

Robust 2D Ear Registration and Recognition Based on SIFT Point Matching

John D. Bustard, Mark S. Nixon

Department of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK

(jdb07r, msn)@ecs.soton.ac.uk

Abstract— Significant recent progress has shown ear recognition to be a viable biometric. Good recognition rates have been demonstrated under controlled conditions, using manual registration or with specialised equipment. This paper describes a new technique which improves the robustness of ear registration and recognition, addressing issues of pose variation, background clutter and occlusion. By treating the ear as a planar surface and creating a homography transform using SIFT feature matches, ears can be registered accurately. The feature matches reduce the gallery size and enable a precise ranking using a simple 2D distance algorithm. When applied to the XM2VTS database it gives results comparable to PCA with manual registration. Further analysis on more challenging datasets demonstrates the technique to be robust to background clutter, viewing angles up to ± 13 degrees and with over 20% occlusion.

I. INTRODUCTION

Ears offer an exciting new approach to non-contact biometrics. They have a number of advantages over other recognition features. In particular, ears are suitable for use at a distance and have the advantage of being relatively constant over a person's life. Also, in comparison with faces, ears do not suffer from variation due to expressions.

An overview of existing ear recognition techniques, by Hurley et al. [1], shows that some of the best results use 3D object matching [2] [3] [4]. With this approach, ears can be recognised under varying lighting conditions and poses (out of plane rotations). One limitation of the technique, however, is that a specialised camera is required to capture the 3D data. Also, these cameras need controlled lighting to produce accurate results [5]. Where these conditions cannot be met, as, for example, in 'ID at a distance' situations where there are restricted data sources, such as grey scale video from security cameras, 2D techniques have to be used. This paper proposes enhancements to the current 2D approach.

Essentially, 2D ear recognition has three stages: detection, registration and classification. Here *detection* refers to the finding of an ear in a probe image, *registration* as the aligning of a potential gallery ear with the probe and *classification* as the ranking of gallery ears to identify the most likely person in the probe. Most existing research has concentrated on the classification stage, with ears being identified and registered manually. Good recognition has been obtained with manual registration, even in the presence of occlusion [6]. However, there is currently no well-established scheme for automatic 2D detection and registration. Several techniques have been proposed but many rely on controlled imaging conditions,

such as assuming that the image is a single head profile in front of a flat background.

The main contribution of this paper is to propose an improved ear registration technique based on the object recognition algorithm of Brown et al. [7]. Their technique attempts to create a homography transform between a gallery object image and a probe image using SIFT (Scale-Invariant Feature Transform) point matches. The probe is considered to include an image of the gallery object, if a homography can be created. In addition, the homography defines the registration between the gallery and the probe. This creates a very accurate registration. Brown demonstrated good results for various objects but is insufficiently discriminating to rank ear images. The work described in this paper extends their technique with an image distance algorithm to obtain a precise ranking. To calculate the image distance accurately, gallery ears are segmented using a mask. These masks are semi-automatically created as a preprocessing step on the gallery.

Collectively, these developments create an automated, accurate, ear recognition technique that is robust to location, scale, pose, background clutter and occlusion. Effectively, the technique is a step towards achieving the accuracy of 3D ear recognition with unconstrained 2D data.

The paper describes the proposed technique and its evaluation, with four datasets used to assess its robustness and accuracy. Section II discusses existing automated registration algorithms and reviews their strengths and weaknesses. Following this, Section III describes the stages of the technique, including the semi-automatic creation of gallery masks. The registration calculation and its theoretical justification are also described, as well as an overview of the distance measure for accurate ranking. In Section IV the paper then discusses the evaluation of the proposed technique. This includes both a traditional, controlled environment, recognition test as well as more challenging datasets that evaluate the techniques robustness to occlusion, background clutter and pose variation. The paper concludes with suggestions for future work.

II. RELATED WORK

A number of approaches to ear recognition in 2D have been proposed. Of these, PCA (Principal Components Analysis) is often used as a baseline comparison because of its good performance in controlled conditions [2] [8]. Unfortunately, it is very sensitive to occlusion and misregistration [8]. Occlusion, in particular, is a key problem as the ear is frequently obscured by hair or earrings. Some progress has

been made to address this issue by, for example, using ear models [8] or by adapting the PCA algorithm [6].

In terms of registration, a number of techniques have been suggested. Broadly they can be categorised as *edge shape matching* and *area matching* approaches.

For edge shape matching (usually based on finding the outer ear curve), Ansari et al. [9] propose a method based on completing convex curved edge regions to find the outer ear. Despite producing precise registrations, this approach can generate many false positives by matching non ear convex regions. Also occlusion is likely to invalidate the convex assumption.

Arbab-Zavar et al. [10] have proposed an enrolment technique exploiting the elliptical shape of the outer ear. This has produced good results with occlusion, but the accuracy of registration is much less than can be achieved manually. Also, it makes the assumption that the ear is the principal elliptical shape in the image. This restricts its use to controlled settings, as the presence of background objects can produce false positives.

The remaining approaches involve area matching. These techniques can have very fast implementations but often have lower registration accuracy, especially when the objects are occluded. One approach, originally developed for face recognition, is the use of a Haar-like feature object detector, as proposed by Viola et al. [11]. This is a fast and robust technique but suffers from inaccuracy in localisation. A refinement, for ear detection, by Abate et al. [12] uses the edge centre of mass for localisation but this is sensitive to occlusion.

Abdel-Mottaleb et al. [13] use Hausdorff edge template matching between an example ear helix edge and edges identified on skin coloured regions of an image. This relies on relatively constrained lighting conditions (to detect the skin region accurately) and is sensitive to outer ear edge occlusion by hair.

Finally, a real-time technique has been developed by Laszlo et al. [14]. This uses edge orientation pattern matching followed by an active contour. By combining the speed of template matching with the accuracy of active contours accurate registration can be achieved. This process is robust to significant pose variation but the pattern matching localisation is sensitive to occlusion, leading to poor active contour fitting.

This paper approaches ear registration from a new perspective. By matching sets of points, rather than areas or edge shapes, the registration transform can be precisely calculated even under occlusion, background clutter and pose variation. This is now described.

III. TECHNIQUE

Before any probe images can be tested, the gallery images are processed to segment the ears. Each gallery image is then analysed to determine its SIFT feature points. Once this is complete a probe image can be recognised.

The first step is to identify feature points in the probe. For each of these points the gallery is searched to find correspondences. If four points can be matched between the

probe and the gallery, they are used to calculate a perspective transformation that registers the probe. Once the two images are aligned, the distance between the images is calculated. The nearest gallery image identifies the person.

Each stage of this process is described in the sub-sections that follow.

A. Building the gallery database

Images of the same ear taken at different times can vary significantly due to changes in hair length and colour. This variation can create many false point matches and significantly reduces the accuracy of image distance measurements. For this reason, gallery ears are masked to segment the ear from the surrounding skin and hair, as illustrated in Figure 1.

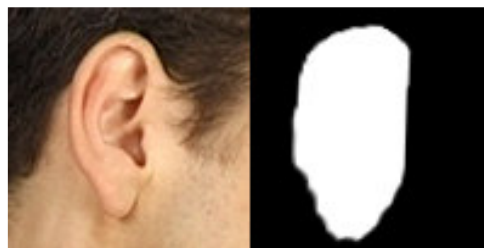


Fig. 1 A gallery ear image and its associated mask

By assuming that ear variations can be achieved through a series of smooth local deformations, these masks can be semi-automatically created. Under this assumption, and by using a sufficiently large gallery, each ear is likely to share at least four points with an ear from a different subject. Some evidence for this hypothesis has been provided by Arbab-Zavar's model-based ear recognition algorithm [8], which describes six growth factors that define an ear's shape.

The masks are created through a bootstrapping process as follows:

A seed ear is selected and a mask manually created for that ear. The rest of the gallery ears are then matched against the seed (following the same technique used for probe recognition). Each match defines a mask for that ear. All the masked ears now form a larger seed, against which the remaining gallery images are tested. This process is repeated until there are no more matches.

If there are any gallery images remaining, a new mask is created manually and that image added to the seed. This is repeated until all gallery images have masks.

B. Feature detection

SIFT [15] was used for the detection of features. It is an effective feature detector, robust to scale in plane rotation and to lighting, and with some robustness to pose (out of plane rotation).

To make the matching of features against a large gallery more efficient the Approximate Nearest Neighbours [16] algorithm was used. This enables efficient 128 dimensional point matches in $O(\log(n))$ where n is the number of feature points in the gallery.

C. Registration calculation

Eight non planar point correspondences between two images provide enough information to calibrate two cameras, thereby fully registering a three-dimensional solid object between two views. Unfortunately finding eight non-planar point correspondences reliably is too tight a constraint for ears. However, if all the points lie in a plane, only four point correspondences are needed [17]. These correspondences can be used to define the transformation of the plane from one image to the other. This transformation is known as an *homography* and its calculation is as follows.

Let \mathbf{x} be a homogeneous point in the probe image and \mathbf{x}' be a homogeneous point in the gallery image, then the homography \mathbf{H} is defined by

$$\mathbf{x}' = \mathbf{H}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad \mathbf{x}' = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix}$$

This can be expressed as

$$\mathbf{x}' \times \mathbf{H}\mathbf{x} = \mathbf{0}$$

By considering \mathbf{H} as a matrix of row vectors \mathbf{h}^{iT}

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}^{1T} \\ \mathbf{h}^{2T} \\ \mathbf{h}^{3T} \end{bmatrix}$$

the cross product can be expanded to give

$$\mathbf{x}' \times \mathbf{H}\mathbf{x} = \begin{pmatrix} y'\mathbf{h}^{3T}\mathbf{x} - \mathbf{h}^{2T}\mathbf{x} \\ \mathbf{h}^{1T}\mathbf{x} - x'\mathbf{h}^{3T}\mathbf{x} \\ x'\mathbf{h}^{2T}\mathbf{x} - y'\mathbf{h}^{1T}\mathbf{x} \end{pmatrix}$$

Since $\mathbf{h}^{iT}\mathbf{x} = \mathbf{x}^T\mathbf{h}^i$ this can be rewritten as:

$$\begin{bmatrix} \mathbf{0}^T & -\mathbf{x}^T & y'\mathbf{x}^T \\ \mathbf{x}^T & \mathbf{0}^T & -x'\mathbf{x}^T \\ -y'\mathbf{x}^T & x'\mathbf{x}^T & \mathbf{0}^T \end{bmatrix} \begin{pmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{pmatrix} = \mathbf{0}$$

This is a linear equation in \mathbf{h} of the form $\mathbf{A}\mathbf{h} = \mathbf{0}$, where \mathbf{A} is a 3x9 matrix and \mathbf{h} is a 9 vector. \mathbf{A} has only two linearly independent equations as the third row is the sum of $-x'$ times the first row and $-y'$ times the second. By omitting this equation the remaining set becomes

$$\begin{bmatrix} \mathbf{0}^T & -\mathbf{x}^T & y'\mathbf{x}^T \\ \mathbf{x}^T & \mathbf{0}^T & -x'\mathbf{x}^T \end{bmatrix} \begin{pmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{pmatrix} = \mathbf{0}$$

This shows that each point correspondence adds two independent equations in the entries of \mathbf{H} . By combining these equations into a single matrix, four point

correspondences create a matrix with a size 8x9 and rank 8. This matrix has a 1-dimensional null-space which can be solved to produce a solution to \mathbf{H} up to a non-zero scale. As these points are homogeneous, if the transformed points are normalised by dividing through by their third component, this scale factor will be removed.

By making the simplification that an ear is a planar structure, ears can be registered accurately. The SIFT matching distance is quite generous to enable large variations in pose and lighting which produces a significant number of false positives in the point correspondences. Performance and accuracy were improved by using an evidence gathering strategy. Feature matches contain position, scale and rotation information and therefore each point provides an estimate of the in-plane affine transform of the probe to the gallery. Correctly matching points will have approximately the same in-plane affine transform (the greater the out of plane rotation the less this will be true). By grouping points into bins based on their affine transform, many false positives can be excluded.

The potential space of affine transforms was subdivided into four dimensions: two for position, one for logarithm of the scale, and one for rotation. Each of these dimensions was then partitioned into bins: eight for scale and rotation and one for every 128 pixels in width and height. A low resolution of bins was used to ensure the matching is robust to pose variation. Each point match is placed in the appropriate bin and in its closest neighbour (sixteen bin entries per point). If any bin contains four or more point matches its points are passed to the next stage.

Even after this process, false positive point matches remain. To address this, a RANSAC algorithm was used: random sets of four points are selected from the list of point correspondences and an homography calculated. The homography that matches the most points within some threshold, in this case 1% of the ear mask size, is selected as the best match.

Gallery images that provide valid homographies are then passed to the distance measure. The combination of Generalised Hough Transform and valid homography greatly reduces the set of potential gallery matches.

D. Distance measure

Once the gallery images have a good registration they are matched against the probe. The distance is calculated as the robust sum of the squared pixel error after normalisation. The distance measure is made robust to occlusion by thresholding the error. Pixels that differ by more than half the maximum brightness variation are considered to be occluded and so excluded.

Normalisation involved adjusting the scale and offset of the intensity values to achieve a defined mean and standard deviation before comparison. This removed variation in brightness and contrast due to different lighting conditions and camera properties.

IV. EVALUATION

Four datasets were used for evaluation. The first provided a straight test of recognition accuracy on a relatively constrained dataset. For this, a subset of the XM2VTS [18] face-profile database was chosen. It consists of 63 subjects with relatively unoccluded ears. This is the same dataset used by Hurley et al. [19] and Arbab-Zavar et al. [8].

The second and third datasets were synthesised from the XM2VTS images to test the effects of occlusion and background clutter. The fourth and final dataset was created by recording 20 subjects from a range of angles to test the technique's robustness to pose variation.

A. Recognition evaluation

Comparison implementations

For the constrained gallery set, two comparison implementations were created. The first used manually registered ear images, applying the technique described by Yan et al. [2]. This involved defining the Triangular Fossa and Incisure Intertragica of each ear manually. These landmarks were then used to standardise the scale and rotation of all gallery and pose images. The resulting normalised images were segmented with a rectangular mask in the centre of the image capturing the inner ear features.

The second technique applied the algorithm described by Arbab-Zavar [10] to register the ear automatically, using the outer ear ellipse. In both cases the intensity values had their mean and standard deviation normalised. These registered images were ranked using the PCA technique giving the results shown in Table I.

Each technique used the 'leave one out' strategy, with each image removed from the gallery and tested against the rest of the dataset in turn.

TABLE I
RECOGNITION RATE FOR DIFFERENT REGISTRATION TECHNIQUES

Registration	Technique	% Rank 1
Manual	PCA	96%
Automatic using outer ellipse	PCA	75%
Automatic using homography	Image distance	96%

Mask creation

The bootstrapping process, using the first ear, matches over 75% of the gallery. In total, 22 masks were created manually to cover 252 gallery images.

Generally, the masks are not a precise fit for the ears but the accuracy is sufficient to obtain enough feature points for the registration and distance measures.

Registration calculation

It can be seen from Table II that the homography registration is the primary point at which the ears are recognised, going from almost the entire gallery down to four candidate images. The registration calculation is also the cause of 4% of the probe images remaining unclassified. All of these ears failed to produce a valid homography because of insufficient SIFT point matches.

TABLE II
NUMBER OF FEATURES AT EACH STAGE XM2VTS DATASET

Feature	Count
Number of gallery images	251
Number of gallery SIFT points	14,234
Average number of SIFT points on XM2VTS image (720x576)	4,659
Average number of SIFT matches	20,834
Average number of images with SIFT matches	250
Average number of images with valid homographies	4

B. Robustness evaluation

Gallery

The second dataset was created by randomly placing XM2VTS masked ear images on a set of complex background images. These images more closely represent the type of unconstrained environment present with covert biometrics. The third dataset was built by adding varying sized solid black rectangles over the top or side of the original gallery images. This reflects the areas of the ear that are most frequently occluded by hair. Finally, to generate the fourth dataset, 20 subjects were recorded turning in front of a camera. Each person had a camera calibration grid affixed to a hat that was worn as they were photographed. This grid enabled the camera intrinsics and pose angles to be calculated accurately. These calculations were performed using the standard camera calibration algorithms provided with the OpenCV [20] libraries. Figure 2 shows examples from each of these datasets.

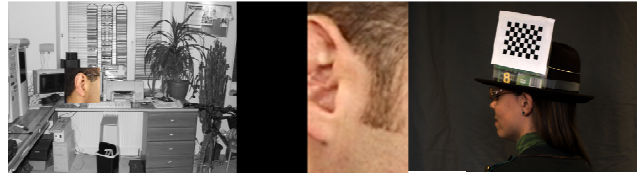


Fig. 2 Examples of more challenging probe images. From left to right background clutter, occlusion and pose variation







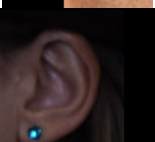
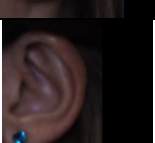
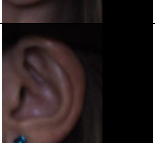
Results

Table III summarises the results of these recognition tests. Background clutter was found to have little effect on the recognition rate as was up to 20% occlusion from above and 10% occlusion from the side. However, any greater occlusion significantly reduced the technique's accuracy. Once again, this was due to failing to find sufficient SIFT matches to calculate the homography.

Figure 3 shows the average recognition rate for 40 ears with varying pose. The technique maintains 100% recognition rate up to ± 13 degrees. As an experiment to improve this technique's robustness to pose variation, additional gallery images were synthesised at novel poses. This was achieved by treating the ear image as a plane photographed at an estimated distance with an approximated field of view. The plane was then rotated in the image plane x and y axes and re-rendered to simulate different poses. This increased the number of SIFT matches but also the number of false positives. As the ears are

not completely planar the image distance increases with angle resulting in incorrect ears having a shorter image distance and so no significant increase in robustness was observed.

TABLE III
AVERAGE RECOGNITION RATES FOR MORE CHALLENGING DATASETS

Technique	% Rank 1 Recognition	Examples
Base recognition rate	96%	
Background clutter	93%	
20% occlusion from above	92%	
30% occlusion from above	74%	
10% occlusion from the side	92%	
20% occlusion from the side	66%	
0 degrees pose variation	100%	
13 degrees pose variation	100%	
22 degrees pose variation	33%	

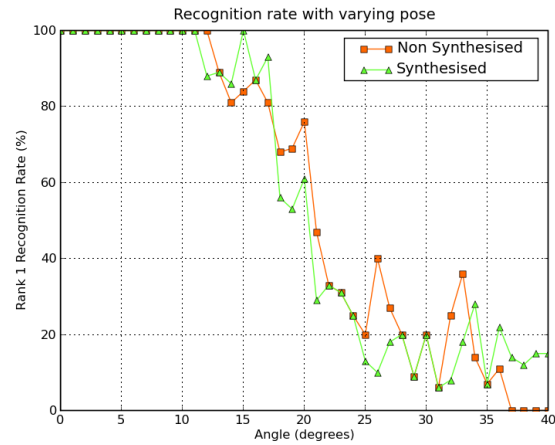


Fig. 3 Recognition rate with varying pose, with and without synthesised ear images

The approach described is relatively successful in identifying ears under different conditions but as is evident from Table III it would be desirable to increase the degree of pose variation over which recognition can be achieved. One strategy would be to record subjects at multiple angles, either at gallery creation or as probes. Alternatively, if this were not possible, the synthesis algorithm could be improved through the use of a morphable model [21]. Further work will explore these possibilities.

Another limitation of the approach is the increased computation time required to achieve the accuracy of the algorithm. Despite the use of the ANN library, the processing of each 720x576 probe image takes over a minute on a 2.4Ghz Dual Core PC. Further work will explore performance improvement through a generic ear model, such as the Viola-Jones classifier [11] trained on ear images. The model would identify regions where an ear is likely to be found, thereby reducing the number of SIFT points that need to be matched. Further improvement might be achieved through a histogram pyramid matching technique. Typically, this enables efficient comparisons between sets of high dimensional features and can be scaled to very large datasets.

In addition, the current system uses image pixel difference as a distance measure. Further work will investigate the benefits of more invariant measures such as Hausdorff edge distances [22].

V. CONCLUSIONS

This paper describes a new technique for ear recognition in 2D images using homographies calculated from SIFT point matches. When applied to the XM2VTS database the technique gives results comparable to PCA with manual registration. In addition, when used on more challenging datasets, it shows robustness to background clutter, 20% occlusion and over ± 13 degrees of pose variation. Further work will focus on performance improvement and increased robustness.

Overall, this paper has demonstrated that automatic, unconstrained 2D ear recognition can be achieved effectively with the proposed homography approach.

VI. REFERENCES

- [1] D. J. Hurley, B. Arbab-Zavar, "The Ear as a Biometric," *Handbook of Biometrics.*, pp. 131-150, Oct. 2007.
- [2] P. Yan and K. W. Bowyer, "ICP-Based Approaches for 3D Ear Recognition", *Biometric Technology for Human Identification II, Proc. of SPIE*, vol. 5779, pp. 282-291, 2005.
- [3] H. Chen and B. Bhanu, "Human Ear Recognition in 3D", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 718-737, Apr. 2007.
- [4] T. Theoharis, G. Passalis, G. Toderici and I.A. Kakadiaris. Unified 3D Face and Ear Recognition using Wavelets on Geometry Images. *Pattern Recognition*, 2007.
- [5] K. W. Bowyer, K. Chang, P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Computer Vision and Image Understanding*, Vol. 101, pp. 1-15, Jan. 2006.
- [6] L. Nanni, A. Lumini, "A multi-matcher for ear authentication", *Pattern Recognition Letters*, vol. 28, issue 16, pp. 2219-2226, Dec. 2007.
- [7] M. Brown, D. G. Lowe, "Invariant Features from Interest Point Groups", *Proceedings of the 13th British Machine Vision Conference*, 2002, pp 253-262, 2002
- [8] B Arbab-Zavar, M. S. Nixon, D. J. Hurley, "On Model-Based Analysis of Ear Biometrics", *Biometrics: Theory, Applications, and Systems*, pp.1-5, 2007.
- [9] S. Ansari, P. Gupta, "Localization of Ear Using Outer Helix Curve of the Ear", *Proceedings of international conference on computing: Theory and Applications*, pp. 688-692, 2007.
- [10] B Arbab-Zavar, M. S. Nixon. "On shape-mediated enrolment in ear biometrics", *Advances in visual computing*, Lecture Notes in Computer Science, vol. 4842, pp. 549-558, 2007.
- [11] P. Viola, M. Jones, "Robust Real-time Object Detection", *International Journal of Computer Vision*, Vol. 57, pp. 137-154, 2004.
- [12] A. F. Abate, M. Nappi, D. Riccio, S. Ricciardi, "Ear Recognition by means of a Rotation Invariant Descriptor", *Proceedings of the 18th International Conference on Pattern Recognition*, Vol.. 4, pp. 437-440, 2006
- [13] M. Abdel-Mottaleb, J. Zhou, "Human Ear Recognition from Face Profile Images", *Advances in Biometrics*, Vol. 3832, pp. 786-792, 2005.
- [14] E. Jeges, L. Máté, "Model-Based Human Ear Localization and Feature Extraction", *International Journal of Intelligent Computing in Medical Sciences and Image Processing*, Vol. 1, pp. 101-112, 2007.
- [15] D. G. Lowe, "Object recognition from local scale-invariant features", *Proceedings of international conference on computer vision*, pp. 1150-1157, 1999.
- [16] S. Arya, D. M. Mount., N. S. Netanyahu., R. Silverman, A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions". *Journal of the ACM*, 45(6), pp. 891-923. 1998
- [17] R. Hartley, A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2000.
- [18] K. Messer, J. Matas, J. Kittler, J. Luetttin, and G. Maitre. "Xm2vtsdb: The extended m2vts database". *Proc. AVBPA*, Washington D.C., 1999.
- [19] D. J. Hurley, M. S. Nixon, and J. N. Carter. "Force field feature extraction for ear biometrics", *Computer Vision and Image Understanding*, 98, pp. 491-512, 2005.
- [20] OpenCV. www.intel.com/technology/computing/opencv/index.htm.
- [21] B. Weyrauch, J. Huang, B. Heisele and V. Blanz, "Component-Based Face recognition with 3D Morphable Models", *IEEE Workshop on Face processing in Video*, FPIV04, 2004
- [22] M. P. Dubuisson, A. K. Jain, "Modified Hausdorff distance for object matching", *Proc. of IAPR Int. Conf. on Pattern Recognition*, pp. 566-568, 1994.