Toward Unconstrained Ear Recognition From Two-Dimensional Images

John D. Bustard, *Student Member, IEEE,* and Mark S. Nixon, *Associate Member, IEEE*

Abstract—Ear recognition, as a biometric, has several advantages. In particular, ears can be measured remotely and are also relatively static in size and structure for each individual. Unfortunately, at present, good recognition rates require controlled conditions. For commercial use, these systems need to be much more robust. In particular, ears have to be recognized from different angles (poses), under different lighting conditions, and with different cameras. It must also be possible to distinguish ears from background clutter and identify them when partly occluded by hair, hats, or other objects. The purpose of this paper is to suggest how progress toward such robustness might be achieved through a technique that improves ear registration. The approach focuses on 2-D images, treating the ear as a planar surface that is registered to a gallery using a homography transform calculated from scale-invariant feature-transform feature matches. The feature matches reduce the gallery size and enable a precise ranking using a simple 2-D distance algorithm. Analysis on a range of data sets demonstrates the technique to be robust to background clutter, viewing angles up to $\pm 13°$, and up to 18% occlusion. In addition, recognition remains accurate with masked ear images as small as 20×35 pixels.

Index Terms—Biometrics, computer vision, ear recognition.

I. INTRODUCTION

EACH biological feature has different strengths and weaknesses as a biometric. Fingerprints and irises, for example, are largely unique and, by using controlled sensors, can be measured accurately [1], [2]. Such features, however, require subjects to interact cooperatively with a device. Other features, such as face [3], gait [4], and ears [5], can be measured from a distance. This makes measurement more convenient and allows remote recognition, with the potential for more frequent and covert use, thereby reducing the opportunity for evasion.

Ears are a particularly appealing approach to noncontact biometrics because they are relatively constant over a person's life and are unaffected by expressions, unlike faces. Also, reported levels of recognition are promising [5]. Unfortunately, however, such results are typically achieved in conditions that are significantly more favorable than those found in many recognition-at-a-distance scenarios. The purpose of this paper is to consider how such limitations might be reduced to make progress toward unconstrained ear recognition.

The following are the five main factors that affect accurate ear recognition:

1) *background*: the difficulty of finding the ear in a specific context that may be cluttered by other objects;
2) *occlusion*: the difficulty of finding the ear when partly obscured, for example, by hair, a hat, or earrings;
3) *lighting*: the amount of light on an ear and the direction and color of that light;
4) *pose*: the angle at which the ear is viewed (out-of-plane rotations);
5) *camera*: the particular attributes of the camera, including its field of view, sensing resolution, color sensitivity, and any noise in the image produced.

An overview of existing ear recognition techniques by Hurley and Arbab-Zavar [5] shows that some of the best results use 3-D object matching [6]–[9]. With this approach, ears can be recognized under varying lighting conditions and poses. One limitation of these techniques, however, is that a specialized camera is required to capture the 3-D data. Also, these cameras need controlled lighting to produce accurate results [10]. For use in noncontact scenarios, ear recognition needs to work with much more restrictive data sources, such as surveillance photographs or security cameras. In practice, this means that ears must be recognized from 2-D data sources. This paper therefore focuses on possible approaches to 2-D ear recognition.

The main technical contribution of this paper is to propose an improved ear registration technique based on the object recognition algorithm of Brown and Lowe [12]. Their technique attempts to create a homography transform between a gallery object and a probe image using scale-invariant feature-transform (SIFT) 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 and Lowe demonstrated good results for various objects, but their approach 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 semiautomatically 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, lighting, background clutter, and occlusion. Effectively, the technique is a step toward achieving the accuracy of 3-D ear recognition with unconstrained 2-D data.

This paper describes the proposed technique and its evaluation, with eight data sets being used to assess its robustness and accuracy. Section II gives an overview of ear recognition and discusses existing automated registration algorithms, reviewing their strengths and weaknesses. Section III then describes the stages of the technique, including the semiautomatic creation of gallery masks. The registration calculation and its theoretical justification are also presented, along with an overview of the distance measure for accurate ranking. The proposed technique is evaluated in Section IV. This includes both a traditional controlled-environment recognition test and more challenging data sets that evaluate the technique's robustness to occlusion, background clutter, resolution, and pose variation. This paper concludes with suggestions for future work.

II. RELATED WORK

Ears were first suggested as a means for identification by Bertillon as early as 1890 [13], but it was not until 1955 that Iannarelli developed a practical process for their measurement [14]. This involved gathering and analyzing over 10 000 ear photographs to demonstrate ear uniqueness and viability as a biometric. His technique was first applied to ear prints in 1967, where they were used as a key piece of evidence in a criminal case [15]. Ear-print forensics has continued to be used in prosecutions as recently as 2008. However, at least one conviction has been overturned on appeal due to insufficient ear-print quality [16].

In 1998, Burge and Burger proposed one of the first computerized ear recognition systems [17]. Their technique used an adjacency graph built from Voroni regions of ear-curve segments. Although their paper had no recognition results, it prompted a range of further studies into the effectiveness of ears as a biometric. Force fields [18], neural networks [19], genetic algorithms [20], and a variety of geometric features [21] have all been used to produce rank-1 recognition rates between 90% and 100%. However, in 2002, the relative value of ears was challenged by Victor *et al.* [22]. They analyzed face and ears using principal-component analysis (PCA) and concluded that face produced a consistently higher recognition rate. In a similar study, however, Chang *et al.* [23] achieved different results indicating that there was no significant difference between the features and suggested that an increase in the presence of earrings, occluding hair, and lighting variation may have caused the discrepancy. The work of Chang *et al.* also showed the effect of pose and lighting variation on recognition rate, with pose variation leading to a rank-1 recognition rate below 30% for both face and ears. This result highlights the importance of a careful examination of the constraints imposed on the data sources in any measurement of biometric performance.

The difficulty of addressing variations in pose and lighting has led to studies of the use of range cameras to extract and match 3-D surface shape. This includes a variety of 3-D ear recognition algorithms [6], [7], as well as studies of 2-D, 3-D, face, and ear fusion strategies [8]. All demonstrated improved results with fusion, with their best recognition rates, again, being between 90% and 100%, although, this time, on much more challenging images.

Range cameras, however, do not provide complete lighting and pose invariance because they require controlled conditions to produce accurate results. This has led to a work using "shape-from-motion" and "shape-from-shading" techniques [24] that extract the 3-D information prior to processing with a 3-D recognition algorithm.

Other recent ear publications have explored new 2-D approaches, such as the use of active contours [25], generic Fourier descriptors [26], and active shape models [27]. These techniques have generally been complete recognition systems with an automated ear enrolment procedure. In addition, three techniques have been developed to improve robustness to occlusion: one using SIFT feature-point models [28], another using modular PCA [11], and the third using nonnegative matrix factorization [29]. However, these systems do not address the more challenging variations such as pose or lighting.

Essentially, 2-D ear recognition has three stages: *detection*, *registration*, and *classification*. Here, detection is referred to as 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 [11]. However, there is currently no well-established scheme for automatic 2-D 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.

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

For edge shape matching (usually based on finding the outer ear curve), Ansari and Gupta [30] 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 nonear convex regions. Also, occlusion is likely to invalidate the convex assumption.

Arbab-Zavar and Nixon [31] 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 the one that 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, particularly 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 and Jones [32]. This is a fast and robust technique but suffers from inaccuracy in localization. A refinement, for ear detection, by Abate *et al.* [26] uses the edge center of mass for localization, but this is sensitive to occlusion.

Abdel-Mottaleb and Zhou [33] use Hausdorff edge template matching between an example ear helix edge and edges

identified on skin-colored 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 Jeges and Máté [34]. This uses edge-orientation pattern matching, followed by an active contour. By combining the speed of template matching with the accuracy of active contours, registration can be achieved successfully. This process is robust to significant pose variation, but the pattern-matching localization is sensitive to occlusion, leading to poor active-contour fitting.

The technique described in this paper uses a combination of approaches to achieve robustness. The initial registration process uses SIFT feature-point matching. These features have been shown to be robust under many variations in real-world environments [35]. By using feature points, the registration is inherently robust to occlusion, as any four point matches are sufficient to register the ear. Also, by modeling the ear as a planar surface and registering using a homography transform, the ear can be recognized across variations in pose and camera properties. To rank registered ears, a distance measure is used, which performs both normalization and outlier detection. This makes the ranking step robust to both lighting and occlusion. Finally, the combination of feature matches, homography registration, and image distance is sufficiently discriminating to successfully detect and recognize ears within cluttered environments. This technique 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 analyzed to determine its SIFT feature points. Once this is complete, a probe image can be recognized.

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 detail in the following sections.

A. Building the Gallery Database

Images of the same ear taken at different times can vary significantly due to changes in hair length and color. 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 shown in Fig. 1.

Ideally, these masks would be created automatically, thus enabling the efficient enrolment of subjects from large existing data sources such as criminal mug shots. Unfortunately, without a model of all possible appearances, new ears cannot be automatically detected. However, the number of manually created masks can be greatly reduced by using a bootstrapping process.

Fig. 1.   Gallery-ear image and its associated mask.

Fig. 2.   Set of gallery ears that partially match the seed ear.

Fig. 3.   Masks automatically created from the homography registration of seed to the gallery.

This approach exploits the fact that while different people's ears vary significantly, they often have regions of local similarity (see Fig. 2).

One explanation for this similarity is that ear variations can be modeled as a set of independent smooth local deformations. Some evidence for this hypothesis has been provided by the model-based ear recognition algorithm of Arbab-Zavar *et al.* [28], which describes six growth factors that define an ear's shape. When different ears have local similarity, matches can be made between their SIFT points. If four SIFT matches are detected, the ears can be registered with one another. These registrations can then be used to transfer the masks (see Fig. 3).

These newly masked ears can then be matched against the rest of the unmasked gallery. These ears may have other local regions that are similar, and so, more masks can be transferred. This process is repeated until no further matches can be made. At this point, one of the unmasked ears must be selected and manually processed. This seed can then be used to bootstrap the remaining gallery. This process repeats until all ears have masks. In this way, only a subset of the gallery requires manual masks to be created. In addition, as the gallery size increases, it becomes more likely that ears will form matches and that a smaller percentage of manual masks will be required.

B. Feature Detection

SIFT [36] was used for the detection of features. It is robust to scale, in-plane rotation, and lighting and has some robustness to pose (out-of-plane rotation). The main parameters to the original SIFT algorithm are of the resolution of the Gaussian image pyramid. Where possible, default values were used, with the number of octaves based on the image size, with the lowest

octave being of size 8×8. Each octave had three intermediate Gaussian blurred versions. To ensure robustness to lighting contrast and brightness, the features were normalized.

To make the matching of features against a large gallery more efficient, the approximate nearest neighbor (ANN) algorithm [37] was used. This enables efficient 128-D point matches in $O(\log(n))$, where n is the number of feature points in the gallery. SIFT points were considered a potential match if their squared Euclidean distance was less than 0.45, with a maximum of 1024 matches being returned (closest first).

C. Registration Calculation

By making the simplification that an ear is a planar structure, ears can be registered accurately. If ears are enrolled with the ear plane facing the camera, the image produced can be used to approximate the ear appearance with varying poses. By finding four point matches between an enrolled gallery image and a probe, a homography can be calculated [38]. This homography can be used to transform the gallery image to match the position, rotation, scale, and pose of the probe ear. This transformed image can then be used to accurately compare the two images.

The *homography* is calculated as follows.

Letting $\mathbf{x}$ be a homogeneous point in the probe image and $\mathbf{x}'$ be a homogeneous point in the gallery image, 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

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

By considering $\mathbf{H}$ as a matrix of row vectors $\mathbf{h}^{j\mathrm{T}}$

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

the cross product can be expanded to give

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

Because $\mathbf{h}^{j\mathrm{T}}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{h}^{j}$, this can be rewritten as

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

This is a linear equation in $\mathbf{h}$ of the form $\mathbf{A}\mathbf{h} = \mathbf{0}$, where $\mathbf{A}$ is a 3×9 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 row. By omitting this equation, the remaining set becomes

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

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 size 8×9 and rank 8. This matrix has a 1-D null space, which can be solved to produce a solution to $\mathbf{H}$ up to a nonzero scale. As these points are homogeneous, if the transformed points are normalized by dividing through by their third component, this scale factor will be removed.

The SIFT-matching distance is quite generous to enable large variations in pose and lighting. However, this will result in a significant number of false positives in the point correspondences. To reduce such errors, an affine consistency constraint was applied. This constraint groups the SIFT matches into sets of points that have an approximately equal in-plane affine transform. This constraint is reasonable under small pose variations where the homography will be close to affine.

As part of the SIFT detection process, there is a search for interest points across locations and scales. When an interest point is detected, the region surrounding it is used to calculate a canonical orientation. By comparing these values between the probe and the gallery, each point can be used to calculate an approximate affine transform between the two images. 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 that the matching is robust to pose variation. Each point match is placed in the appropriate bin and its closest neighbor (16 bin entries per point). If any bin contains four or more point matches, its points are passed to the next stage.

This process greatly reduces false positives, but some invalid point matches remain. To address this problem, a RANdom SAmple Consensus (RANSAC) algorithm was used. Random sets of four points were selected from the list of point correspondences and a homography calculated. The homography that matches the most points within some threshold, i.e., in this case, 1% of the ear mask size, was selected as the best match.

Gallery images that have four affine matching feature points are then passed to the distance measure. The combination of SIFT-matching and affine constraints greatly reduces the set of potential gallery matches. This process is sufficiently accurate to prevent false matches, both with the majority of incorrect ears and with background clutter.

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 errors after normalization. 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, thus, excluded.

Normalization 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

$$G(I, x, y)$$
$$= (r(I(x,y)) + g(I(x,y)) + b(I(x,y)))/3$$
$$\forall x \in [1, \ldots, w]$$

$$mean(G, I)$$
$$= \left(\sum_{y=1}^{h} \sum_{x=1}^{w} G(I, x, y) \right) \bigg/ (wh)$$

$$scale(G, I)$$
$$= \sqrt{\sum_{y=1}^{h} \sum_{x=1}^{w} (G(I, x, y))^2 - mean(G, I)^2 \bigg/ (wh)}$$

$$N(I, x, y)$$
$$= (G(I, x, y) - mean(G, I)/scale(G, I))$$
$$\forall x \in [1, \ldots, w] \quad \forall y \in [1, \ldots, h]$$

$$notoutlier(I_1, I_2, x, y)$$
$$= \left\{ \begin{array}{l} 0 \; \|G(I_1, x, y) - G(I_2, x, y)\| >= 0.5 \\ 1 \; \|G(I_1, x, y) - G(I_2, x, y)\| < 0.5 \end{array} \right\}$$

$$ndistance(I_1, I_2)$$
$$= \sum_{y=1}^{h} \sum_{x=1}^{w} notoutlier(x, y) . (N(I_1, x, y) - N(I_2, x, y))^2$$

where G is a function that returns the grayscale values of an image, N is a function that returns the normalized values of an image, and w and h are the width and height of those images, respectively. In addition, $r()$, $g()$, and $b()$ are functions that return the magnitude of the red, green, and blue components, respectively.

IV. EVALUATION

Eight data sets were used for evaluation. The first provided a straight test of recognition accuracy on a relatively constrained data set. For this, a subset of the XM2VTS [39] face-profile database was chosen. It consists of 63 subjects with relatively unoccluded ears. This is the same data set used by Hurley *et al.* [18] and Arbab-Zavar *et al.* [28].

The second data set was created by recording those ears of 20 subjects from a range of angles to test the technique's robustness to pose variation. The remaining data sets were synthesized from these XM2VTS images to test the effects of

TABLE I
RECOGNITION RATES FOR DIFFERENT REGISTRATION TECHNIQUES

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

TABLE II
NUMBER OF FEATURES AT EACH STAGE XM2VTS DATA SET

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 affine constrained homographies	4

occlusion, background clutter, resolution, noise, contrast, and brightness.

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 and Bowyer [6]. This involved defining the triangular fossa and incisura intertragica of each ear manually. These landmarks were then used to standardize the scale and rotation of all gallery and pose images. The resulting normalized images were segmented with a rectangular mask in the center of the image capturing the inner ear features.

The second technique applied the algorithm described by Arbab-Zavar and Nixon [31] to register the ear automatically, using the outer ear ellipse. In both cases, the intensity values had their mean and standard deviation normalized. These registered images were ranked by using the PCA technique giving the results shown in Table I.

Each technique used the "leave-one-out" strategy, with each image being removed from the gallery and being tested against the rest of the data set in turn.

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 recognized, 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.

B. Robustness Evaluation

Gallery: The clutter data set was created by randomly placing XM2VTS masked ear images on a set of complex
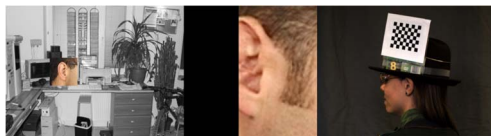
Fig. 4. Examples of more challenging probe images. (From left to right) Background clutter, occlusion, and pose variation.

TABLE III
AVERAGE RECOGNITION RATES FOR CONTROLLED
AND CLUTTERED ENVIRONMENTS

Technique	% Rank 1 Recognition	Examples
Base recognition rate	96%	
Background clutter	93%	

TABLE IV
AVERAGE RECOGNITION RATES FOR POSE, OCCLUSION, RESOLUTION,
NOISE, CONTRAST, AND BRIGHTNESS

Technique	% Rank 1 Recognition	Examples
0 degrees pose variation	100%	
13 degrees pose variation	100%	
22 degrees pose variation	33%	
30% occlusion from above	92%	
40% occlusion from above	74%	
18% occlusion from the side	92%	
37% occlusion from the side	66%	
50% of original resolution	93%	
40% of original resolution	87%	
30% of original resolution	66%	
Gaussian noise with standard deviation at 10% maximum channel magnitude	93%	
Gaussian noise with standard deviation at 20% maximum channel magnitude	75%	

background images. These images more closely represent the type of unconstrained environment that was present with covert biometrics. The occlusion data set 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. To determine the percentage of occlusion that each rectangle represents, the occlusion of each mask was calculated and then averaged across the gallery. The resolution data set was created by linearly downsampling and then bicubically upsampling the probe images. The contrast data set was constructed by subtracting the mean pixel color, scaling the result, and then adding back the mean. Similarly, the brightness data set added an offset to each pixel's channel. Finally, to generate the pose data set, 20 subjects were recorded by letting them turn in front of a camera. Both sides of the head were recorded to obtain 40 unique ears. For the purposes of evaluation, each ear was treated as an independent subject. 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 [40] libraries. Fig. 4 shows examples from some of these data sets.

Results: Tables III and IV summarize the results of these recognition tests.

Background clutter, as well as up to 30% occlusion from above and 18% occlusion from the side, was found to have little effect on the recognition rate. 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. With resolution changes, images remained recognizable at 50% of their original size (i.e., when reduced from 40×70 to 20×35 pixels, depending on the mask size). The contrast results show that the approach maintains 90% recognition accuracy with 80% of the contrast. The approach is sensitive to brightness, however, with a 20% increase almost

TABLE IV
(Continued.) AVERAGE RECOGNITION RATES FOR POSE, OCCLUSION, RESOLUTION, NOISE, CONTRAST, AND BRIGHTNESS

|		
Gaussian noise with standard deviation at 30% maximum channel magnitude	54%	
10% reduction in contrast	94%	
20% reduction in contrast	90%	
30% reduction in contrast	81%	
10% maximum channel magnitude increase in brightness	92%	
20% maximum channel magnitude increase in brightness	58%	
30% maximum channel magnitude increase in brightness	31%	

Fig. 5. Recognition rate with varying pose, with and without synthesized ear images.

halving the recognition rate. In both cases, recognition failures are primarily due to failing to find SIFT matches.

Fig. 5 shows the average recognition rate for 40 ears with varying pose. The technique maintains 100% recognition rate

TABLE V
COMPARISON OF RANK-1 RECOGNITION RATES WITH THOSE OF OTHER PUBLISHED APPROACHES

Technique	No. of Ears	Recognition Rates under different conditions		
		Optimum	Pose	Occluded
SIFT Homography (this paper)	63	*96%*	-	*Top: 92% Side: 92%*
Theoharis [9]	830	95%	-	-
Cadavid [24]	25	84%	-	-
Hurley [18]	63	99.2%	-	-
Arbab-Zavar [28]	63	92%	-	Top: 80% Side: 88%
Yuan [29]	24	91%	-	Top: 85%
SIFT Homography (this paper)	40	*100%*	*13 Degrees 100% 22 Degrees 33%*	-
Chang [23]	197	72%	23 Degrees 22%	-
Lu [27]	56	93.3%	5 Degrees 93.3%	-

up to $\pm 13°$. However, this performance is dependent on enrolled gallery ears being recorded with the ear plane facing the camera. If gallery ears are protruding, e.g., if they are recorded with the ear plane tilted by $30°$ from the camera, the pose invariance is reduced to $\pm 10°$. It should also be noted that the pose robustness is approximately equal for both forward and backward yaw rotations of the head.

As an experiment to improve this technique's robustness to pose variation, additional gallery images were synthesized 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 rerendered to simulate different poses. This increased not only the number of SIFT matches but also the number of false positives. As the ears are not completely planar, the image distance increases with the angle, resulting in incorrect ears having a shorter image distance, and so, no significant increase in robustness was observed.

Table V compares the rank-1 recognition rates of the SIFT-based approach described in this paper with that of other published approaches. Direct comparison is not possible in any of the cases described because of the use of different data sets and variations in the way that evaluation is performed. In particular, occlusion is calculated differently in a SIFT-based approach from that used in nonnegative matrix factorization work. Thus, for each study in the table, the closest corresponding occlusion measure has been estimated. The recognition rates are calculated using the largest occlusion factors for which the homography approach remains accurate, namely, 30% from above and 18% from the side.

Generally, the existing work has concentrated on demonstrating high accuracy on controlled data sets; therefore, many of the comparison techniques do not evaluate the robustness to quality factors, such as noise and resolution. In addition, all of the approaches constrain the probe image to a single head profile, removing the challenging problem of background clutter.
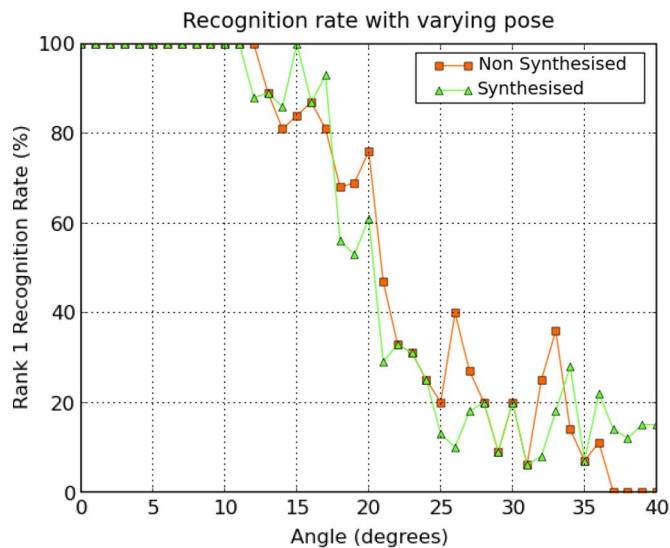
The first comparison technique in Table V, developed by Theoharis *et al.* [9], uses range data and achieves accurate results on a large data set. By using 3-D data, the technique is likely to have some pose and lighting invariance, but this analysis is not presented. The main disadvantages are the need for manual ear detection and the dependence on a specialized range camera. Other techniques based on 3-D data achieve similar recognition rates and have similar restrictions [6], [7].

The small study by Cadavid and Abdel-Mottaleb [24] has a relatively low recognition rate. This may be due to the inherent sensitivity of the technique, particularly when applied to low resolution data, or it may reflect the more challenging nature of their data set.

The approach developed by Hurley *et al.* [18] uses the force-field transform and has one of the highest published recognition rates. It uses the same 63-subject data set as that in this paper. However, as the transform effectively performs a large blur operation on the image, it is likely to be sensitive to clutter and occlusion.

The work of Arbab-Zavar *et al.* [28] uses a SIFT-based model and is evaluated using the same data set as that in this paper. The recognition rates in the table are based on manual registration. When automatic registration is used, recognition rates fall from 92% to 87%. Their approach also has some occlusion robustness, but it is less than that achieved by the homography technique described in this paper.

The work of Yuan *et al.* [29] has an 85% rank-1 recognition rate with 30% occlusion, which is less than that of the homography approach. However, this may be due to the less pose-constrained nature of their data set.

The study by Chang *et al.* [23] uses PCA to compare ear images using the eigenear approach. The published results of this technique are much lower than those of other approaches. However, this may be affected by the precision of their manual registration.

In contrast, the shape model produced by Lu *et al.* [27] achieves high recognition rates with a $5°$ pose variation. However, like many of these initial approaches, the technique is likely to be sensitive to occlusion.

In summary, the existing work demonstrates many viable alternative approaches to ear recognition. However, they currently lack the robustness and automated detection that are necessary to be used as passive recognition systems. This paper describes a complete and accurate technique that is a step toward achieving such a system. As demonstrated in Tables III and IV, this approach is more robust to a much wider range of variations than any existing approach.

V. Further Work

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 [41].

Another area for improvement is the computation time of the algorithm. Despite the use of the ANN library, the processing of each 720×576 probe image takes over 4 min on a 2.4-GHz PC. The majority of this time is spent in calculating and measuring the image distance and the RANSAC homographies. Each image requires over $10\,000$ of these calculations on average. In total, this takes over 3.5 min. The next most expensive stage is the SIFT-matching process, which takes around 1 min. The remaining calculations, such as the detection of SIFT points in the probe, take seconds and have a relatively small impact on performance. Further work will explore the improvement of these times through a generic ear model, such as the Viola–Jones classifier [32] 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 data sets.

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 [42].

To fully automate the enrolment process, there is a need to construct a model of ear variation so that novel ears can be detected. The current system requires precise matches between feature points and is therefore limited in its capacity for generalization. An alternative approach would be to train a classifier for a set of ear feature points that are common across all ears. This is similar to the approach used in face recognition to detect features such as the corners of eyes and lips. The difference measure could also be generalized by constructing an active appearance model for the ear. By matching against a single model rather than every ear in the gallery, the recognition system would not suffer from the same performance issues for large data sets. Also, by adding a final validation step using SIFT points and the robust distance measure, there is a potential to achieve the same accuracy and robustness of recognition with improved performance and fully automatic enrolment.

VI. Conclusion

This paper has described a new technique for ear recognition in 2-D images using homographies calculated from SIFT point matches. When applied to the XM2VTS database, the technique has given results that are comparable to that of PCA with manual registration. In addition, when used on more challenging data sets, it shows robustness to background clutter, 18% occlusion, and over $\pm13°$ of pose variation. Further work will focus on performance improvement and increased robustness.

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

References

[1] J. R. Matey, D. Ackerman, J. Bergen, and M. Tinker, "Iris recognition in less constrained environments," in *Advances in Biometrics*. Berlin, Germany: Springer-Verlag, Oct. 2007, pp. 107–131.

[2] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. New York: Springer-Verlag, Jun. 2003.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[4] M. S. Nixon and J. N. Carter, "Advances in automatic gait recognition," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recog.*, 2004, pp. 139–144.

[5] D. J. Hurley and B. Arbab-Zavar, "The ear as a biometric," in *Handbook of Biometrics*. Berlin, Germany: Springer-Verlag, Oct. 2007, pp. 131–150.

[6] P. Yan and K. W. Bowyer, "Biometric recognition using three-dimensional ear shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1297–1308, Aug. 2007.

[7] H. Chen and B. Bhanu, "Human ear recognition in 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 718–737, Apr. 2007.

[8] P. Yan and K. W. Bowyer, "Empirical evaluation of advanced ear biometrics," in *Proc. IEEE Workshop Empirical Eval. Methods Comput. Vis.*, Jun. 2005, vol. 3, p. 41.

[9] T. Theoharis, G. Passalis, G. Toderici, and I. A. Kakadiaris, "Unified 3D face and ear recognition using wavelets on geometry images," *Pattern Recognit.*, vol. 41, no. 3, pp. 796–804, Mar. 2008.

[10] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Comput. Vis. Image Underst.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[11] L. Nanni and A. Lumini, "A multi-matcher for ear authentication," *Pattern Recognit. Lett.*, vol. 28, no. 16, pp. 2219–2226, Dec. 2007.

[12] M. Brown and D. G. Lowe, "Invariant features from interest point groups," in *Proc. 13th Brit. Mach. Vis. Conf.*, 2002, pp. 253–262.

[13] A. Bertillon, *La Photographie Judiciaire, Avec un Appendice sur la Classification et l'identification Anthropometriques*. Paris, France: Gauthier-Villars, 1890.

[14] A. Iannarelli, *Ear Identification*. Amarillo, TX: Paramount, 1989.

[15] J. W. Osterburgh, *Crime Laboratory*. Amarillo, TX: Paramount, 1968.

[16] "State v. David Wayne Kunze," *Court of Appeals of Washington Division 2*, 1999.

[17] M. Burge and W. Burger, "Ear biometrics," in *Biometrics: Personal Identification in Networked Society*. Berlin, Germany: Springer-Verlag, 1998, pp. 271–286.

[18] D. J. Hurley, M. S. Nixon, and J. N. Carter, "Force field feature extraction for ear biometrics," *J. Comput. Vis. Image Underst.*, vol. 98, no. 3, pp. 491–512, Jun. 2005.

[19] B. Moreno and A. Sanchez, "On the use of outer ear images for personal identification in security applications," in *Proc. IEEE 33rd Annu. Int. Conf. Security Technol.*, 1999, pp. 496–476.

[20] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama, "Study on individual recognition for ear images by using genetic local search," in *Proc. Congr. Evol. Comput.*, 2002, vol. 1, pp. 237–242.

[21] M. Choras, "Ear biometrics based on geometrical feature extraction," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 5, no. 3, pp. 84–95, 2005.

[22] B. Victor, K. W. Bowyer, and S. Sarkar, "An evaluation of face and ear biometrics," in *Proc. Int. Conf. Pattern Recog.*, 2002, pp. 429–432.

[23] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1160–1165, Sep. 2003.

[24] S. Cadavid and M. Abdel-Mottaleb, "Human identification based on 3D ear models," in *Proc. IEEE Conf. Biometrics, Theory, Appl., Syst.*, 2007, pp. 1–6.

[25] L. Máté, "Localizing feature points on ear images," in *Proc. Joint Hungarian-Austrian Conf. Image Process. Pattern Recog.*, 2005, pp. 57–63.

[26] A. F. Abate, M. Nappi, D. Riccio, and S. Ricciardi, "Ear recognition by means of a rotation invariant descriptor," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 437–440.

[27] L. Lu, X. Zhang, Y. Zhao, and Y. Jia, "Ear recognition based on statistical shape model," in *Proc. 1st Int. Conf. Innovative Comput., Inf. Control*, 2006, vol. 3, pp. 353–356.

[28] B. Arbab-Zavar, M. S. Nixon, and J. N. Carter, "On model-based analysis of ear biometrics," in *Proc. IEEE Conf. Biometrics, Theory, Appl., Syst.*, 2007, pp. 1–5.

[29] L. Yuan, Z. Mu, Y. Zhang, and K. Liu, "Ear recognition using improved non-negative matrix factorization," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 501–504.

[30] S. Ansari and P. Gupta, "Localization of ear using outer helix curve of the ear," in *Proc. ICCTA*, Mar. 2007, pp. 688–692.

[31] B. Arbab-Zavar and M. S. Nixon, "On shape-mediated enrolment in ear biometrics," in *Advances in Visual Computing*. Berlin, Germany: Springer-Verlag, 2007, pp. 549–558.

[32] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[33] M. Abdel-Mottaleb and J. Zhou, "Human ear recognition from face profile images," in *Advances in Biometrics*. Berlin, Germany: Springer-Verlag, 2005, pp. 786–792.

[34] E. Jeges and L. Máté, "Model-based human ear localization and feature extraction," *Int. J. Intell. Comput. Med. Sci. Image Process.*, vol. 1, pp. 101–112, 2007.

[35] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006.

[36] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[37] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, Nov. 1998.

[38] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[39] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Audio- Video-Based Biometric Person Authentication*, Washington, DC, 1999, pp. 72–77.

[40] *OpenCV*. [Online]. Available: www.intel.com/technology/computing/opencv/index.htm

[41] B. Weyrauch, J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3D morphable models," in *Proc. IEEE Workshop Face Process. Video*, 2004, pp. 1–5.

[42] M. P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th Int. Conf. Pattern Recog.*, 1994, vol. 1, pp. 566–568.

John D. Bustard (S'08) received the B.A. degree in computer science from the University of Cambridge, Cambridge, U.K., in 2000. He has been working since 2007 toward the Ph.D. degree in unconstrained ear recognition (the topic of this paper) in the School of Electronics and Computer Science, University of Southampton, Southampton, U.K.

From 2000–2007, he was in the computer games industry, working on Microsoft Xbox and Sony PlayStation platforms, in areas ranging from the design and pitching of next-generation game concepts to the development of core components of commercial game systems. These included such areas as physics and collision systems, artificial intelligence, graphical effects, user interface development, and character control. His research interests include the robust recognition of objects and the reconstruction of shape from images.

Mark S. Nixon (A'05) received the Ph.D. degree in 1983.

He is currently a Professor in computer vision with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. His team were early workers in face recognition, who later came to pioneer gait recognition and more recently joined the pioneers of ear biometrics. His team develops new techniques for static- and moving-shape extraction, which have found applications in automatic face and gait recognition and in medical image analysis. Among research contracts, he was the Principal Investigator, together with John Carter, of the Defense Advanced Research Projects Agency-supported project Automatic Gait Recognition for Human ID at a Distance. His vision book, cowritten with Alberto Aguado, is entitled *Feature Extraction and Image Processing* (Butterworth, 2002), and his new book, cowritten with Tieniu Tan and Rama Chellappa, is entitled *Human ID Based on Gait* (Springer, 2005), which is part of the new International Series on Biometrics. His research interests are in image processing and computer vision.

Dr. Nixon chaired the 1998 British Machine Vision Conference, cochaired (with Josef Kittler) the 2003 Audio- and Video-Based Biometric Person Authentication, and was the Publications Chair for the 2004 International Conference on Pattern Recognition and the Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG2006).