Millimetre Wave Person Recognition: Hand-crafted vs Learned Features

Ester Gonzalez-Sosa*, Ruben Vera-Rodriguez*, Julian Fierrez* and Vishal M. Patel†
*BiDA Lab, EPS, Universidad Autonoma de Madrid
{ester.gonzalezs, ruben.vera, julian.fierrez}@uam.es
†Rutgers University, Piscataway, New Jersey vishal.m.patel@rutgers.edu

Abstract

Imaging using millimeter waves (mmWs) has many advantages including ability to penetrate obscurants such as clothes and polymers. Although conceal weapon detection has been the predominant mmW imaging application, in this paper, we aim to gain some insight about the potential of using mmW images for person recognition. We report experimental results using the mmW TNO database consisting of 50 individuals based on both hand-crafted and learned features from Alexnet and VGG-face pretrained CNN models. Results suggest that: i) mmW torso region is more discriminative than mmW face and the entire body, ii) CNN features produce better results compared to hand-crafted features on mmW faces and the entire body, and iii) hand-crafted features slightly outperform CNN features on mmW torso.

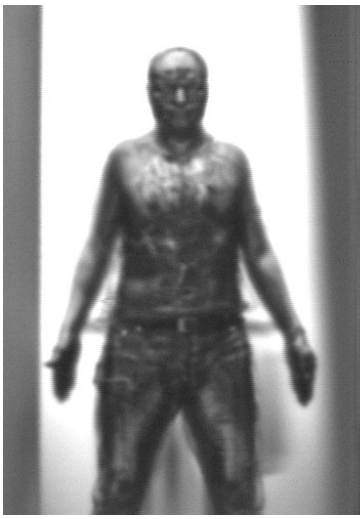

Fig. 1: A sample mmW image from the mmW TNO database [2].

I. Introduction

Millimeter waves (mmW) are high-frequency electromagnetic waves usually defined to be in the range of $30 - 300$ GHz with corresponding wavelengths between 10 to 1 mm. Since radiation at these frequencies is non-ionizing, it is considered to be safe for human exposure. Imaging using mmW has gained the interest of the security community [3], [12], [10], mainly due to its low intrusiveness and the ability to pass through clothing and other atmospheric obscurants such as cloud cover, fog, smoke, rain and dust storms. The predominant application of mmW images in the literature has been concealed weapon detection (CWD) or contraband. Indeed, the majority of international airports currently use mmW scanners for detecting concealed objects.

Exploration of mmW images for other purposes such as person recognition has been scarcely addressed in the literature. The privacy concerns of mmW images and the high cost of data acquisition are the two main obstacles that have prevented researchers from discovering new applications of mmW imaging. A sample mmW image corresponding to an individual in the mmW TNO dataset is shown in Figure 1. Since millimeter waves can penetrate through clothing, in mmW images, we are able to see things that can not be seen in a visible image. As a result, information collected in a mmW image can be used for person recognition in addition to its traditional use of CWD.

To this end, Alefs *et al.* [2] developed one of the first reported efforts for person recognition using real mmW passive images acquired in outdoors scenarios. They exploited the texture information contained in the torso region of the image through multilinear eigenspaces techniques. However, the experimental protocol carried out in that work was really optimistic, being far away from a realistic verification scenario in which a traveler would enter in a mmW scanner to verify his identity

with a previous enrolled sample. On the other hand, the works by Moreno-Moreno *et al.* [8] and by Gonzalez-Sosa *et al.* [5] proposed and analyzed a biometric person recognition system using shape information, based on the idea that shape information retrieved from mmW images may be more robust to clothes variations than if it were extracted from visible images. In the mentioned works, shape information is extracted from BIOGIGA database [8], which contains synthetically generated mmW images. They exploited geometrical measures between different silhouette landmarks and features based on contour coordinates, respectively. In all cases, images were extracted in the range of 94 GHz.

In this work, we present further insight about the potential of using mmW images for person recognition. As mmW waves have the ability to pass through clothes, person recognition may be achieved not only through face information, but also through other parts of the body such as the torso or even the whole body. We empirically show the discrimination capabilities of different mmW body parts using both hand-crafted features and deep convolutional neural networks (CNN) features.

Rest of the paper is organized as follows. Section II introduces the different mmW body parts considered in this work and some preprocessing techniques applied to them. Section III describes the selected hand-crafted and CNN features analyzed and compared throughout the paper. Section IV presents the mmW TNO database, while Section V gives details about the experimental protocol for both person verification and identification. Experimental results are presented in Section VI. Finally, Section VII concludes the paper with a brief summary and discussion

II. Preprocessing

In this work, three mmW body parts are considered: mmW face, mmW torso and mmW whole body (see the first column of Fig. 2). These mmW body parts are extracted from the original mmW images by manually defining the corresponding bounding boxes. The approximate size of these bounding boxes is 70×90 for mmW face; 120×170 for mmW torso and 250×450 for the body ($width \times height$ format). Then, mmW faces are histogram equalized using the INface toolbox v2.0 [11]. Millimeter torso and body are not histogram equalized.

III. Feature Representation

In this section, we provide details of the different features employed in this work. Features are first extracted from conventional hand-crafted approaches. Then, some details regarding the feature extraction from the cutting-edge deep learning approaches are given.

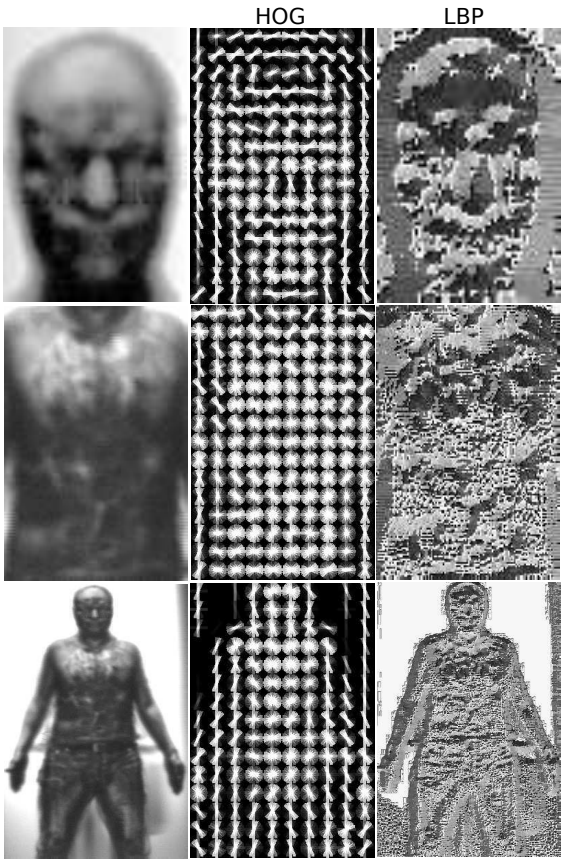

Fig. 2: Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) features from the different mmW body parts: face, torso and the whole body.

A. Hand-Crafted Features

Among the wide variety of hand-crafted features presented in the literature for biometric recognition, we select two of the most widely used ones for mmW person recognition: i) Local Binary Patterns (LBP) and ii) Histogram of Oriented Gradients (HOG).

1) Local Binary Patterns features: Local Binary Patterns provide discriminatory texture information that has been proved to be robust against illumination variations [7]. In this case, first images are resized to 100×150 ($width \times height$ format). The image is then divided into non overlapping 10×10 blocks. For each block the LBP histogram feature is computed with radius 1, 8 neighbours and uniform patterns, resulting in a 59-length vector. The final feature vector of each image is the concatenation of all the histograms from all blocks. Extracted LBP features from the face, torso and body parts are shown in the third column of Fig. 2. Note though that the final feature vector is in terms of histograms of LBP from each block. LBP features are extracted based on the implementation

TABLE I: Alexnet and VGG-face configurations. Size follows $width \times height \times depth$ format.

Description	Alexnet [6]	VGG-face [9]
# of layers	21	39
# of conv-relu layers	5	16
# of parameters	60 M	135 M
Input Size	$227 \times 227 \times 3$	$224 \times 224 \times 3$
Output Size of conv1	$55 \times 55 \times 96$	$224 \times 224 \times 64$
Output Size of conv2	$27 \times 27 \times 256$	$112 \times 112 \times 128$
Output Size of conv3	$13 \times 13 \times 384$	$56 \times 56 \times 256$

provided by [1].

2) Histogram of Oriented Gradients features: Histogram of Oriented Gradients are able to retain both shape and texture information. Likewise, first images are resized to the same dimension used for LBP features. HOG features are also computed block-wise. They comprise histogram calculation followed by histogram normalization. Each 10×10 block is described by a histogram of gradients with 8 number of orientations, with each gradient quantized by its angle and weighted by its magnitude. Then, four different normalizations are computed using adjacent histograms, resulting in 8×4-length feature vector for each block. The final feature vector of a given image is the vectorization of the HOG features from all blocks. Extracted HOG features from the face, torso and body parts are shown in the second column of Fig. 2. One can clearly see the different gradient magnitudes of each of the 150 blocks in this figure. Each of the HOG features are extracted using the implementation provided in [4].

B.Convolutional Neural Network features

In recent years, features obtained using deep CNNs have yielded impressive results on various computer vision and biometrics recognition problems such as face recognition. Recent studies have shown that in the absence of massive datasets or hardware infrastructure, transfer learning can be effective as it allows one to introduce deep CNNs without having to train it from scratch. This is possible because the lower layers (the ones closest to the input layer) in CNNs learn low-level features, and the layers closer to the output learn high-level features. One can think of the lower layers as learning things like edges, and the higher layers as learning more complex shapes. As a result, higher layers can be tuned to the task at hand. Therefore, one can use the lower layers of commonly used deep CNNs such as AlexNet [6] or VGG-face [9] to extract general features, that can then be used to train other classifiers or matchers.

In this paper, we use two pretrained CNN models, namely AlexNet and VGG-face to extract features for mmW person recognition by finetuing these model on a subset of the mmW TNO dataset.

1) Alexnet: Alexnet [6] is a CNN that was trained for the ISLVRC competition using a dataset of 1.2 million images of 1000 classes (animals, objects, etc.) from the Imagenet dataset. During the training we resized images into the standard 227×227 input size. All images are also scaled into $[0, 1]$ and subtracted from their mean value. As we only have 2 images per class, we augment the number of images per class up to 10 by creating mirrored versions of the original samples followed by adding some noise with different standard deviations (4 additional images per original sample). We use the stochastic gradient descent to learn the parameters, with a momentum of 0.9, a number of epochs of 100, and a batch size of 32 samples. The learning rate and regularization parameter for the mmW person recognition task were set to 10^{-5} (without decay) and 10^{-3}, respectively for all mmW body parts-based fine-tuned networks. The bias and weight learning rate factor of the non fine-tuned layers is set equal to their default value, 1, while the learning rate factor for bias and weights of the fine-tuned layers are set equal to 20 and 100, respectively.

2) VGG-face: The VGG-face network [9] was inspired by the previous VGG-Very-Deep-16 CNN network. It has been trained using a dataset of 2.6 million faces and 2622 classes (people). We resized images into the standard 224×224 input size with the average face image subtracted. We also perform data augmentation here following the same strategy that we followed with the Alexnet. Optimization is also achieved by stochastic gradient descent using mini-batches of 32 samples and momentum coefficient of 0.9. The VGG-face model is regularized using dropout layers after fully connected layers with a rate of 0.5. The learning rate and regularization parameter for the mmW person recognition task were set equal to 10^{-6} and 10^{-4}, respectively (no decay). The bias and weights learning rate factors for non fine-tuned and fine-tuned layers are set equal to the same values as with the Alexnet.

In both networks, we set the negative slope to 0 in ReLU. The softmax loss layer computes the multinomial logistic loss of the softmax of its inputs. After fine-tuning the AlexNet or VGG-face on the target dataset, we extract the deep feature as the output of the $fc7$ layer, which is a 4096 dimension vector (for feature extraction), or feed forward the sample until the last layer of the CNN, which will give a class score (for classification purposes).

Table I describes the configuration of both CNNs. As can be seen, VGG-face has a larger number of layers, hence a larger number of learnable parameters. Notice also from both networks that as it goes deep in the network, the number of filters ($depth$ dimension) in a convolutional layer increases (see e.g. for Alexnet 96, 256 and 384 filters for conv1, conv2 and conv3 respectively).

Besides, even if the input size of both networks is very

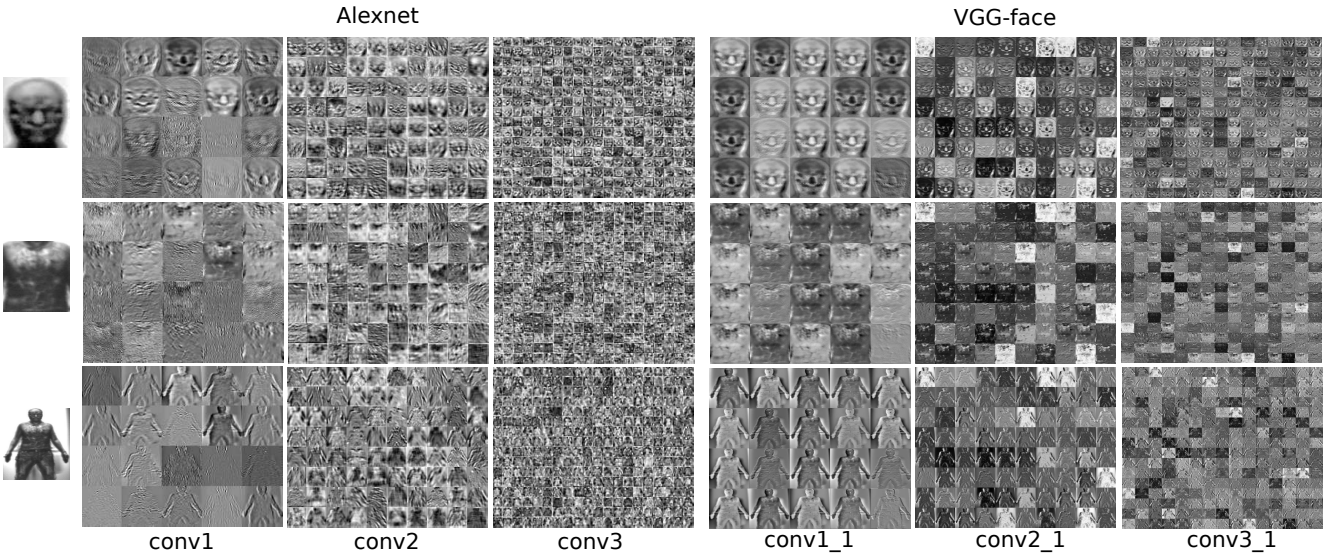

Fig. 3: Feature maps from the resulting Alexnet and VGG-face fine-tuned nets for the different mmW parts: face (first row), torso (second row) and whole body (last row). We show Alexnet features from conv1, conv2 and conv3 layers; while VGG-face features are shown from conv1_1, conv2_1 and conv3_1 layers. Please notice that VGG-face feature maps have a higher dimension than Alexnet feature maps, they have been arranged withe similar sizes in order to be display properly.

similar, the output size of the first three convolutional layers are quite different between Alexnet and VGG-face. As the latter contains more conv-relu layers, it may reduce the dimensionality of convolutional outputs more gradually than in Alexnet. Figure 3 depicts the outputs or feature maps of conv1, conv2 and conv3 layers from Alexnet and conv1_1, conv2_1 and conv3_1 from VGG-face for the mmW face, mmW torso and mmW whole body. As can be seen from this figure, the layers closest to the input layer contains a lower number of filters with low-level features such as edges. Conversely, deeper convolutional layers hold a larger number of filter but with a lower size, addressing high-level features such as complex shapes or fine details. Even if outputs from convolutional layers from both networks appear to be equal in dimensionality, they have been arranged for display purposes. The larger resolution of VGG-face outputs is clearly notable.

IV. The mmW TNO Database

The mmW TNO database (created by the Dutch Research Institute TNO in The Hague) is the only database available for research purposes that contains real images of subjects extracted in the range of mmW specifically designed for person recognition purposes [2]. Images are recorded using a passive stereo radiometer scanner in an outdoor scenario.

The database is comprised of images belonging to 50 different male subjects in 4 different scenarios. These 4 different scenarios derive from the combination of 2 different head poses and 2 different facial occlusions. In the first head pose configuration, the subject is first asked to stand in front of the scanner with head and arms position fixed (*frontal head pose*). In the second pose configuration (*lateral head pose*), the subject is asked to turned his head leftward while the torso is asked to remain fixed (it may suffer some small changes due to the head movement). A second round of images with the first and second head pose configurations were extracted but now a large part of the facial region was occluded using an artificial beard or balaclava (disguised).

As mentioned before, each scanning is a set of two grayscaled images. By dividing this set into single images of 348×499, the TNO database is comprised of 50 subjects $\times$ 2 head pose configurations $\times$ 2 facial clutter configurations $\times$ 2 images per set, making a total of 400 images in the whole mmW TNO database. A sample 348×499 image from this dataset is shown in Fig. 1.

V. Experimental Protocol

We only consider in this work the subset of 200 images which have *frontal head pose* configuration. The set of 4 images per subject is divided randomly and evenly into training and test sets, with 2 images each. By selecting randomly the 4 images, we ensure that disguised images are not predominantly in any of the sets. We report experiments in both identification and verification modes.

A. Identification Mode

In the identification mode, a model is learnt for each subject in the dataset. The model of the hand-crafted features is computed by averaging the features from the 2 images belonging to the same class (subject).

In order to get identification results, all test images are faced against the 50 subject models. For the hand-crafted features, we compute cosine distances between a particular subject model and the test features. For CNN, we just feed forward the corresponding fine-tuned network (Alexnet or VGG-face) with the test images to get the class scores from the classification layer.

B. Verification Mode

Verification implies one-to-one comparisons to find out whether the two given images belong to the same subject or not. Hand-crafted features are extracted as mentioned in Section III-A. In order to extract the CNN features, we feed forward both training and test images in the corresponding fine-tuned networks and extract the features from the next-to-last fully connected layer, resulting in a 4096-feature vector (fc7).

Finally, for both hand-crafted and CNN features, training features are matched against test features using cosine distance to obtain genuine and impostor scores.

VI. Results

In this section, we present the performance results in both verification and identification modes. Identification results are reported in terms of cumulative match curves (CMC), while Verification results are reported in terms of receiving operating curves (ROC). For each mode, we have assessed four different feature approaches, two hand-crafted and two CNN-based. All the approaches are tested with the three considered mmW body parts: mmW face, mmW torso and mmW whole body. All the experiments have been carried out using Neural Network and Parallel Computing toolboxes from Matlab 2016*b*, along with CUDA7.5 and GeForce GTX TITAN X.

A. Verification

Fig. 4 shows the quantitative results in terms of ROC curves and Equal Error Rate (EER) corresponding to mmW face, mmW torso and mmW whole body. For each body part, we report ROC curves for all feature approaches.

From Fig. 4 left, one can see that person recognition through mmW faces achieve 30% of average EER for all approaches, which are far from being comparable to the state-of-the-art results in the visible face recognition. CNN features are able to reach results which are slightly better than hand-crafted features, obtaining a 27% of relative improvement between the average EER of hand-crafted and CNN features. These poor results may be due to several reasons: low resolution and insufficient number of samples per class to fine-tune the CNN pretrained models, among others.

It is also worth noting the importance of applying histogram equalization to mmW faces. We empirically proved a reduction of EER by an average of 7% for the different approaches when using histogram equalized images.

Superior performance of mmW torsos can be seen from From Fig. 4 center, in which all considered features achieve outstanding results, reaching the best result of 4.5% EER with the HOG features. This may be due to the fact that torso is somewhat more stable than other body parts such as face in terms of pose or expression variations. It is also surprising the fact that hand-crafted features are performing better than the learned features. It may be due to the robust nature of torsos, which allow spatial features to be pseudo aligned effortlessly.

Hand-crafted features with mmW whole body are performing similarly as with mmW faces (see Fig. 4 right). In this case, a previous alignment of the body may have helped to improve performance of this spatial-dependent features. CNN features are performing reasonably, being more discriminative than faces but less than torsos. It is interesting to note that the VGG-face features perform slightly better than the Alexnet features.

B. Identification

Fig. 5 presents the CMC curves for all mmW body parts and feature approaches. Rank-1 (R1) indicates the percentage of probe samples where the system has assigned the right identity in the first place among the 50 possible candidates.

Similar conclusions can be drawn from the CMC curves: the best performance of the CNN features over the hand-crafted features for mmW faces and mmW whole body, the superiority of torsos regardless of the feature approach, being hand-crafted feature also superior here. The best performance of hand-crafted features for either verification and identification may be due to the aforementioned lack of samples per class for properly fine-tuning pre trained models, but also due to the distance between source and target dataset. Bear in mind than Alexnet has been trained using images of objects and VGG-face has been trained using images of visible faces.

It is also worth noting that not always better EERs in verification produce better R1. Such is the case of VGG-face and Alexnet for mmW faces, in which VGG-face achieves a better rank-1 (49%) than Alexnet (40%), although Alexnet achieve less EER in verification mode. This is happening also the other way around between

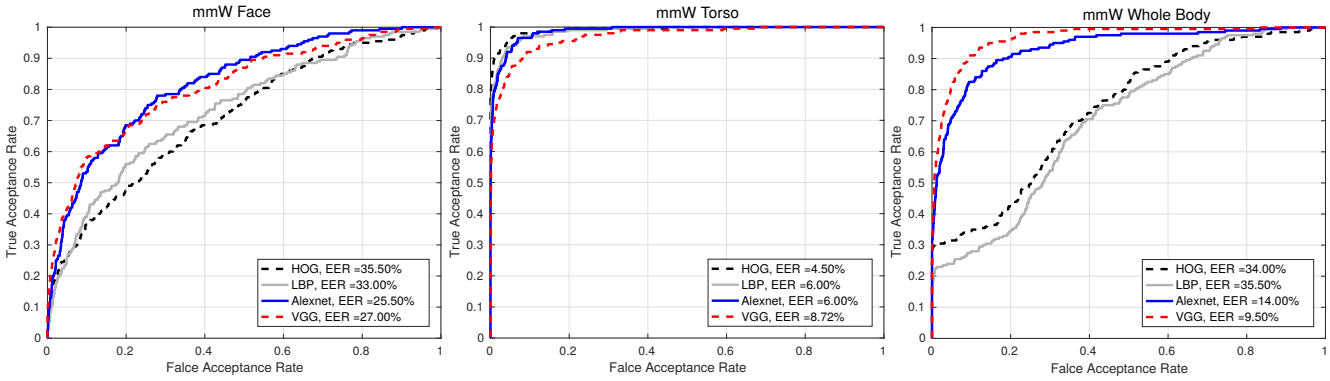

Fig. 4: Verification results. ROC Curves are drawn for mmW face, mmW torso and mmW whole body (left, center and right, respectively).

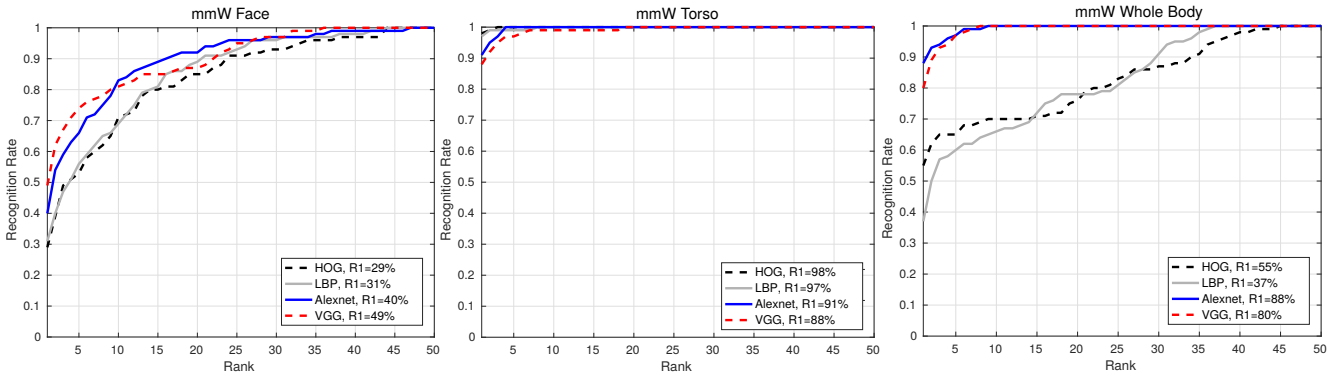

Fig. 5: Identification results. CMC Curves are drawn for mmW face, mmW torso and mmW whole body (left, center and right, respectively).

VGG-face and Alexnet for mmW whole body. Now the approach with a larger EER is achieving better R1 results. When working in identification mode, more insight can be drawn also from confusion matrices. Fig. 6 plots the confusion matrices of all mmW body parts and features considered in this paper. As can be seen, the confusion matrices of torso experiments are almost perfect, especially those from the HOG features. Confusion matrices from the mmW faces are the worst ones, while there is a remarkable difference between the confusion matrices corresponding to CNN features and hand-crafted features for mmW whole body. It is also worth noting that each feature approach confuses in a different manner, suggesting us that fusion schemes between different approaches may play a key role in the search of performance improvement and robustness.

VII. Conclusions

This paper has presented one of the very first works addressing person recognition using mmW images. Different mmW body parts have been considered in our work: face, torso and whole body. We have carried out experiments with several hand-crafted features and some state-of-the-art CNN features. Some of the findings from the experiments are: i) mmW torsos are the most discriminative body parts in mmW images and mmW faces are the least discriminative ones; ii) CNN features overcome hand-crafted features with faces and whole body parts; iii) hand-crafted features achieve outstanding results for torso-based person recognition.

We believe one of the main drawbacks that prevent us from exploiting more CNN learning capabilities in our current application is the lack of a reasonable number of samples per class (there are only 2 real images per class), and the lower resolution of some of the body parts with which we are working.

In our future work, we will consider different fusion schemes at different levels to exploit the best of the different feature approaches. We will also investigate the possibility of fusing shape and texture information jointly.

In order to move towards real applications (e.g. border control), we should consider performing person recognition using active mmW images (those which include artificial illumination in order to increase the resolution

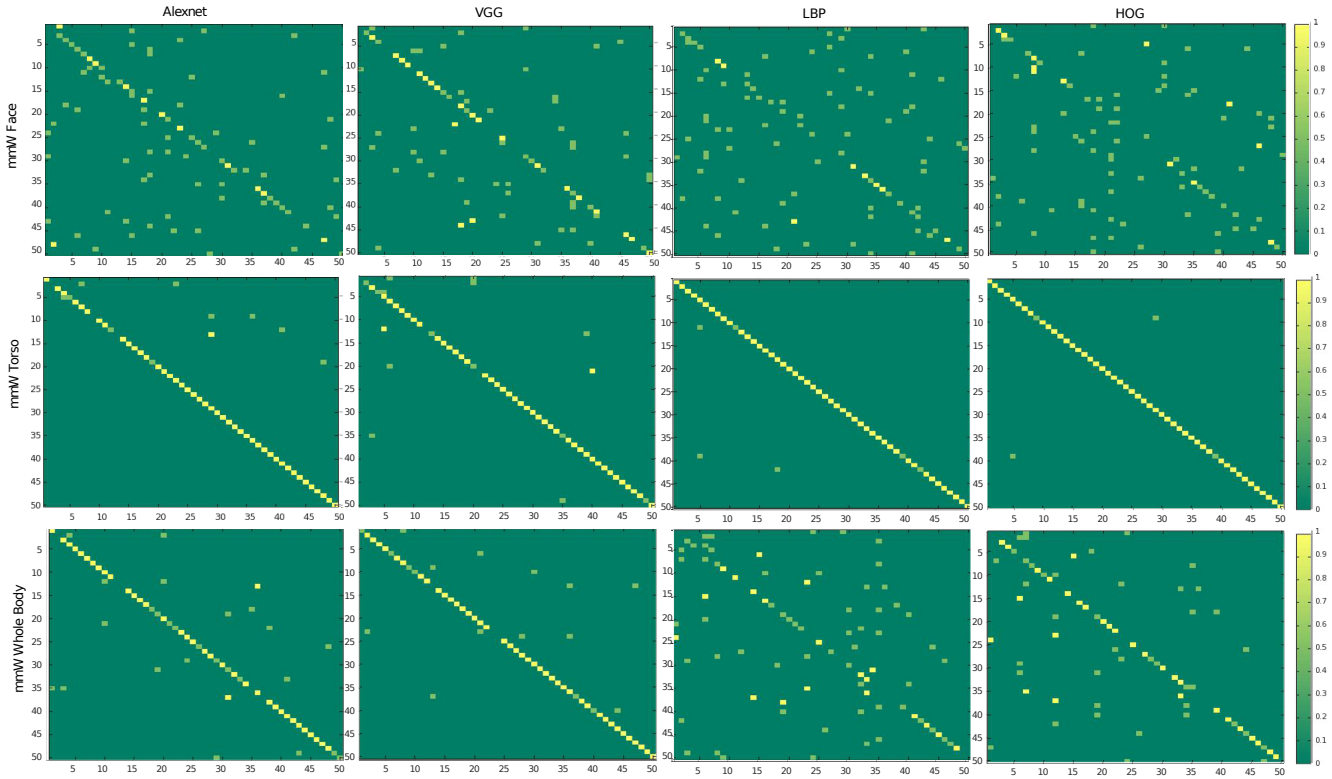

Fig. 6: Identification results. Confusion Matrices for the different mmW body parts and the 4 different feature extractor considered.

of the resulting image), which are actually the ones that are used in commercial mmW scanners. To this aim, we would need datasets of active mmW images from different subjects and a larger number of subjects, an issue that due to privacy concerns is nowadays a real challenge.

Acknowledgment

This work has been partially supported by project CogniMetrics TEC2015-70627-R (MINECO/FEDER), and the SPATEK network (TEC2015-68766-REDC). E. Gonzalez-Sosa is supported by a PhD scholarship from Universidad Autonoma de Madrid. Vishal M. Patel was partially supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134. Authors wish to thank also TNO for providing access to the database.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] B. Alefs, R. den Hollander, F. Nennie, E. van der Houwen, M. Bruijn, W. van der Mark, and J. Noordam. Thorax Biometrics from Millimetre-wave Images. *Pattern Recognition Letters*, 31(15):2357–2363, 2010.

[3] R. Appleby and R. N. Anderton. Millimeter-wave and Submillimeter-wave Imaging for Security and Surveillance. *Proc. of IEEE*, 95(8):1683–1690, 2007.

[4] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox.

[5] E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, M. Moreno-Moreno, and J. Ortega-Garcia. Feature Exploration for Biometric Recognition using Millimetre Wave Body Images. *EURASIP Journal on Image and Video Processing*, (1):1–13, 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[7] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination Invariant Face Recognition using Near-infrared Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007.

[8] M. Moreno-Moreno, J. Fierrez, R. Vera-Rodriguez, and J. Parron. Simulation of Millimeter Wave Body Images and its Application to Biometric Recognition. In *Proc. of SPIE, Biometric Technologies for Human Identification*, volume 8362, 2012.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *British Machine Vision Conference*, 2015.

[10] V. M. Patel, J. N. Mait, D. W. Prather, and A. S. Hedden. Computational Millimeter Wave Imaging. *IEEE Signal Processing Magazine*, 1053(5888/16), 2016.

[11] V. Struc et al. The INface Toolbox v2.0 The Matlab Toolbox for Illumination Invariant Face Recognition.

[12] L. Yujiri. Passive Millimeter Wave Imaging. In *IEEE MTT-S International Microwave Symposium Digest*, pages 98–101, 2006.