Improved RGB-D-T based Face Recognition

Oliu Simon, Marc; Corneanu, Ciprian; Nasrollahi, Kamal; Guerrero, Sergio Escalera; Nikisins, Olegs; Sun, Yunlian; Li, Haiqing; Sun, Zhenan; Moeslund, Thomas B.; Greitans, Modris

Published in:
IET Biometrics

DOI (link to publication from Publisher):
10.1049/iet-bmt.2015.0057

Publication date:
2016

Document Version
Accepted manuscript, peer reviewed version

Improved RGB-D-T based Face Recognition

Marc Oliu Simón[1], Ciprian Corneanu[1], Kamal Nasrollahi[2*], Olegs Nikisins[3], Sergio Escalera[1], Yunlian Sun[4], Haiqing Li[4], Zhenan Sun[4], Thomas B. Moeslund[2], and Modris Greitans[3]

[1] Human Pose Recovery and Behavior Analysis (HuPBA) Group, University of Barcelona, Computer Vision Center, 08193 Bellaterra (Cerdanyola), Barcelona, Spain
[2] Visual Analysis of People (VAP) Laboratory, Aalborg University, Rendsburggade 14, 9000, Aalborg, Denmark
[3] Institute of Electronics and Computer Science, Dzerbenes 14, LV-1006, Riga, Latvia
[4] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), 95 Zhongguancun East Road, Beijing, 100190, China
[*]kn@create.aau.dk

Abstract: Reliable facial recognition systems are of crucial importance in various applications from entertainment to security. Thanks to the deep-learning concepts introduced in the field, a significant improvement in the performance of the unimodal facial recognition systems has been observed in the recent years. At the same time a multimodal facial recognition is a promising approach. This paper combines the latest successes in both directions by applying deep learning Convolutional Neural Networks (CNN) to the multimodal RGB-D-T based facial recognition problem outperforming previously published results. Furthermore, a late fusion of the CNN-based recognition block with various hand-crafted features (LBP, HOG, HAAR, HOGOM) is introduced, demonstrating even better recognition performance on a benchmark RGB-D-T database. The obtained results in this paper show that the classical engineered features and CNN-based features can complement each other for recognition purposes.

1. Introduction

Biometric systems, including DNA, face, iris, palmprint, fingerprint, to name a few, are systems that use human behavioural or physical characteristics for identification purposes. Facial biometrics, as opposed to many other biometrics, can be collected contact-free. Face recognition algorithms, however similar to other biometrics systems, have to deal with a wide spectrum of challenges, like: intra-class variations, noise in the system, spoofing attacks and constantly increasing performance requirements. Several algorithms have been developed for face recognition to deal with these challenges. Examples of such algorithms, include but are not limited to, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [24, 27, 28], Local Binary Patterns (LBP) [1], Histograms of Oriented Gradients (HOG) [5], Haar-like rectangular features (HAAR) [16], and Histograms of Gabor Ordinal Measures (HOGOMs) [4].

Just recently a feature learning approach based on deep learning techniques has received enormous attention in computer vision generally, and in recognition systems specifically, including face recognition [25]. Systems based on features learned by deep learning have reported outstanding results outperforming most of the other existing face recognition methods. These systems are getting close to fill up the gap between human and machine performances in recognition tasks. This phenomenon can be explained by the large amount of data currently available for learning and growing computational power of the devices used for training. Companies like Google, Facebook, and Baidu are building face recognition software based on deep learning concepts [25].

Parallel to the development of better features and classifiers for face recognition, multimodality has also been used as a solution for developing systems [22]. The overall idea is that different modalities should complementary contribute to the recognition process. Additionally Depth and Thermal information can increase robustness to head pose variations and changing illumination contexts. Depth also contributes with additional information about subtle changes in the geometry of the face. Multimodal face recognition systems have usually used two modalities [8, 14, 23, 29, 32]. Similar approaches for different, but related problems, like full body gender recognition have also been proposed [33]. Recently a database containing synchronised RGB, Depth, and Thermal (hereafter, RGB-D-T) images has been introduced in [18], including facial images of different rotations, illuminations, and expressions. The authors have highlighted the complementarity of different modalities motivating further research in this direction.

The novelty of the proposed system in this paper is twofold:

- Advancing the recent successful deep learning-based facial features, by combining them with classical hand-crafted facial features, like LBP, HOG, HAAR, and HOGOM for the purpose of finding complementary nature of these features through conducting extensive experiments. These experiments show that at least one of these classical facial features, specifically HOGOM [4], can complementarily contribute to the recognition performance obtained by deep learning-based facial features.
- The above mentioned fusions of facial features, have been applied to the multimodal RGB-D-T facial database of [18], highlighting the effect of the proposed feature fusion in advancing the recognition results.

To the best of our knowledge, none of the above mentioned points, i.e., fusion of deep learning based features and classical appearance based facial features, specifically with HOGOM, and using a tri-modal facial database with such a fusion, have been studied before. The rest of this paper is organized as follows: the related work is reviewed in Section 2. The details of the proposed multimodal face recognition system are given in Section 3. The experimental results are then elaborated in Section 4 and finally Section 5 concludes the paper.

2. Related Work

This related work section has been written in three parts: first, we review some of the multimodal face recognition systems. Then, the related state-of-the-art deep learning-based systems are discussed. Next, the details of the HOGOM features which are shown to contribute positively to the fusion with deep learning-based facial features are revisited.

2.1. *Multimodal Face Recognition Systems*

Generally, face modalities are robust against some facial appearance variations, while sensitive to others. For example, depth images can handle changes caused by different poses to some extent, but they are susceptible to expression variations. For thermal images, they are sensitive to temperature changes of the surrounding environment. Thus, a reasonable way to utilize these modalities is to fuse them for the reduction of diverse corrupting factors, which usually affect different modalities in different ways [2].

Depth information has been widely used for face recognition using different techniques [3]. A 3D model can capture the geometry information of a face, which cannot be well represented in a 2D image. Range image is a common representation used in 3D applications. 3D based techniques do offer a more suitable description of facial features than 2D models, increasing the robustness to viewpoint and lighting variations. However, the high cost of specialized 3D sensors limits their use in practical applications. With the progress in sensor technology, low cost sensors capable of capturing 3D information have been developed paving the way for multimodal systems based on RGB and depth (RGB-D). Such RGB-D cameras usually provide synchronised images of both color and depth. The color image characterizes the appearance and texture information of a face, while the depth image provides the distance

of each pixel from the camera, thus representing the geometry of the face to a certain degree. Exploiting RGB-D images has become popular in tackling various computer vision problems [9, 10, 21]. In [8, 14, 23], such RGB-D images have been used for face recognition. Another less commonly used modality is the infrared imagery [12]. A thermal infrared image records the amount of infrared radiation emitted by an object. The amount of radiation increases with temperature, therefore, this imagery allows us to see variations in temperature. When viewed through a thermal imaging camera, humans and other warmblooded animals become easily visible against the cool environment, with or without visible illumination. It has been shown that thermal images offer a promising alternative to intensity images for handling facial variations in face appearance due to changes of illumination [2]. In [18], a RGB-D-T face database was collected to perform face identification using synchronised RGB, depth and thermal images. In most cases, using such three modalities provides better identification performance than isolated and bi-modal approaches.

2.2. *Deep Learning-based Face Recognition Systems*

There are different deep learning techniques for face recognition. The main one is based on the idea of employing deep Convolutional Neural Network (CNN). CNN has been used for face recognition about two decades ago [13], however, it has achieved significant decrease in the recognition error rates just in recent years. A bright example of deep learning success in facial recognition has been introduced in [25] in which a deep network containing more than 120 million parameters has been trained on the largest available database of four million facial images of 4000 individuals. The proposed performance is stated to be approaching the human-level performance on the Labeled Faces in the Wild (LFW) database [11]. Another interesting video face recognition method based on deep learning is introduced in [7] wherein a stack of denoising autoencoders and deep Boltzmann machines are used. The fact of selecting significant frames for facial feature extraction and matching, and employing deep learning has resulted in state-of-the-art performance on the Youtube Faces and Point and Shoot Challenge in [7].

In [26] the authors enhanced the learning principles of deep convolutional networks and applied these discoveries to face recognition task. First, they have identified and explained the role of the bottleneck of the CNN as an important regularizer between training-set specificity and generality. Then, they have replaced the naive random subsampling of the training set with

a more advanced bootstrap process. Based on the proposed ideas authors have managed to improve face recognition accuracy on the LFW database [11].

2.3. *HOGOM-based Face Recognition*

Discriminative facial features should describe the dissimilarities between different individuals and remain stable during the variation of the same individual. However, discriminability and robustness are difficult to be satisfied at the same time. Gabor filters can extract discriminative local texture of faces, but are not robust against large expression and illumination variations. In order to enhance the robustness of Gabor features, Chai et al. very recently encoded the Gabor filtering responses by using Ordinal Measures (OMs), which is named Gabor Ordinal Measures (GOMs) [4]. Unlike Gabor filters which calculate the quantitative values of filtering results, OMs encode the qualitative relationship between signals in different locations.

GOMs first convolve a facial image with Gabor filters, and then binarize the magnitude, phase, real and imaginary filtering response using OMs. After that the OM binary codes are encoded into integers which can be treated as texture primitives. The statistical distributions of these texture primitives in different blocks are concatenated to form the final GOMs feature, and are used in form of Histograms of GOMs (HOGOM). HOGOMs combine the discriminability of Gabor filters and the robustness of OMs, and have achieved state-of-the-art performance on publicly available face image databases, such as FERET [19], AR [15] and large scale FRGC v2.0 [20].

The proposed system in this paper fuses these two types of state-of-the-art features: the hand-crafted feature of HOGOM, and the the facial features obtained by deep learning. The experimental results show that such a fusion advances the performance of both of these types of features when applied to a multimodal face recognition dataset.

3. The Proposed System

To find proper hand-crafted feature to be fused with deep learning-based facial features, we have extracted four well-known such features: LBP, HOG, HAAR, and HOGOM. Following the block diagram of the proposed system, shown in Figure 1, these hand-crafted features are extracted from each modality and independently normalized and, in the case of

HOGOM, reduced. These features are then concatenated into a single feature vector for training the Weighted Nearest Neighbour Classifier (WNNC) identical to the one in [18]. At the same time, each modality is also processed by modality-specific CNN, performing a stratified sampling and augmentation at each training epoch. Lastly, the final classifier is obtained by fusing WNNC and CNN classifiers with different weights.

The following subsections cover the HOGOM feature descriptors, the CNN architecture and training procedure, and the fusion between multiple modalities and classifiers. HOGOM and CNN have been further discussed in the following subsections because their fusion produces the best results, the details of the other descriptors can be found in their related references: LBP in [1], HOG in [5], and HAAR in [17].

3.1. *HOGOM*

For extracting the HOGOM descriptor, faces were cropped using the ground truth provided in the dataset [18]. Ground truth data includes parameters of the face bounding box, which is based on the depth images. It provides reasonable stability of the region of interest (ROI) on a frame to frame level in the input sequences of the images and does not generate false detections. The first stage of the ROI detection algorithm is binarization of the depth image allowing to suppress the background. Next, the pixels corresponding to the body region are set to zero (the part exceeding the predetermined width limit). In the last step the rectangular region growing algorithm is applied on the binary image. The algorithm iteratively increments the size of the rectangular region, which is centered in the position of the mean of the binary image. The cost function is the ratio of sum of pixel values to the number of pixels in the current ROI. The iteration is terminated when the cost value is below the specified limit determining the parameters of the face bounding box. This box is used as a ground-truth for the database.

The cropped faces were then resized to 128×96 pixels. The HOGOM descriptor has a dimensionality $D_{HOGOM} = N_{scales} \times N_{orientation} \times N_{bins} \times N_{blocks}$. In our experiments we have used N_{scales} =5, $N_{orientations}$ =8, N_{bins}=16 and block of size is 32×32 pixels which divides the images in 12 blocks. The feature vector is then first normalized. Then, its dimensionality is reduced by half using PCA. Final dimension of the feature vector is 3840. Finally, following [18] two classification methods were used for testing the performance of the descriptor when used alone: WNNC and a linear SVM. Choosing these classifiers allows fair comparison of the results of the proposed system against those in the original paper. In the multimodality case the feature

vectors for each modality are concatenated before training the multimodal classifiers. The reader is referred to Tables 1-7 and Figures 4-5 for a presentation of the results and to Section 4 for further discussion.

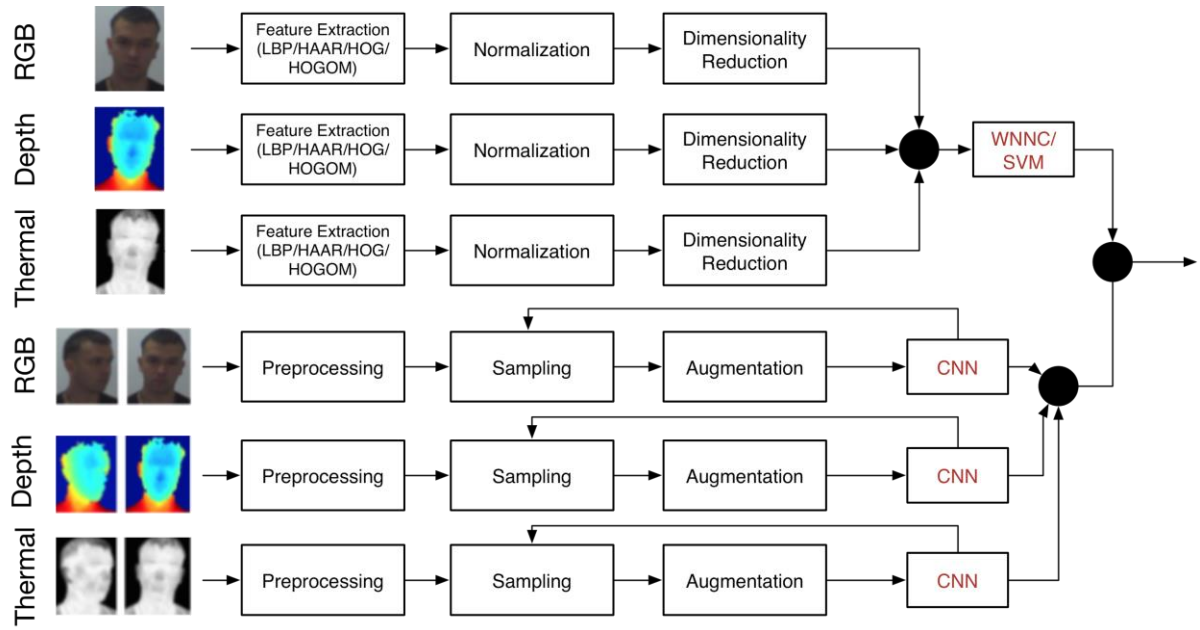

Figure 1: The block diagram of the proposed system. RGB, Depth and Thermal captures of the face are used for training modality specific CNNs for deciding if two samples are from the same person or not. The results are fused with a HOGOM trained WNNC and SVM.

3.2. *Convolutional Neural Networks*

The proposed CNN approach is implemented as a binary classification problem predicting whether two input frames correspond to the same person or not. The approach takes as input the rescaled facial regions of the two modality-specific frames to be compared as separate channels, training a separate CNN for each modality. The topology selected for the CNN is shown in Figure 2. It consists of three convolutional layers each followed by a 2×2 max-pooling. The first convolutional layer C_1 has 8 convolutional masks of size 9×9, the second layer C_2 has 16 masks of size 7×7, and the third layer C_3 has 16 masks of size 5×5. Finally, there is a 100-unit fully-connected layer, and an output layer with a single neuron, which predicts whether the two images are from the same subject or not. All neurons in the topology are Rectified Linear Units (ReLu) [6].

Preprocessing of the images is done first by transforming the facial bounding boxes of each modality into squared regions. The narrower dimension is expanded in order to match the

other one, afterwards cropping the squared region and rescaling it to a size of 100 x 100 pixels are performed. For the color and thermal modalities, the resulting images are normalized by dividing the intensity values by 255, also transforming the color images to grayscale. In the case of depth, before normalizing the depth maps, these are brought to the camera by subtracting the value of the nearest pixel, afterwards setting the maximum distance to 200 mm by thresholding the values that are further than this distance. This distance has been manually chosen in order to keep only the pixels corresponding to the head based on the two depth training set frames (first and last one) of each rotation sequence that are used for training regardless of the training partition size. The mean for each modality, calculated over the training instances, is subtracted from each frame.

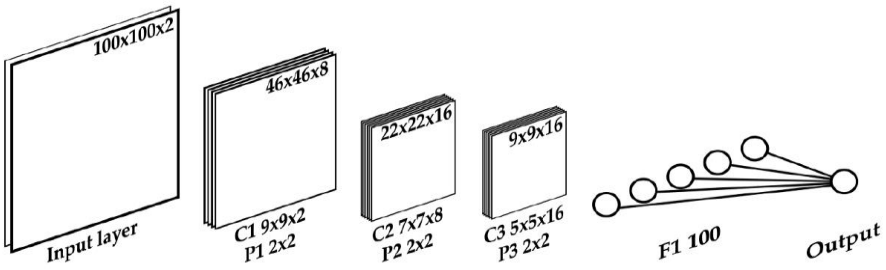

Figure 2: Topology of the proposed CNN. It consists of three convolutional layers, each followed by a 2×2 max pooling layer, and finally, a fully connected layer.

Stratified sampling is applied at each epoch to randomly select a subset of training instances, such that there is an equal amount of positive and negative instances, and an equal number of instances from each subject for both positive and negative instances. This kind of sampling is required as only around 4% of all possible pairs correspond to positives.

Augmentation is used to increase the number of training instances available. Although all pairs of training frames can be used as training samples, providing a maximum of $(51\times25)^2$ = 1,625,625 possible training pairs for the database in [18], the variation between contiguous frames is very small. In order to address this problem, at each epoch the sampled pairs are randomly flipped horizontally with a probability of 0.5. Furthermore, each individual frame for each pair is randomly displaced vertically and horizontally by ±5% and random noise in the ±0.05 range is introduced.

3.3. *Multimodal and Model Fusion*

Fusion of different modalities is done following two different approaches for WNNC and SVM by one side, and CNN. In the case of WNNC/SVM, early fusion of the features is applied to obtain a single multimodal model by concatenating the feature vectors of different modalities and training a single model using these features. In the case of CNNs, weighted late fusion is used, finding the optimal weights combining the outputs of each modality through least squares. While early fusion, both at the input image and the last hidden layer levels, was also considered, the resulting models were prone to overfitting in both cases.

Weighted late fusion is used again when combining WNNC and CNN models, evaluating all possible pairs for each mode of variation. When fusing a WNNC with a CNN classifier, the three CNNs corresponding to each modality are directly combined with the target multimodal WNNC classifier instead of using the multimodal CNN classifiers. This is done in order to prevent restricting the weights assigned to each CNN modality, while at the same time allowing for early fusion of the modalities in the WNNC classifier, which should convey more flexibility.

4. Experimental Results

Experiments were conducted on the multimodal RGB-D-T face dataset first introduced in [18]. This dataset consists of 51 persons, mostly Caucasian men between 20 and 40 years old. Faces were captured under varying rotation, illumination and facial expression conditions in a controlled laboratory context (see Figure 3). Microsoft Kinect was used to capture RGB and Depth and an AXIS Q1922 sensor was used for capturing Thermal images. In total the dataset consists of 100 images for each variation context in each of the three modalities. This amounts to 900 images for each subject and a total of 45900 images for the whole dataset.

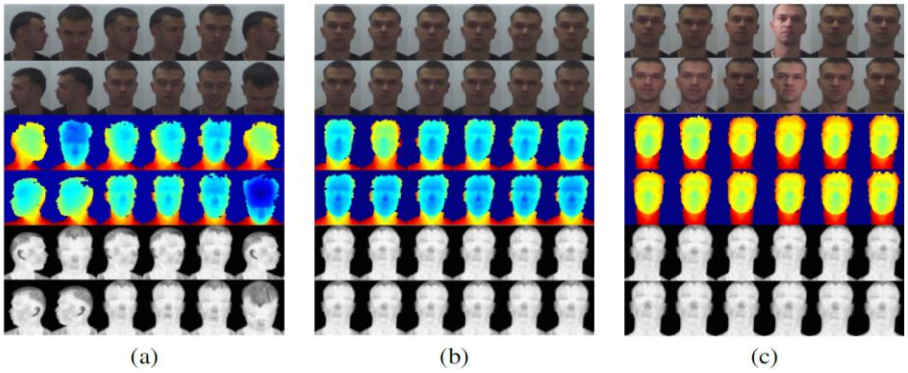

Figure 3: Example images from the RGB-D-T dataset used for experimenting: (a) Rotation. (b) Facial Expression. (c) Illumination variation of the same person captured in RGB, Depth and Thermal [18].

During the experiments we have used the same sample splitting protocol between training and testing samples as provided in [18]. Training was performed using 2, 5, 10 or 25 evenly spaced samples. 50 samples were used for testing. The ground truth ROI around the face provided by the dataset was used for cropping the facial regions of the samples.

The results of the the WNNC and the SVM classifier trained with HOGOM were relatively modest compared with the descriptors previously used in [18]. For brevity and ease of understanding we present only the results obtained with the highest number of training samples. In Tables 1 to 4 we show Equal Error Rate (EER) for each modality and for all of them combined (early fusion is used as depicted in Figure 1). The results were obtained by using 25 training samples and extracting several different hand-crafted descriptors (LBP in Table 1, HOG in Table 2 and HOGOM in Table 3). In all cases we show the classification error obtained both by WNNC and SVM. In Table 4 we show results when training a CNN in similar conditions. In Figures 4 and 5 we show how the error dynamic depending on the number of training samples. For a complete set of results with different numbers of training samples, the interested reader is referred to Tables 8 and 9 in the Appendix.

Table 1: Error over test data in % for LBP when trained with 25 samples per class. Results from [18].

		RGB	D	T	All
WNNC	ROT.	23.6	21.8	31.5	18.5
	EXP.	0.7	1.9	1.2	0.7
	ILLUM.	10	4.8	3.8	3
SVM	ROT.	0.1	0.2	0.3	0
	EXP.	0.1	0	0.1	0
	ILLUM.	0	0	0	0

Table 2: Error over test data in % for HOG when trained with 25 samples per class. Results from [18].

		RGB	D	T	All
WNNC	ROT.	21.2	35.5	26.4	22.4
	EXP.	0.6	6.5	0.8	0.6
	ILLUM.	6.5	14.3	4.5	4
SVM	ROT.	0.1	1.6	0.3	0
	EXP.	0	0.1	0.1	0
	ILLUM.	0.2	0.4	0	0.1

Table 3: Error over test data in % for HOGOM when trained with 25 samples per class.

		RGB	D	T	All
WNNC	ROT.	29.7	35.3	28.4	33.0
	EXP.	3.6	5.9	2.2	2.2
	ILLUM.	13.2	12.4	5.2	6.3
SVM	ROT.	0.7	0.3	0.2	0
	EXP.	0.1	0	0	0
	ILLUM.	0	0.2	0	0

Table 4. Error over test data in % for CNN when trained with 25 samples per class.

		RGB	D	T	All
CNN	ROT.	5	18.6	12.1	4.3
	EXPR.	0.1	2.1	0.4	0
	ILLUM.	1.9	2.9	1.1	0.5

From Tables 1 to 4, and also Figure 4 and Figure 5, one can see that the highest EER is produced in the case of rotation because of large variability of the face appearance. Regarding modalities, error is generally higher for the depth modality (except in the case of LBP representation, see Table 1) and modality fusion does not systematically decrease the error below all the other individual modalities. In the case of the linear SVM trained with HOGOM when increasing the number of training samples to 25, classification error reduces practically to zero (Table 3).

In the case of CNNs, the individual modality predictions outperform in most cases all the descriptors extracted with WNNC, but have a higher error when compared with SVM methods (see Table 4). It is however important to note that the One-Versus-All approach used for the SVM models is not capable of working with new unseen subjects. This highly constrains their applicability and was the main reason for proposing new descriptors obtaining better results in a more general scenario. Weighted late fusion has been found to significantly reduce the error for the three modes of variation when using CNNs, especially for illumination changes.

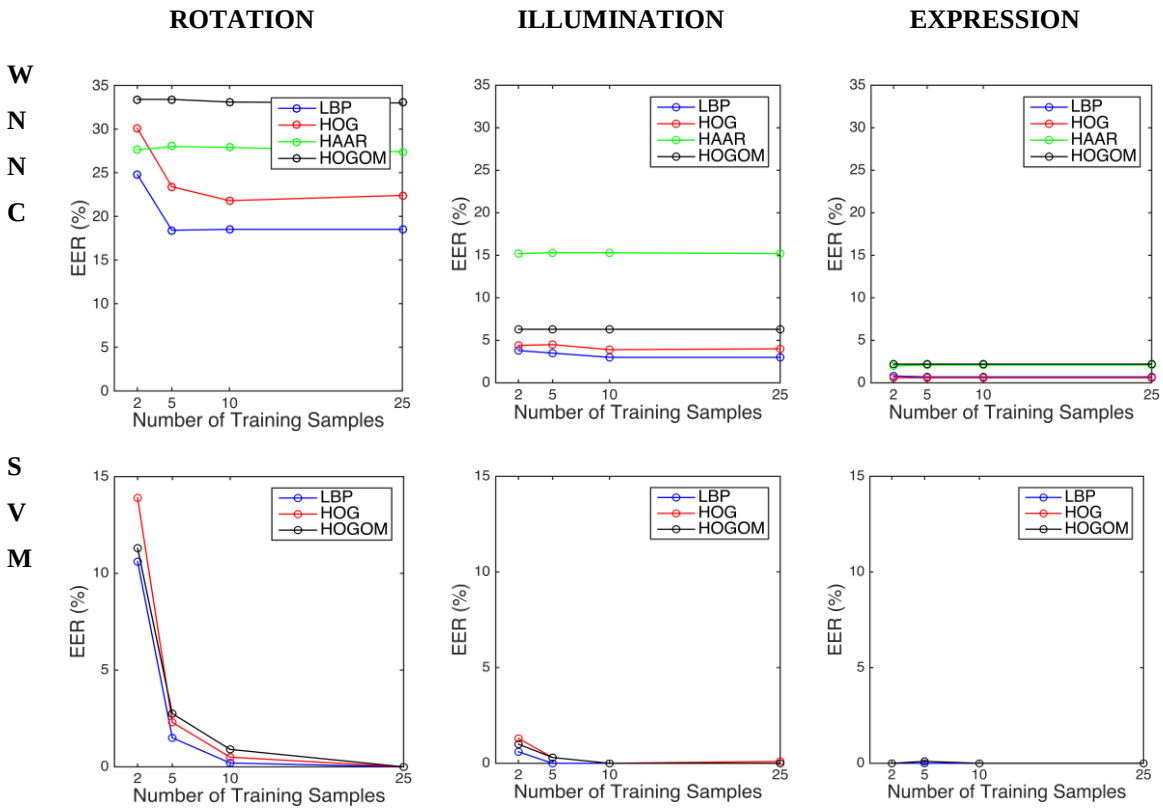

Figure 4. WNNC and SVM Equal Error Rate (EER) when trained with all modalities.

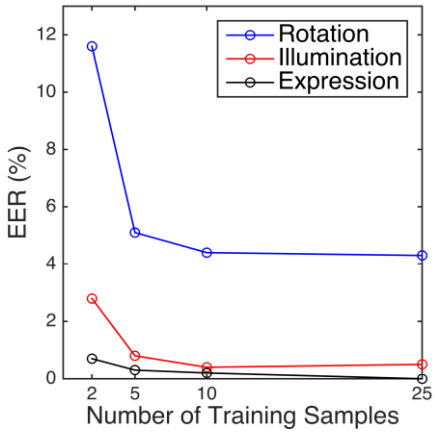

Figure 5. CNN Equal Error Rate (EER) when trained with all modalities.

The weight assigned to each modality is shown in Table 5. In the case of CNNs, these values are obtained by first calculating the absolute value of each modality weight and afterwards normalizing the resulting weights vector. For the WNNC classifiers, where weights are always positive, it corresponds to the sum of weights assigned to the features of a given

modality divided by the sum of the weights for all features. In the case of CNN, visible and thermal images make a significant contribution to the final prediction for all modalities, with gray-scale information being especially important for rotation, and thermal information for illumination. Depth information, on the other hand, makes no contribution to the rotation case, and makes a small but important contribution to the expression and illumination modes of variation. When compared to the LBP descriptor, which obtains the best overall results for the WNNC method, CNN has a different weight distribution between modalities. Depth information has an overall importance similar to that of visual information in the case of LBP, while CNN assigns to it a much lower influence. The HOGOM features, on the other hand, give more importance to the thermal modality.

Table 5: Late fusion weights for CNN and WNNC multimodal classifiers.

	CNN			LBP			HOGOM		
	ROT.	EXPR.	ILLU.	ROT.	EXPR.	ILLU.	ROT.	EXPR.	ILLU.
RGB	0.754	0.515	0.318	0.37	0.35	0.06	0.530	0.101	0.020
D	0	0.065	0.117	0.37	0.15	0.25	0.158	0.002	0.006
T	0.246	0.419	0.565	0.26	0.50	0.69	0.311	0.897	0.974

To further improve the accuracy of the WNNC and CNN approaches, which in contrast to SVM have the ability to generalize to unseen individuals, late fusion has been attempted between all pairs of models. This means the space dimensionality does not change with higher populations. The same fusion scheme is valid for larger populations during training, and the fused models are applicable to unseen individuals. The results of fusion of all pairs of models for each mode of variation for WNNC and CNN approaches are shown in Table 6. While HOGOM is the worst performing descriptor for both rotation and expression modes of variation, when fused with the modality-specific CNNs, the lowest error is obtained. CNNs alone obtain an EER of 4.3, 0, and 0.5 for variation modes 1 to 3 (see Table 4), but when combined with HOGOM the EERs go down to 3.8, 0 and 0.4 (see Table 6). The fusion of the other WNNC models using different descriptors with the CNNs also slightly reduce the EER, except in the case of HOG features, and are in all cases better than the fusion of WNNC models.

Table 6: EER over test data in % when applying late fusion to pairs of models using 25 training samples.

	HOG	HAAR	HOGOM	CNN

	ROT	EXP	ILL.	ROT	EXP	ILL.	ROT	EXP	ILL.	ROT	EXP	ILL.
LBP	20.3	0.6	3.7	18.6	0.7	3.5	17.1	0.8	2.5	4.1	**0**	0.5
HOG	-	-	-	23.3	0.6	3.6	19.2	0.5	2.2	4.3	**0**	0.5
HAAR	-	-	-	-	-	-	25.4	1.8	6.1	4.2	**0**	0.5
HOGOM	-	-	-	-	-	-	-	-	-	**3.8**	**0**	**0.4**

We hypothesize the reason why HOGOM better complements CNNs is its ability to both perform texture description and its multi-scale approach. Because the CNN models have a restricted amount of convolutions in the first two layers to prevent over-fitting, it is not capable of correctly describing textures. This likely makes HOGOM a good complement to the CNN models. The fusion of LBP (another texture descriptor) and CNNs further supports this argument, being the second best combination of models. Furthermore, HOGOM is a multi-scale descriptor, which provides a certain degree of scale invariance which CNNs don't possess except when using specific layers designed for that purpose [30, 31].

When trying to combine more than one model to CNN (Table 7), no further improvements are found. In fact, doing so produces a higher error than two of the models alone for some combinations. This is likely due to both over-fitting and the three-model combinations not providing further useful information compared to two individual models.

Table 7: EER over test data in % when applying late fusion to CNN and two other models using 25 training samples.

	HOG			HAAR			HOGOM		
	ROT	EXP	ILL.	ROT	EXP	ILL.	ROT	EXP	ILL
LBP	4.3	0	0.5	0.4	0	0.5	3.8	0	0.5
HOG	-	-	-	4.4	0	0.5	4.0	0	0.5
HAAR	-	-	-	-	-	-	**3.8**	**0**	**0.4**

5. Conclusion and future works

In this work the performance of classical image descriptors has been compared to that of Convolutional Neural Networks and the recent descriptor Histograms of Gabor Ordinal Measures over a multimodal face recognition database. We have found that 1-vs-All SVM classifiers have the highest accuracy for face recognition, performing especially well with texture descriptors such as LBP or HOGOM. This demonstrates both the power of texture descriptors and 1-vs-All SVM classifiers, which is a good option when discriminating between a small set of individuals. This technique, though, cannot generalize to unseen individuals, which drives us to further analyse the other considered approaches.

We have shown that, while CNNs is by far the best approach capable of generalizing to unseen individuals when directly compared to pre-designed descriptors, their accuracy can be improved using a late-fusion approach. Even more interestingly, we have found that the best performing methods are not necessarily the best ones under a late fusion scheme. While HOGOM obtained a bad overall performance when compared to other descriptors, it had the greatest impact when combined with CNNs, obtaining the lowest EER on a multimodal benchmark facial database (see bottom-right part in Table 6).

Combining only promising channels rather than all the channels might be a suitable choice for the trade-off between accuracy and efficiency. However, in this paper, we focused on the RGB-D-T based face recognition. That is, we did not consider fusion schemes like 2-modal (RGB-D/RGB-T/D-T) face recognition, because we are interested in building a system whose results can be compared against the original work of [18] which uses three modalities of RGB-D-T. Hence, in all our three fusion schemes (early fusion WNNC, early fusion SVM, late fusion WNNC and CNN features), we combined all the three channels (3-modal). Furthermore, in this work we aimed to find the complementary value of individual hand-crafted features to CNN features for RGB-D-T based face recognition. Therefore, we did not exploit 2-modal fusion schemes. However, it would be very interesting to try these fusion strategies for both efficiency and accuracy. For example, for each modality, we can employ the feature that achieves the lowest EER, and then consider both 2-modal and 3-modal face recognition. We would like to consider this in our future work.

Appendix

Table 8: EER over test data in % for various modalities, classifiers, features for 0 (NCC), 2 and 5 training samples per class. HOGOM and CNN performances are added to previous results from [18] for LBP, HOG, and HAAR.

			Number of training samples											
			0 (No Training (NCC))				2				5			
		Modality	RGB	D	T	All	RGB	D	T	All	RGB	D	T	All
WNNC	ROT.	LBP	31.9	34.6	31.3	31.4	27.4	26.1	30	24.8	23.3	21.5	31.1	18.4
		HOG	30.1	25.2	34.4	32.5	24.9	38.3	31.6	30.1	21.3	36.2	26.9	23.4
		HAAR	24.2	31.8	32.6	27.5	-	-	-	27.6	-	-	-	28
		HOGOM	-	-	-	-	28.9	34.7	28.8	33.4	29.6	35.5	28.4	33.4
	EXPR.	LBP	1.2	1.8	1.5	1.1	0.7	2.3	1.9	0.8	0.7	1.8	1.2	0.7
		HOG	1.6	6.1	2.2	1.7	0.7	6.6	0.9	0.6	0.7	6.5	0.9	0.6
		HAAR	2.6	8.8	3.2	2.8	-	-	-	2	-	-	-	2.1
		HOGOM	-	-	-	-	3.6	5.3	2.3	2.2	3.4	5.8	2.2	2.2
	ILLUM.	LBP	15.2	8.8	9.8	8.4	11.2	5.2	4.7	3.8	12.3	4.9	4.4	3.5
		HOG	15.7	16.3	8.7	11.2	8	24.6	4.9	4.4	8.2	15.7	4.9	4.5
		HAAR	21.3	23.3	16.2	16	-	-	-	15.2	-	-	-	15.3
		HOGOM	-	-	-	-	13.1	13	6	6.3	13	12.6	5.5	6.3
SVM	ROT.	LBP	-	-	-	-	13	14.1	17.1	10.6	1.9	3.9	4.9	1.5
		HOG	-	-	-	-	13.5	22.5	16.7	13.9	2.5	8.2	4.2	2.3
		HOGOM	-	-	-	-	12.5	18.5	14.8	11.3	3.1	5.7	4.0	2.75
	EXPR.	LBP	-	-	-	-	0.1	0.1	0.2	0	0.1	0.1	0.1	0
		HOG	-	-	-	-	0	0.6	0.1	0	0	0.2	0.1	0.1
		HOGOM	-	-	-	-	0	0.3	0	0	0.12	0	0	0
	ILLUM.	LBP	-	-	-	-	1.4	0.9	0.7	0.6	0.4	0.3	0.1	0
		HOG	-	-	-	-	2.3	4.8	1	1.3	0.6	2.4	0.1	0.3
		HOGOM	-	-	-	-	1.6	3.1	0.9	1.0	1.0	1.5	0.25	0.3
CNN	ROT.		-	-	-	-	12.6	21.4	17.5	11.6	6.3	17.6	9.9	5.1
	EXP.		-	-	-	-	1.1	3.7	1.5	0.7	0.3	3.3	0.7	0.3
	ILLUM.		-	-	-	-	7.8	7	3.3	2.8	3.3	3.8	1.5	0.8

Table 9: EER over test data in % for various modalities, classifiers, features for 10 and 25 training samples per class. HOGOM and CNN performances are added to previous results from [18] for LBP, HOG, and HAAR.

			Number of training samples							
			10				25			
		Modality	RGB	D	T	All	RGB	D	T	All
WNNC	ROT.	LBP	23.1	21.3	30.4	18.5	23.6	21.8	30.5	18.5
		HOG	20.5	35.3	25.7	21.8	21.2	35.5	26.4	22.4
		HAAR	-	-	-	27.9	-	-	-	27.4
		HOGOM	29.7	35.1	28.4	33.1	29.7	35.3	28.4	33.0
	EXPR	LBP	0.6	2	1.2	0.7	0.7	1.9	1.2	0.7
		HOG	0.8	7.1	0.8	0.6	0.8	6.5	0.8	0.6
		HAAR	-	-	-	2.1	-	-	-	2.1
		HOGOM	3.5	6.2	2.2	2.2	3.6	5.9	2.2	2.2
	ILLUMI	LBP	9.8	4.7	3.8	3	10	4.8	3.8	3
		HOG	6.6	13.8	4.4	3.9	6.5	14.3	4.5	4
		HAAR	-	-	-	15.3	-	-	-	15.2
		HOGOM	13.3	12.4	5	6.3	13.2	12.4	5.2	6.3
SVM	ROT.	LBP	0.5	1.1	0.9	0.2	0.1	0.2	0.3	0
		HOG	0.6	3.8	1	0.5	0.1	1.6	0.3	0
		HOGOM	1.3	2.2	1.1	0.9	0.7	0.3	0.2	0
	EXPR.	LBP	0.1	0.1	0.1	0	0.1	0	0.1	0
		HOG	0	0.2	0.1	0	0	0.1	0.1	0
		HOGOM	0.2	0	0	0	0.1	0	0	0
	ILLUM.	LBP	0.1	0.1	0.1	0	0	0	0	0
		HOG	0	0.7	0	0	0.2	0.4	0	0.1
		HOGOM	0	0.2	0	0	0	0.2	0	0
CNN	ROT.		5.3	17.1	9.9	4.4	5	18.6	12.1	4.3
	EXP.		0.2	2.5	0.6	0.2	0.1	2.1	0.4	0
	ILLUM.		2.1	2.9	0.8	0.4	1.9	2.9	1.1	0.5

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(12):2037–2041, Dec 2006. doi: 10.1109/TPAMI.2006.244.

[2] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis. Face recognition by fusing thermal infrared and visible imagery. IVC, 24(7):727–742, 2006.

[3] K.W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3D + 2D face recognition. CVIU, 101(1):1–15, 2006.

[4] Zhenhua Chai, Zhenan Sun, H. Mendez-Vazquez, Ran He, and Tieniu Tan. Gabor ordinal measures for face recognition. IEEE Transactions on Information Forensics and Security, 9(1):14–26, Jan 2014.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893 vol. 1, June 2005.

[6] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of International Conference on Machine Learning , 2010.

[7] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa. Mdlface: Memorability augmented deep learning for video face recognition. In Biometrics (IJCB), 2014 IEEE International Joint Conference on, pages 1–7, Sept 2014. doi: 10.1109/BTAS.2014.6996299.

[8] G. Goswami, M. Vatsa, and R. Singh. RGB-D face recognition with texture and attribute features. IEEE TIFS, 9(10):1629–1640, 2014.

[9] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using Kinect style depth cameras for dense 3D modeling of indoor environments. IJRR, 31(5):647–663, 2012.

[10] D. Holz, S. Holzer, R. Rusu, and S. Behnke. Real-time plane segmentation using RGB-D cameras. In Proc. Robot Soccer world Cup XV, 2012.

[11] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] I.A. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis, and N. Murtuza. Multimodal face recognition: Combination of geometry with physiological information. In Proc. CVPR, 2005.

[13] Steve Lawrence, C. Lee Giles, and Ah Chung Tsoi. Convolutional neural networks for face recognition. In 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96), June 18-20, 1996 San Francisco, CA, USA, pages 217–222. IEEE Computer Society, 1996. ISBN 0-8186-7258-7.

[14] B.Y.L. Li, A.S. Mian, W. Liu, and A. Krishna. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In Proc. WACV, 2013.

[15] Aleix M Martinez. The AR face database. CVC Technical Report, 24, 1998.

[16] K. Nasrollahi and T.B. Moeslund. Haar-like features for robust real-time face recognition. In Image Processing (ICIP), 2013 20th IEEE International Conference on, pages 3073–3077, Sept 2013. doi: 10.1109/ICIP.2013.6738633.

[17] Kamal Nasrollahi and Thomas B Moeslund. Are Haar-like rectangular features for biometric recognition reducible? In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pages 334–341. Springer, 2013.

[18] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B. Moeslund. RGB-D-T based face recognition. In 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014, pages 1716–1721. IEEE, 2014.

[19] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The ferret evaluation methodology for face-recognition algorithms. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(10):1090–1104, 2000.

[20] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In

IEEE computer society conference on computer vision and pattern recognition, volume 1, pages 947–954. IEEE, 2005.
[21] A. Ramey, V. González-Pacheco, and M. A. Salichs. Integration of a low-cost rgb-d sensor in a social robot for gesture recognition. In Proc. ACM/IEEE HRI, 2011.
[22] A. Ross and A. K. Jain. Multimodal Biometrics: an overview. In Proceedings of 12th European Signal Processing Conference, pages 1221–1224, 2004.
[23] M.P. Segundo, S. Sarkar, D. Goldgof, L. Silva, and O. Bellon. Continuous 3d face authentication using RGB-D cameras. In Proc. CVRPW, 2013.
[24] Ning Sun, Haixian Wang, Zhen-hai Ji, Cairong Zou, and Li Zhao. An efficient algorithm for kernel two-dimensional principal component analysis. Neural Computing and Applications, 17(1):59–64, 2008. doi: 10.1007/s00521-007-0111-0.
[25] Y. Taigman, Ming Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 1701–1708, June 2014. doi: 10.1109/CVPR.2014.220.
[26] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and LiorWolf. Web-scale training for face identification. CoRR, abs/1406.5266, 2014.
[27] Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Diagonal principal component analysis for face recognition. Pattern Recognition, 39(1):140 – 142, 2006. ISSN 0031-3203.
[28] Haitao Zhao and Pong Chi Yuen. Incremental linear discriminant analysis for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 38(1):210–221, 2008.
[29] Yufeng Zheng and Adel Elmaghraby. A brief survey on multispectral face recognition and multimodal score fusion. In ISSPIT, pages 543–550, 2011.
[30] Xu, Yichong, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-Invariant Convolutional Neural Networks. arXiv preprint arXiv:1411.6369, 2014.
[31] Ngiam, Jiquan, Zhenghao Chen, Daniel Chia, Pang W. Koh, Quoc V. Le, and Andrew Y. Ng. Tiled convolutional neural networks. In *NIPS*, pages 1279-1287, 2010.
[32] Hsu, G. S. J., Liu, Y. L., Peng, H. C., and Wu, P. X. . RGB-D-Based Face Reconstruction and Recognition. *Information Forensics and Security, IEEE Transactions on*, *9*(12) ,pp. 2110-2118, 2014.
[33] Linder, Timm, Sven Wehner, and Kai O. Arras. "Real-time full-body human gender recognition in (RGB)-D data." *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015.