Detection of Age-Induced Makeup Attacks on Face Recognition Systems Using Multi-Layer Deep Features

Ketan Kotwal, Zohreh Mostaani, and Sébastien Marcel

Abstract—Makeup is a simple and easy instrument that can alter the appearance of a person's face, and hence, create a *presentation attack* on face recognition (FR) systems. These attacks, especially the ones mimicking ageing, are difficult to detect due to their close resemblance with genuine (non-makeup) appearances. Makeups can also degrade the performance of recognition systems and of various algorithms that use human face as an input. The detection of facial makeups is an effective prohibitory measure to minimize these problems.

This work proposes a deep learning-based presentation attack detection (PAD) method to identify facial makeups. We propose the use of a convolutional neural network (CNN) to extract features that can distinguish between presentations with age-induced facial makeups (attacks), and those without makeup (*bona-fide*). These feature descriptors, based on shape and texture cues, are constructed from multiple intermediate layers of a CNN. We introduce a new dataset AIM (Age Induced Makeups) consisting of 200+ video presentations of old-age makeups and *bona-fide*, each. Our experiments indicate makeups in AIM result in 14% decrease in the median matching scores of a recent CNN-based FR system. We demonstrate accuracy of the proposed PAD method where 93% presentations in the AIM dataset are correctly classified. In additional testing, it also outperforms existing methods of detection of generic makeups. A simple score-level fusion, performed on the classification scores of shape- and texture-based features, can further improve the accuracy of the proposed makeup detector.

Index Terms—Biometrics, Face Presentation Attack Detection, Convolutional Neural Network, Deep Learning, CNN Embeddings, Texture Descriptor, Shape Descriptor, Score Fusion, Makeup Attacks, Makeup Attack Detection, Old-Age Makeups, AIM

1 Introduction

Presentation attacks (PA) on a face recognition (FR) system can be broadly categorized in *obfuscation* and *impersonation* [1]. Obfuscation attacks can be created by altering one's appearance using a variety of facial disguises, makeups, prosthetics, accessories, etc. This increases intra-class variability of the face recognition system [2]; and thereby, poses a challenging problem for the biometric community. A systematic detection of PAs is critical to ensure better and trustworthy functioning of FR systems.

In recent years, the research in detecting obfuscation PAs has been mainly focused on disguises using accessories such as glasses, goggles, hats, or caps [2], [3], [4]. The datasets collected by the respective works also include obfuscations created from variations in hairstyle, beard, and moustache, while a small fraction of images consists of makeup variations of celebrities. It may be observed that some of the PAs from datasets in [2], [3] either cover a face region too much, or appear quite unnatural. Therefore, such obfuscations may not represent a real world scenario [4]. Also, in a controlled environment, some PAs consisting of accessory-based disguises (e.g, dark glasses, scarfs, hats, etc.) can be averted. An actual threat to an FR system lies in obfuscation attacks that retain the *natural appearance* of a human face, yet facilitate the concealment of true identity of an attacker.

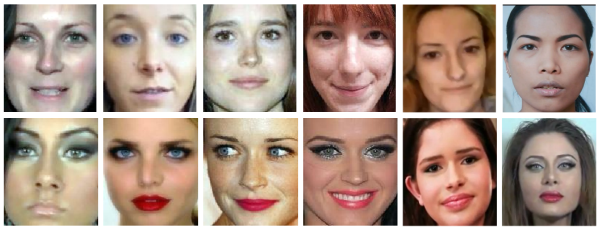


Fig. 1. Examples of makeup and non-makeup images. Top row shows *bona-fide* (non-makeup) images; and bottom row consists of presentations with traditional makeups. The first two columns from left display samples from the YMU dataset [5]; next two columns provide samples from the MIW dataset [6]. Last two columns on right provide samples from the MIFS dataset [7].

Facial makeup is one of the most simple and common mechanisms to alter the appearance of a face– typically to appear more attractive, and to cover facial flaws. Makeup can also be conveniently used to create an aged appearance or even that of a different person as seen in movies and drama. Given its ability and extent to modify the appearance, a makeup can be regarded as an effective instrument for PA on FR systems. Specifically, makeup creates various challenges for an FR system- (*i*) *obfuscation*: Whereas there are various ways to disguise the face, makeup is one of the easiest and low-cost options. Thereby posing consequential concern to the biometric FR system. (*ii*) *impersonation*: It is possible to impersonate someone else's identity through makeup. Although such examples are mostly seen in enter-

- *Authors are with Idiap Research Institute, Martigny, Switzerland.*
- *Manuscript received June 14, 2019; revised August 26, 2019; accepted September 24, 2019.*

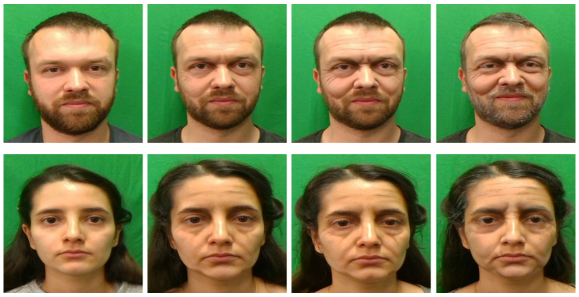


Fig. 2. Samples of age-induced makeup with different levels of makeup intensity from AIM dataset. For each row, the left image is *bona-fide*, and intensity of makeup increases from left to right.

tainment industry, their threat is serious as the modern face matching systems are highly vulnerable to such attacks [7]. (*iii*) *accuracy of FR*: A facial makeup directly impacts the matching accuracies of FR systems. A study by Dantcheva *et al* [5] investigated the performance of 3 different FR algorithms including a commercial face matching system by comparing images of the same subject with and without the use of cosmetics. Their results indicate that equal error rates (EER) of face matching systems can degrade up to 23% when the images with makeup (along with non-makeup ones) were considered.

The aforementioned challenges need to be addressed over a wide range of test scenarios given the fact that makeup is a low-cost, easily accessible, and non-constant PA instrument. Unfortunately, detection of makeup-based PAs has not received due attention despite its serious threat. The task is indeed very hard due to a wide variety of appearances that makeups can create. Additionally, large variations in the skin complexions of genuine (non-makeup) presentations also impacts the detection of makeup PAs. Fig. 1 provides some examples of makeups on female subjects where variety in the makeups and skin complexions can be observed. These makeups use lipstick, eyeliners, and foundation materials. Such makeups are, however, mostly apt for female subjects; and hence, their applicability is limited.

Another way to obfuscate using makeup is to disguise one's face by providing a look of an old-age person. For instance, the entertainment sector across the world has relied on makeups to induce an aged look for portrayal of an older character. Such appearances are realistic, detail-oriented, individualistic; and often so meticulous that audiences fail to recognize the real actor. We refer to such makeups as "age induced" makeups—that have demonstrated the potential to disguise or *obfuscate* the true identity of an individual. The success of age-inducing makeups is an alarming signal for security of FR systems; and countermeasures to these PAs must be developed. The age induced makeups are particularly of interest because- (*i*) This makeup is a simple and effective PA that requires little artistic skills. Relevant tutorials are readily available (although meant for a different purpose). (*ii*) Attack presentations appear extremely similar to genuine aged persons; and, therefore, are realistic and natural. (*iii*) It is universal, and applicable to both genders-male and female. (*iv*) From a technical standpoint, age-induced makeups are difficult to detect as the features of both classes (i.e., makeup attacks v/s genuine aged persons) exhibit significant overlap.

To demonstrate realism of old-age makeups, some samples from our new dataset are presented in Fig. 2. The makeup, performed on both male and female subjects, exhibits progression of makeup intensities for a given subject. The old-age effect is created by adding highlights, shadows, wrinkles, or by emphasizing the existing wrinkles of the subject. Furthermore makeup is used to change the apparent shape of the face specially around the eyes and chin to show the saggy skin effect that can happen during natural aging process.

In this work, we propose a presentation attack detection (PAD) method to identify age-induced makeups. We design a stand-alone *makeup detector* that can act as a pre-filter to the FR system to test the possibility of PA. The proposed PAD method is efficient not only in detecting age-induced makeup attacks; but also in handling other cases of traditional makeups as depicted in Fig. 1. The makeup detector is helpful in non-attack scenarios where makeups degrade the accuracy of FR systems. It is possible to process makeup presentations separately to improve the results of face recognition. Additionally, the makeup detector can act as a preprocessor for algorithms that process facial features.

Deep learning (DL)-based FR systems have shown superior performance over the ones using handcrafted features. Multiple layers in the convolutional neural network (CNN) are capable of learning complex features from input images. The initial convolutional layers of a CNN learn local features; whereas fully connected layers towards the end learn the spatial (or geometrical) arrangements of these features. Therefore, a CNN can be used to obtain local as well as global descriptors of a face image. Considering CNN as a multi-layer feature extractor, we develop a novel CNN-based technique for detection of makeup PAs.

One major challenge in developing a DL-based method is limited amount of training data curated for the specific application. To circumvent this problem, the machine learning community has devised several transfer learning mechanisms where a pretrained DL model (mostly trained with a large data) is adapted for a similar task. At present, the publicly available makeup datasets are very small; and thus, training a deep CNN from scratch could be non-trivial. On the other hand, several CNN models, trained on good amount of data, have demonstrated excellent capabilities at FR. However, it is not certain to what extent the features learnt by FR CNN are useful towards detection of makeup attacks as well; and how can this task be facilitated. We propose a novel method of obtaining shape- and texture-based feature descriptors from multiple layers of pretrained FR CNN for the goal of detecting makeup PAs.

A dataset consisting of custom age-induced makeup PAs is indispensable for research in PAD involving makeups. For this work, we have curated a new dataset of *bona-fide* (non-makeup) presentations and makeup-based PAs. A total of 456 video presentations from 72 subjects constitute our dataset.

The contributions of this work are summarized below.

- To the best of our knowledge, this is the first demonstration of detection of makeup PAs using deep learning. Our PAD method employs a CNN to compute descrip-

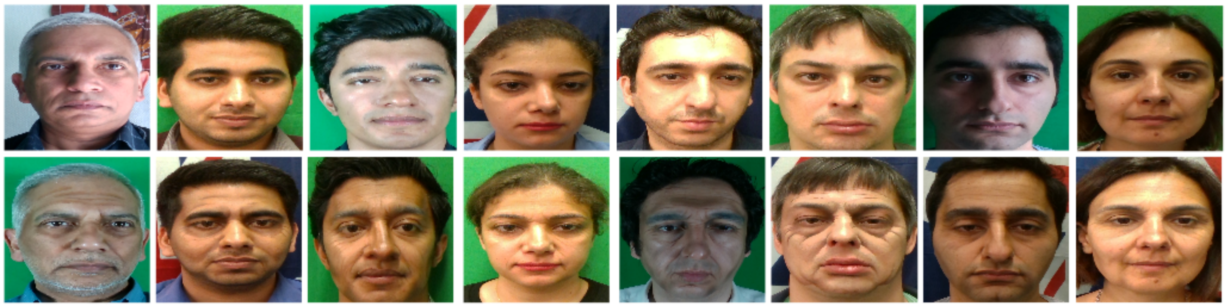


Fig. 3. Examples of makeup PAs and *bona-fide* (non-makeup) samples from AIM dataset. Top row shows *bona-fide* (non-makeup) images; and bottom row consists of makeup presentations for the same subject in respective column.

tors related to texture and shape in the context of age-induced facial makeups.

- We show that a CNN pretrained for FR tasks is an efficient feature extractor for detecting makeups without explicit transfer learning or fine-tuning.
- Since we do not process specific regions in the face, the proposed technique can be employed for location agnostic makeups. Through a simple fusion scheme, we demonstrate the complementary nature of our feature descriptors.
- We analyze the impact of age-induced makeups on FR accuracy of a specific CNN-based FR method.
- A new dataset named AIM consisting of PAs using age induced makeups has been introduced.[1]
- We demonstrate efficacy of the proposed makeup detector on AIM dataset. For its generalization, we also test the proposed PAD method on three publicly available makeup datasets (not targeted toward old age). Our results outperform the existing state-of-the-art methods by $\approx 2\%$.

Section 2 provides an overview of related work towards detection of facial makeups. The details of AIM dataset and other traditional makeup datasets for PAD experiments are provided in Section 3. This section also describes the experimental protocols for each dataset. We describe the proposed method of makeup detection in Section 4. Section 5 presents the experimental results, and discussions on experiments are provided in Section 6. Section 7 summarizes conclusions and future directions.

2 Related Work

The problem of makeup detection has been receiving attention since last few years, although majority of the works focus on accuracy aspect of FR system rather than its security.

Initial work by S. Varshovi [8] explores color and texture features for detection of makeup on female subjects. This work categorizes a makeup into eye shadows, lipstick, and liquid foundation. In [6], authors consider different shape descriptors in addition to color and texture descriptors. Their shape descriptors are based on Gabor wavelets, GIST [9], and edge oriented histogram (EOH); and the color descriptors are computed over eye and mouth regions. The makeup is detected using support vector machine (SVM) and Adaboost classifiers. The use of local binary patterns (LBP) and histogram of oriented gradients (HOG) for texture descriptors has been proposed in [10]. These features are fused for training of SVM and Alligator–based classifiers. Guo *et al* consider smoothness and highlight of facial skin in addition to color and texture as feature descriptors [11]. These features are computed on 12 location-specific patches of face region. They advocate patch-based processing as different makeups (or products) that are applied over different face regions. A low-rank dictionary learning approach for detection and decomposition of makeup is proposed in [12]. Rasti *et al* used biologically inspired features (BIFs) over whole face image for makeup detection [13]. In [14], an entropy image is computed from a local neighborhood of pixels of input image. The features, obtained by stacking the gradient orientation vectors of pyramidal (multi-level) representations of an entropy image, are used to train SVM classifier. It may be observed that most of the existing works have used handcrafted features to describe the texture and shape, and often involving explicit localization to create patches.

Several researchers have considered the problem of makeup purely from the perspective of degradation of matching accuracy. To mitigate the same, they have proposed FR systems that are robust to makeup changes [15], [16], [17], [18]. However, such solutions do not specifically consider makeup as PAs; and thus, are not useful in detecting makeups.

3 Makeup Datasets

The first subsection describes our newly created dataset AIM that consists of age induced makeups. In the next subsection, we also provide the details of other publicly available makeup datasets. The experimental protocol for every dataset is also described. We use these datasets to test the proposed method for the detection of traditional makeups.

3.1 AIM: Age Induced Makeup Dataset

The AIM (Age Induced Makeup) dataset consists of videos of *bona-fide* (BF) and presentation attacks (PA) constructed using old-age makeups. The dataset consists of 240 BF presentations corresponding to 72 subjects; and 216 attack presentations captured from a subset of 20 subjects. For every participant of makeup presentation of AIM, a BF (non-makeup) video is also available. The AIM dataset

[1] AIM dataset: https://www.idiap.ch/dataset/aim

comprises of male and female subjects belonging to different ethnicities. While other publicly available makeup datasets consist of only female subjects. AIM dataset provides a wider perspective by combining presentations from male and female subjects.

The presentations, as shown in Fig. 3, have been captured by varying the background and lighting conditions. The background could either be plain (a green curtain), or complex; whereas illumination was varied by adjusting ceiling light and spotlights in the capturing environment. Makeups were created by professionals using regular makeup materials to compose age-inducing effects like coloring of eyebrows, and creation of wrinkles on cheeks, or forehead. No prosthetic objects or materials were considered.

Each presentation in AIM dataset was captured for $\approx 10\,s$ from the color channel of Intel RealSense SR300 camera.[2] The presentations cover a frontal view of face; however, not explicitly localized. We believe that availability of *bona-fide* for every attacker subject will be helpful in investigating multiple research problems.

For experiments reported in this work, we have created a `grandtest` protocol where the AIM dataset is divided into three fixed, disjoint sets- *train*, *dev*, and *eval*— which are used to conduct training, parameter tuning, and testing, respectively. We have ensured that the subjects of one set are not present in other two. The grandtest protocol utilizes 20 frames from each video selected in a uniform manner. Experiments are conducted and evaluated on a frame-level basis. A total of 3160 frames from 86 BF and 72 PA videos constitute the `train` set. The `dev` set comprises of 1600 BF frames, and 1440 frames of makeup attacks. In the `eval` set, there are 2920 frames from 74 BF and 72 PA videos. The grandtest protocol is summarized in Table 1.

TABLE 1
`grandtest` protocol for the AIM dataset.

Partition	# Videos	# Frames	Split ratio (%)	Total Frames
train BF	86	1720	54.43	3160 (35%)
train PA	72	1440	45.56	
dev BF	80	1600	52.63	3040 (33%)
dev PA	72	1440	47.37	
eval BF	74	1480	50.68	2920 (32%)
eval PA	72	1440	49.32	
Total	456	9120		9120

3.2 Traditional Makeup Datasets

3.2.1 YouTube MakeUp Database (YMU) [5]

This dataset was originally introduced by Dantcheva *et al* [5], where they collected images from YouTube makeup tutorials. The dataset consists of facial images of Caucasian female subjects with and without application of makeups. In [5], the dataset consisted of images of 99 subjects. The extended dataset consists of 604 images of 151 subjects. Each subject has 2 images (or shots) with and without makeups which results in 302 makeup images and 302 non-makeup images. The YMU dataset, thus, is balanced in terms of data from both classes. The authors mention that the dataset has varying degrees of makeups; and it includes variations in facial expression, pose and resolution. These characteristics indicate/result in a relatively unconstrained dataset. The left two columns of Fig. 1 shows samples from YMU dataset.

In [6] and [10], a 5-fold cross validation (CV) protocol has been used to demonstrate the respective works on detection of makeups. We also create a 5-fold CV protocol following the similar guidelines as mentioned in the aforementioned works. We divide the YMU dataset into 5 partitions that are disjoint in terms of the subjects. Each subset or a fold consists of approximately 120 images of 30 unique subjects. We use an aggregate of 4 folds to train, and test the results on the remaining fold. By repeating this process on a different fold, we generate 5 CV protocols or trials (`cv0`, `cv1`, ..., `cv4`). We do not have information on exact folds or partitions used in [6], [10]; however, since our implementation of this protocol is similar, we reasonably assume that the results can be comparable.

3.2.2 Makeup In the Wild Database (MIW) [6]

This dataset consists of makeup and non-makeup images assembled from the Internet. It has been referred to as "in the wild" as the images are obtained from the Internet [6]. The dataset, hence, provides significant variations in the illumination, pose, and degree of makeups— as seen from samples in Fig. 1. With 77 makeup and 77 non-makeup images of 125 subjects, the MIW dataset has been used in the literature to evaluate the generalization capabilities of the makeup detector. For this purpose, the makeup detector trained on the YMU dataset is tested on the MIW dataset in [6], [10]. We also perform a similar experiment to demonstrate the performance of the proposed makeup detector.

3.2.3 Makeup Induced Face Spoofing (MIFS) [7]

The MIFS dataset was used to study impersonation (or spoofing) attacks using makeups. Each subject wears a makeup to *look like* a different person (called target identity). The dataset is collected from YouTube makeup video tutorials. It consists of 3 categories of images- subjects before makeup, subjects after makeup, and target identity (which is being spoofed using makeups). Since we are interested in detection of makeups, we consider only first two categories of images, i.e., before makeup and after makeup.

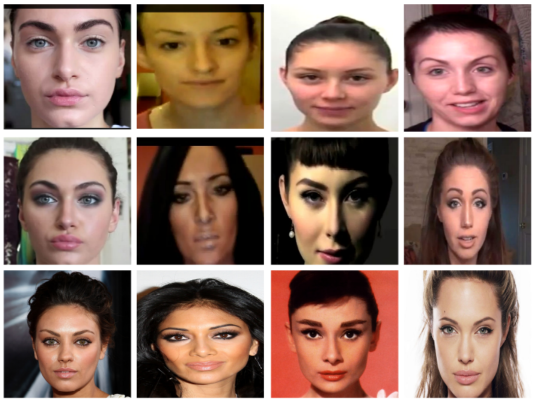


Fig. 4. Examples of cropped images from MIFS dataset. In each column, the top image shows the subject before makeup (*bona-fide*), and the middle image represents the presentation after makeup. The corresponding target identity is shown in the bottom image.

[2] software.intel.com/en-us/realsense/sr300

TABLE 2
Details of datasets used for detection of facial makeups.

Dataset	Makeup Type	Subject Type	Data Sources	Format	# Subjects	# Files	Purpose
YMU [5]	Generic	Female only	Internet	Images	151	604	Obfuscation
MIW [6]	Generic	Female only	Internet	Images	125	154	Obfuscation
MIFS [7]	Generic	Female only	Internet	Images	107	416	Impersonation
AIM (This work)	Age Progressive	Male + Female	Professional custom setup	Video	72	456	Obfuscation

This dataset has 208 non-makeup images and 208 makeup images. In the subsequent discussions, we refer to this dataset simply as MIFS, though it should be noted that it does not use the target identities present in MIFS dataset, as assessing the Impostor Attack Presentation Match Rate (IAPMR) is not the aim of this work- as it was done by [7]. Samples of the subjects before makeup (*bona-fide*), and after makeup are shown in Fig. 4. It also provides the target identity being spoofed.

Table 2 provides salient details of the AIM and other makeup datasets.

4 Proposed Method

Makeup affects shape and texture of face [6], [10]. Despite their predominant color-based alterations, makeups affect performance of FR systems based on grayscale images as well [5]. In this work, we derive our approach from the texture and shape information without exploiting color information. Various PAs can be constructed by applying makeups to specific facial features such as eyes and lips. Alternatively, a major (or entire) face region can be subject to makeup. The first case is a local or region-based activity— where one may localize the regions of interests (ROIs), and then learn the appropriate features to discriminate non-makeup (*bona-fide*) faces from makeup attacks. For the later case, the face PAD requires understanding of texture and shape information at global level (i.e., from entire face image).

A shape, being a global property, can be well-learnt from facial features along with knowledge of their spatial arrangement. On the other hand, texture can be learnt through low-level features of local patches with minimal inference of location information. We formulate our PAD method by incorporating both complementary requirements, yet without completely duplicating the feature computation efforts.

Deep CNN-based methods have shown extremely promising results across wide application areas of image processing. Initial layers of a typical CNN consist of sets of convolutional filters (`conv`); and one or more fully connected (`FC`) layers towards the end. The `conv` layers learn characteristics of the local regions in the input image; wherein the shape related properties are captured through the `FC` layers [19]. An output of intermediate layers can be used to represent the local information in the image. Cimpoi *et al* have proposed texture descriptor by orderless pooling of the last `conv` layer of a CNN [19]. They compute the Fisher vector (FV) from dense features of last `conv` layer which is suitable at describing textural information. In [20], a texture CNN (T-CNN) has been specifically designed for texture detection where the energy pooling of last `conv` layer is passed to subsequent `FC` layers. In both [19], [20], authors have suggested the use of classic CNNs to derive shape information of the object. A pyramidal generative adversarial network (GAN) to learn age-progression of a human face has been proposed in [21]. It utilies features from multiple scales to simulate aging effect; however, it does not derive any explicit texture descriptor.

Inspired by the CNNs' ability to learn texture and shape information of object, we derive our PAD approach as follows. Our interest lies in learning the shape- and texture-related features of face. It can be achieved either through training a CNN from scratch using a large number of makeup and *bona-fide* presentations; or through selective modification of a CNN pretrained for face-related tasks. We select the second approach, and hypothesize that a CNN pretrained for FR activity can be apt for representing features that discriminate makeup (PA) and non-makeup (BF) presentations of face. Although their input requirements are same, the problems of FR and face PAD are quite different from each other. The objective of FR is to correctly identify a specific individual enrolled into the system. For PAD, all such *bona-fide* presentations constitute a single class that needs to be discriminated against PAs (which highly resemble a genuine face). Therefore, use of FR CNN towards detection of makeup attacks is not straightforward, and requires a systematic method to compute relevant feature descriptors. We employ the pretrained FR CNN as a pure feature extractor for shape and texture without any explicit transfer learning or domain adaptation. We propose a mechanism to obtain feature descriptors (from FR CNN) that are suitable for detection of makeup PAs. The overall framework of proposed PAD method is illustrated in Fig. 5.

Shape descriptor (FI-CNN): We obtain shape descriptor as the output or *embeddings* of the pre-final `FC` layer of an FR-CNN which provides a compact representation of the input face image. Since the descriptor is generated from `FC` layer, it embeds information pertaining to the overall shape. The input to the CNN is a cropped face image matching to the specifications of FR-CNN (The preprocessing steps are explained in Sec. 5). The shape descriptor is produced from a single forward pass of the CNN. Based on its extraction process, we designate this as full image CNN (FI-CNN) descriptor.

Texture descriptor (P-CNN): The texture can be captured from one or more `conv` layers of FR-CNN. The texture descriptor in [19] uses VLAD (Vector of Locally Aggregated Descriptors) pooling of last `conv` layer features to generate a $65k$-dimensional output. Since we are dealing with a two-class problem of detection of makeup attacks, we design a simple and computationally inexpensive texture descriptor. The last `conv` layer of an FR-CNN generates a $(C\times M\times N)$-dimensional output where M and N are the spatial dimensions of each of the C feature maps. It represents densely

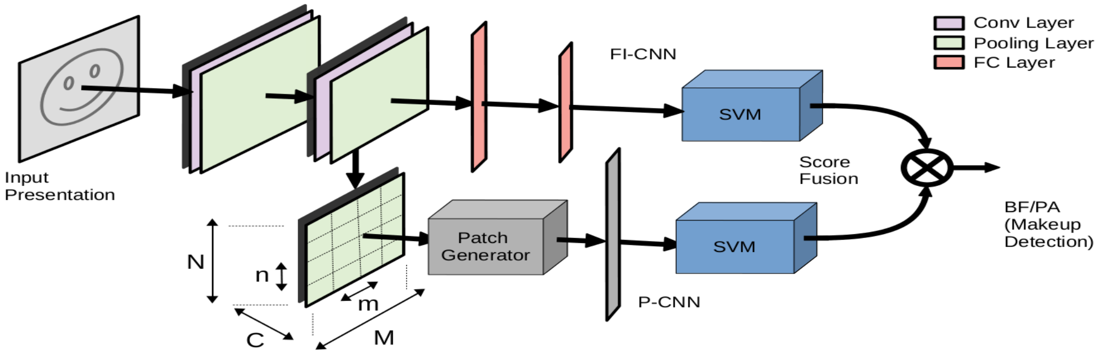


Fig. 5. Framework of the proposed PAD method for makeup attacks.

pooled local features of an input image through weight sharing across previous layers. We divide each feature map in this output into spatially non-overlapping patches of $m \times n$ dimensions, such that $M = k_1\, m$, and $N = k_2\, n$; where k_1, k_2 are some positive integers. This process produces tessellated feature maps of ($C \times m \times n$)-dimensions for each patch which are subsequently linearized. At this stage, we obtain $k_1\, k_2$ linearized feature descriptors for a given face image. The final texture descriptor is computed as the average of ($k_1\, k_2$) patch-level descriptors. This setup, to obtain a ($C\, m\, n$)-dimensional texture descriptor, is referred to as patch CNN (P-CNN).

Our texture descriptor is not independent of spatial information in strict sense. Since we linearize patch-level features, the local geometry of the patch is implicitly encoded in the descriptor. However, as the patch represents a small fraction of total image area, the effect of this information is nominal on the shape of overall image. Additionally, these features are averaged over the entire image. Therefore, the final descriptor represents the mean-local features of the image, rather than spatial information at global level. In other words, we consider multiple patches (or crops) of the input image to learn local features, and perform an `average pooling` over this stack of features.

Our texture descriptor suffices the desired objective while it is easy to calculate, and does not require separate parametric modelling or dimensionality reduction. It may be noted that the features prior to the patch generation are computed through neighbourhood-oriented operations such as convolution and pooling; and thus, the patches are not mutually exclusive in terms of information content.

Classifiers: The sets of both aforementioned feature descriptors can be classified separately to analyse their efficacy towards detecting makeup attacks. We use a two-class support vector machine (SVM) classifier in both cases. The SVM employs a radial basis function (RBF) kernel. When the PA consists of makeup applied to specific facial features, the FI-CNN is expected to perform better than P-CNN. For a general makeup PA, the solution is likely to be opposite. Hence, combining both features or corresponding scores can be the subsequent step. In this work, we propose a simple score fusion for CNN-based detection of makeup PAs.

Score Fusion: The scores of both SVM classifiers (FI-CNN and P-CNN) can be fused to improve robustness of the overall PAD system. The score-fusion rule can be defined based on specific application or *a priori* knowledge of the makeup conditions. A single score so obtained can be thresholded to generate a binary decision. Here, we employ a simple average of both classification scores to obtain a result.

5 Experimental Results

We conduct various experiments to test the performance of the proposed PAD method.[3] We analyze the impact of age-induced makeups on accuracy of the CNN-based FR method, followed by our main experiment that evaluates the proposed PAD method on our AIM dataset. Ideally, we would like to compare our work against established baselines for a similar problem. However, we did not come across any relevant research in detecting old-age makeups. Since the AIM dataset is first of its kind, there are no equivalent datasets to test the proposed method of makeup detection. Therefore, we have chosen some of the publicly available datasets of generic makeups to evaluate our work for a couple of reasons. First, availability of existing works on selected datasets facilitates comparison of our method (though on generic makeups). Second, it automatically tests efficacy of the proposed method towards detection of a different class of makeups, and hence, its adaptability.

Our next experiment is, thus, aimed at cross-validation testing on generic makeups from the YMU dataset. This combination of experiment and dataset has also been used earlier to test makeup detectors. Later, we describe new experiments that evaluate the cross-dataset performance of the proposed method on generic makeups from YMU, MIW, and MIFS datasets. In two separate instances, we used YMU and MIFS datasets to train the P-CNN and FI-CNN classifiers of the proposed method, and testing was conducted on other two datasets. The MIW, being a small dataset, was never used for training.

We begin with description of our experimental setup and evaluation measures. It follows by the preprocessing details—that are the same for all of our PAD experiments. Subsequently, each experiment is discussed along with the results.

5.1 Experimental Setup

Our proposed PAD method derives features from multiple layers of an FR-CNN. In this work, we utilize the 9-layer

[3] Python code for all experiments described in this work is available at: https://gitlab.idiap.ch/bob/bob.paper.makeup_aim

LightCNN [22] as a feature extractor. LightCNN is one of the state of the art FR systems that has demonstrated $\approx 98.8\%$ accuracy on the LFW dataset [23]. First, we preprocess the input image to obtain a face image to match the specifications of LightCNN. This CNN has two fully connected (`FC`) layers termed as `MFM_fc1` and `MFM_fc2`. We use the output (or *embeddings*) of `MFM_fc1` layer to obtain shape descriptor (FI-CNN). The patch-level descriptor (P-CNN) is computed from the output of final `conv` layer (`MFM5` as per [22]). These features are used to accomplish the task of classification; alternatively, the classification scores of FI-CNN and P-CNN descriptors are fused to obtain a final decision. The classifier generates a score for every input image or frame of every video presentation.

5.2 Performance Evaluation

We use the equal error rate (EER) on the dev set to compute the score threshold, τ_{EER} where the number of incorrectly classified *bona-fide* presentations is approximately equal to the number of incorrectly classified PAs (i.e., makeups). For CV experiments, the score threshold τ_{EER} is computed on the test set.

As per ISO/IEC 30107-3:2017 standard, a face PAD system can be evaluated using the following three measures [24].

- **APCER** (Attack presentation classification error rate) is defined as the proportion of presentation attacks (PA) incorrectly classified as *bona-fide* for a specific species of attack for a given PAD system.
- **BPCER** (Bona-fide presentation classification error rate) is defined as the proportion of *bona-fide* presentations incorrectly classified as attacks for a given PAD system.
- **ACER** (Average classification error rate) is often provided as a single measure of overall performance of a PAD system. It is calculated as the average of APCER and BPCER, i.e., $\text{ACER} = 0.50 \times (\text{APCER} + \text{BPCER})$.

In this work, we are presenting the binary classification problem of detection of makeup PAs. We designate a *bona-fide* or non-makeup presentation with *positive* or true label; and an attack presentation using makeup is designated with *negative* or false label. This convention is commonly followed by the biometric community—where the main focus is identification of *bona-fide* presentations.

The same binary classification problem can be interpreted as a generic makeup detection problem by interchanging the true/false labels. (by associating makeup presentations with positive label). This change also needs to be accompanied by appropriate changes in the performance measures that are computed on specific positive or negative labels. The overall accuracy of the classification, however, is unaffected by these interpretations as it refers to the correct identification of samples from both classes.

The relevant literature discussed in Sec. 2 follows this later approach and reports their results accordingly. Mostly, the performance of their makeup detectors is reported using classification rate (CR) which indicates the percentage of overall correct classifications. To facilitate direct comparison, we calculate its equivalent ACER value as: $ACER = 100 - CR$. This equivalence is valid only for the datasets with equal number of attack and *bona-fide* samples. The datasets- YMU, MIW, and MIFS- are balanced in terms of samples in both classes; and hence, we can calculate the corresponding ACER. Note that we have discussed the PAD problem considering *bona-fide* (non-makeup) presentations as positive.

5.3 Preprocessing

We have used the same preprocessing steps across all our PAD experiments. Since the FR-CNN acts as our feature extractor, we prepared the input image to match the specifications of the given CNN (in this case, LightCNN). It requires a grayscale face image of 128×128 pixels, specifically aligned using 5 facial landmarks.

We have used the Multi-Task Cascaded Convolutional Network (MTCNN) [25] to detect the facial region. The images were appropriately aligned and resized after obtaining the facial landmarks using Menpo [26].

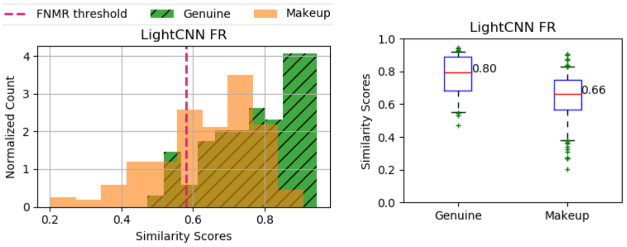
5.4 Results

First we analyze the impact of obfuscation attempts using old-age makeups on the accuracy of LightCNN FR, followed by results of the proposed PAD method on AIM dataset. Subsequently, we describe the results of other makeup detection experiments, and their comparisons against existing works wherever appropriate.

5.4.1 *Experiment I*

We analyzed the performance of LightCNN FR for *bona-fide* (non-makeup) and makeup presentations from the AIM dataset. This experiment was conducted on the subset of 20 identities from the AIM dataset—where makeup as well as *bona-fide* presentations were available. One *bona-fide* video of each subject was used for enrollment to the FR system. Other presentations of a subject, including their makeups, were compared against the enrolled presentation of the same subject. Similar to the other experiments, this experiment was performed on frame level by choosing 20 frames from each video. To probe (or verify), a total number of 73 samples were used for *bona-fide*, and 216 samples for makeup which amounts to 1460 and 4320 frames for *bona-fide* and makeups, respectively. The *embeddings* of pre-final layer of LightCNN were considered as compact representations of the input image. We chose the *cosine similarity* metric to compare the features of probe and enrollment images.

The distribution of similarity scores for the aforementioned experiment is shown in Fig. 6a. Higher similarity score indicates a better match in the range [0, 1]. We observed a general drop in the similarity scores of makeup presentations compared to the scores of *bona-fide*. The degree of reduction in scores indicates a possibility of the subject not being matched to their enrolled (true) identity; and thus, of successful obfuscation. If this FR system employs a threshold of FNMR 10% (i.e. at most 10% of *bona-fide* presentations can be falsely rejected); then 32% makeup presentations will be unverified as indicated by the corresponding FNMR threshold at 0.58. A boxplot representation in Fig. 6b indicates that the median similarity score for makeups dropped by 14% over a total range of [0, 1]. We obtained median scores of 0.80 and 0.66 for *bona-fide*



(a) (b)

Fig. 6. Similarity scores for LightCNN FR on *bona-fide* and age-induced makeups from AIM dataset. (a) Score histograms where the vertical line indicates score threshold at FNMR 10%. (b) Boxplots of scores where the median for each class is marked with red line; and the boxes indicate 1st and 3rd quartiles of scores. Whiskers are provided at 5 and 95 percentiles of the corresponding scores.

and makeup presentations, respectively. From the boxes indicating quartiles, it can be observed that similarity scores of 75% *bona-fide* presentations were higher than a similar fraction of makeup presentations. A few makeup presentations resulted in the similarity scores as low as 0.25 which is extremely unlikely to be above score threshold for a reasonable FR system. This experiment, thus, confirms that age-induced makeup impacts FR matching, and creates a powerful PA instrument.

5.4.2 Experiment II

We evaluated the performance of proposed PAD method on the AIM dataset using `grandtest` protocol. The train set is used to extract P-CNN and FI-CNN feature descriptors—which are then used to train independent SVM classifiers. The score thresholds τ_{EER} for P-CNN, FI-CNN, and score fusion are computed on the dev set. Table 3 provides performance evaluation of the proposed makeup detector. The ACER of 8.27% was obtained from the classification of patch-level features (P-CNN). The error rate was nominally reduced to 7.54% when image-level features (FI-CNN) were alone considered. A simple score-level fusion of these classifiers, however, improved the combined performance by 25%; and the ACER dropped to 6.65%. The APCER of fused scores did not improve over the use of image-level classifier; but, incorrect classification of *bona-fide* presentations reduced by a factor of three when the score fusion was applied. The increased accuracy of score fusion ascertains that texture and shape–based features are complementary towards detection of age induced makeups.

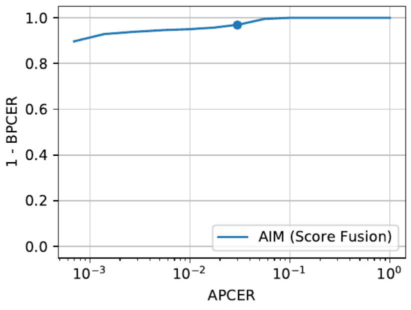


Fig. 7. ROC curve of the proposed method on the `dev` set of AIM using grandtest protocol. The encircled point indicates EER threshold.

Fig. 7 illustrates the receiver operating characteristics (ROC) curves for dev set of the AIM dataset. (All ROCs are plotted in a *semi-log* format.) The BPCER of dev set was always below 0.2 for the entire range of APCER.

5.4.3 Experiment III

Here, we tested the proposed method on the YMU dataset using a 5-fold CV protocol. Makeups in this dataset are aimed towards attractive or stylish look, and not towards old-age appearances.

Results of this experiment are provided in Table 4. When only patch-based features of CNN (P-CNN) were considered, the proposed detector yielded ACER below 10% for each of the CV protocol/trial. Similarly, we obtained accuracy of 90+% by using features from fully connected layer of CNN (FI-CNN), except for one of the five trials. The mean-level fusion of these scores resulted in ACER of 6.1% for the YMU dataset— which is 1.15% better than using P-CNN alone; and 2.50% better than that of FI-CNN. The decrease in average error rate of score fusion, similar to the previous experiment, follows the inference that P-CNN and FI-CNN features are complementary towards detection of generic makeups too. It may be noted that, except for one of five trials, score fusion has outperformed the results of using a single set of features (either P-CNN or FI-CNN). Score fusion has reduced the average classification error by nearly 20% when compared against the ACER obtained from either P-CNN or FI-CNN.

On average, 7–8 image samples from the test partition of YMU dataset were misclassified. For the same experiment, Dantcheva *et al* have reported average classification rate of 91.2% where a best trial provided 92% correct results [6]. These numbers indicate the ACER of 8.8% when averaged over 5 trials. Our best trial provided 96.67% correct classification as the corresponding ACER was 3.33%. We do not have the exact details of their CV protocol; however, we have followed the similar constraints to create a 5-fold CV protocol. In [14], the entropy-based method has achieved an ACER of 8.3% on a similar CV protocol on the YMU dataset. The accuracy of 98% has been claimed in [10] for a similar CV experiment. Since their implementation of feature-fusion method was not available, we attempted to reproduce their work. The overall experiment consists of several processing blocks (preprocessing, feature extraction, fusion, etc.), wherein the performance of each processing block can significantly vary based on its parameters, configurations, and actual implementation. We ran multiple experiments based on the details from [10] and using reasonable assumptions

TABLE 3
Performance Evaluation of the proposed method on the AIM dataset using grandtest protocol. All measure rates are in %. The numbers in parenthesis indicate the number of incorrectly classified samples for total samples in the given class.

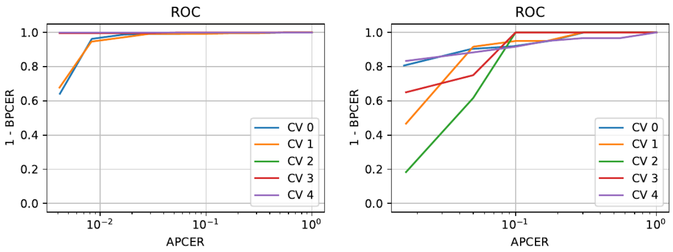
Features	ACER	APCER	# False Accept	BPCER	# False Reject
P-CNN	8.27	11.94	(172/1440)	4.59	(68/1480)
FI-CNN	7.54	10.42	(150/1440)	4.66	(69/1480)
Score Fusion	6.65	11.88	(171/1440)	1.42	(21/1480)

TABLE 4
Performance Evaluation of the proposed method on the YMU dataset for trials of cross validation protocol. All measure rates are in %. The numbers in parenthesis indicate the number of incorrectly classified samples for total samples in the given trial.

Protocol	P-CNN	FI-CNN	Score Fusion		
	ACER	ACER	ACER	APCER	BPCER
`cv0`	9.68	8.06	6.45	4.84 (3/62)	8.06 (5/62)
`cv1`	5.00	11.67	8.33	3.33 (2/60)	13.33 (8/60)
`cv2`	6.67	8.33	5.00	10.00 (6/60)	0.00 (0/60)
`cv3`	6.67	5.00	3.33	6.67 (4/60)	0.00 (0/60)
`cv4`	8.33	10.00	7.50	6.67 (4/60)	8.33 (5/60)
Average	7.27	8.61	6.12	6.30	5.94

wherever the details were lacking. We were able to obtain the ACER in the range of 12–15% for the feature-fusion based method on the YMU dataset.

The ROC curves for each trial on the training and testing sets is shown in Fig. 8. For training sets, the nature of the ROC curves demonstrate a near-perfect behavior of the proposed PAD method on YMU dataset.



(a) Training folds (b) Testing fold

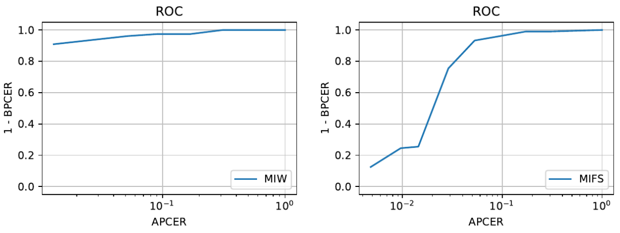
Fig. 8. ROC curves of the proposed method on the YMU dataset using a 5-fold cross validation protocol.

5.4.4 Experiment IV

In this experiment, we evaluated the cross-dataset efficiency of the proposed makeup detector on generic makeups. We trained SVM classifiers from P-CNN and FI-CNN using the entire YMU dataset; and tested them over the MIW and MIFS datasets. Since our fusion averages the scores obtained from these individual classifiers, no explicit model needs to be trained for fusion. The P-CNN pretrained on YMU obtained an ACER of 6.5% on MIW dataset; and the average classification error of FI-CNN was found to be 10.4%. On fusing both scores, the ACER of makeup detector significantly decreased below 4% where only 6 of 154 images were incorrectly classified. In [6], a classification rate of 95.4% has been reported for a similar experiment which refers to ACER of 4.6%. Other works in [10], [14] reported accuracies in the range of 98% on MIW dataset when the classifiers were trained on the YMU dataset. When we conducted the same experiment using our implementation of their code, the ACER of 13% was obtained. The results in Table 6 demonstrate *(i)* cross-dataset generalization of the proposed method; and *(ii)* efficacy of combining P-CNN and FI-CNN features. The BPCER of this cross-dataset experiment is consistently low over a large range of APCER as indicated by ROC in Fig. 9a.

TABLE 5
Performance comparison of makeup detectors over the YMU and MIW datasets. All measure rates are in %. ∗ refers to the results reported in corresponding publications; and † indicates reproduced results from our implementation.

Method	Expt III: YMU CV		Expt IV: MIW test	
	Accuracy	ACER	Accuracy	ACER
Dantcheva *et al* [6]*	91.20	8.80	95.45	4.55
Liu *et al* [14]*	91.72	8.28	98.05	1.95
Kose *et al* [10]*	98.50	1.50	99.35	0.65
[10] reproduced†	85.07	14.93	87.01	12.99
Proposed work	93.88	6.12	96.10	3.90



(a) Test dataset: MIW (b) Test dataset: MIFS

Fig. 9. ROC curves of the proposed method over different test datasets in cross-dataset experiment. The classifiers were trained on the entire YMU dataset.

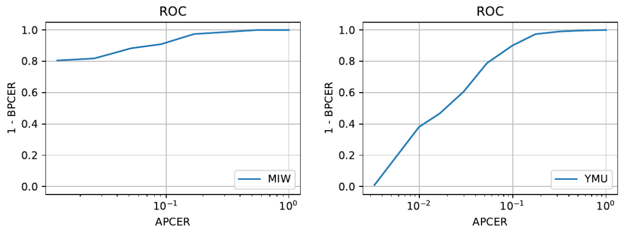
In a subsequent experiment, where the MIFS dataset was tested using the same pretrained SVM models (using YMU), we obtained the ACER of 10.3% and 7.7% for P-CNN and FI-CNN features, respectively. The score fusion resulted in a nearly 1% improvement with ACER of 6.7% (28 incorrect classifications out of 416) on MIFS dataset. The corresponding ROC curve is shown in Fig. 9b.

In an additional cross-dataset experiment, similar to previous one; here we have used the MIFS dataset to train SVMs from the P-CNN and FI-CNN features; and test the performance over YMU and MIW datasets. Table 7 provides the evaluation of these experiments. We obtained the ACER of 9% on the MIW dataset after fusing scores from P-CNN and FI-CNN classifiers, which is same as the performance of P-CNN classification. A total of 14 images were misclassified, which is higher number than the previous experiment where the classifiers were trained on a different dataset. When the entire YMU dataset was tested in this setup, we obtained ACER of 8.8% on fused results. In both cases (cross-dataset with MIFS to train), the patch-level features provided 2% lesser error rates than those corresponding to image-level features.

ROCs of these experiments are provided in Fig. 10. For MIW dataset, the ROC curve is similar to the one obtained in

TABLE 6
Performance Evaluation of the proposed method in cross-dataset experiment. The classifiers were trained on the entire YMU dataset. All measure rates are in %. The numbers in parenthesis indicate the number of incorrectly classified samples for total samples in the given dataset.

Test	P-CNN	FI-CNN	Score Fusion		
Dataset	ACER	ACER	ACER	APCER	BPCER
MIW	6.49	10.39	3.90	2.60 (2/77)	5.19 (4/77)
MIFS	10.34	7.69	6.73	7.21 (15/208)	6.25 (13/208)



(a) Test dataset: MIW (b) Test dataset: YMU

Fig. 10. ROC curves of the proposed method over different test datasets in cross-dataset experiment. The classifiers were trained on the MIFS dataset.

the previous experiment with YMU-pretrained classifiers. A relatively poorer ROC and results of fusion for YMU dataset indicate that perhaps a higher order fusion scheme needs to be developed.

TABLE 7
Performance Evaluation of the proposed method in cross-dataset experiment. The classifiers were trained on the MIFS dataset. All measure rates are in %. The numbers in parenthesis indicate the number of incorrectly classified samples for total samples in the given dataset.

Test	P-CNN	FI-CNN	Score Fusion		
Dataset	ACER	ACER	ACER	APCER	BPCER
MIW	9.09	11.04	9.09	6.49 (5/77)	11.69 (9/77)
YMU	9.60	11.92	8.77	11.59 (35/302)	5.96 (18/302)

6 Discussions

Our experiment on accuracy of LightCNN FR shows that age induced makeups lead to considerable decrease in recognition scores. For FNMR of 10%, one of three makeup PAs remain undetected by one of the state-of-the-art FR system. A similar observation was reported in [5] for impact on FR accuracies due to traditional generic makeups. Our study reinforces the threat of makeups, the ones mimicking older look in particular, in obfuscating modern FR systems.

CNNs have delivered promising results across a wide range of image recognition problems. However, limited amounts of training data often restrict the use of CNNs. In the proposed method, we hypothesize that the shape- and texture-based features derived from an FR CNN can be directly utilized towards detection of age induced makeup attacks. The results of our experiment indicate that such multi-layer deep features are indeed effective at discriminating makeups. Thereby, this work demonstrate that a rich hierarchy of features in FR CNN can be exploited for different problems related to face biometrics. This work also introduces a simple and computationally inexpensive method to obtain a texture descriptor from a `conv` layer of a CNN.

From the experiments with traditional makeups, it can be seen that the P-CNN and FI-CNN features are successful across different classes of makeup PAs as well. For most experiments, the proposed method surpassed the performance of existing methods of traditional makeup detection. This is an important observation since it confirms that the proposed method generalizes well to different classes of makeup attacks; and is not biased towards simply learning age-related features when applied on AIM dataset.

When the scores of individual P-CNN and FI-CNN classifiers were fused, we observed overall improvement in PAD results. Although the degree of such improvement is not consistent across experiments, fusion has resulted in improvements in APCER, BPCER, as well as ACER in most cases.

A natural extension to the aforementioned experiments is to train the proposed PAD method on a certain class of makeup attacks (either old age, or generic ones); and test its performance on another class of makeups. This will evaluate the generalizability of the proposed method across a variety of makeups. We conducted few experiments where the P-CNN and FI-CNN classifiers were trained on the AIM dataset; and tested these models on the YMU and MIW datasets. It was observed that the method performs very poorly. We conclude that, in its present formulation, the proposed PAD method adapts extremely well when trained and tested on the similar class of makeups; however it does not generalize when the class of makeups in training is different from the testing.

7 Conclusion

We have proposed a CNN-based face PAD method to detect obfuscation using age-induced facial makeups. The makeup detector computes patch-based (P-CNN) and image-based (FI-CNN) features to extract texture and shape related information, respectively. Improved accuracy from a simple score-level fusion illustrates complementary nature of P-CNN and FI-CNN features. We have also demonstrated that a pretrained face recognition CNN (FR CNN) can be directly used for detection of makeups without any fine tuning or domain adaptation.

The proposed makeup detector has been tested on newly created AIM dataset. It consists of PAs based on age induced makeups with variety in terms of gender, ethnicities, and capture conditions. Our experiments indicate that median matching scores for FR using LightCNN were degraded by 14% for these makeups. The proposed PAD method provided an ACER of 6.6% on the grandtest protocol of the AIM dataset. We have also tested the proposed makeup detector on three datasets of unconstrained traditional facial makeups. Our experiments include in-dataset (using cross validation) and cross-dataset testing. We obtained ACER of 6.1% on the YMU dataset using cross validation protocol; and cross dataset ACER values in the range of 4–9% on different combinations of training and testing datasets. These classification results are comparable, if not better, than other state-of-the-art methods of makeup detection.

This work opens up several new research problems. The classifier models for detection of old-age makeups (from AIM) and other generic makeups are different. The detection of attacks encompassing a mix of these makeups is further challenging. A PAD method that generalizes across different categories of makeups needs to be investigated. It is also important to test such models on genuine presentations of older persons to evaluate their capability of correctly identifying makeups. We have demonstrated our proposed approach using LightCNN. Given the availability of different FR CNNs, it will be useful to compare their capabilities at identifying makeups. A generic CNN consists of several convolutional (conv) layers, and often multiple fully connected (FC) layers. The choice of layers to extract P-CNN and FI-CNN features needs to be experimented further. Similarly, improvement in fusion scheme can be another possible extension.

Acknowledgments

This work has been supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The work has also been supported by the Swiss Center for Biometrics Research and Testing.

References

[1] ISO, "Information Technology—Biometric presentation attack detection—Part 1: Framework," International Organization for Standardization, Geneva, CH, Standard, 2016.

[2] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa, "Disguised faces in the wild," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[3] T. Dhamecha, R. Singh, M. Vatsa, and A. Kumar, "Recognizing disguised faces: Human and machine evaluation," *PLOS ONE*, vol. 9, pp. 1–16, 07 2014.

[4] T. Wang and A. Kumar, "Recognizing human faces under disguise and makeup," in *Proc. Int. Conf. Identity, Security and Behavior Analysis (ISBA)*, Feb. 2016, pp. 1–7.

[5] A. Dantcheva, C. Chen, and A. Ross, "Can facial cosmetics affect the matching accuracy of face recognition systems?" in *Proc. Int. Conf. Biometrics: Theory, Applications and Systems*, Sep. 2012, pp. 391–398.

[6] C. Chen, A. Dantcheva, and A. Ross, "Automatic facial makeup detection with application in face recognition," in *Proc. Int. Conf. on Biometrics*, June 2013, pp. 1–8.

[7] C. Chen, A. Dantcheva, T. Swearingen, and A. Ross, "Spoofing faces using makeup: An investigative study," in *Proc. Int. Conf. Identity, Security and Behavior Analysis*, Feb. 2017, pp. 1–8.

[8] S. Varshovi, "Facial makeup detection using hsv color space and texture analysis," Master's thesis, Concordia University, September 2012.

[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Computer Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[10] N. Kose, L. Apvrille, and J. Dugelay, "Facial makeup detection technique based on texture and shape analysis," in *Proc. Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, vol. 1, May 2015, pp. 1–7.

[11] G. Guo, L. Wen, and S. Yan, "Face authentication with makeup changes," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 814–825, May 2014.

[12] S. Wang and Y. Fu, "Face behind makeup," in *Proc. AAAI Conference on Artificial Intelligence*, ser. AAAI'16, 2016, pp. 58–64.

[13] S. Rasti, M. Yazdi, and M. A. Masnadi-Shirazi, "Biologically inspired makeup detection system with application in face recognition," *IET Biometrics*, vol. 7, no. 6, pp. 530–535, 2018.

[14] K. Liu, T. Liu, H. Liu, and S. Pei, "Facial makeup detection via selected gradient orientation of entropy information," in *Proc. Int. Conf. Image Processing*, Sep 2015, pp. 4067–4071.

[15] E. Derman, C. Galdi, and J. Dugelay, "Integrating facial makeup detection into multimodal biometric user verification system," in *Proc. Int. Workshop on Biometrics and Forensics*, Apr 2017, pp. 1–6.

[16] J. Hu, Y. Ge, J. Lu, and X. Feng, "Makeup-robust face verification," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, May 2013, pp. 2342–2346.

[17] Y. Sun, L. Ren, Z. Wei, B. Liu, Y. Zhai, and S. Liu, "A weakly supervised method for makeup-invariant face verification," *Pattern Recognition*, vol. 66, pp. 153–159, 2017.

[18] Y. Li, L. Song, X. Wu, R. He, and T. Tan, "Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification," in *Proc. AAAI Conf. Artificial Intelligence*, 2018.

[19] C. Mircea, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, June 2015.

[20] V. Andrearczyk and P. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016.

[21] H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning Face Age Progression: A Pyramid Architecture of GANs," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2018, pp. 31–39.

[22] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[23] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[24] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys*, vol. 50, no. 1, pp. 8:1–8:37, 2017.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *CoRR*, vol. abs/1604.02878, 2016.

[26] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 679–682.

Ketan Kotwal received the M.Tech and Ph.D. degrees in electrical engineering from the Indian Institute of Technology Bombay (IIT Bombay, Mumbai, India). His current research interests include various topics in image processing, machine learning, and data processing. He has also been actively involved in technology consulting in relevant areas.

He received the Excellence in Ph.D. Thesis Award from the IIT Bombay, and the Best Ph.D. Thesis Award from the Computer Society of India for his doctoral work. Dr. Kotwal is a co-author of research monograph "Hyperspectral Image Fusion" (Springer, US). At present, he is a member of the Biometrics Security and Privacy Group at the Idiap Research Institute.

Zohreh Mostaani obtained the B.Sc. in Electrical Engineering from University of Tehran, Iran, and M.Sc. in Electrical and Electronics Engineering from Ozyegin University, Turkey. She is currently working as a Research and Development Engineer in the Biometrics Security and Privacy group at Idiap Research Institute. Her research intrests are Computer vision, Machine learning, and Biometrics.

Sébastien Marcel received the Ph.D. degree in signal processing from Université de Rennes I, Rennes, France, in 2000 at CNET, the Research Center of France Telecom (now Orange Labs). He is currently interested in pattern recognition and machine learning with a focus on biometrics security. He is a Senior Researcher at the Idiap Research Institute (CH), where he heads a research team and conducts research on face recognition, speaker recognition, vein recognition, and presentation attack detection (anti-spoofing). He is a Lecturer at the Ecole Polytechnique Fédérale de Lausanne (EPFL) where he teaches a course on "Fundamentals in Statistical Pattern Recognition." He is an Associate Editor of IEEE Signal Processing Letters. He has also served as Associate Editor of IEEE Transactions on Information Forensics and Security, co-editor of the "Handbook of Biometric Anti-Spoofing," Guest Editor of the IEEE Transactions on Information Forensics and Security Special Issue on "Biometric Spoofing and Countermeasures," and co-editor of the IEEE Signal Processing Magazine Special Issue on "Biometric Security and Privacy." He was the Principal Investigator of international research projects including MOBIO (EU FP7 Mobile Biometry), TABULA RASA (EU FP7 Trusted Biometrics under Spoofing Attacks), and BEAT (EU FP7 Biometrics Evaluation and Testing).