Covariate analysis of descriptor-based ear recognition techniques

Conference Paper · July 2017

DOI: 10.1109/IWOBI.2017.7985520

CITATIONS
11

READS
136

4 authors:

Žiga Emeršič
University of Ljubljana
39 PUBLICATIONS **404** CITATIONS
SEE PROFILE

Blaž Meden
University of Ljubljana
19 PUBLICATIONS **115** CITATIONS
SEE PROFILE

Peter Peer
University of Ljubljana
128 PUBLICATIONS **1,693** CITATIONS
SEE PROFILE

Vitomir Štruc
University of Ljubljana
151 PUBLICATIONS **1,938** CITATIONS
SEE PROFILE

Covariate Analysis of Descriptor-Based Ear Recognition Techniques

Žiga Emeršič*, Blaž Meden*, Peter Peer* and Vitomir Štruc†
*Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana, Slovenia
Email: {ziga.emersic, blaz.meden, peter.peer}@fri.uni-lj.si
†Faculty of Electrical Engineering, University of Ljubljana
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
Email: vitomir.struc@fe.uni-lj.si

***Abstract*—Dense descriptor-based feature extraction techniques represent a popular choice for implementing biometric ear recognition system and are in general considered to be the current state-of-the-art in this area. In this paper, we study the impact of various factors (i.e., head rotation, presence of occlusions, gender and ethnicity) on the performance of 8 state-of-the-art descriptor-based ear recognition techniques. Our goal is to pinpoint weak points of the existing technology and identify open problems worth exploring in the future. We conduct our covariate analysis through identification experiments on the challenging AWE (Annotated Web Ears) dataset and report our findings. The results of our study show that high degrees of head movement and presence of accessories significantly impact the identification performance, whereas mild degrees of the listed factors and other covariates such as gender and ethnicity impact the identification performance only to a limited extent.**

1. Introduction

Ear recognition represents a sub-field of biometrics with important applications in security, surveillance and forensics. Many techniques have been proposed in the literature for ear recognition ranging from geometric and holistic techniques [40], [11], [41], [2], [4] to more recent descriptor-based methods [20], [29], [6], [8], [23]. Especially the latter have proven highly successful and represent the existing state-of-the-art in this area as identified by recent surveys and comparative evaluations [28], [1], [34], [16].

Descriptor-based techniques typically extract identity cues from local image areas and use the extracted information for identity inference. As emphasized by Emersic et al. [16], two groups of techniques can in general be considered descriptor-based: *i)* techniques that first detect interest points in the image and then compute descriptors for the detected interest points, and *ii)* techniques that compute descriptors densely over the entire images based on a sliding window approach (with or without overlap). Examples of techniques from the first group include [3], [9] or more recently [33]. A common characteristic of these techniques is the description of the interest points independently one from the other, which makes it possible to design matching techniques with robustness to partial occlusions of the ear area. Examples of techniques from the second group include [10], [5], [19], [38]. These techniques also capture the global properties of the ear in addition to the local characteristics which commonly results in a higher recognition performance, but the dense descriptor-computation procedure comes at the expense of the robustness to partial occlusions. Nonetheless, recent trends in ear recognition favor dense descriptor-based techniques primarily due to their computational simplicity and high recognition performance.

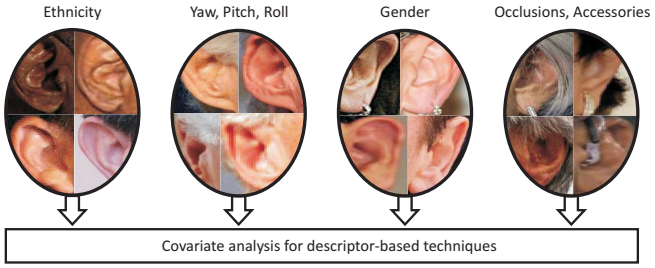


Figure 1: Sample images from the experimental dataset with different labels. The dataset is annotated according to ethnicity, head rotation (yaw, tilt and roll angles), gender and presence of occlusions and accessories. These labels serve as the basis for studying the sensitivity of descriptor-based ear recognition techniques to various covariates in this work.

Descriptor-based ear recognition technology has advanced over the last decades thanks to the introduction of powerful new image descriptors that helped to discriminate better between identities. Many ear recognition techniques were presented in the literature exploiting these new descriptors, however, studies focusing on the strengths and weaknesses of these techniques are still limited in the literature. In this paper, we try to fill this gap and present a covariate analysis of (dense) descriptor-based ear recognition techniques. Our goal is to identify which factors (or covariates) influence descriptor-based techniques the most and, hence, contribute the greatest to recognition errors. A detailed understanding of the strengths and weaknesses of state-of-the-art recognition technology is extremely important, because

978-1-5386-2064-9/17/$31.00 © 2017 IEEE

it allows us to devise more effective recognition techniques and helps identify future research trends in this area.

We conduct our covariate analysis on the AWE (Annotated Web Ears) dataset [16] (http://awe.fri.uni-lj.si), which represents one of the most challenging datasets available for ear recognition research. The dataset ships with ground truth annotations with respect to various characteristics of the ear images, such as ethnicity, head rotation (in terms of yaw, roll and tilt angles), gender and presence of occlusions and accessories (as shown in Figure 1) and therefore represents a perfect choice of our work. We include 8 state-of-the-art descriptor-based ear recognition techniques in our analysis and show the effects of different factors on the performance of the identification.

The main contributions of this paper are:

- We conduct a comprehensive covariate analysis of several state-of-the-art ear recognition techniques on a challenging dataset of ear images gathered from web with the goal of studying unconstrained ear recognition,
- We identify the most important covariates with the biggest impact on ear recognition performance and provide directions for future research in this area,
- We evaluate 8 descriptor-based techniques for ear recognition and establish a relative ranking of the assessed techniques as a byproduct of our covariate analysis.

The rest of the paper is structured as follows. In Section 2 we review the existing work related to our paper and further motivate our analysis. We describe the ear recognition techniques considered in this work in Section 3 and introduce the experimental dataset and protocol in Section 4. We present the results of the covariate analysis and discuss its implications in Section 5 and finally conclude the paper with some final comments and directions for future work in Section 6.

2. Motivation and Related Work

Understanding the characteristics of biometric recognition technology is of paramount importance to the advancement of the field. What properties of the input data make the recognition process difficult? What properties make it is easy? Are certain techniques better suited for specific data characteristics than others? Answers to questions like these make it possible to target weak points of existing techniques and provide directions for research needed in this area.

In the field of biometric ear recognition some of these question outlined above are (partially) discussed in recent survey papers, such as [28], [1], [34], [16], where structured comparisons of existing ear recognition techniques are presented. The comparisons in these papers are based on previously reported results and summarize recognition experiments on different datasets with different experimental protocols. While general trends about the advancement of ear recognition technology over the years are presented in these surveys and some of the strengths and weaknesses are identified, no detailed information about the performance of the existing techniques with respect to different covariates is given.

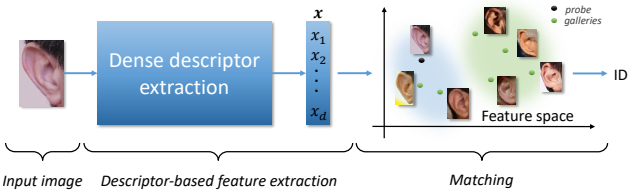


Figure 2: The identification pipeline used in our experiments. Descriptors are computed from the given input ear image in a dense manner and stacked into a d-dimensional feature vector $\mathbf{x}$. The feature vector is then matched against all gallery feature vectors and the ID of the most similar vector is returned as the output.

Similarly to our work, the survey by Emersic et al. [16] also presents a comparison of some descriptor-based feature extraction techniques from the literature using a challenging dataset and predefined experimental protocol, but unlike this paper does not focus on the impact of image characteristics on the recognition performance of the tested techniques.

Pflug et al. [28] compare the performance of various texture and surface descriptors for ear biometrics, but different from our work uses the descriptors in combination with subspace projection techniques. The reported experiments are conducted on a dataset of ear images with laboratory-like quality, but no ablation study is presented.

The study from [31] is likely the closest to our work, as it also compares descriptor-based ear recognition techniques with respect to different covariates. However, the focus here is on image characteristics, such as noise and blurring, and not on ear-related covariates such as in our work.

3. Descriptor-based Techniques

In this section we present the descriptor-based ear recognition techniques considered in our analysis. We commence the section with the description of the overall recognition pipeline used for the experiments (which is the same for all techniques) and then proceed with a brief description of the descriptors relevant for our work.

3.1. Descriptor-based Ear Recognition

For our experiments, we use a simple identification pipeline illustrated in Figure 2. The pipeline takes an ear image as input, converts it to gray-scale and computes descriptors from the gray-scale image in a dense matter. The computed descriptors are then stacked into a d-dimensional feature vector, which is matched against the gallery vector feature vectors. The identity of the most similar feature vector is ultimately assigned to the input image. Based on this identity it is possible to generate performance metrics for the identification process.

At the core of our identification pipeline is the descriptor-based feature extraction technique, which computes a feature vector from the input image. This part of the pipeline is implemented in this paper with 8 different descriptors that are briefly discussed in the remainder of this section.

3.2. Local Binary Patterns

Local Binary Patterns (LBP) represent powerful texture descriptors that achieved competitive recognition performance in various areas of computer vision [32]. The use of the LBP descriptor for ear recognition is mainly motivated by its computational simplicity and the fact that the texture of the ear is highly discriminative. Many successful ear recognition techniques have been presented in the literature exploiting LBPs either as stand-alone texture representations or in combination with other techniques, e.g., [30], [17], [7].

LBPs encode the local texture of an image by generating binary strings from circular neighborhoods of points thresholded at the gray-level value of their center pixels. The generated binary strings are interpreted as decimal numbers and assigned to the center pixels of the neighborhoods. The number of sampling points P used to generate the binary strings depends on the radii R of the circular neighborhoods and results in the following encoding [32]:

$$LBP_{P,R} = \sum_{p=0}^{P-1} 2^p s(g_p - g_c), \tag{1}$$

where $LBP_{P,R}$ stands for the computed binary pattern of some center pixel, g_c and g_p denote the gray-level values of the center pixel and the p-th pixel from the neighborhood, respectively, and the thresholding function $s(\cdot)$ stands for:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In practice, not all binary patterns returned by Eq. (1) are useful for texture representation. Typically, only binary strings with at most two bitwise transitions from 0 to 1 (or vice versa) are considered in the final descriptor. For a 8-pixel neighborhood and a consequent 8-bit binary string, for example, exactly 58 such patterns (called uniform patterns) can be computed. Most methods exploiting LBPs with a 8-pixel neighborhood for texture description, therefore, compute 59-bin histograms from local image blocks and then concatenate the computed histograms over all blocks into a global texture descriptor (our d-dimensional feature vector $\mathbf{x}$) that can be used for recognition. A similar procedure is also used in our experiments in Section 5.

3.3. (Rotation Invariant) Local Phase Quantization

Local Phase Quantization (LPQ) features [26] are very similar in essence to LBPs, as the local image texture is again encoded using binary strings, histograms are again computed from the binary strings of local image blocks and concatenated into the final representation of the given image. LPQ features are computed from the Fourier phase spectrum of an image and are known to be invariant to blurring under certain conditions. This feature makes LPQs an attractive alternative for ear recognition (see, e.g., [30]), where blurred and low-resolution images represent a problem for the existing technology.

With LPQ, the local neighborhoods of every pixel in the image are first transformed into the frequency domain using a short-term Fourier transform. Local Fourier coefficients are computed at four selected frequency points and the local phase information contained in these (complex) coefficients is then encoded. Here, a similar quantization scheme is used as in iris recognition systems, where every complex Fourier coefficient contributes two bits to the final binary string. The result of this coding procedure is a 8-bit binary string for every pixel in the image from which the local 256-bin histograms are computed and later concatenated into a global descriptor of the image.

An extension of this technique to rotation invariant local phase quantization features (RILPQ) was presented in [27]. The idea here is similar to the original LPQ technique with the difference that a characteristic orientation is first estimated for the given local neighborhood and then this orientation is used to compute a directed version of the binary descriptor. The binary descriptor is computed with the same procedure as the original LPQ, but every local neighborhood is first rotated in accordance with its characteristic orientation. RILPQ descriptors are not only blur invariant, but also exhibit a certain degree of robustness towards image rotation.

3.4. Binarized Statistical Images Features

Binarized Statistical Images Features (BSIF) [18] represent a more recent tool for texture description. Here, binary strings (encoding texture information) are again constructed for each pixel in the image, but this time by projecting image patches onto a subspace, whose basis vectors are learnt from natural images. The subspace coefficients are then binarized using simple thresholding. This procedure is equivalent to filtering the input image with a number of pre-learned filters and binarizing the filter responses at each pixel location. Each filter contributes 1 bit to the binary string of a pixel making the length of the binary string dependent on the number of filter used. Similar to LBP and LPQ, the binary string of each pixel is interpreted in decimal form and a global histogram-based representation (our d-dimensional feature vector $\mathbf{x}$) is constructed for the given images by concatenating histograms constructed from smaller image blocks.

The main characteristic that makes BSIF features so appealing is the the fact that the binary strings are not constructed based on heuristic operations, but on the basis of statistics of natural images. The idea behind BSIF-based texture description is in line with recent feature learning approaches, which produced competitive results for many computer vision problems in recent years. The use of BSIF

features for ear recognition was advocated by Pflug et al. in [30], where excellent performance was reported.

3.5. Histograms of Oriented Gradients

Descriptors exploiting Histograms of Oriented Gradients (HOG) were originally introduced for the problem of human detection by Dalal and Triggs [12], but have since been successfully applied to various fields of computer vision, including ear recognition [30], [13]. HOG descriptors have excellent texture description properties and are considered robust towards moderate illumination changes. This fact makes them highly suitable for problems, such as ear recognition, where illumination-induced variability is one of the major problems.

HOGs are computed based on a simple procedure. The computation starts by calculating the gradient of the image using 1-dimensional convolutional masks, i.e., $[-1, 0, 1]$ and $[-1, 0, 1]^T$. In the next step, the image is divided into a number of cells and compact histograms of quantized gradient orientations are computed for each cell. Here, a voting procedure is used during histogram construction, so that pixels with higher gradient magnitudes contribute more to the histogram bins than pixels with lower magnitudes. Neighboring cells are then grouped into larger blocks and normalized to account for potential changes in contrast and illumination. This normalization procedure is applied in a sliding-window manner over the entire image with some overlap between neighboring blocks. Ultimately, all normalized histograms are concatenated into the final HOG descriptor (our feature vector $\mathbf{x}$) that can be used for matching and recognition.

3.6. Dense Scale Invariant Feature Transform

The Scale Invariant Feature Transform (SIFT), introduced by Lowe in [22], represents one of the most successful techniques for image description in computer vision. The original approach to SIFT calculation includes both a keypoint detector, capable of finding points of interest in an image, as well as a local descriptor that can effectively represent the local neighborhood around the detected keypoints. As indicated in the introductory section, early techniques to ear recognition relied on the SIFT keypoint detector as well as the SIFT descriptor, e.g., [15], [3] and [9], and, therefore, demonstrated a high degree of robustness towards partial occlusions.

More recent techniques, on the other hand, compute dense SIFT (DSIFT) representations from the images and do not rely on the keypoint detector. Here, the keypoints are simply arranged uniformly into a grid that is placed over the image. Techniques based on DSIFT (e.g., [24], [19]) have reported excellent recognition performance as well as robustness to partial occlusions similar to techniques based on the original SIFT formulation. We evaluate a DSIFT-based technique in the experimental section and, thus, discuss here only the SIFT keypoint descriptor. The reader is referred to [22] for a detailed description of the keypoint detection procedure.

The SIFT descriptor shares similarities with the HOG descriptor. For every point of interest, SIFT considers a local neighborhood of 16×16 pixels. This neighborhood is partitioned into sub-regions of 4×4 pixels and for each sub-region an 8-bin histogram is computed based on the orientations and magnitudes of the image gradient in that sub-region. The gradients are also weighted by a Gaussian function to give more importance to image gradients closer to the point of interest and normalized by the dominant gradient orientation to achieve rotation invariance. The final dimensionality of the SIFT descriptor is 128 for a single keypoint, so care needs to be taken when computing DSIFT representations from the image. The dimensionality of final feature vector can easily become computationally prohibitive if too many grid points are chosen for DSIFT calculation.

3.7. Gabor Wavelets

2D Gabor wavelets were originally introduced by Daugman [14] for the problem of iris coding, but due to their ability to analyze images at multiple scales and orientations, they have been successfully employed in other problem areas as well. In the spatial domain, Gabor wavelets are defined with the following expression[35], [36]:

$$\psi_{u,v}(x,y) = \frac{f_u^2}{\pi\gamma\eta} e^{-(\frac{f_u^2}{\gamma^2}x'^2 + \frac{f_u^2}{\eta^2}y'^2)} e^{j2\pi f_u x'}, \tag{3}$$

where

$$\begin{aligned} x' &= x\cos\theta_v + y\sin\theta_v, \\ y' &= -x\sin\theta_v + y\cos\theta_v \end{aligned} \tag{4}$$

and the parameters f_u and θ_v represent the center frequency and orientation of the complex sinusoidal from Eq. (3), respectively. γ and η define the ratio between the center frequency and the size of the Gaussian and ensure that all generated wavelets share some specific properties [37]. For feature extraction, a family of wavelets is typically created and used to extract features from the processed image. This family commonly consist of wavelets of 5 scales and 8 orientations, i.e., $f_0, f_1, ..., f_7$ and $\theta_0, \theta_1, ..., \theta_4$.

To extract Gabor features from an image, the image is convolved with the entire family of Gabor wavelets (filters), the magnitude responses of the convolution outputs are retained (the phase responses are discarded), down-sampled and concatenated into a global feature vector encoding multi-resolution, orientation-dependent texture information of the input image.

Techniques based on the outlined procedure and its modifications (e.g., using log-Gabor wavelets) are among the most popular techniques for ear recognition [21], [20], [25], [39], [23]. Their advantages lie in their excellent discriminative properties, however, Gabor features are computational relatively complex to compute, as the input image needs to be filtered with an entire family of filters.

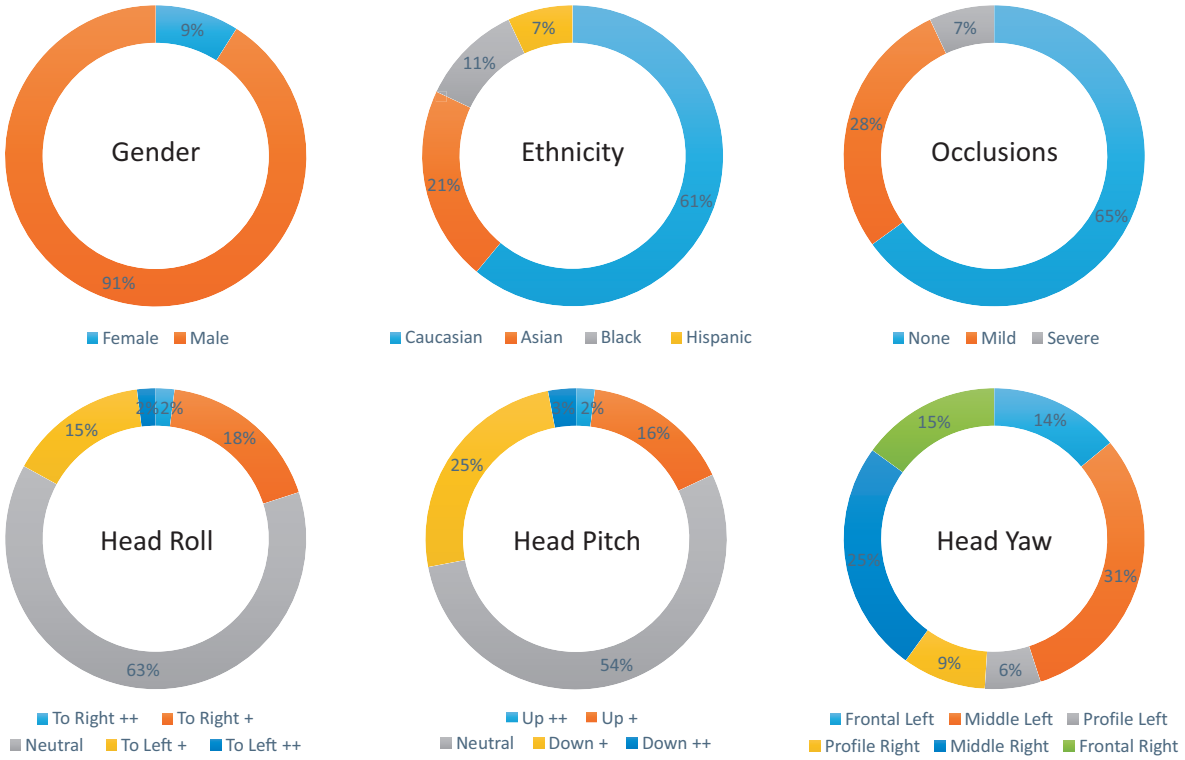


Figure 3: The graphs show the distribution of covariates (labels) of the images of the AWE dataset. The dataset contains 1000 images of 100 subjects. Gender and Ethnicity are lebeled on a *per subject* basis, whereas occlusions, head roll, head pitch, and head yaw vary for each image in the dataset. Accessories are not shown explicitly here, but from the 1000 AWE images, 91% have no accessories, 8% have some accessories and 1% (or 9 images) has a significant amount of accessories.

3.8. Patterns of Oriented Edge Magnitudes

Patterns of Oriented Edge Magnitudes (POEM) [38] represent another popular approach to texture description that combines ideas from LBP and HOG descriptors as well as Gabor wavelets.

The POEM construction procedure starts by computing the gradient of the input image and building magnitude-weighted histograms of gradient orientations for every pixel in the image. This histogram is computed from local pixel neighborhoods referred to by the authors as cells. In this regard, POEM shares similarities with the HOG descriptor, which also relies on gradient directions to encode an image, but different from HOG, POEM computes the histograms densely in a sliding window-manner over the entire image. After this step, every pixel in the image is represented by a local histogram of quantized gradient orientations, or in other words, the image is decomposed into m oriented gradient images, where m is the number of discrete orientations of the local histograms. Each of these images is then encoded using the LBP operator and a global image descriptor is constructed by concatenating all block histograms computed from the oriented gradient images.

The POEM descriptor has demonstrated impressive performance for face recognition [38] and exhibits desirable properties, such as orientational-selectivity, robustness to moderate illumination changes and low-computational complexity, which make it appealing for image representation in ear recognition systems.

4. Dataset and Experimental Protocol

For our experiments, we use the recently introduced Annotated Web Ears (AWE) dataset, which contains 1000 ear images of 100 distinct subjects (with 10 images per subject). The dataset was gathered from the web with a semi-automatic two-step procedure. In the first step candidate images for the dataset were collected from the web using web-crawlers that looked for appropriately tagged imagery on Flickr and Google's image search. The candidate images were then manually screened and curated in the second step to ensure that ears were indeed present in all images. This approach ensured that the appearance variability of the images was not artificially reduced through automatic ear-detection techniques and resulted in a challenging dataset of ear images captured in unconstrained settings [16].

The images of the AWE dataset contain ground truth annotations in terms of gender, extent of head pitch, roll, and yaw rotations, ethnicity and presence of occlusions, and thus provide a perfect starting point for our covariate analysis. The labels/annotations were assigned to the images by a trained annotator and validated by the authors of the

Figure 4: Sample images from the AWE dataset.

dataset. Because the image acquisition procedure was not controlled, each image from the dataset typically exhibits variations across several attributes (e.g., large pitch, roll and yaw angles at the same time) and is annotated with multiple labels, so attribute cross-talk effects need to be taken into account when interpreting the results presented in the next section. The distribution of the individual label categories is presented in Figure 3 and some sample images from the dataset are shown in Fig. 4.

To assess the impact of the different covariates, we conduct identification experiments with the AWE dataset and observe various performance metrics, such as the rank 1 recognition rate (rank-1), the rank 5 recognition rate (rank-5) and the Area Under the Cumulative Match Score Curves (AUC). For each of the experiments the probes consist of all images with a specific label (e.g., severe head yaw), while the galleries represent all images from the AWE dataset. With this setup, the gallery size is fixed for all experiments, while the number of probes (and consequently number of conducted identification experiments) depends on the label distribution (shown in Fig. 3) and differs from experiment to experiment. Related covariates are merged for the experiments: mild head yaw from both left and right are merged into one group of mild yaw, the same for the severe yaw rotation and the other head rotations (roll and pitch).

For the descriptor-based feature extraction methods we use the implementations that ship with AWE toolbox [16] and make no change to the default parameters.

5. Experiments and Results

In this section we report the performances of 8 state-of-the-art feature descriptors for each of the covariate factors. A visual comparison of the rank-1 recognition rates for all experiments is presented in Figure 5 and more detailed comparison including rank-1, rank-5 and AUC values for the evaluation is given in Table 1.

The results show that head rotation negatively impacts the identification performance. Gender and ethnicity have the smallest impact on identification performance - the results for all subgroups of these covariates are very close, while the minor performance differences are likely a consequence of the different number of probes in each subgroup. Surprisingly, occlusions which consist mostly of hair have a limited impact on performance. The reason for this, we argue, is that the occlusions are more or less consistent throughout all ear images for a selected subject, whereas accessories significantly vary from image to image and have therefore a bigger performance impact.

The impact of accessories requires a more in-depth analysis. In the most severe cases where accessories represent a significant part, the performance is degraded the most among all tested covariates. This can be attributed to the fact that samples that fall into this category contain large hearing aids, headphones or some large ornaments, which may not be present in the gallery images. The rank-1 recognition rates of 0%, 11.1% and 22.2% need to be considered with reservation since only 8 samples were available for this experiment. Nevertheless, we believe, that the low performance can still be ascribed to the presence of large accessories and cannot be explained away with the small sample size. More experiments are needed thought to validate this result.

In Figure 6 a comparison of the rank-1 recognition rate for all assessed techniques with respect to the considered covariates is given in the form of radar graphs. Here, the most challenging subgroup is selected and plotted for each technique and each covariate. For gender, the male group was selected for all 8 extractors, for ethnicity Black was chosen for LBP, BSIF, RILPQ, POEM and HOG, and Caucasian for LPQ, DSIFT and Gabor. For accessories, pitch and roll rotations, the most severe cases were the most challenging for all the cases and these values are plotted in the radar graphs. However, for yaw rotations for POEM and DSIFT the neutral poses were selected, whereas for others the most severe cases were used. For occlusion the most challenging proved to be completely non-occluded images for LPB, BSIF and LPQ, while for the remaining techniques the most challenging were the most heavily occluded ears. The graphs show that none of the assessed techniques has a clear advantages over the others, except for BSIF, which covers a slightly larger area than the other techniques in the radar graphs.

6. Conclusion

We have evaluated 8 popular dense descriptor-based feature extraction methods for ear recognition with different covariates. The results show that gender and ethnicity with some exceptions do not impact identification performance significantly. However, severe angles at which ear images are taken (head poses) and severe use of accessories all negatively impact recognition performance. Furthermore, we showed that hair occlusions negatively impact performance to a much more limited extent than other factors. The reason for this, we argue, is that hair that belongs to a specific person is similar throughout all (or most) ear images.

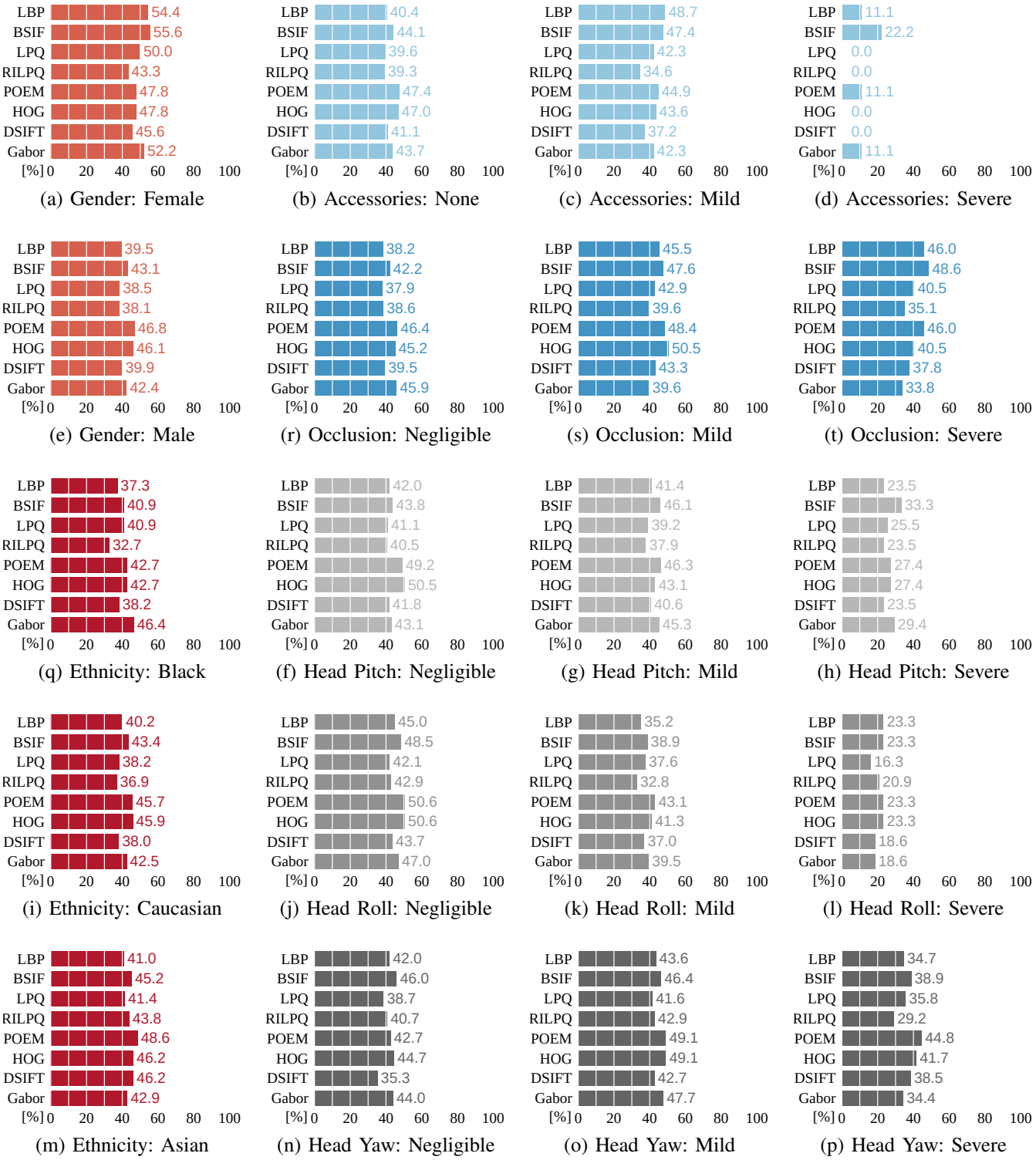


Figure 5: The plots show a comparison of 8 state-of-the-art descriptor-based ear recognition techniques for all considered covariates in terms of rank-1 recognition rates. Due to the small number of subjects for the Hispanic ethnicity subgroup, the results for this subgroup were omitted from the comparison. The small number of images also contributes to the higher rank-1 recognition rates for women compared to men. Here, only 9 classes of female subjects are present in the AWE dataset compared to 91 male subjects. The most severe cases of pitch, roll and tilt angles show a negative impact on the performance of all assessed techniques. The biggest impact is observed with large accessories, but the results for this test are also generated with a small number of probe images. The results are best viewed in color.

Table 1: Comparative assessment of 8 descriptor-based techniques considered in this work. Results were generated on the whole AWE dataset of 1000 images of 100 subjects for rank-1 and rank-5 recognition rates, and the Area Under the Cumulative match score curve (AUC). All results are given in perecentages, *RIL.*, *PO.*, *DSI.* and *Gab.*, denote RILPQ, POEM, DSIFT and Gabor, respectively.

Perf. metric	Rank-1 [%]								Rank-5 [%]								AUC [%]							
Method	LBP	BSIF	LPQ	RIL.	PO.	HOG	DSI.	Gab.	LBP	BSIF	LPQ	RIL.	PO.	HOG	DSI.	Gab.	LBP	BSIF	LPQ	RIL.	PO.	HOG	DSI.	Gab.
Female	54.4	55.6	50.0	43.3	47.8	47.8	45.6	52.2	70.0	71.1	67.8	61.1	67.8	66.7	64.4	71.1	92.8	90.3	90.6	90.1	90.8	89.8	88.8	93.4
Male	39.5	43.1	38.5	38.1	46.8	46.2	39.9	42.4	61.8	62.0	60.4	58.1	65.7	70.2	59.0	67.9	87.8	89.5	88.8	86.3	89.9	92.2	87.5	93.6
Asian	41.0	45.2	41.4	43.8	48.6	46.2	46.2	42.9	63.3	64.3	61.4	63.3	69.5	67.1	61.4	68.6	89.8	92.4	91.3	87.9	91.5	91.9	89.6	94.1
Caucasian	40.2	43.4	38.2	36.9	45.7	45.9	38.0	42.5	60.8	62.3	60.3	56.7	65.6	70.2	58.5	67.4	87.4	88.4	87.7	86.5	89.5	92.1	86.9	92.9
Black	37.3	40.9	40.9	32.7	42.7	42.7	38.2	46.4	60.0	57.3	58.2	51.8	59.1	70.0	56.4	70.0	86.5	89.1	88.1	83.7	86.7	90.1	85.9	94.5
Accessories /	40.4	44.1	39.7	39.3	47.4	47.0	41.1	43.7	62.4	62.8	61.2	59.0	66.4	70.7	60.0	68.7	88.4	89.7	89.1	86.7	90.1	92.1	87.9	93.8
Accessories +	48.7	47.4	42.3	34.6	44.9	43.6	37.2	42.3	66.7	65.4	64.1	55.1	65.4	65.4	56.4	66.7	87.6	88.4	87.7	87.4	89.3	91.4	86.3	91.8
Accessories ++	11.1	22.2	0.0	0.0	11.1	0.0	0.0	11.1	33.3	44.4	22.2	22.2	22.2	33.3	33.3	33.3	82.0	82.9	85.7	80.3	83.6	80.5	72.3	82.8
Pitch /	42.0	43.8	41.1	40.5	49.2	50.5	41.8	43.1	63.4	62.6	62.3	59.5	66.5	73.5	62.1	68.1	88.5	90.2	89.5	87.0	90.5	93.4	88.6	93.8
Pitch +	41.4	46.1	39.2	37.9	46.3	43.1	40.6	45.3	63.8	64.8	61.1	58.4	67.0	67.2	58.6	70.2	88.5	89.2	88.8	86.6	90.3	91.1	87.3	93.6
Pitch ++	23.5	33.3	25.5	23.5	27.5	27.5	23.5	29.4	43.1	49.0	49.0	47.1	51.0	52.9	39.2	52.9	83.5	85.5	84.3	83.4	81.7	84.0	79.3	90.5
Roll /	45.0	48.5	42.1	42.9	50.6	50.6	43.7	47.0	66.4	67.4	65.0	61.6	70.2	75.7	63.5	73.6	90.1	91.3	90.9	87.9	91.8	94.0	89.7	94.6
Roll +	35.2	38.9	37.7	32.8	43.1	41.3	37.1	39.5	59.0	57.5	56.9	56.3	61.1	63.3	55.7	61.5	86.4	88.0	86.9	85.8	88.8	89.7	85.2	92.2
Roll ++	23.3	23.3	16.3	20.9	23.3	23.3	18.6	18.6	32.6	37.2	37.2	27.9	39.5	37.2	30.2	41.9	75.8	76.5	76.3	75.2	72.0	79.4	75.6	88.0
Yaw /	42.0	46.0	38.7	40.7	42.7	44.7	35.3	44.0	60.7	68.0	60.0	59.3	64.7	70.0	61.3	74.0	88.7	89.6	90.0	85.7	88.6	94.1	87.7	94.7
Yaw +	43.6	46.4	41.6	42.9	49.1	49.1	42.7	47.7	65.5	63.9	63.5	61.7	68.7	72.8	59.8	70.8	89.5	90.7	89.8	88.0	91.0	92.3	88.8	94.4
Yaw ++	34.7	38.9	35.8	29.2	44.8	41.7	38.5	34.4	57.6	58.0	56.9	51.4	61.1	64.2	58.0	60.1	85.6	87.3	86.8	84.5	88.7	90.3	85.2	91.3
Occlusion /	38.3	42.2	37.9	38.6	46.4	45.2	39.5	45.9	61.4	61.9	60.4	57.6	65.0	67.7	58.2	69.9	87.6	89.1	88.5	85.6	89.5	91.9	87.3	93.8
Occlusion +	45.5	47.6	42.9	39.6	48.4	50.6	43.3	39.6	64.7	65.1	63.3	61.5	66.6	75.3	62.6	65.8	88.9	90.4	89.4	88.8	90.6	92.5	87.8	93.1
Occlusion ++	46.0	48.7	40.5	35.1	46.0	40.5	37.8	33.8	63.5	62.2	59.5	54.1	71.6	68.9	59.5	62.2	91.2	90.9	91.1	88.6	92.3	91.1	89.3	92.8

We hope that the findings of this paper help with the development of new ear recognition algorithms – our results show there is need for pose normalization techniques and unwanted-objects (accessories) segmentation.

Acknowledgements

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Programme P2-0250 (B) Metrology and Biometric Systems, the ARRS Research Programme P2-0214 (A) Computer Vision.

References

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. Nixon. A Survey on Ear Biometrics. *ACM Computing Surveys*, 45(2):1–22, 2013.

[2] M. Alaraj, J. Hou, and T. Fukami. A neural network based human identification framework using ear images. In *Proceedings of the International technical conference of IEEE Region 10*, pages 1595–1600. IEEE, 2010.

[3] B. Arbab-Zavar and M. S. Nixon. Robust log-gabor filter for ear biometrics. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[4] Z. Baoqing, M. Zhichun, J. Chen, and D. Jiyuan. A robust algorithm for ear recognition under partial occlusion. In *Proceedings of the Chinese Control Conference*, pages 3800–3804, 2013.

[5] A. Basit and M. Shoaib. A human ear recognition method using non-linear curvelet feature subspace. *International Journal of Computer Mathematics*, 91(3):616–624, 2014.

[6] A. Benzaoui, N. Hezil, and A. Boukrouche. Identity recognition based on the external shape of the human ear. In *Proceedings of the International Conference on Applied Research in Computer Science and Engineering*, pages 1–5. IEEE, 2015.

[7] A. Benzaoui, A. Kheider, and A. Boukrouche. Ear description and recognition using ELBP and wavelets. In *Proceedings of the International Conference on Applied Research in Computer Science and Engineering*, pages 1–6, 2015.

[8] H. Bourouba, H. Doghmane, A. Benzaoui, and A. H. Boukrouche. Ear recognition based on Multi-bags-of-features histogram. In *Proceedings of the International Conference on Control, Engineering Information Technology*, pages 1–6, 2015.

[9] J. D. Bustard and M. S. Nixon. Toward unconstrained ear recognition from two-dimensional images. *Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 40(3):486–494, 2010.

[10] T.-S. Chan and A. Kumar. Reliable ear identification using 2-D quadrature filters. *Pattern Recognition Letters*, 33(14):1870–1881, 2012.

[11] M. Choraś. Perspective methods of human identification: ear biometrics. *Opto-electronics review*, 16(1):85–96, 2008.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the International Conference on Computer Vision and Patten Recognition*, pages 886–893. IEEE, 2005.

[13] N. Damar and B. Fuhrer. Ear recognition using multi-scale histogram of oriented gradients. In *Proceedings of the Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 21–24, 2012.

[14] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

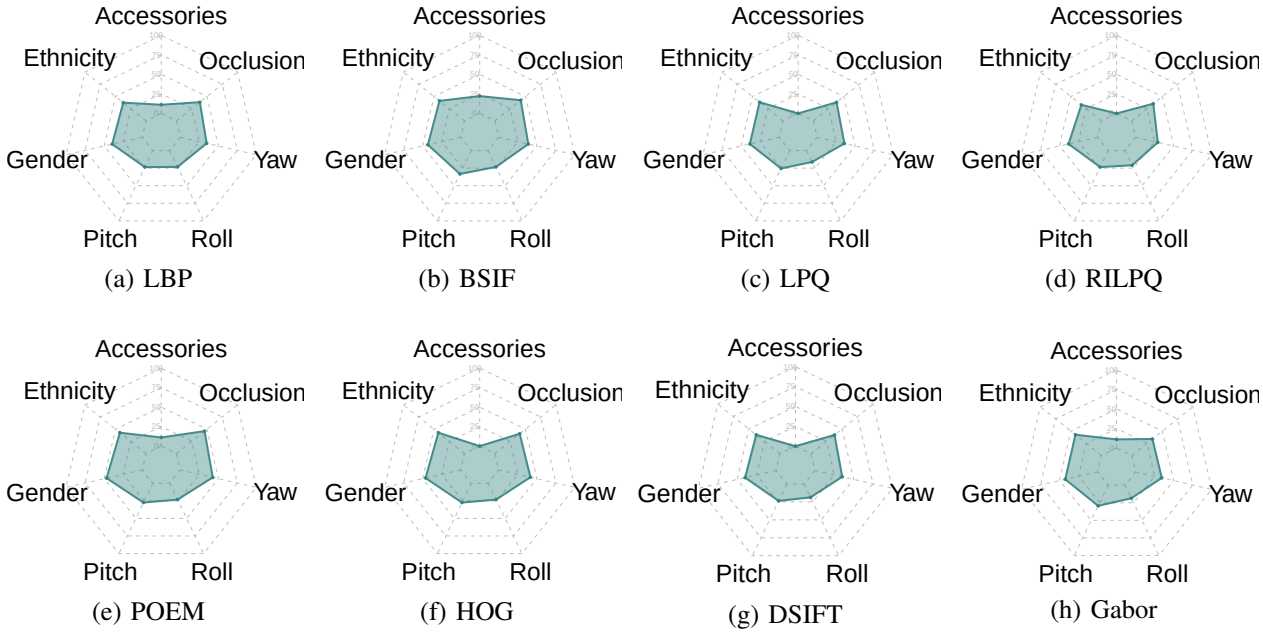
[15] K. Dewi and T. Yahagi. Ear photo recognition using scale invariant keypoints. In *Proceedings of the Computational Intelligence*, pages 253–258, 2006.

[16] Z. Emersic, V. Struc, and P. Peer. Ear recognition: More than a survey. *Neurocomputing*, in press:1–22, 2017.

[17] Y. Guo and Z. Xu. Ear recognition using a new local matching approach. In *Proceedings of the International Conference on Image Processing*, pages 289–292. IEEE, 2008.

[18] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *Proceedings of the International Conference on Pattern Recognition*, pages 1363–1366. IEEE, 2012.

[19] J. Krizaj, V. Struc, and N. Pavesic. Adaptation of SIFT features for robust face recognition. In *Proceedings of the Image Analysis and*



(a) LBP (b) BSIF (c) LPQ (d) RILPQ

(e) POEM (f) HOG (g) DSIFT (h) Gabor

Figure 6: The radar graphs show a comparison of the rank-1 recognition rates of the evaluated feature extraction methods with respect to the covariates. The axes show values from 0 to 100%. For each covariate the most challenging covariate factor for each feature extraction method was selected for the graph, i.e. the most severe head pitch, the most occluded images etc. The graphs show that all 8 descriptors perform similarly, with BSIF showing some robustness to accessories.

Recognition, pages 394–404. Springer, 2010.

[20] A. Kumar and C. Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, 2012.

[21] A. Kumar and D. Zhang. Ear authentication using log-gabor wavelets. In *Proceedings of the Symposium onDefense and Security*, page 65390A. International Society for Optics and Photonics, 2007.

[22] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[23] A. Meraoumia, S. Chitroub, and A. Bouridane. An automated ear identification system using Gabor filter responses. In *Proceedings of the International Conference on New Circuits and Systems*, pages 1–4. IEEE, 2015.

[24] A. Morales, M. Ferrer, M. Diaz-Cabrera, and E. Gonzalez. Analysis of local descriptors features and its robustness applied to ear recognition. In *Proceedings of the International Carnahan Conference on Security Technology*, pages 1–5. IEEE, 2013.

[25] L. Nanni and A. Lumini. Fusion of color spaces for ear authentication. *Pattern Recognition*, 42(9):1906–1913, 2009.

[26] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, pages 236–243. Springer, 2008.

[27] V. Ojansivu, E. Rahtu, and J. Heikkilä. Rotation invariant local phase quantization for blur insensitive texture analysis. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[28] A. Pflug and C. Busch. Ear biometrics: a survey of detection, feature extraction and recognition methods. *Biometrics*, 1(2):114–129, 2012.

[29] A. Pflug, C. Busch, and A. Ross. 2D ear classification based on unsupervised clustering. In *Proceedings of the International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[30] A. Pflug, P. N. Paul, and C. Busch. A comparative study on texture and surface descriptors for ear biometrics. In *Proceedings of the International Carnahan Conference on Security Technology*, pages 1–6. IEEE, 2014.

[31] A. Pflug, J. Wagner, C. Rathgeb, and C. Busch. Impact of severe signal degradation on ear recognition performance. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1342–1347. IEEE, 2014.

[32] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Computer Vision Using Local Binary Patterns*. Computational Imaging and Vision. Springer, 2011.

[33] S. Prakash and P. Gupta. An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems*, 52(3):1435–1448, 2013.

[34] R. Purkait. Role of External Ear in Establishing Personal Identity - A Short Review. *Austin Journal of Forensic Science and Criminology*, 2(2):1–5, 2015.

[35] V. Struc, R. Gajsek, and N. Pavesic. Principal Gabor filters for face recognition. In *Proceedings of the Conference on Biometrics: Theory, Applications and Systems*, pages 1–6. IEEE, 2009.

[36] V. Struc and N. Pavesic. Gabor-based kernel partial-least-squares discrimination features for face recognition. *EURASIP Journal on Advances in Signal Processing*, 20(1):115–138, 2009.

[37] V. Struc and N. Pavesic. The complete gabor-fisher classifier for robust face recognition. *EURASIP Journal on Advances in Signal Processing*, 2010:1–26, 2010.

[38] N.-S. Vu and A. Caplier. Face recognition with patterns of oriented edge magnitudes. *Computer Vision*, pages 313–326, 2010.

[39] W. Xiaoyun and Y. Weiqi. Human ear recognition based on block segmentation. In *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 262–266. IEEE, 2009.

[40] Z. Xie and Z. Mu. Ear recognition using lle and idlle algorithm. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[41] Z. Zhang and H. Liu. Multi-View ear recognition based on b-spline pose manifold construction. In *Proceedings of the World Congress on Intelligent Control and Automation*, 2008.