressive power of CNNs. The use of CLAHE instead of ORG increases the EER for all classifiers. Remarkably, the patch size has a positive influence on the EER of CNNs, whereas in the case of non CNN classifiers this relationship is reversed.

We identify several extensions to this work. First, we deliberately selected simple CNNs; another complementary



a) CNN A b) CNN B c) CNN C

d) CNN D e) CNN E e) CNN F

Fig. 7. ROC curves of CNN classifiers.

approach is to use a more complex pre-trained CNN and apply transfer learning. Also, we can evaluate our approach multiple times and on more datasets to create performance confidence intervals and measure differences between sets. Finally, although our approach was to detect facial marks, we could augment face recognition systems in challenging cases such as the comparison of facial images of mono zygotic twins.

REFERENCES

[1] Srinivas, Nisha; Flynn, Patrick J.; Vorder Bruegge, Richard W.: Human Identification Using Automatic and Semi-Automatically Detected Facial Marks. Journal of Forensic Sciences, 61:117–130, 2016.

[2] Zeinstra, C.; Veldhuis, R.; Spreeuwers, L.: Grid-Based Likelihood Ratio Classifiers for the Comparison of Facial Marks. IEEE Transactions on Information Forensics and Security, 13(1):253–264, Jan 2018.

[3] Bertillon, A.: Identification anthropométrique: instructions signalétiques. 1893.

[4] Finn, Jonathan: Capturing the Criminal Image: From Mug Shot to Surveillance Society. University of Minnesota Press, New edition, 2009.

[5] Becerra-Riera, Fabiola; Morales-Gonzlez, Annette; Mndez-Vzquez, Heydi: Facial marks for improving face recognition. Pattern Recognition Letters, 2017.

[6] Dantcheva, Antitza; Elia, Petros; Ross, Arun: What else does your biometric data reveal? A survey on soft biometrics. IEEE Transactions on Information Forensics and Security, 11(3):441–467, 2016.

[7] Park, U.; Jain, A. K.: Face Matching and Retrieval Using Soft Biometrics. IEEE Transactions on Information Forensics and Security, 5(3):406–415, Sept 2010.

[8] Gogoi, Usha Rani; Bhowmik, Mrinal Kanti; Saha, Priya; Bhattacharjee, Debotosh; De, Barin Kumar: Facial Mole Detection: An Approach towards Face Identification. Procedia Computer Science, 46:1546 – 1553, 2015. Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace and Island Resort, Kochi, India.

[9] Jain, A. K.; Klare, B.; Park, U.: Face Matching and Retrieval in Forensics Applications. IEEE MultiMedia, 19(1):20–20, Jan 2012.

[10] Srinivas, N.; Aggarwal, G.; Flynn, P. J.; Bruegge, R. W. V.: Facial marks as biometric signatures to distinguish between identical twins. In: CVPR 2011 WORKSHOPS. pp. 106–113, June 2011.

[11] Srinivas, N.; Aggarwal, G.; Flynn, P. J.; Bruegge, R. W. Vorder: Analysis of Facial Marks to Distinguish Between Identical Twins. IEEE Transactions on Information Forensics and Security, 7(5):1536–1550, Oct 2012.

[12] Nurhudatiana, Arfika; Kong, Adams Wai-Kin: On criminal identification in color skin images using skin marks (RPPVSM) and fusion with inferred vein patterns. IEEE Transactions on Information Forensics and Security, 10(5):916–931, 2015.

[13] Nurhudatiana, Arfika; Kong, Adams Wai-Kin; Matinpour, Keyan; Chon, Deborah; Altieri, Lisa; Cho, Siu-Yeung; Craft, Noah: The individuality of relatively permanent pigmented or vascular skin marks (RPPVSM) in independently and uniformly distributed patterns. IEEE Transactions on Information Forensics and Security, 8(6):998–1012, 2013.

[14] Nurhudatiana, Arfika; Kong, Adams Wai-Kin; Craft, Noah; Tey, Hong Liang: Relatively Permanent Pigmented or Vascular Skin Marks for Identification: A Pilot Reliability Study. Journal of Forensic Sciences, 61(1):52–58, 2016.

[15] Loy, G.; Zelinsky, A.: Fast radial symmetry for detecting points of interest. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(8):959–973, Aug 2003.

[16] Bhanu, Bir; Kumar, Ajay: Deep learning for biometrics. Springer, 2017.

[17] Zuiderveld, Karel: Contrast limited adaptive histogram equalization. Graphics gems, pp. 474–485, 1994.