Chapter 6
Unconstrained Ear Processing: What is Possible and What Must Be Done

Silvio Barra, Maria De Marsico, Michele Nappi and Daniel Riccio

Abstract Ear biometrics, compared with other physical traits, presents both advantages and limits. First of all, the small surface and the quite simple structure play a controversial role. On the positive side, they allow faster processing than, say, face recognition, as well as less complex recognition strategies than, say, fingerprints. On the negative side, the small ear area itself makes recognition systems especially sensitive to occlusions. Moreover, the prominent 3D structure of distinctive elements like the pinna and the lobe makes the same systems sensible to changes in illumination and viewpoint. Overall, the best accuracy results are still achieved in conditions that are significantly more favorable than those found in typical (really) uncontrolled settings. This makes the use of this biometrics in real world applications still difficult to propose, since a commercial use requires a much higher robustness. Notwithstanding the mentioned limits, ear is still an attractive topic for biometrics research, due to other positive aspects. In particular, it is quite easy to acquire ear images remotely, and these anatomic features are also relatively stable in size and structure along time. Of course, as any other biometric trait, they also call for some template updating. This is mainly due to age, but not in the commonly assumed way. The apparent bigger size of elders' ears with respect to those of younger subjects, is due to the fact that aging causes a relaxation of the skin and of some muscle-fibrous

S. Barra
University of Cagliari, Via Ospedale, 72 - 09124 Cagliari, Italy
e-mail: silvio.barra@unica.it

M. De Marsico (✉)
Sapienza University of Rome, Via Salaria, 113 - 00198 Rome, Italy
e-mail: demarsico@di.uniroma1.it

M. Nappi
University of Salerno, Via Ponte Don Melillo, 84084 Fisciano, SA
e-mail: mnappi@unisa.it

D. Riccio
University of Napoli Federico II, Via Cintia 21, 80126 Napoli, Italy
e-mail: daniel.riccio@unina.it

J. Scharcanski et al. (eds.), *Signal and Image Processing for Biometrics*,
Lecture Notes in Electrical Engineering 292, DOI: 10.1007/978-3-642-54080-6_6,

structures that hold the so called pinna, i.e. the most evident anatomical element of the ear. This creates the belief that ears continue growing all life long. On the other hand, a similar process holds for the nose, for which the relaxation of the cartilage tissue tends to cause a curvature downwards. In this chapter we will present a survey of present techniques for ear recognition, from geometrical to 2D-3D multimodal, and will attempt a reasonable hypothesis about the future ability of ear biometrics to fulfill the requirements of less controlled/covert data acquisition frameworks.

6.1 Introduction

For thousands years humans have used physical traits (face, voice, gait, etc.) as a primary resource to recognize each other. As a matter of fact, the earliest documented example of related strategies goes back to ancient Egyptians, who used height to identify a person during salary payment. Recognizing an individual through unique biometric traits can be used in many different applications. Authentication is a classical example. Relying on something that the subject must possess (cards, keys, etc.) is subject to loss or theft, while relying on something that the subject must remember (passwords, pins, etc.) is subject to forgetfulness or sniffing. Therefore, it is sometimes impossible to surely distinguish the real subject from an impostor, especially in a remote session. Biometric traits are more difficult to steal or reproduce, so that their integration in the authentication processes can reinforce it. Moreover, biometrics can be exploited in further settings such as forensics, video-surveillance or ambient intelligence. Ear biometrics is not among the most widely explored at present, though it gained some credit for its reliability. The aim of this essay is to explore the available approaches to extract a robust ear template, providing an accurate recognition, and to discuss how these techniques can be suited to uncontrolled settings, which are typical of real world applications. Even if we will try to describe the overall scenario of the main techniques, an extensive and complete presentation of all methods presented in literature is out of our scope. For this, we suggest the comprehensive works by Pflug and Busch [74] and by Abaza et al. [5].

As an introductory consideration, it is worth noticing that quite few people are aware of the uniqueness of the individual ear features. Actually, the external ear "flap", technically defined as pinna, presents several morphological components. These make up a structure that, though relatively simple, varies significantly across different individuals, as shown in Fig. 6.1.

The chapter will proceed as follows. The present section will introduce ear anatomy and some preliminary considerations on its use as a biometric trait in automatic recognition systems. Section 6.2 will present the most popular preprocessing techniques for ear recognition, from acquisition to detection and feature extraction, and beyond. Section 6.3 will summarize techniques for ear recognition from 2D images, while Sect. 6.4 will do the same for ear recognition from 3D models. Section 6.5 will present some relevant multimodal biometric approaches involving ear, and finally Sect. 6.6 will draw some conclusions.

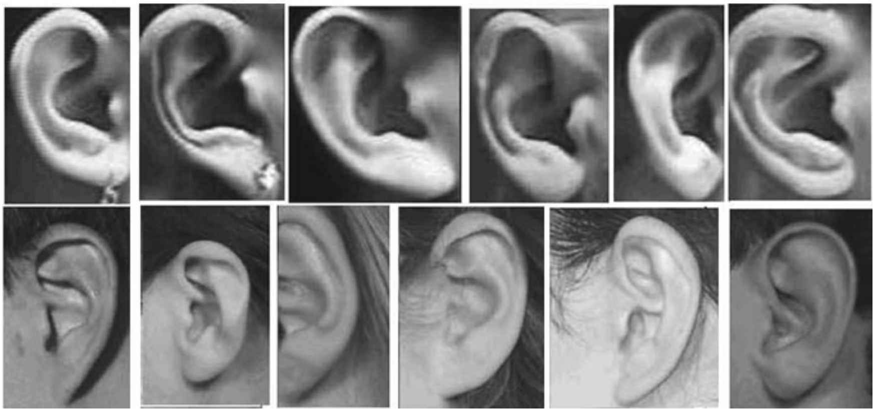

Fig. 6.1 Differences in the morphology of the pinna in different individuals

6.1.1 Ear Anatomy

Recognition systems based on ear analysis deal with the variations that characterize its structure. Among the fist researchers to understand and investigate the potential of ear recognition, we can mention the Czech doctor R. Imhofer, who in 1906 noted that he was able to distinguish among a set of 500 ears using only four features. Further studies followed quite later. First of all, it is worth mentioning the pioneering work by Iannarelli in 1949 [48]. The developed anthropometric recognition technique is based on a series of 12 measurements and on the information concerning sex and race. Obviously, among the anatomical structures that identify the so-defined external, middle and inner ear, only the visible part of the first ones, namely those composing the pinna, are relevant for recognition (Fig. 6.2). Actually, humans are not able to recognize a person from the ears. However, this might be merely due to the lack of training, since Iannarelli's studies and a number of following ones have demonstrated that they have a good discriminating power. A further outcome of these studies is that ear variations are very limited along time. Growth is proportional from birth to the first 4 months, then the lobe undergoes a vertical elongation from 4 months to the age of 8 years due to gravity force. The further growth in size, until about the age of 20, does not alter the shape, which remains constant until about the age of 60/70, and finally undergoes a further elongation. According to a widely diffused belief, ears continue growing all life long, therefore elders' ears are usually bigger than those of younger subjects. However, this is rather due to the fact that aging causes a relaxation of the skin and of some muscle-fibrous structures, in particular those that hold the so called pinna, i.e. the most evident anatomical element of the external part of the ear.

An interesting remark regards soft biometrics used for recognition, which are somehow related to ear. Akkermans et al. [7] exploit a particular acoustic property of the ear: due to the characteristic shape of the pinna and ear canal, when a sound signal

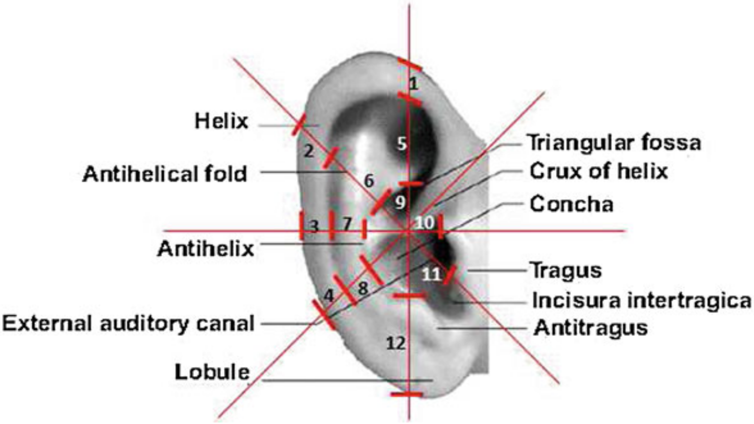

Fig. 6.2 Anatomy of the pinna of external ear and Iannarelli's measures

is played into it, it is reflected back in a modified form. This can be considered to produce a personal signature, which can be generated, for example, by a microphone incorporated into the head-phones which catches the reflected sound signal for a test sound produced by the speaker of a mobile phone.

Grabham et al. [40] present a thorough assessment of otoacustic emissions as a biometric trait. Otoacustic emissions are very low level sounds that the human ear normally emits. They can be classified in those spontaneously occurring (transient evoked otoacoustic emissions—TEOAE), and those triggered by stimulus (distortion product otoacoustic emission—DPOAE). Although they only allow to recognize people wearing an headset, they can be used for continuous re-identification during interaction with a protected system. The authors demonstrate that stability is generally limited to a period of six months. For this reason, a suitable setting may require an initial identification of the user through a stronger biometrics at the beginning of each session, and the continuous re-identification through this light-weight trait. Since parameters such as OAE type, equipment type and use, and background noise, may significantly affect recognition, this biometrics requires further exploration.

6.1.2 Positive and Negative Aspects

It is important to summarize pros and cons of ear recognition according to the basic requirements for a biometric feature [50]:

- *universality*: every subject must present that feature;
- *uniqueness*: no pair of subjects share the same value for that feature;
- *permanence*: the value of the that feature must not change in time;
- *collectability*: the value of that feature must be quantitatively measurable;

More parameters regard performance in terms of accuracy and computational requirements, acceptability in terms of subjects' level of acceptance, and circumvention in terms of robustness to cheat.

Derived from the above elements, we can first list positive aspects of ear recognition:

- universality: as a part of the body, every subject normally has them;
- uniqueness: they are sufficient to discriminate among individuals;
- efficacy/efficiency: good accuracy/low computational effort due to the small area;
- stability:
  - low variations except for lobe length;
  - uniform distribution of colors;
  - absence of expression variations;
- conctactless acquisition: the image can be acquired at a distance;
- cheap acquisition: low cost of capture devices (CCD camera);
- easy and quick measurability.

The main negative aspects are:

- the small area is more sensible to occlusions by hair, earrings, caps;
- ear image can be confused by background clutter;
- the prevalent 3D nature of ear features makes captured images very sensible to pose and acquisition angle;
- acquisition at a distance introduces heavy limits due to the combination of small area and possibly low resolution.

As a consequence, the best results are presently obtained thanks to a strict profile pose, absence of hair or other occlusion and relatively short distance; in other words, user cooperation is required.

6.1.3 A Summary of Different Techniques Applied to Ear Biometrics

A first rough classification of human ear identification techniques using outer anatomical elements distinguishes two approaches: the first one exploits detection and analysis of either feature points or relevant morphological elements, and derives geometric measures from them; the second one relies on global processing [68].

Taking into further consideration the dimensionality of captured data and the kind of feature extraction, we can also identify the following categories:

1. Recognition from 2D images
   - Geometrical approaches (interest points)
   - Global approaches
   - Multiscale/multiview approaches
   - Thermograms

2. Recognition from 3D models
3. Multimodal systems

We will briefly sketch some examples here and return on more details in the following chapters.

Recognition from 2D images

In geometrical approaches, a set of measures are computed over a set of interest points and/or interest contours, identified on a 2D normalized photo of the ear. The pioneering work related to ear recognition, namely "Iannarelli system" of ear identification, is a noticeable example along this line. In Iannarelli's system, the ear image is normalized with respect to dimensions. The point named *Crux of Helix* is identified (see Fig. 6.2) and becomes the center of a relative space. All measures are relative to it, so that a wrong identification of such point compromises the whole measurement and matching processes.

Burge and Burger [18, 19] are among the first to try more advanced techniques. They use Voronoi diagrams to describe the curve segments that surround the ear image, and represent the ear by an adjacency graph whose distances are used to measure the corresponding features (Fig. 6.3). The method has not been extensively tested, however we can assume that Voronoi representation suffers from the extreme sensitiveness of ear segmentation to pose and illumination variations.

Global approaches consider the whole ear, instead of possible relevant points.

Hurley et al. [45, 46] address the problem of ear recognition through the simulation of natural electromagnetic force fields. Each pixel in the image is treated as a Gaussian attractor, and is the source of a spherically symmetric force field, acting upon all the other pixels in a way which is directly proportional to pixel intensity and inversely proportional to the square of distance. The ear image is so transformed into a force field. This is equivalent to submit it to a low pass filter transforming it into a smooth surface, where all information is still maintained. The directional properties of the force field can support the location of local energy peaks and ridges to be used to compose the final features. The technique does not require any explicit preliminary description of ear topology, and the ear template is created just following the force field lines. It is invariant to the translation of the initializing position and to scaling, as well as to some noise. However, the assessment of the recognition accuracy has been attempted only with images presenting variations on the vertical plane and without hair occlusion [47].

Since face and ear present some common feature (e.g., sensitiveness to pose and illumination, sensitiveness to occlusions) it is natural to inherit face related techniques. In particular, Principal Component Analysis (PCA) has been widely used for face template coding and recognition. In [26], PCA is applied to both biometrics for comparison purposes, demonstrating similar performance as well as limitations. Despite the apparently easiest task, even for ear recognition images must be accurately registered, and extraneous information must be discarded by a close cropping. Last but not least, this method suffers from very poor invariance to those factors which also affect face recognition, in particular pose and illumination.

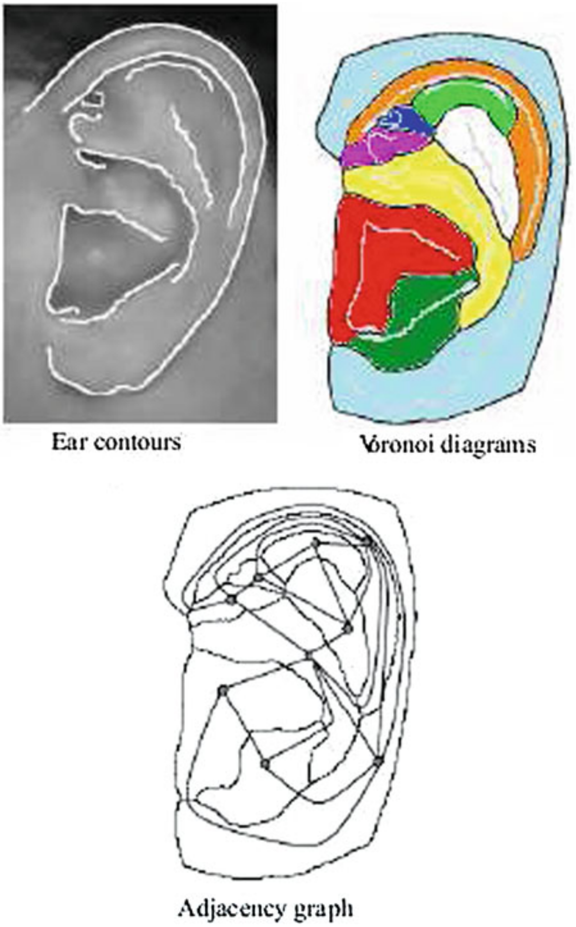

Fig. 6.3 Relevant edges on ear external surface and derived Voronoi diagram and adjacency graph

Further global approaches, still borrowing from face recognition, are presented in [68, 84]. The former applies neural network strategy to ear recognition, testing Compression Networks, Borda Combination, Bayesian and Weighted Bayesian Combinations techniques. The latter exploits Haar wavelets transformation.

In multiscale/multiview approaches, the set of features used for recognition is enriched by considering more scales for the same image, or more acquisitions, possibly from slightly different points of view, for the same trait of the same subject. While multiview approach is often used to obtain a 3D model of the anatomical element, it can be also used in 2D techniques.

Though acquired by different equipment and containing information different from pixel intensities, also thermograms are 2D images. Through a thermographic camera, a thermogram image captures the surface heat (i.e., infrared light) emitted by the subject. These images are not sufficiently detailed to allow recognition, but can rather be used for ear detection and segmentation, especially in those cases where the ear is partially covered and passive identification is involved (the user does not cooperate and might be unaware of the acquisition). In such cases, texture and color segmentation should allow to discard hair region. As an alternative, Burge and Burger [19] propose to use thermogram images. The pinna usually presents an higher temperature than hair, so that the latter can be segmented out. Moreover, if the ear is visible, the Meatus (i.e., the passage leading into the inner ear) is the hottest part of the image, which is clearly visible and allows to easily detect and localize the rest of the ear region. Disadvantages include sensitiveness to movement, low resolution and high costs.

Recognition from 3D models

Like other techniques, 3D processing for ear has followed the success of three-dimensional techniques applied to face. As a matter of fact, they solve similar problems of sensitiveness of 2D intensity images to pose and illumination variations, including shadows which may sometime play a role similar to occlusions. On the other hand, the outer part of the ear presents even richer and deeper 3D structure than face, with a very similar discriminating power, which can be profitably modeled and used for recognition purposes. Among the first and most significant works along this direction, we mention the one by Chen and Bhanu [16, 29], and the one by Yan and Bowyer [108, 111]. Actually, both research lines have an articulated development in time, and we will only mention the main achievements. Both use acquisitions by a range scanner. In the first approach, ear detection exploits template matching of edge clusters against an ear model; the model is based on the helix and antihelix, which are quite extended and well identifiable anatomical elements; a number of feature points are extracted based on local surface shape, and a signature called Local Surface Patch (LSP) is computed for each of them. This signature is based on local curvature, and is used together with helix/antihelix to compute the initial translation/rotation between the probe and the gallery model. Refined transformation and recognition exploit Iterated Closest Point (ICP). ICP is quite simple and accurate, therefore it is widely used for 3D shape matching. The reverse of the medal is its high computational cost. The authors also test on 3D models of the ear a more general approach presented in [30], which integrates rank learning by SVM for efficient recognition of highly similar 3D objects.

In the second mentioned approach, an efficient ICP registration method exploits enrollment data, assuming that biometric applications include a registration phase of subjects before they can be recognized. Moreover, ear extraction exploits both 2D appearance and 3D depth data. A detailed comparison of the two methods can be found in [29]. It is worth mentioning that a number of attempts are made to reduce the computational cost of ICP. As an example, in [112] Yan and Bowyer use a k-d tree structure for points in 3D space, decompose the ear model into voxels, and extract surface features from each of these voxels. In order to speed up the alignment process, each voxel is assigned an appropriate index so that ICP only needs to align voxel pairs with the same index.

Multimodal systems

Multimodal systems including ear biometrics range from those that jointly exploit 2D images and 3D models, like those tested in [110], to those that exploit ear with a variety of different biometrics. Due to physical proximity allowing an easy combined capture, ear is very often used together with face. Results presented in all works on multibiometric fusion demonstrate that performance of multimodal systems is affected by both those of single involved recognition modules, but also by the adoption of a suitable fusion rule for the obtained results.

6.1.4 Some Available Datasets

In this section, we will only summarize the main features of the public ear datasets which are used in the works considered herein. The aim of this presentation is to provide a way to better compare such works, and in no way to give an exhaustive list. On the other hand, it is to consider that the kind of approaches discussed here address variable settings. Early datasets contain images acquired in well-controlled pose and illumination conditions, and therefore rise too trivial problems to provide a significant evaluation of recognition performance. Moreover, this section will only describe either databases which are most frequently used and well described in literature, or databases which are available in different well-identifiable compositions, which are alternatively used in different works. For home-collected databases or for those lacking a consistent description, we will include details within the presentation of approaches using them. For a more detailed list and description, we again refer to the extensive review works by Pflug and Busch [74] and by Abaza et al. [5].

Carreira-Perpiñan database

Carreira-Perpiñan database is among those containing images with less distortion; it includes 102 grey-scale images (6 images for each of 17 subjects) in PGM. The ear images are cropped and rotated for uniformity (to a height/width ratio of 1.6), and slightly brightened.

IIT Kanpur ear database

IIT Kanpur ear database is one of the first made available to the research community. It seems not available anymore, or no complete description is available at present. Different works report a different composition, or at least different subsets used for experiments, and it is not possible to identify the exact content. One version contains 150 side images of human faces with resolution 640 × 480 pixels. Images are captured from a distance of 0.5–1 m. A second version contains images of size 512 × 512 pixels of 700 individuals. In general, it seems that even in its "hardest" composition, this database mainly contains images affected by rotation and illumination. Moreover, images are mostly filled by the ear region, so that detection is not that challenging. When this database is exploited, we will report the composition used from time to time.

University of Notre Dame (UND)

UND offers five databases containing 2D images and depth images for the ear, which are available under license (http://www.cse.nd.edu/~cvrl/CVRL/Data_Sets.html). The databases are referred as Collections with different features, acquired in different periods.

- Collection E (2002): about 464 right profile images from 114 subjects. From three to nine images were taken for each user, on different days and under varying pose and illumination conditions.
- Collection F (2003–2004): about 942 3D (depth images) and corresponding 2D profile images from 302 subjects.
- Collection G (2003–2005): about 738 3D (depth images) and corresponding 2D profile images from 235 subjects.
- Collection J2 (2003–2005): about 1800 3D (depth images) and corresponding 2D profile images from 415 subjects.
- Collection NDOff-2007: about 7398 3D and corresponding 2D images of 396 subjects. The database contains different yaw and pitch poses, which are encoded in the file names.

USTB ear database

USTB ear database is distributed in different versions (http://www1.ustb.edu.cn/resb/en/subject/subject.htm).

- USTB I (2002)—180 images of 60 subjects; each subject appears in three images: (a) normal ear image; (b) image with small angle rotation; (c) image under a different lighting condition.
- USTB II (2003–2004)—308 images of 77 subjects; each subject appears in four images; the profile view (0°) is identified by the position of the CCD camera perpendicular to the ear; for each subject the following four images were acquired: (a) a profile image (0°, i.e. CCD camera is perpendicular to the ear plane); (b) an image with −30° angle variation; (c) an image with 30° angle variation; (d) an image with illumination variation.
- USTB III (2004)—79 volunteers; the database includes many images for each subject, divided in regular and occluded images:
  - Regular ear images: rotations of the head of the subject towards the right side span the range from 0° to 60°, and images correspond to the angles 0°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, 45°, 50°, and 60°; the database includes two images at each angle resulting in a total of 22 images per subject; in a similar way, but with less acquisitions, rotations of the head of the subject towards the left side span the range from 0° to 45° and images correspond to the angles 0°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, 45°; even in this case, the database includes two images at each angle resulting in a total of 18 images per subject;

    the left rotation directory also contains face images: subjects labeled from 1 to 31 have two frontal images, labeled as 19 and 20; subjects labeled from 32 to 79 have two frontal face images and also two images in each of four

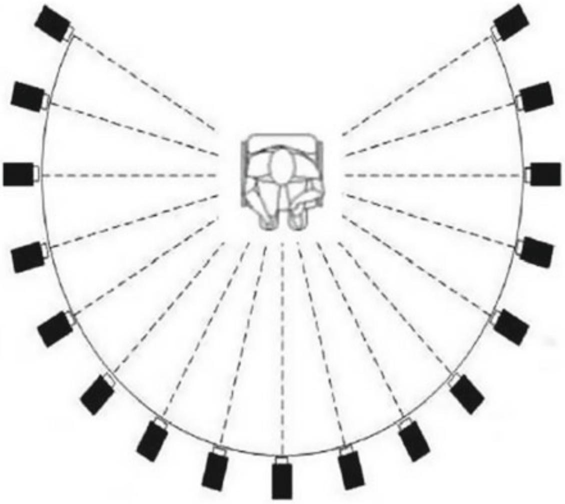

Fig. 6.4 Capture angles for images in USTB IV

other views, namely 15° rotation to right, 30° rotation to right, 15° rotation to left and 30° rotation to left, therefore each one has 10 face images labeled as 19–28.

- Ear images with partial occlusion: 144 ear images with partial occlusion belong to 24 subjects with 6 images per subject; three kinds of occlusion are included, namely partial occlusion (disturbance from some hair), trivial occlusion (little hair), and regular occlusion (natural occlusion from hair).

- USTB IV (2007–2008)—17 CCD cameras at 15° interval and bilaterally symmetrical are distributed in a circle with radius 1 m (see Fig. 6.4); each of 500 subjects is placed in the center, and images correspond to the subject looking at eye level, upwards, downwards, right and left; each pose is photographed with the 17 cameras simultaneously to capture the integral image of face and ear.

XM2VTS database

XM2VTSDB [64] collects multimodal (face and voice) data. The database includes 295 subjects, each recorded during four sessions. Each session corresponds to two head rotation shots and six speech shots (three sentences twice). Datasets from this database include high-quality color images, sound files, video sequences, and a 3D model. The XM2VTS database is publicly available but not for free.

FERET database

The large FERET database of facial images was gathered from 1993 to 1997 (http://www.itl.nist.gov/iad/humanid/feret/feret_master.html); the images were collected in a semi-controlled environment. The database includes 1564 sets of images for a total of 14126 images, corresponding to 1199 subjects and 365 duplicate sets of images.

Each individual has images at right and left profile (labeled pr and pl). The FERET database is available for public use.

IIT Delhi ear database
It is among the last presented ones. It is publicly available on request. It has been collected by the Biometrics Research Laboratory at IIT Delhi since October 2006 using a simple imaging setup. Images come from the students and staff at IIT Delhi, New Delhi, India, and are acquired in indoor setting. The database currently contains images from 121 different subjects and each subject has at least three ear images. The resolution of these jpeg images is 272 × 204 pixels. Recently, a larger version of ear database (automatically cropped and normalized) from 212 users with 754 ear images was also made available on request (http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm).

UBEAR database
This database is available after registration at http://ubear.di.ubi.pt/ and can be considered among the first sets of images actually captured in a realistic setting [79]. Data was collected from subjects on-the-move, under changing lighting. The subjects were not required any special care regarding ears occlusions or poses. For this reason the UBEAR dataset is a valuable tool allowing to assess the robustness of advanced ear recognition methods. Both ears of 126 subjects were captured starting from a distance of 3 m and facing the capture device. Then the subjects were asked to move their head upwards, downwards, outwards, towards. Afterwards, subjects stepped ahead and backwards. For each subject two different capture sessions were performed, giving a total of four videos per subject. Each video sequence was manually analyzed and 17 frames were selected:

- 5 frames while the subject is stepping ahead and backwards
- 3 frames while the subject is moving the head upwards
- 3 frames while the subject is moving the head downwards
- 3 frames while the subject is moving the head outwards
- 3 frames while the subject is moving the head towards.

6.1.5 Brief Discussion

When addressing real-world applications, uncontrolled settings can affect accurate ear recognition through many factors: (1) the ear might be cluttered/occluded by other objects, and due to the small size this is a more serious problem than with face; (2) the amount, direction and color of light on an ear can affect its appearance, due to its rich 3D structure; (3) out-of-plane head rotations can hinder recognition, again due to the 3D structure of the pinna; (4) the resolution and field of view of the camera can significantly affect recognition, especially if acquisition is at a distance.

Literature shows that the best results are achieved by 3D model matching alone or in combination with, e.g., 2D images. This allows recognizing ears under varying illumination and poses. However, a specialized equipment is required, which further needs controlled illumination. Moreover, in many non-contact/non-collaborative applications, it is required to work with surveillance photographs or with frames captured by surveillance videos, and this means that ears must be most often recognized from 2D data sources. The next sections will attempt a comprehensive overview of present techniques that potentially lend themselves to the above scenario.

6.2 Ear Image Preprocessing

Biometric recognition deals with the identification of a subject by characteristic biometric traits (e.g., ears), and requires to extract discriminative information from such traits and to organize it in a template. Settings and conditions for biometric sample acquisition may vary according to the trait at hand and to the dimensionality of the captured sample (2D or 3D). Afterwards, the acquired sample must be possibly enhanced and segmented: it is necessary to detect and locate the relevant biometric elements, and to extract them from the non-relevant context. The extracted region possibly undergoes a normalization/correction phase, in order to obtain a "canonical" representation. These preprocessing tasks are often independent from the chosen recognition approach. On the other hand, the following feature extraction and biometric template (key) construction, are always performed consistently with the methodology chosen for the recognition phase. For sake of consistent comparison, the same set of procedures for sample acquisition and template extraction is applied during both enrolling of interesting subjects, and recognition of probes. During enrollment the obtained template is stored in a database (gallery), while during recognition it is compared with the stored ones. In any real world application, robust enrollment is as important as robust recognition: quality of the former may significantly affect performance of the latter. The overall process implemented by any biometric system can be facilitated and enhanced, through the execution of appropriate preprocessing steps. We will survey a number of preprocessing techniques for ear biometrics which are found in literature. It is worth noticing that a correct ear location/segmentation is a necessary condition for a reliable recognition, though this may be hindered anyway by further factors like pose and illumination. In all cases, if the ear is poorly acquired or segmented, the possibilities of a correct recognition decrease dramatically. This especially holds for real world applications.

6.2.1 Acquisition Process

In most biometric systems, the acquisition process allows to capture an image of the biometric trait (e.g., fingerprints, face, iris, and ear). Relevant factors that influence the following process are the quality of the acquisition device and the kind of capture environment. For instance, the quality of a 2D ear image can be affected by a dirty lens, which may produce image artifacts, or by a low sensor resolution, which may cause a loss of sufficient image details. Illumination is an important environmental factor, which may produce uneven image intensity and shadows. Finally, further elements to consider are occlusions, like hair, earrings and the like, which may distort relevant ear regions, as well as pose changes, affecting the 2D projection of the inherently 3D ear anatomical structures. Pose and illumination variations can be dealt with using 3D models. At present, biometric modalities that rely on 3D information are constantly spreading, thanks to the increasing availability and decreasing costs of 3D capture devices, e.g., 3D scanners. In particular, biometric recognition algorithms based on 3D face and, more recently, 3D ear data are achieving growing popularity. In both cases, one can control the pose of the acquired subject and the distance from the capturing device, as well as the illumination, using an appropriate set and disposition of light sources. However, this is not possible in real world applications, and neither 2D nor 3D models are able to provide full invariance to any kind of distortion. Of course, acquisition in 2D and 3D differ not only for the kind of device but also for the structure of the obtained samples.

Acquisition in 2D requires digital cameras or video cameras, to capture still images or video sequences from which single frames can be extracted. In this case, elements that can change are the quality of lens and resolution of sensor. The acquired sample is usually a 2D intensity image, where the value of each pixel represents the intensity of the light reflected in that point by the surface. The color depends on how the light is reflected. The features and quality of an intensity image also depend on the camera parameters, which can be divided into extrinsic and intrinsic ones. The extrinsic parameters denote the coordinate system transformations from 3D world coordinates to 3D camera coordinates, so that they are used to map the camera reference frame onto the world reference frame. The intrinsic parameters describe the optical, geometric and digital characteristics of the camera, for example focal length, image format, and the geometric distortion introduced by the optics. As discussed above, intensity information can be substituted by heat information, if a thermografic camera is used.

Acquisition in 3D can be performed in more ways, and therefore can produce different kinds of samples. The most common ones are range images, also referred to as 2.5D images, which are 2D images where the value associated to each pixel represents the distance between that point and the capture device, and shaded models, which are structures including points and polygons connected to each other within a 3D space. Shaded models can also be enriched by mapping a texture on their surface, which can be derived by a 2D image of the same subject. Figure 6.5 shows examples of the most used kinds of ear sample.

One of the earliest techniques to obtain a 3D model relies on stereo vision, i.e. on stereo image pairs. It is a passive scanning technique (see [34]), since the scanning device does not emit an “exploratory” beam towards the object. Stereoscopic systems usually employ two video cameras, looking at the same scene from two slightly translated positions. These cameras are calibrated. This process allows to know their extrinsic and intrinsic parameters. In practice, this methodology allows to reconstruct a 3D model from a pair of 2D images. The critical point of the model construction

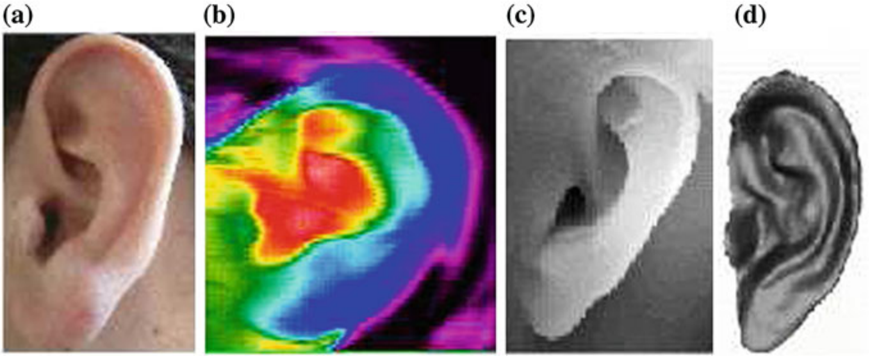

Fig. 6.5 2D and 3D ear samples: **a** intensity image; **b** thermogram; **c** range image; **d** shaded model

procedure is the search for relevant corresponding features in the images captured by each camera. The differences between such features make it possible to determine the depth of each point inside the images, according to principles similar to those underlying the human stereoscopic vision. The final result might be influenced by illumination, but the overall quality of the sample is medium. A number of recent proposals aim at overcoming the current limitations. In [121] a 3D ear reconstruction method is presented which is based on binocular stereo vision. A matching phase based on Scale Invariant Feature Transform (SIFT) is used to obtain a number of seed matches. This gives a set of sparse correspondences. Then the authors apply an adapted match propagation algorithm exploiting epipolar geometry constraints, to obtain a quasi-dense set of correspondences. Finally, the 3D ear model is reconstructed by triangulation.

In [88], the same authors directly exploit epipolar geometry, after cameras calibration and after ear location using AdaBoost in the pair of captured images (we will address location techniques later). In most reconstruction techniques, feature points are extracted independently in the two images, and then correspondences are searched. Along an alternative line, in the mentioned work Harris corner detector is used to extract feature points from the first image. For every such point, epipolar geometry constraints guide the identification of an epipolar line in the second image. When two cameras view a 3D scene from two distinct but well known positions, a number of geometric relations hold between the 3D points and their projections onto the two 2D images. These relations constrain correspondences between pairs of image points. Following the consideration that the corresponding point in the second image should be on, or very near to, the epipolar line, the search can be narrowed. The 3D ear model can be reconstructed by the triangulation principle.

Photometric systems are a further passive alternative for 3D scanning. They usually exploit a single camera, but take multiple images under varying illumination. Using such block of overlapped images, they invert the image formation model in the attempt to recover the surface orientation at each pixel. In particular, in Close-Range

Photogrammetry (CRP) the camera is close to the subject and can be hand-held or mounted on a tripod. This type of photogrammetry is often called Image-Based Modeling.

All passive scanning techniques require some illumination control. An alternative is to use active scanners instead. They emit some kind of light (or radiation), including visible light, ultrasound or x-ray, and detect the light reflection (or the radiation passing through the object). Laser scanners project a single laser beam on the object; a 3D extension of triangulation (knowing the position of the laser source and that of the capture device) is the most used technique for surface reconstruction when close objects reconstruction is involved [33]. Laser illumination has two main advantages over incandescent or fluorescent light: (a) a laser beam can be focused tightly, even over long distances, therefore allowing to improve the scan resolution, and (b) laser light has a very narrow radiation spectrum, so that it is robust to ambient illumination. As a consequence, the results are of high level. However, this technique is quite invasive, since the laser beam may damage the retina. On the contrary, structured light scanners (see [66]) use ordinary light, so that there is no danger for the retina. In order to improve robustness, they project a light pattern (e.g., a grid, or an arc) on the object to acquire: distortions of the pattern caused by the 3D structure of the object give information on its 3D surface. The final result is still less accurate than that obtained with laser. In both cases the reconstruction of the full 3D model of an object requires to scan it from different points of view.

The recent work in [61] presents a cheap equipment for the acquisition of 3D ear point clouds, including a 3D reconstruction device using line-structure light, which is made of a single line laser, a stepping motor, a motion controller and a color camera. The laser is driven by the step motor and continually projects beams onto the ear. The camera captures and stores the ear images projected by the laser. These serial strips are used as a basis for triangulation to recover the 3D coordinates of the scene points and finally to acquire 3D ear data.

6.2.2 Detection

While the reduced size of the ear can be considered as an advantage for biometric processing, at the same time it may hinder its detection, and makes its recognition less robust to occlusions. The detection of the ear, i.e. the location of the relevant head region containing it, should allow to separate it from the non-relevant parts of the image, e.g., hair and the remaining part of the face. Less recent approaches localize the ear after (manually) defining the small portion of the side of the face around it, to which the search can be reasonably limited. However, non-intrusive, real-world applications require to be able to detect the ear from a whole side face image, where, as noticed, hair and earrings cause hindering occlusions, and where ear position might not be frontal with respect to the acquisition device. Moreover a good resolution is often required. We will briefly summarize some more recent approaches, were some of these problems are better addressed. They can exploit some anatomical information, like skin color or the helix curve, or variations of AdaBoost methodology.

6.2.2.1 Skin Color and Template Matching

Skin color has been used in template based techniques, or together with contour information. In fact, due to the similar color or the ear and surrounding face region, a first rough identification of the candidate area is used to reduce the search space, and is almost always followed by a refinement through different techniques.

In [6] ear detection is preceded by a step of skin location to identify the face region. Skin-tone detection follows Flek's method [38], which uses a skin filter based on color and texture. A morphological dilation operation is applied to the output of the skin filter, to re-include zones possibly filtered out due to the color variation caused by 3D structure (shadows, over illumination). Canny edge detection is applied only to the skin region and an edge size filter removes short and/or isolated edges. Finally, the ear region is detected using template matching. Actually, the used ear template consists only of an edge representing the helix.

The approach presented in [116] uses skin color together with contour information, which is particularly rich in the ear region. In particular, the paper presents a tracking method for video data which combines a skin-color model and intensity contour information to detect and track the human ear in a sequence of frames. Due to the presented setting, this method appears to be suited to real-world applications. In a first step, Continuously Adaptive Mean Shift (CAMSHIFT) algorithm is used for roughly tracking the face profile in the frame sequence, and therefore identify the Region of Interest (ROI) for the next step. Afterwards, a contour-based fitting method is used for accurate ear location within the identified ROI. Since the mean shift algorithm operates on probability distributions, the color image must be represented by a suitable color distribution. The original CAMSHIFT algorithm requires to pre-select part of the face as the calculation region of the probability distribution to build the skin-color histogram; of course, this is not permitted by the addressed settings. Therefore the original procedure is modified by computing a suitable skin-color histogram off-line. Once the skin pixels have been extracted from the side face, edge detection and contour fitting are used to refine ear location. Since the shape of the ear is, almost always, more or less similar to an ellipse, ellipse fitting is used to locate it. The technique gives good results even with slight head rotation.

In [75] ear location in a side face image is carried out in three steps. First, skin segmentation eliminates all non-skin pixels from the image; afterwards, ear localization performs ear detection on the segmented region, using a template matching approach; finally, an ear verification step exploits the Zernike moments-based shape descriptor to validate the detection result. The image is mapped onto chromatic color space, where luminance is removed. Skin segmentation exploits a Gaussian model to represent the color histogram of skin-color distribution, since it can be approximated in this way in the chromatic color space. Skin pixels are selected according to a corresponding likelihood measure. It is interesting to notice that ethnicity may influence this process [25]. Skin chromas of different races may show some overlaps, but certain chromas only belong to certain races. Therefore, a compromise must be decided according to the kind of application. In [25], a chroma chart is prepared in a training phase to determine pixel weights (likelihood to be skin pixels) in test images. If the goal is to locate ears of a particular race in an image, the chroma chart should be created using samples from only that race. This will increase the skin detection accuracy. On the other hand, if the goal is to locate skin in images from different races, as is more likely, the chroma chart must contain information about the skin colors of various races, but this will select more image pixels as skin, including a higher number of false positives. While these considerations might not be relevant in an academic scenario, where enrolled subjects usually belong to the main ethnic community of the country where the laboratory is located, they become of paramount importance in real world applications, where they suggest appropriate design choices. At present, the influence of demographics on biometric recognition is a still emerging topic (see for example [56, 81] for face recognition).

Returning to [75], refinement of ear location within the skin region is performed through an appropriate ear template. The template is created during a training phase by averaging the intensities of a set of ear images, and considering various typical ear shapes, from roughly triangular to round ones. The template is appropriately resized during detection, according to the size of the bounding box of the skin region identified in the side face, and to the experimentally measured typical ratio between this region and the ear region. The normalized template is moved over the test image, and normalized cross-correlation (in the range -1, 1) is computed at each pixel, with high values (close to 1) indicating the presence of a ear. Such presence is verified in the last step. The reported accuracy of localization, defined in (6.1), is 94 %.

$$\text{accuracy} = (\text{genuine localization/total sample}) \times 100 \tag{6.1}$$

As can be expected, failures are reported especially for poor quality images and hair occlusion.

The work [82] uses mathematical morphology for automated ear segmentation. Opening top hat transformation, which is defined as the difference between the input image and its morphological opening, allows to detect bright pixels on a surrounding dark area, and to suppress dark and smooth areas. As a matter of fact, the ear is usually surrounded by darker pixels (hair) and smooth areas (cheek), and it contains dark regions whose effect is to further emphasize the edges after the transformation. The filtered image is thresholded, and 8-connected neighborhood is used to group the pixels in connected regions and to label them. Smaller and bigger components (with area less or greater than a threshold) are discarded, and the geometric properties of the remaining ones are analyzed. The system is tested on three datasets: the profile visible part of the database collected at the University of Notre Dame (UND) for the purpose of face and ear recognition, the CITeR face database, collected at West Virginia University to study the performance of face recognition, and the database of the Automated 3-D Ear Identification project, collected at West Virginia University and containing 318 subjects. The first set contains 100 profile images with different background and variable illumination conditions, the second set contains 150 images from 50 video frames for 50 subjects with different views and variable heights, and the third set contains 3500 images from 226 video frames for 226 subjects with different views. This is one of the most extensive testbeds reported in literature. The method achieves 89 % correct segmentation on the first set, 92 % on the second one, and 92.51 % on the third one. The authors further devise an automated technique to assess the segmentation result and avoid to pass a wrong region to the next recognition steps. They define five subclasses for the ear segmentation outcome, based on the viewing angle and on the quality of segments. A proper segment is a rectangular area containing a correctly segmented ear. An improper segment is a rectangular area that either contains only part of an ear or does not contain an ear at all. Two subclasses of proper segments include areas under viewing angles between $[-10°, 40°]$ and $[41°, 75°]$, two sub-classes of improper segments contain part of the ear in areas under viewing angles between $[-10°, 40°]$ and $[41°, 75°]$, and one subclass of improper segments does not contain an ear. An Automated Segmentation Evaluator (ASE) is implemented using low computational-cost, appearance-based features, and a Bayesian classifier to identify the subclass for a given segmentation.

6.2.2.2 Skin Color and Contour Information

As already noticed, era contour information is extremely rich, even if possibly influenced by illumination and pose. A number of proposed detection techniques are based on this.

The work reported in [9] considers the helix, with its characteristic pattern of two curves moving parallel to each other. Ear edges are extracted using Canny edge detector. Afterwards, they are thinned and junctions are removed, so that each edge has only two end points. Edges are divided into concave and convex ones. In this process, curves may become segmented in more edges that must be reconnected. The outer helix curve is identified by searching, for each candidate, another curve running parallel to it in the direction of concave side, at a distance from it characterized by a standard deviation below a given threshold. Even after this step, one may have more possible curves representing helix. In order to find the right one, perpendicular lines along each curve are drawn. If these lines intersect any other curve from convex side, the inner curve cannot be the outer helix and hence is rejected. Among the remaining curves, the longer one is selected. Once the outer helix has been identified, its end points are joined by a straight line to complete ear detection.

The work [76] presents a number of differences with respect to the above mentioned one. While Canny is used again for edge detection, when a junction is detected the edge is simply split in different branches. Spurious (short) edges are discarded, and the remaining ones are approximated by line segments. All the ear edges present some curvature, therefore their representation would need more line segments, and so more than two points. On the other hand, linear or almost linear edges can be represented by only two points, and since they cannot be a part of the ear they can be removed. An edge connectivity graph is created, where edges are vertices (points) and two vertices are connected if the convex hulls of the corresponding edges intersect. Ear localization takes into account the connected components in the graph. A core observation is that most ear edges are convex, and in general outer edges contain inner ones. As a consequence, convex hulls of outer edges must intersect those of inner edges. This implies that their corresponding vertices are connected in the graph. After computing the connected components for the graph created after the edge map, the bounding box of the edges corresponding to the largest connected component is claimed to be the ear boundary. The accuracy reaches 94.01 %, and falls to 90.52 % when moving away from frontal view.

In [32] the authors use the image ray transform, based upon an analogy to light rays. This particular transform can highlight tubular structures. In particular, one is interested in detecting the helix of the ear, and in possibly distinguishing it from spectacle frames. After Gaussian smoothing to reduce noise, and thresholding to produce an image with a strong helix, a simple matching with an elliptical template is used, across different rotations and scales; finally, the matched section is normalized and extracted.

The work presented in [96] exploits Jet space similarity, i.e. similarity of Gabor Jets. The original definition of Jet dates to [59], where they are used in the context of an object recognition application based on the Dynamic Link Architecture. The latter is a kind of hierarchical clustering into higher order entities of the neurons of a Neural Network. In this approach, the image domain I contains a two-dimensional array of nodes $\mathbf{A}_x^I = \{(x, \alpha) | \alpha = 1, \ldots, F\}$. Each node at position x consists of F different feature detector neurons (x, α), where the label α denotes different feature types, from simple local light intensities, to more complex types derived for example by some filter operation. The image domain I is coupled to a light sensor array, given that the same model may apply to the eye or to a camera. When such array is put on an input, this leads to a specific activation $s_{x\alpha}^I$ of the feature neurons (x, α) in the image domain I. After this, each node A_x^I contains a set of activity signals $\mathbf{J}_x^I = \{s_{x\alpha}^I | \alpha = 1, \ldots, F\}$. $\mathbf{J}_x^I$ is the feature vector named "jet". Given this structure, images are represented as attributed graphs. Their vertices are the nodes. Attributes attached to the vertices are the activity vectors of local feature detectors, i.e. the "jets." The links are the connections between feature detector neurons. In [96] a number of kernel functions determining Gabor filters with orientation selectivity are convoluted with the image. This produces features vectors defined as Gabor Jets. Gabor Jet function is used as a visual feature of the gray scale image $I(v)$ at each image point v. The presented approach introduces an "ear graph" whose vertices are labeled by the Gabor Jets at the body of the antihelix, superior antihelix crus, and inferior antihelix crus. These Jets are stored as ear graphs in the gallery, and PCA is used to obtain the eigenear graph; an ear is detected using the similarity between sampled Jets and those reconstructed by the probe.

In [57], exact detection relies on two steps: identification of ear ROI, and contour extraction from ROI. The identification of ear ROI in turn consists of three steps: first, skin detection is performed using Gaussian classifiers, then edge detection exploits LoG (Laplacian of Gaussian), and finally labeling of the edges uses intensity clusters. Afterwards, contours extraction is performed by a localized region-based

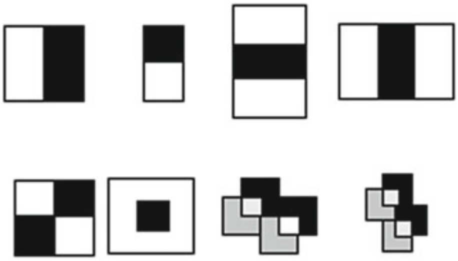

Fig. 6.6 Haar features used in [49]

active contours model, which is robust against initial curve placement and image noise. Features are extracted by SIFT and by log-Gabor for comparison. On a database of 100 users, SIFT features are superior as they achieve Genuine Authentication Rate (GAR) = 95 %, at False Authentication Rate (FAR) = 0.1, while log-Gabor features achieve GAR = 85 % at the same FAR. The significant limit of this work is that, to capture the ear images, volunteers are requested to keep their ear in the central hole of a wooden stand. Therefore this system is not suited for most real world settings.

The work in [58] exploits morphological operators for ear detection. Ear images undergo smoothing with a Gaussian filter which helps suppressing noise; this filtering is followed by histogram equalization. Ear shape is extracted by a series of gray scale morphological operations. The obtained gray scale image is binarized to extract the ear shape boundaries. This image is combined with the silhouette obtained by a thresholding operation, in order to obtain a mask able to eliminate all the skin region around the ear. Fourier descriptors are used to smooth the obtained boundary.

6.2.2.3 AdaBoost

AdaBoost algorithm for machine learning, as well as its variations and related methodologies, have proven very useful and powerful in addressing many practical problems. In particular, it gives good results in real-time, and is able to address poor quality settings. After its adoption for face recognition in [91], its use has been extended to many other kinds of objects.

The work presented in [49] is among the first systematic attempts to devise a specific instantiation of the general AdaBoost approach, specialized for ear detection. As in the original algorithm, Haar-like rectangular features representing the grey-level differences are used as the weak classifiers, and are computed as in [91] by using the Integral Image representation of the input image. AdaBoost is used to select the best weak classifiers and to combine them into a strong one. The final detector is made up by a cascade of classifiers. The differences between the structure of face and ear, and in particular the curves of the helix and of the anti-helix and the ear pit, suggest to use the eight types of features shown in Fig. 6.6.

Features with two rectangles detect horizontal and vertical edges, as usual; features with three or four rectangles detect different types of lines and curves;

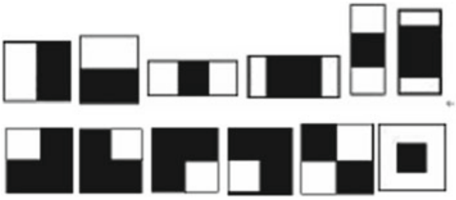

Fig. 6.7 Haar features used in [119]

the centre-surround feature (the second from left in the second row in Fig. 6.6) detects the ear pit. In order to address the same problem, a different set of features is used in [119], as shown in Fig. 6.7, and a 18 layer cascaded classifier is trained to detect ears. The number of features in each layer is variable.

One of the drawbacks of AdaBoost is the time required for training. In [3] a modification of the Viola-Jones method proposed in [99] is exploited for the ear, reducing the complexity of the training phase.

The ear detection algorithm proposed in [87] attempts to exploit the strengths of both contour-based techniques and AdaBoost. Ear candidate extraction adopts the edge-based technique that the authors call the arc-masking method, together with multilayer mosaic processing, and an AdaBoost polling method for candidate verification only within the pre-selected regions.

6.2.2.4 Detection in 3D

At the best of our knowledge, all the attempts found in literature to capture a 3D model of the ear are related to its detection from a range image of the side face. In [27] ears are extracted exploiting a template matching-based detection method. The model template is built by manually extracting ear regions from training images and is represented by an average histogram (obtained from the histograms computed for the single training images) of the shape index of this ear region. A shape index is a quantitative measure of the shape of a surface at a point p, and is defined by (6.2):

$$S(p) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \tag{6.2}$$

where k_1 and k_2 are maximum and minimum principal curvatures. For sake of compactness, the approach uses the distribution (histogram) of shape indexes as a robust and more compact descriptor than shape index 2D image. During detection, the obtained template must be appropriately matched with the test range image of a side face. To this aim, matching is preceded by preliminary processing. Step edge detection allows to identify regions with higher variation in depth, which usually correspond to the boundary of pinna. Thresholding is applied to get a binary image, and dilation allows to fill gaps. The analysis of connected components allows to determine the best candidate ear region, which is then matched to the template.

According to the authors themselves, this method cannot always identify the ear region accurately, therefore in [28] they propose a variation. The exploited ear template is the main difference with the preceding work. Ear helix and anti-helix parts are extracted by running a step edge detector with different thresholds, choosing the best result and thinning the edges, and finally performing a connected component labeling. After this, step edge detection and binarization are performed on the test image as above. The rest of the methodology is very similar, except for edge clustering in place of the analysis of connected components, and for the procedure for template matching.

The work in [127] exploits 3D local shape cues and defines new concise as well as discriminative features for 3D object recognition. Even in this case, such features are applied to the task of ear detection in range images. Again, the aim is to find a set of features, which are robust to noise and pose variations and at the same time can adequately characterize 3D human ear shape and capture its discriminative information. The used method for feature encoding is inspired by the Histograms of Oriented Gradients (HOG) descriptor. To locate the ear in a range image, the image is scanned with a fixed-size window. The presented feature is named Histograms of Categorized Shapes (HCS), and is extracted in each window position. To build HCS descriptors, each pixel in the image is assigned a shape index based on its shape category (e.g., spherical cup, saddle, etc.). The shape index of a rigid object is independent of its position and orientation in space, as well as of its scale. Moreover, each pixel is also assigned a magnitude, based on its curvedness values. HCS feature vector is extracted and used to train a binary ear/non-ear SVM classifier. Such classifier determines whether the current window contains an ear (positive sample) or non-ear (negative sample). To allow ear detection at different scales, the image is iteratively resized using a scale factor, and the detection window is applied on each of the resized images. Multiple detections in overlapping windows are fused.

Ear detection in 3D is addressed in [78] with an approach following the line in [75] (already presented above). It starts from the consideration that in a 3D profile face range image, ear is the only region containing maximum depth discontinuities, and is therefore rich in edges. Moreover edges belonging to an ear are curved in nature. A 3D range image of a profile face is first converted into a depth map image. The depth map image from a range image contains the depth value of a 3D point as its pixel intensity. The obtained depth map image undergoes the edge computation step using Canny edge operator. A list of all the detected edges is obtained by connecting the edge pixels together into a list, and when an edge junction is found, the list is closed and a separate list is created for each of the new branches. An edge length-based criterion is applied to discard edges originated by noise. Edges are then approximated by line segments. Even in the case of the depth map image, an edge connectivity graph is created, with edges being vertices which are connected if the convex hulls of the corresponding edges intersect. Tests performed assess that the method is scale and rotation invariant without adding any extra training. The technique achieves 99.06 % correct 3D ear detection rate for 1604 images of UND-J2 database, and is also tested with rotated, occluded and noisy images.

6.2.3 More Preprocessing

After a correct detection, it is often the case that ear must be transformed into a canonical form, which should allow extracting templates that can be matched despite different sizes and rotations. Related techniques might be exploited during enrolling, in order to store each template in its new form, and in this case the same procedure must be applied to the probe image. As an alternative, registration and alignment might be performed separately for each match. This topic is addressed in few works in literature.

In [8] the contour of an ear in an image is fitted by applying a combination of a snake technique and an ovoid model. The main limit of this approach is that a sketch of the ear must be manually drawn to start the procedure, so that it cannot be used in real-time, online, massive applications. After manual sketching, a snake model is adapted to ear contour characteristics to improve the manual sketch. Finally, the parameters of a novel ovoid model are estimated, that better approximate the ear. The estimation of ovoid parameters achieves a twofold goal: it allows to roughly compare two ears using the ovoid parameters, and allows to align two ear contours in order to compare them irrespective of different ear location and size.

Active Shape Model (ASM) is proposed in [117] for ear normalization. An offline training step exploits ASM to create the Point Distributed Model. To this aim, the landmark points on the ear images of the training set are used. The training samples are appropriately aligned by scaling, rotation and translation, so that the corresponding points on different samples can be compared. The model is then applied on the test ear images to identify the outer ear contour. Finally, the ear image is normalized to standard size and direction according to the long axis (the line crossing through the two points which have the longest distance on the ear contour) of outer ear contour. After normalization, the long axes of different ear images will have the same length and same direction.

The ear registration technique proposed in [20] is based on an algorithm that attempts to create a homography transform between a gallery image and a probe image using matching of feature points found by applying SIFT. Figure 6.8 shows an example of correspondences that can be found by SIFT.

As a result of the original algorithm in [17], the probe includes an image of the object to detect, if an homography can be created. Furthermore, the homography defines the registration parameters between the gallery and the probe. The work in [20] extends the original technique with an image distance algorithm, which requires to segment gallery ears using a mask. These masks are semiautomatically created as a preprocessing step when populating the gallery. The resulting ear recognition technique is robust to location, scale, pose, background clutter and occlusion, and appears as a relevant step towards the accuracy of 3D ear recognition using unconstrained 2D data.

3D point clouds are exploited in [42] to implement a method based on projection density, used to obtain the normalization of 3D ear coordinate direction. The method stems from the observation that there are specific projective directions in a 3D model,

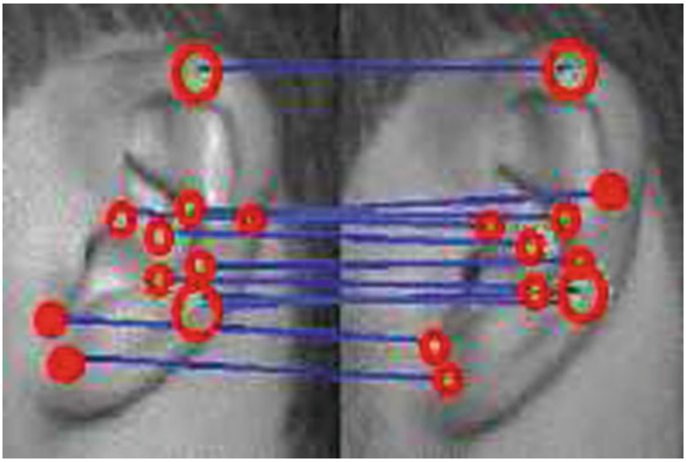

Fig. 6.8 Examples of point correspondence computed by SIFT

along which points projected onto the low dimensional spaces are distributed most sparsely. Optimal projection tracking is used to retrieve such directions and to revise the posture of 3D surface until all the projective points are distributed as sparsely as possible on the projection plane. Then PCA is used to obtain the principal axes of the projective points and adjust the axes to vertical direction. Among the claimed advantages of this method, compared with traditional ones, the authors mention that it is not necessary to organize the 3D model in mesh grid, but a point cloud can be used; moreover it is robust for noise and holes, and it generates a relatively uniform posture, which speeds up registration by, say, ICP.

6.3 Recognition of 2D Images

We refer to the rough classification of 2D methods sketched in Sect. 6.1.3, with a more detailed characterization of techniques. Geometrical approaches imply an often demanding processing to identify, with an acceptable reliability, some characteristic points or curves in the ear, able to drive the following feature extraction and recognition. Global methods aim at avoiding this preliminary step, by extracting the overall characterization of the ear. However, due to the rich spatial structure, this may lead to high dimensional feature vectors, so that dimensionality reduction techniques may be required, as well as different alternatives able to support a lighter computation. It is to say that most methods are hybrid, in the sense that they exploit the cooperation among more different techniques. When this is the case, we will try to consider the main approach. We will return on already presented works only when further detail is worth mentioning.

6.3.1 Geometrical Approaches

As reported in Sect. 6.1.3, one of the first works following Iannarelli's pioneering research was presented by Burge and Burger in [18]. In that first system implemented to assess the viability of ear biometrics, the authors first experiment two different techniques for ear location. In the first one, generalized Hough transform is used to find the characteristic shapes of helix rim and lobule, resulting in two curves loosely bounding the ear. The second method uses deformable templates of the outer and inner ear according to an active contour approach. The methodology is refined in [19], where the ear is located by using deformable contours on a Gaussian pyramid representation of the gradient image. Within the located region, edges are computed by Canny operator, and then edge relaxation is used to form larger curve segments. Since differences in illumination and positioning would undermine this method, the authors improve it by describing the relations between the curves in a way invariant to affine transformations and to small shape changes caused by different illumination. The chosen description is based on the neighborhood relation, represented by a Voronoi neighborhood graph of the curves. The matching process searches for subgraph isomorphisms also considering possibly broken curves. No experimental results have ever been reported regarding this technique.

In [86], after edge detection and appropriate pruning, the extracted features are all angles. The angles are divided into two vectors. The first vector contains angles corresponding to edges in the outer shape of the ear, and the second vector is composed examining all other edges. The approach is based on the definition of max-line and normal lines. Max-line is defined as the longest line with both endpoints on the edges of the ear. If edges are correctly identified and connected, the max-line has both its end points on the outer edge of the image. Features corresponding to each possibly detected max-line are to be extracted. Normal lines are those which are perpendicular to the max-line and whose intersections divide the max-line into $(n + 1)$ equal parts, with n a positive integer.

Given the points where the outer edge intersects the normal lines, each angle stored in the first vector is formed by the segment connecting one such point with the center of the max-line, and the segment connecting the upper end point of the max-line and its center. The second feature vector is calculated similarly, and all the points where all the edges of the ear intersect the normal lines are considered, except for those already used for the first feature vector, i.e. those belonging to the outer curve. Figure 6.9 gives and example of line pattern and of the two classes of angles.

The distance between two samples is computed in a hierarchical way, in two steps: the first vector is used first, and the second one later. In the first step, two distance measures are computed. The first one is given by the sum of absolute difference between corresponding vector elements, while the second one counts the number of similar elements, i.e. those whose difference is under a threshold. In order to pass the first check, both measures must comply with the respective thresholds. In the second step, two points are considered to match if their difference is under a threshold and they also belong to corresponding normal lines. However, in this case

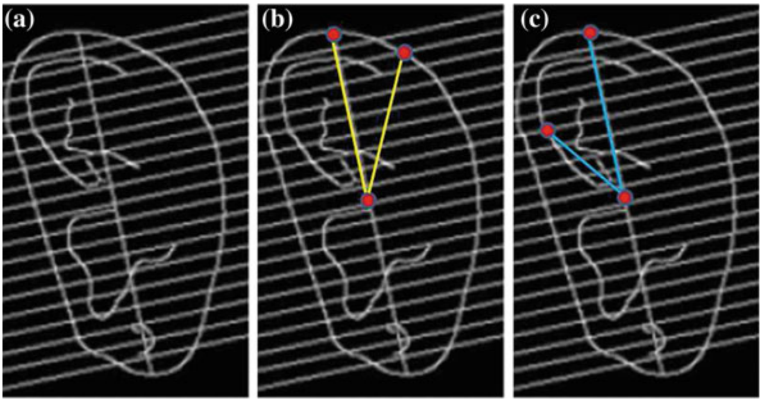

Fig. 6.9 **a** Example of max-line and normal lines; **b** example angle in the first vector; **c** example angle in the second vector

the vectors may have a different length, so that the final number of matching points must be appropriately normalized as a percentage of the lower dimension between the two vectors. Two images are said to match if they match with respect to the first feature vector and the last percentage of matching points is greater than some threshold. During an identification operation (1:N matching) a given query image is first tested against all the images in the gallery using the first feature vector. Only the images that are matched in this first stage are considered for the second stage of comparison. Such division of the identification task into two stages significantly reduces the time required. Therefore, this approach is especially appropriate for massive applications, and is also scale and rotation invariant. However, it may suffer for pose and illumination. The reported Equal Error Rate (EER) is 5 %. After the first stage of classification the search space reduces by 80–98 %. Due to lack of details on the test database, it is not possible to fully appreciate such performance.

Active Shape Model (ASM) technique is applied to ears in [63]. The aim is to model the shape and local appearances of the ear in a statistical form. Steerable features are also extracted from the ear image before applying the procedure. The magnitude and phase of such features are able to encode rich discriminant information regarding the local structural texture, and to better support shape location, though entailing a computational cost lower than the popular Gabor filtes applied at different scales. Eigenearshapes are used for final classification. The dataset used for experiments contains 10 images of size 640 by 480 from each of 56 individuals. In more detail, 5 images are for the left ear and 5 for the right ear, corresponding to 5 different poses, with a difference of 5° between adjacent poses. Both left and right ear images are used and the work shows that, as it might be expected, their fusion outperforms single ears (95.1 % recognition rate for double ears). This can be therefore also considered as an example of multibiometric system, where multiple istances of the same trait are used.

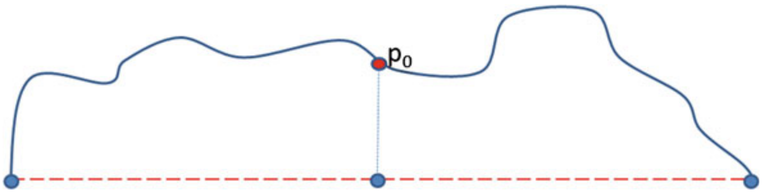

Fig. 6.10 Identification of point p_0 for each contour

Choraś and Choraś [31] introduce a number of different methods based on the processing of ear geometric features after normalization:

- Concentric circles based method—CCM: the centroid of (binary) ear contour is assumed as the (0, 0) origin of a system of coordinates and becomes the center of the concentric circles c_i of increasing radius r_i. Radial partitions are also identified using the circle circumscribed to the contour image. Features rely on suitably computed points of intersection of ear contours with the created concentric circles
- Contour tracing method—CTM: for each contour line in the binary ear image a number of characteristic points are considered, including contour ending points, contour bifurcations, points of contour intersection with the concentric circles identified by the previous method
- Angle based contour representation method—ABM: each extracted contour is considered independently and is represented by two sets of angles, corresponding to the angles between the vectors centered in a point p_0, that is characteristic for each contour; this point is the intersection between the perpendicular to the segment identified by the contour ending points in its central point, and the contour itself (see Fig. 6.10); the point p_0 becomes the center of a system of polar coordinates as well as the center of a set of concentric circles. The intersections of the contour with such circles are used to build the contour representation
- Geometric Parameters Method—GPM:
  - Triangle ratio method—TRM: finds the maximal chord of the longest contour and the intersection points of such contour with the longest line perpendicular to the maximal chord
  - Shape ratio method—SRM (linear vs. circular contours): the shape ratio for each contour is computed between the contour length, and the length of the line connecting the ending points of the contour
  - Contour Complexity—CC: the number of intersections between each maximal chord and corresponding contours.

The best results are achieved by the GPM method. However, the work lacks of details about the composition of the exploited dataset. Moreover, it would be interesting to try a combination of the proposed features.

The use of active contours is found again in [52], where ear is used for identity tracking through images from a live camera. All works along the same line rely on the basic concepts which characterize this methodology. It implies to define a model whose internal forces prevent arbitrary shape deformations, while external forces attract and deform the shape segments to reach the best fit with the edges on an underlying image. Internal forces are defined as constraints on the characteristic parameters of the curve, while direction and intensity of the external forces depend from the pixels of the edges that are detected on the tracked image. The curve segments tend to an equilibrium position where internal and external forces balance each other. The shape in motion is tracked by the deforming model. In the cited work, after the edges have been detected and made more evident through morphological operators, the active contours technique is applied to superimpose the lines of a standard ear model to them. This process is performed in two steps: the first one tries to match the outer contour of the ear model to the bordering edge of the ear; the second one executes a more precise matching of the inner contour lines afterwards. The authors consider the first step as a refinement of the localization of the ear, used to normalize the rest of the image. In this process, pixels on a segment representing an edge exert a greater attraction force on a line segment of the model if they are parallel. Once an active contour (model) has fitted the probe ear image, the last step is to analyze this modified model and collect features from it. Two sets of features are defined. The first one is derived from the distortion of the modified model with respect to the original one, the second is derived using three appropriate axes of measurement and 21 points selected appropriately on the model, each of which is associated with one of the axes. A feature is given by the distance measured between the projection of one such point on its associated axis, and the intersection of the three axes. The proposed method is tested on images chosen from video frames presenting twenty-eight different people. It is interesting to notice that the view angle of the ears varied up to about 50°, a variation which is quite realistic, for instance, in videosurveillance applications. Therefore the method is promising for real settings, given that the reported EER is below 10 % (7.6 %), although it is not usable as a single authentication measure in a high security setting.

The authors of [10] propose one of the first model-based approaches to recognition. The claim is that, being an abstract form of the corresponding object, a model capitalizes on the specific structures and thus naturally discards unnecessary detail. If well conceived, it can be robust to noise and occlusion and also potentially viewpoint-invariant. Moreover, the authors choose to work with 2D images, since they seem much more realistically appropriate to real world applications. The model includes a number of ear parts, each having its specific typical appearance. Each ear image is represented by a set of features. SIFT is used to automatically extract potential interest points (keypoints) in images, which describe neighborhoods of pixels. During a training phase, these features are clustered across a training dataset to denote the common ones. Model learning requires detecting the clusters and expressing them by suitable statistical properties. SIFT produces features which are naturally sufficiently invariant to illumination and viewpoint, and here they are also normalized with respect to scale and rotation. When a probe image is submitted, the ear is located by approximating the relevant region by an ellipse. An initial feature vector is extracted as the set of keypoints that are detected using SIFT. At this point, the model acts as a mask, since only keypoints common with the model are used for recognition, for which the mean of Euclidean distances between corresponding keypoints is used. The approach is tested on 63 subjects from XM2VTS database [64], choosing images without air occlusion, and then also occluding them with a rectangle from the top. However, it is to say that a sharp occlusion like a synthetic shape is easier to detect than a natural one, say unevenly diffused hair or a multicolor scarf. The worst result in the tests is obtained with automatic ear registration (in contrast to manual one) and with such occlusion, and achieves a recognition rate of 80.4 %. Therefore, it seems appropriate for settings with medium acquisition control.

SIFT methodology and its modifications is a technique often found in literature regarding ear recognition. For instance, it is adopted in [21, 122] . The first approach has been already partially described in Sect. 6.2.3 in relation to [20]. As part of the SIFT detection process, interest points are searched across locations and scales. When an interest point is detected, its canonical orientation is calculated. 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. However, a serial procedure is used to avoid false positives. To this aim, the potential space of affine transforms is subdivided into two dimensions for position, one for logarithm of the scale, and one for rotation. Each dimension is further divided into bins: eight for scale and rotation and one for every 128 pixels in width and height. Each point match is placed in the appropriate bin and also in its closest neighbors (16 bin entries per point). The choice of bins and this multiple classification of a point match should ensure robustness to pose variations. Only points contained in bins attaining four or more point matches pass to the next stage. A RANdom SAmple Consensus (RANSAC) algorithm is used. Random sets of four points are selected from the list of point correspondences and a homography is calculated. The homography that matches the most points within 1 % of the ear mask size, is selected as the best match. Gallery images that have four affine matching feature points are passed to the distance measure. This process greatly reduces false positives. The distance between gallery images and probe is the sum of the squared pixel errors after a suitable normalization. The distance measure is made robust to occlusion by thresholding the error. In practice, pixels that differ by more than half the maximum brightness variation are considered to be occluded and therefore are not included in the distance computation. The approach is tested on subsets of XM2VTS, some of which were processed to test the effect of occlusion, background clutter, resolution, noise, contrast, and brightness. The results highlight a good robustness to background clutter, occlusion up to 18 %, and over ±13° of pose variation. Due to the interesting yet highly detailed discussion, we address the interested readers to the original work.

The ear recognition approach in [122] uses the SIFT descriptor with global context (SIFT+GC) and some projective invariants to obtain ear features. The addition of context to compute matching points is deemed useful due to the presence of multiple local ear regions with similar texture. Recognition features are constructed using the number of the matching points, and by five projective invariants, which are obtained by computing the cross ratios of five collinear points on the longest ear axis. Experiments are carried out on the USTB II ear database. We remind that it contains 308 images from 77 subjects (4 each). For each subject, the first image is captured under uniform illumination condition and side pose, the other ones are captured under various poses or illumination conditions. The result using the first image as gallery and all the other three as probes is 91.34 % accuracy

SIFT approach is also used in [53]. In order to make it more robust, SIFT feature extraction is only performed in regions having color probabilities in certain ranges. The color model for ear skin is formed by Gaussian Mixture Model (GMM) and clustering of the ear color pattern using vector quantization. K-L divergence is applied to the GMM framework for recording the color similarity in the specified ranges. After segmentation of ear images in color slice regions, SIFT keypoints are extracted from single regions (slices) and an augmented vector of extracted SIFT features is created for matching. Since the number of points is not fixed, the SIFT feature points are varying in each slice region. After their extraction, they are gathered together by concatenation to form an augmented group for each reference model and probe model. The presented results come from tests on the IITK Ear database.

The version of the database used here consists of 800 ear images from 400 individuals taken in a controlled environment; the frontal view images are considered, so that the ear viewpoints are consistently kept neutral; the ear images are downscaled to 237 $\times$ 125 pixels with 500 dpi resolution.

The experiments entail two sessions. In the first session, the ear verification is performed with SIFT features without color segmentation; in the second session, verification is performed with the SIFT keypoint features detected from segmented slice regions. When nearest neighbor is used for verification with these slice regions, the system achieves the best performance of 96.93 % recognition accuracy. An evolution of the method presented in [55] uses the Dempster-Shafer decision theory to merge the detected keypoint features from individual slices, by combining the evidences obtained from different sources in order to compute the final distance. The performance increases to 98.25 %.

An improvement of the technique in [96] is proposed in [97]. It is based on Gabor Jet similarity and has been already presented in Sect. 6.2.2.2. A higher robustness to rotations in depth allows a good recognition accuracy. Data training not only exploits the Gabor Jets from the registration image, but also the estimated Gabor Jets of different poses obtained through linear jet transformation. Similarity is determined through the correlation between the Gabor Jet of the registered images and the input image.

The authors of [41] argue that the use of monolithic neural networks presents slow learning and other limitations that can be addressed by devising a modular neural network (MNN). They divide the ear images from USTB II into three parts containing relevant ear regions: the helix, the concha and the lobule. Each subimage is decomposed by wavelet transform and then fed into a modular neural network. Such network is made of three higher-level modules, and each of them includes three lower-level modules, one for each ear region. The difference among higher level modules is that they are trained with different subjects. The authors test Sugeno measures and Winner-Takes-All for integrating the lower-lever modules, while for higher level ones integration of decisions exploits Gating Network. The learning function is chosen between scaled conjugate gradient (SCG), or gradient descent with momentum and adaptive learning rate (GDX). Depending on the combination between integrator and learning function, the results vary between 88.4 and 97.47 % rank-1 performance on the USTB II database. The highest rank-1 performance is achieved with Sugeno measure and conjugate gradient.

6.3.2 Global Approaches

The approaches described in this section are defined as "global" in the sense that they do not consider specific points of the ear or specific parts of its contours, but rather assign a potentially identical role to every pixel in the image. The first and most cited one is based on force fields, and has been already presented in Sect. 6.1.3. We will present further such approaches, attempting some classification based on the main underlying concept.

6.3.2.1 Algorithms Based on Space Dimensionality Reduction

PCA is a very popular technique for feature space dimensionality reduction, even though, when applied to biometric recognition, it presents a number of limitations due to sensitivity to pose, illumination and expression (PIE) modifications. Nonetheless, many variations to the basic technique and further strategies have been presented in literature. As we already reported in Sect. 6.1.3, in [26] PCA is applied to both face and ear for comparison. The obtained results show that given that limitations are the same, also performance are comparable.

In the already cited work in [117], after ear detection ad normalization, recognition relies on a full-space linear discriminant analysis (FSLDA) algorithm. Tests are performed on the USTB ear database, which is chosen because it also contains ear images with rotation variations. In principle, FSLDA makes full use of the discriminant information in both the null space and in the non-null space of the within-class scatter matrix. Results are slightly different if head turns left or right. When the head turns left, recognition rates at 5, 10, 15 and 20° are above 90 %. After 20°, the recognition rate drops dramatically. When the head turns right, the recognition rates at 5 and 10° are above 90 % as well. The recognition rate at 15 and 20° is worse, but still above 80 %, and even in this case drops dramatically after 20°.

A kind of hierarchical approach, defined as Compound Structure Classifier System for Ear Recognition (CSCSER), is adopted in [125]. After an appropriate segmentation, the ear probe is first roughly classified in one of five classes, based on the height/width ratio. Afterwards Independent Component Analysis (ICA) is used, and finally a Radial Basis Function Network (RBFN) returns the result. Experiments on a home-collected image library demonstrate higher performance of ICA with respect to PCA, especially when the proposed classification structure is adopted. However, the method heavily depends on the first classification step, which in turn heavily depends on the assumption of a frontal pose. Therefore, methods of this kind, though indicating a possible line, cannot be adopted in under-controlled settings.

In order to address the problem of pose variation, a different space reduction technique is exploited in [101]. The authors rather start from Locally Linear Embedding (LLE), which is a nonlinear dimensionality reduction deemed more able to find the intrinsic structure of data points. This is achieved by constructing an embedded space which preserves their topological relationship, though transforming the nonlinear problem into a linear one. The LLE algorithm selects k (nearest) neighbors for each input vector, according to Euclidean distance, and then expresses the original vector as a combination of such neighbors. Finally, new vectors in a space of reduced dimension are computed using such weight, subject to some constraints. In the mentioned work, the authors propose an optimization procedure to compute k. Moreover, they further improve LLE by applying LDA to the obtained vectors. The approach is tested on USTB III database. The used dataset contains two groups of 19 images each, corresponding to 19 different poses, for each subject out of 79 individuals. Head rotations in the right side cover angles of 0°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, 45° and 60°. Rotations in the left stop at 45°. On the other hand, illumination is constant. While LLE achieves a 40.03 % recognition rate, using the proposed improvements achieves a significantly higher 60.75 %. Notice that the exploited database is one of the most challenging for pose variations.

The same authors in [102] further observe that, when the distribution of data points is highly sparse, the choice of neighbors might be unstable. For this reason, they propose to use a different distance function to select the “nearest” neighbors, defined in (6.3), obtaining the Improved Locally Linear Embedding (IDLLE):

$$Hsim(X, Y) = \frac{\sum_{i=1}^{d} \frac{1}{1+|x_i - y_i|}}{d} \tag{6.3}$$

Experiments are performed on the same dataset from USTB III database again. They show a better performance of IDLLE with respect to LLE, but it is not possible to compare them with the preceding ones since results are disaggregated according to different in-depth rotation angles. Of course, the best results are achieved in the range [−15°, +25°] with negative rotations being towards left, and positive towards right, starting from 0° which is the up-right frontal position.

A different approach to improve LLE aiming at multi-pose recognition is proposed in [106]. In traditional LLE, the neighbors for a point are chosen according to k-nearest neighbor algorithm or, as an alternative, according to ε-neighbor algorithm. In the latter, ε is a fixed radius around a point where neighboring points are selected, causing different points to possibly have a different number of neighbors. Instead of computing LLE many times with different values of k to find the optimal one, the proposed approach uses both algorithms together. In this way each point has a neighborhood including a set with a fixed number of neighbors, plus a further set with a variable number of neighbors depending on the computation of a minimal reconstruction error. Compared with the k-nearest neighbor algorithm, this method allows different numbers of neighbors for different points, thus avoiding unconnected areas and misrepresentations caused by clustering in other areas. Compared with the ε-nearest neighbor algorithm, this method uses an ε-ball with changing radius for each different point, thus avoiding an excessive difference in the number of neighbors for different points. Results of tests performed on USTB III pass from an accuracy of 66.17 % with LLE to 78.14 % with this approach.

ICA is the starting point for the work in [100]. To address the intrinsic limitations of ICA linearity, the authors adopt for ear recognition the Kernel Independent Component Analysis (KICA) proposed by [12], and use SVM for recognition with Gaussian Radial Basis Function (GBBF). Unfortunately, the authors only report results on noise (Gaussian, salt, Poisson) obtained by adding noise to images from Carreira-Perpiñan database. Images in such database are all frontal and with uniform illumination, therefore assessment with respect to variations in pose and illumination is lacking.

6.3.2.2 Algorithms Based on Wavelets, Transforms and Hybrid Systems

Generic Fourier Descriptor (GFD) [123] has been defined for image indexing, and is adapted in [1] to ear recognition. Fourier Transform (FT) is quite robust to noise and to some kinds of image distortion which only affect high frequencies. For this reason, it has been widely used in image processing. As an example, 1D FT can be applied to contour shapes. However, applying it for shape indexing involves the knowledge of the object boundary, and unfortunately this requirement can be seldom satisfied. 2D FT overcomes this by considering the whole image information. However, 2D FT has its limitations too when used for shape indexing. As a matter of fact, when applied to the Cartesian representation of an image, it produces descriptors which are not rotation invariant. On the other hand, rotation invariance is among the main requirements for a shape descriptor. In [123] the problem is solved by transforming the image in polar coordinates, and than treating the polar image in polar space as it was a normal two-dimensional rectangular image in Cartesian space. In practice, the 2D FT on this rectangular image is similar to the normal 2D discrete FT in Cartesian space. In [1] this consideration is applied to the segmented ear image, once the center of the ear has been located, as shown in Fig. 6.11.

The GFD descriptor is made robust to rotation by ignoring phase information in the coefficients, and only retaining their magnitudes; the first magnitude value is normalized by the area of the circle containing the polar image, and all the other magnitude values are normalized by the magnitude of the first coefficient. However, in transforming the input image from Cartesian to Polar space, the correct location of the center of the axes $C(0, 0)$ plays a crucial role. GFD is not robust to translations, therefore even small shifts of this point can cause significant modifications in the coefficient values. In [1] translation invariance is added by choosing the center of mass $MC(x_c, y_c)$ of the edge map of the input shape as $C(0, 0)$. Results are presented

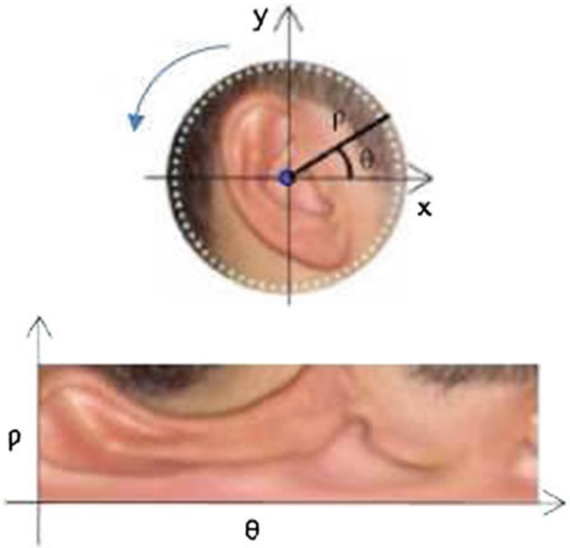

Fig. 6.11 Transformation of the ear region from Cartesian to polar space

for a home-collected database, and demonstrate high robustness to rotations with respect to the horizontal plane.

The representation of an image obtained by the Gabor filter exploits its orientation-selective properties and is robustness to PIE variations, and can provide high recognition performance. However, such performance decreases with the increase of classification space, and also computational complexity plays against this technique. On the other hand, the feature space dimension can be reduced by any of the classical methods. The approach presented in [94] uses General Discriminant Analysis (GDA), which is designed for nonlinear classification based on a kernel function. Results obtained on USTB II database demonstrate that the combination of the two techniques achieves better results than the two separately, and also in a shorter time. The recognition rate of Gabor algorithm is the lowest one (96.73 %) out of the three algorithms the recognition rate of GDA reaches 98.35 %, but the total time of GDA is much longer (2.1203 s vs. 0.5731 s); finally the recognition rate of Gabor+GDA is the highest one (99.1 %), achieved in the shortest total time (0.5689 s).

An improvement of the technique presented in [10] and discussed in Sect. 6.3.1 exploits log-Gabor filtering after SIFT feature clustering [11]. The wavelet approach aims at extracting the frequency content of the fluctuating surface of the Helix and the Anti-helix of an ear. Since the wavelet process is localized, it can provide uncorrupted information in presence of occlusions. Given an estimate of the Helix location, log-Gabor filtering is applied to this region to describe this curve. The model parts identified by the approach in [10] are used to vote for the position of the Helix. A template is built by the sampled image intensities in a semi-circular region which includes the identified Helix. This semi-circle is centered in the point where the Crus of helix curves inwards. In general, this point is almost the midpoint of the overall ear height and is situated on the outermost part of the ear, opposite to the Helix. The template is formed by sampling the image intensities along the lines starting from the identified center, which are mostly normal to the Helix curve. In this template the columns exhibit the variation between the ridge of the Anti-helix and the Helix at each specific angle. A one dimensional log-Gabor filter is just used on these columns. The presented tests use 252 images from 63 individuals, which are those in XM2VTS database whose ear is not obscured by hair. The 4 images per individual are taken in 4 different sessions over five months. Recognition using the original nearest-neighbour algorithm on the model parts, obtains 91.5 % correct recognition. The new log-Gabor coefficients alone achieve 85.7 % recognition rate. Substituting the Euclidean distance with a more robust distance metric, the latter performance reaches 88.4 %. Combination of log-Gabor and model, using the simple sum of the normalized scores, reaches 97.4 % recognition rate. This suggests that, although the log-Gabor performs worse than the model, it contains novel and independent information which enhances the recognition.

The approach in [72] aims at extracting edge information along three independent directions. To this aim, after histogram equalization and normalization, 2D wavelets are computed on the input image in horizontal, vertical and diagonal directions separately, to obtain three feature matrices. The filter bank exploits Daubechie wavelet class, which is a family of orthogonal wavelets. The points with high or low intensity correspond to significant changes and represent the extremes for wavelet coefficients. PCA applied to the single matrices of wavelet coefficients reveals that Horizontal matrix has larger recognition rate than Vertical and Diagonal matrices (90.19, 86.27, and 82.35 % respectively). Therefore, in combining them to obtain a single feature matrix, different weights are assigned to the original ones. Finally, the integrated matrix undergoes PCA to reduce the feature space dimension. Recognition tests are preformed on USTB II Database and Carreira-Perpiñan database. Accuracy on the first one, which is the “hardest” one, reaches 90.5 % and outperforms PCA alone and ICA.

The authors of [113] propose to use the HMAX model, introducing a new robust set of features. Each element of this set is actually a complex feature, representing the combination of position- and scale-tolerant edge detectors over neighboring positions and multiple orientations. The system relies on a quantitative model of visual cortex. In its simplest version, such model includes four layers of computational units. Simple S units combine their inputs with Gaussian-like tuning to increase object selectivity; they alternate with complex C units, which pool their inputs through a maximum operation, providing gradual invariance to scale and translation.

The system first simulates S1 responses by applying to the input image a battery of Gabor filters, which can be described by (6.4):

$$G(x, y) = exp(-\frac{X^2 + \gamma^2 Y^2}{2\sigma^2}) \times cos(\frac{2\pi}{\lambda}X) \tag{6.4}$$

where $X = x \cos\theta + y \sin\theta$, $Y = -x \sin\theta + y \cos\theta$, and θ (orientation), σ (width) and λ (wavelength) are the filter parameters which allow to reproduce the behavior of cortical cells. After appropriate selection, 16 scales and 4 orientations are chosen and organized in 8 scale bands. The next C1 stage is built by following the proposal by Riesenhuber and Poggio to adopt a max-like pooling operation for building position- and scale-tolerant C1 units. The subsequent S2 stage is where learning occurs and is exploited during training, where each S2 unit behaves like an RBF classifier on a number of image patches at random positions and in all orientations. Input images processed by C1 are convolved at different scales with units in S2, and a C2 feature vector is produced whose final size depends only on the number of patches extracted during learning and not on the input image size. This C2 feature vector is passed to a classifier for final analysis. The authors test k-nearest neighbor (KNN) and support vector machine (SVM) approaches for this last step. Experiments exploit USTB I database (60 subjects with 3 images each with slight rotation and some illumination variation), i.e. controlled conditions. Therefore this very interesting method, though being scale and rotation invariant, should be further tested for position invariance.

An improvement for the classical force field approach in [45] is proposed in [36], again to address the multi-pose recognition problem. In the original method, the feature extraction is driven by some test pixels following the force field lines due to the variation of potential energy, and getting the channels and wells. The latter are stable and unique for each person, but this holds only for pose-fixed images, otherwise they become unstable and may even disappear. To avoid this problem and address multi-pose settings, the presented technique computes the force field transformation without extracting the potential well feature. Feature description is rather performed using Kernel Fisher Discriminant Analysis (KFDA), which is a nonlinear subspace analysis method combining the kernel technique with LDA. As force is a vector, the absolute value of force replaces the intensity in the ear image which becomes a force image. Then column vectors of the two-dimensional image are concatenated to obtain a one-dimensional augmented vector which is used as a sample in NKFDA. USTB III database is used for experiments. While the best results achieved with wells reach only 75.3 %, the proposes method approaches a 90 % recognition rate with a dimension less than 20.

A further work inspired by physics of electricity is presented in [14]. Here, the adopted pattern recognition approach relies on Edge Potential Function (EPF), which models the attraction force generated by edge structures contained in an image over similar curves. The model exploits the joint effect of single edge points in complex structures, determined by edge position, strength, and continuity, and builds a corresponding edge map. Good results in the correct matching are obtained even in presence of noise and partial occlusions. To obtain robustness to rotation, horizontal and vertical translation, and scale modification, a probe template is compared to the templates in the database exploiting a genetic algorithm, which is recursively run to translate, scale, and rotate the template until a fitness function returns a value greater than a pre-defined threshold. The proposed method achieves a rank-1 recognition rate of 98 %, but the datasets used for testing are USTB II and UMIST (face images with rotations of 5°, 10°, and 15°) which are not suited to significantly assess robustness to pose variations.

Different Gabor-related techniques are tested in [58] for ear identification: the extraction of phase information using either 1D log-Gabor and 2D Gabor filters, or the complex Gabor filters, and the extraction of orientation information using a bank of even Gabor filters. The best results in terms of rank-1 recognition accuracy are obtained using a pair of log-Gabor filters, and achieve 96.27 and 95.93 %, respectively, on a database of 125 subjects and on an extended version with 221 subjects. Images contained in the used dataset are acquired in a strictly controlled setting, and in frontal position.

6.3.2.3 More Hybrid Techniques

The system presented in [124] combines Independent Component Analysis (ICA) and an RBF network. The ear image is decomposed into a linear combination of several basic images. Then the corresponding coefficients are fed up into an RBF network. Though being linear, ICA presents significant advantages over PCA: a better probabilistic model of the data, which gives a better identification of clusters in the n-dimensional space, a better unmixing ability, and the ability to handle higher order data. In some experiments images are pre-filtered to enhance contours, using either Wiener filtering or Laplacian-Gaussian filtering. The approach is tested on Carreira-Perpiñan database and on a home-collected one. Both sets of images present ears in a frontal pose.

The work in [35] adopts an approach that can be considered as global in the definition used here, but is local with respect to the way ear image is processed. The feature extraction process is based on the fractal technique of PIFS (Partitioned Iterated Function Systems). Such process is made local by dividing the normalized image of the ear region in four quadrants. They have no special relation with specific anatomical elements inside, but rather represent usual zones possibly affected by occlusion (upper ones by hair, lower ones by earrings). PIFS is based on building a map of self-similarities within the image, and dividing this process into the four quadrants allows to better address occlusions, since corrupted information in one quadrant can be compensated by that provided by the other ones. Some computational optimizations are adopted to avoid the usual computational weight of fractal techniques. Experiments are performed on subsets of Notre Dame ear database and of side images from FERET. Images in the latter set are mostly affected by illumination variations, while those in the former one also undergo pose changes. Obtained results show that both the system in [35] and the compared ones (PCA, LDA, KDA, OLPP) are less robust to pose than to illumination, but the former is much more robust to occlusions. Moreover, since an illumination variation affecting only a region may be compared to an occlusion, even this problem is better addressed.

A multimatcher approach is adopted in [69–71]. In the first work, each matcher is trained using features extracted from a single-subwindow (SW) of the entire 2D image. The ear is segmented using two landmarks, namely Triangular Fossa and

Incisura Intertragica, according to one of the approaches in [109]. The whole image (150 × 100) is divided into SWs of dimension 50 × 50 with a step of translation of 11 pixels to obtain 50 SWs for each image. The features are extracted by the convolution of each SW with a bank of Gabor filters, and afterwards their dimensionality is reduced using Laplacian eigen-maps. The best matchers correspond to the most discriminative SWs, and are selected by running Sequential Forward Floating Selection (SFFS). Experiments use subjects from the UND database (collection E) and the sum rule is employed for fusing the results from the selected matchers at the score level. The rank-1 recognition rate is about 84 %. A quadratic discriminant classifier (QDA) is trained in the later work in [71] to improve discrimination between genuine and impostors in verification tasks, based on the same feature extraction technique. Finally, in [70], the multimatcher approach is used along very similar lines yet combining different color spaces instead of different SWs, according to the ability of different color spaces of bringing different information. The input images are in the RGB color space, and are then mapped onto the 12 spaces: YPbPr, YCbCr, YDbDr, JPEG-YCbCr, YIQ, YUV, HSV, HSL, XYZ, LAB, LUV, LCH. The three image components of each of the 13 spaces are preprocessed. Feature extraction is performed by using a bank of Gabor filters of different scales and orientations applied on a regular grid superimposed to the image. Even in this case selection is performed by SFFS. Identification results reach about 84 %. The authors' conclusion is that pose variations are still better approached through 3D approaches. However, it might be interesting to test the combination of their two sets of matchers.

The work in [120] starts from a similar multimatcher approach to investigate the problem of partial occlusion. In this work, the implementation of the different steps is different from [69]. During training, sub-windows of the same position are combined to form a sub-space, and then Neighborhood Preserving embedding (NPE) is applied to get the projection matrix of each sub-space, so that the feature vectors of training images are extracted according to the same technique. For each sub-space, a sub-classifier uses the nearest neighbor rule. Sub-spaces are ranked according to to their recognition rates. In the test stage, after ear detection and sub-window division, the feature vectors of each sub-window are extracted, and the top ranking sub-classifiers are selected for recognition. The final decision is produced by score level fusion exploiting weighted sum rule. Experiments on the same USTB set show that this method is more robust to occlusion than [69], but pose variation may still be a harder problem to address.

The work presented in [77] approaches the problem of robustness to pose variations by considering different aspects. The first step entails image enhancement. Three image enhancement techniques are applied in parallel to neutralize the effect of poor contrast, noise and illumination (histogram equalization, non-local means filter and steerable filter), and the three obtained images are then processed separately. Each of them separately undergoes a local feature extraction technique, namely Speeded-Up Robust Features (SURF) [15], which is deemed able to minimize the effect of pose variations and poor image registration. It makes use of Hessian matrix for key-point detection. Haar wavelet responses *dx* and *dy* in horizontal and vertical directions are computed in a circular region around the detected points. These responses are used to obtain the dominant orientation in the circular region, and then feature vectors are measured relative to the dominant orientation. In this way the resulting vectors are invariant to image rotation. Afterwards the method considers a square region around each key-point, which is aligned along the dominant orientation and divided into 4 × 4 sub-regions. Haar wavelet responses are computed for each sub-region, and the sum of the wavelet responses in horizontal and vertical directions for each subregion are used as feature values. The SURF feature vector of a key-point is obtained by concatenating the feature vectors from all sixteen sub-regions around it, resulting in a vector of 64 elements. Three sets of local features are obtained, one for each enhanced image. Three separate nearest neighbor ratio classifiers are trained on these sets of features. After computing nearest neighbor, its validity is assessed by computing the ratio of the distance from the closest neighbor to the distance of the second closest neighbor. Matching score between two ear images depends on the number of matched feature points. These matching scores are normalized using min-max normalization technique and are then fused using weighted sum rule. Tests are performed using IITK database and UND database (Collection E). The composition used here of IIT Kanpur database includes two data sets. Data Set 1 contains 801 side face images of 190 subjects. For each subject, 2 to 10 images are acquired. Data Set 2 includes 801 side face images from 89 individuals. For each of them, 9 images are captured containing three rotations (person looking straight, looking approximately 20° down, and looking approximately 20° up), and three scales (camera at a distance of approximately 1 m and digital zoom of the camera at 1.7x, 2.6x and 3.3x) for each rotation. For a detailed discussion of results, the reader can refer to the original work. However, it is worth noticing that results on UND-E database, which presents pose variations, are worst that those on IITK, which mainly presents rotation and illumination variations.

6.3.3 Algorithms Based on Multiscale/Multiview

Robustness to illumination and pose can be significantly improved by adopting techniques which address the problem by separately considering regions which are usually affected at a different extent by illumination and pose variations. Multi-scale/multiview approaches can be considered among these. In particular, multi-view imaging aims at acquiring richer geometric and structural features for the ear.

In the already mentioned work in [63], multi-resolution is applied in an early stage to select the best set of landmarks for each image. As a matter of fact, ASM can suffer from changes in illumination and local minima in optimization, and the result depends on initialization. The authors generate a pyramid of images with different resolutions and obtain the profile statistics for each landmark and each pyramidal level.

The technique proposed in [93] focuses on the textural features of ear image. It stems from the Local Binary Patterns (LBP) approach, but exploits a multi-resolution variation allowing to capture both smaller and larger structural information. Circular LBP is more flexible ans uses a circular region with discretionary radius and number of considered neighbors. In practice this is equivalent to acquire different resolution LBP operators. This solution can be affected by aliasing effects in the obtained representation of the image, when large radii are exploited. The authors propose to solve the problem by wavelet low-pass filters. Haar wavelet decomposition is first performed, in order to obtain different filtered versions of the original image, in a number depending on the chosen scale factor. Besides this, they adopt Uniform (Circular) LBP to reduce the size of the derived histogram. A local binary pattern is uniform if it contains at most two bitwise transitions from 0 to 1 or vice-versa. In practice, uniform binary patterns occur more commonly in texture images than others. In the computation of the LBP labels, a separate label is assigned to each uniform pattern, while all the non-uniform patterns share a same label. For example, when using a 8-neighborhood, there are a total of 256 patterns: since 58 of them are uniform, we can reduce to 59 different labels. A further contribution to pose and illumination invariance is given by applying a block division method joined with multi-resolution. First the original image is divided into several overlapping blocks of arbitrary size. Then the method calculates the ULBP distributions for each of the blocks and chains the histograms into a single vector representing the image. Multi-resolution is achieved by applying ULBP with two different radii and number of neighbors, and combining the obtained histograms. In summary, ULBPs are combined simultaneously with block-based and multi-resolution methods to describe together the texture features of the filtered ear images, which are obtained by Haar wavelets. Matching is performed by using Canberra distance measure [89]. Experiments are performed on a subset of images from USTB III, namely on right rotations of 0°, 5°, 20°, 35° and 45°. Results demonstrate that the combination of the proposed techniques allows to pass, at a rotation of 35°, from 8.86 % recognition rate of PCA to 62.66 %.

A similar but much simpler approach is presented in [95]. In this work, LBP is applied to images at different scales, which are obtained by wavelet transform. The eigenvectors of the obtained LBP images undergo Linear Discriminant Analysis (LDA) algorithm to extract feature. In test phase, Euclidean distance is used. Tests show a 100 % recognition rate with a feature space dimension of 60, but are performed on USTB II which is not the most challenging one.

Different views of the ear contain shape features of different orientations. In [126] a subspace of feature space termed ear manifold is exploited. Ears can be projected onto feature space, and if this is done for multiple views, it is possible to obtain a smooth trajectory. Ear pose manifold is deemed to play an important role in recognition. The authors first construct a multi-view dataset of manually segmented images, where for each out of 60 subjects there are eight different views (−60°, −50°, −40°, −20°, 0°, +20°, +40°, +60°). However, the capturing device is placed in a dark room and is illuminated with fixed lighting, and moreover, as it can be observed in Fig. 6.12, some frame seems to be further used to make segmentation easier. Therefore, it is arguable if results fully hold in real-world settings.

Null space kernel discriminate analysis (NKDA) [60] is implemented on a subset of training images to form ear multi-view discriminative feature space. Training samples are projected onto the feature space to generate points at different positions.

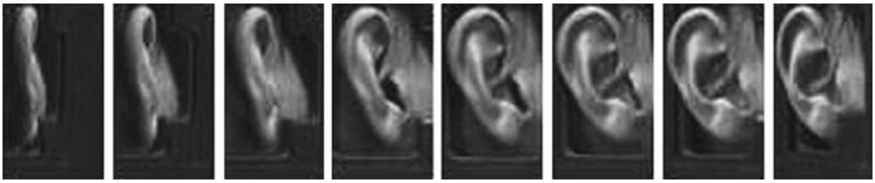

Fig. 6.12 Multiple view of a ear which are projected onto the ear manifold

Then B-Spline is used to interpolate views for all projected points, and to acquire the pose curve representing the ear pose manifold used for recognition. It is assumed that ears of different subjects give raise to different pose B-Spline curves, which in turn represent different pose manifolds. During test phase, a test sample is projected onto the discriminative feature space, and the distance to all existing points of B-Spline pose manifolds is calculated. The subject corresponding to the closest pose manifold from the test point is recognized as its identity. During the experiments, orientations (−60°, −50°, −20°, +20°, +60°) are used for training, while orientations (−40°, 0°, +20°) are taken for test. The best achieved performance is 97.7 % rank-one recognition rate.

A different multi-view approach is adopted in [61]. The work investigates multi-pose ear recognition using Partial Least Square Discrimination (PLSD). Partial Least Square Regression (PLSR) is a statistical regression technique that allows to relate several dependent variables (called responses) to a large number of independent variables (called predictors). PLSR models the relationship between the dependent and the independent variables by creating latent vectors, that account for as much of the covariance between the dependent and independent variables as possible. In this way PLS can be used as a dimensionality reduction technique. Classification problems under linear representation framework, use the partial least square regression model (6.5):

$$Y = XB + E \tag{6.5}$$

where Y are the responses, X are the predictors, B is the regression coefficient matrix and E stands for the residual matrix. In this case the matrix X of independent variables is treated as the samples training matrix and the matrix Y of dependent variables encodes the class membership of the independent variables. Since this is a multi-view approach, many training samples belong to the same subject, therefore Y includes vectors of all "1" on the diagonal, and of "0" elsewhere. In this way, the matrix B of regression coefficients becomes the weighting matrix for class membership, which can be used to combine the columns of X to represent the different class membership in Y. Matrix B can be computed by the classical nonlinear iterative partial least squares (NIPALS) algorithm, and after this the regression equation can be explained as an approximation equation from the training data. In PLSD approach, PLS is first used for data dimension reduction and components extraction, then classification is performed through methods such as logistic discrimination, SVM, ANN, etc. The aim of the cited work is to study the actual classification performance of partial least square representation, after the extraction of partial least square latent components. When a new test sample is submitted for classification, the system computes its regression response estimation for all classes, and then the residual between the estimation and the expected class membership response "1". The test sample is assigned to the class producing the smallest residual.

It is interesting to consider the results in [4] about the experimental assessment of ear symmetry, and therefore related to the multi-view approach. It is well known that the human body, and in particular the face, are not perfectly symmetrical. Asymmetry can reach different levels, and also be exaggerated for artistic purposes, like the famous artist Antonio De Curtis (Totò). Along the same direction, a further result from research is that human irises from the same individual are completely different. Therefore it is straightforward to wonder if this holds for ears too. The results in the mentioned study aim at a twofold effect: to understand the possibility of matching the left and right ears of an individual, and to assess the feasibility of reconstructing portions of the ear that may be occluded in a surveillance video. Both symmetry operators and Iannarelli's measurements are exploited. All experiments use the WVU Ear Database. The WVU ear database consists of 460 video sequences of 402 different subjects, and has multisequences for 54 subjects. Ear images are extracted from video sequences. The starting frame of each video captures the left profile (0°) of the subject, and takes about 2 min to end at the right profile (180°). The database includes 55 subjects with eyeglasses, 42 subjects with earrings, 38 subjects with partially occluded ears, and 2 subjects with fully occluded ears. It is not listed in Sect. 6.1.4 since it is not still available. As for the first kind of experiments, the authors concatenate the left and right ears and then apply the symmetry transform suggested by Reisfeld et al. in [80]. Such transform assigns a symmetry magnitude to every pixel in the image to construct a symmetry map. In [4] symmetry maps are built for concatenated left and right ears, and also perfect symmetry maps, by concatenating each right ear and its mirror image. Both visual comparison and distance measurement of the two maps for the same subject show that they are very similar, and this suggests symmetry for most individuals. On the other hand, experiments with Iannarelli's measures highlight that asymmetries are mostly located in specific areas, in particular they are related to the width of the helix rim of the concha. The authors also present several experiments to assess the ear symmetry from a biometric recognition system perspective. A small preliminary experiment with a similar setting had also been presented in the earlier work in [109]. It exploits images from 119 subjects under two poses. The right ear of the subject is used as the gallery and the left ear is used as the probe. The results highlight that for most people the left and right ears are quite symmetric, but for a number of subjects the left and right ears have different shapes. In the experiments in [4] the authors compare the performance obtained by matching right ear images in the gallery with right ear images used as probes, and the performance obtained by matching left ear images in the gallery with reflected right ear images. From the results of these two experiments, they estimate that the effect of ear asymmetry affects the recognition rate by a magnitude between 5.94 and 7.84 % in verification mode. Experiments in identification mode show that

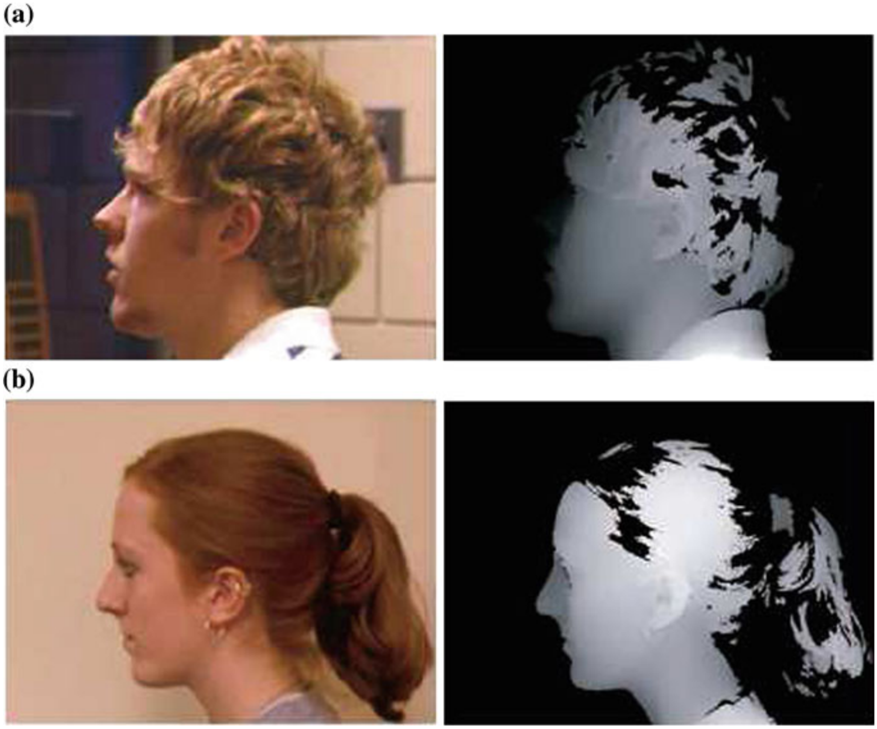

Fig. 6.13 Examples of aligned color and range images of the same subject

the asymmetry of the two ears does affect the identification performance, where the rank-1 accuracy drops by an absolute value of ~35 %.

6.4 Recognition of 3D Models

We already mentioned in Sect. 6.1.3 the seminal works by Chen and Bhanu [16, 29], and by Yan and Bowyer [108, 111] on 3D ear recognition. Figure 6.13 shows two examples of aligned color and range images used for feature extraction and fusion.

We will describe here some different techniques. However, before going further, it is interesting to mention the results reported in [98] and regarding a series of tests aiming at a comparison of the recognition performance of 3D face, 3D ear, and 3D finger surface. Identification experiments exploit the Iterative Closest Point algorithm on a multi-modal biometric dataset of multiple range images collected from 85 individuals. The images are part of the multi-modal biometric database of the University of Notre Dame. As one may expect, 3D face achieves the best performance with a rank-1 recognition rate of approximately 93 % followed by both 3D ear and

3D finger with 80 %. However, looking at the CMC curve for ear one can notice that it stays constantly lower than the CMC curve for finger: at rank 20 finger reaches about 97 %, while ear hardly exceeds 90 %. On the other hand, finger recognition is supported in these tests by more information. Finger experiments employ score-level fusion of the three separate matching scores obtained by comparing probe images of the index, ring, and middle fingers to corresponding gallery images.

An important issue to consider is that real-life biometric applications, especially online ones, require algorithms that are both robust, like 3D-based ones, but also efficient, in order to scale with the increasing size of involved databases. The biometric keys extracted from 3D ear models must be easily comparable, and this essentially requires that possibly computationally expensive steps should be only performed during preprocessing and not during normal operation.

6.4.1 Methods Based on 3D Acquisition

The recognition scheme presented in [62] aims at satisfying fast and low-cost practical applications. It exploits a fast slice curve matching for identification based on 3D ear point clouds. Acquisition is based on sequential laser strips, as described in Sect. 6.2.1. Unfortunately, the acquisition is performed by putting the ear into an oval hole, so that any segmentation problem is avoided. For this reason, it is impossible to exploit the approach as is within an under-controlled setting. Isolated points are pruned, based on cloud density and on point-to-point distances. After further noise removal and Gaussian smoothing, PCA is applied to the point cloud to extract the principal directions of the three largest spreads of distribution of these points. The three principal directions are matched to a set of fixed specific directions, so obtaining a pose normalization. A plane is used to cut the 3D ear models. Such plane is perpendicular to the principal axis of ear point clouds, and cutting will provide a slice of the contour of the ear. Thanks to pose normalization, it is easy to determine the principal axis in a consistent way. A series of slices at different locations will reflect different characteristic parts of the 3D ear shape. If two 3D models are similar, then the slice curves at the same locations will be similar too. After some experimentation, the chosen number of slices is 17. The curvature information extracted from each slice is stored in a feature vector together with an index indicating the slice position within the 3D model. A similarity measure is computed by searching the longest common sequence between two slice curves with similar indexes. On a home-collected dataset of 50 subjects, with 4 samples each, the proposed slice curve matching approach can reach 94.5 % rank-one recognition rate.

In order to achieve fast 3D matching during biometric recognition, the authors of [73] propose to exploit an annotated ear model (AEM) that is representative of human ears. This purely geometrical model is used to register each ear in a dataset and to acquire its shape through a fitting process. The 3D information extracted from the fitted model is stored as metadata. This representation is compact and directly comparable, thus making the method robust and efficient. The AEM needs to be created only once and is based on statistical data. Moreover, the used polygonal 3D mesh does not represent the whole ear, but only the inner region around concha, since this is the one that is most often free from hair. During enrollment, raw captured data are preprocessed and segmented, and then registered to the AEM. This expensive as well as critical (for future recognition) step is performed only once, when a new template is added to the database, and not for all matching operations. A better accuracy is achieved by applying two different registration algorithms in sequence: the Iterative Closest Point (ICP) first and a fine tuning one afterwards. The annotated model is fitted to the data, and then a biometric key is extracted using geometry information and stored in the database as metadata. During authentication metadata retrieved from the database are directly compared using a light distance metric. However, it is to say that such point is not completely clear, because in real applications one needs to process probes in order to obtain matchable features, and this step seems excluded by the way the approach is presented. Test for both time and accuracy performance are conducted on Notre Dame database (UND) augmented with a further home-collected dataset. The reported processing time is about 30 s for each enrollment operation, but falls to 1 ms per comparison during authentication, which was one of the goals for the approach. For the UND database, the rank-one rate is 93.9 %.

6.4.2 Video-Based Methods

The first approaches to 3D ear recognition all use 3D range data. The work in [22] is an attempt to use video to derive 3D structure. In real-world applications, video is of course much more feasible than range data due to the lack of special equipment, and can be exploited also in normal video-surveillance settings. For this reason we will devote some more space to describe this approach. The authors present two different systems for 3D ear recognition from video. The first one is based on structure from motion (SFM), and relies on four steps: preprocessing, three-dimensional reconstruction, post-processing, and recognition. The second system is based on shape from shading (SFS), and relies on three steps only: preprocessing, three-dimensional reconstruction, and recognition (post-processing is absent). The preprocessing step is performed in the same way in both systems. Its implementation starts from the consideration that video frames are generally of lower quality than still images, since they are affected by motion blur, compression and interlacing. Therefore image enhancements is of even greater importance. In particular, artifacts must be eliminated or sufficiently smoothed to avoid tracking false features, which would lead to incorrect 3D reconstruction. In order to smooth the image artifacts, this step exploits the force field transformation method presented in [44]. The principal curvatures of the image intensities are used to detect both ridges and ravines located along the ear. Among the possible features, the location of the helix is particularly useful, since it is generally the outermost detected ridge, and is an important reference element to create a boundary for the valid ear region. After this preprocessing, the two techniques diverge. Structure from motion (SFM) differs from stereopsis because the latter attempts to reconstruct a 3D model from two slightly translated images of an object captured simultaneously (same time, different space). SFM does the same by using a sequence of video frames of the same object (different time, same space). On the other hand, both techniques try to locate corresponding features in the two or more available images, to be able to determine their position in space. Kanade-Luca-Tomasi feature tracker is used to track features across the sequence of frames. Distance between a feature in the initial frame and the corresponding candidate in a following one considers both linear translations and affine transformations, and is obtained using the root mean square (RMS) residue; when the distance raises above a given threshold the feature is abandoned. Post-processing relies on outlier filtering, to eliminate artifacts (3D points whose nearest neighbor is too far in space). Matching is divided into two steps: correspondence matching and depth matching. Correspondence matching is in turn divided into two procedures: computing a similarity between the shapes and then geometrically transforming the shapes using 2D-ICP in order to maximize the similarity. If a sufficient number of corresponding matches are found between two shapes, depth matching is invoked. It compares the similarities in terms of depth of the retained features from both point sets. It is worth noticing that the authors find that correspondence matching is very useful to filter out insufficiently matching database models in the (x, y) domain before performing depth matching, since they observe that the locations of features are quite unique to each individual. For this reason, a significant number of database models would not reach the second step. The approach using Shape From Shading (SFS) might be less interesting for fully automatic systems, since it requires to perform the 3D reconstruction process on a single image, manually chosen among those with an optimal view of the entire ear region. As a matter of fact, SFS is defined as a process which recovers the 3D (visible) surface of a 3D object from the shading in its 2D intensity image, by using information about the reflectance and illumination properties of the scene. However, an automatic though computationally demanding procedure to select the most stable 3D model is proposed later in [23], which implies to reconstruct and compare 3D models from a number of subsequent frames. Returning to [22], in the second system, the recognition phase first involves finding a closest match between a database and test model via 3D-ICP and then computing the similarity measure of the two models using an RMS difference. Experiments exploit the above mentioned collection of video-clips by West Virginia University (WVU). With SFM, performance by using from 10 to 50 frames for reconstruction, results in an optimal recognition rate of about 95 % using 16 frames. As a possible explanation, using less frames may cause missing information, while using more may cause more noise. These experiments use the same video clip, dividing the set of frames between training and testing. Unfortunately, when incorporating a testing set from a completely different set of videos, the obtained recognition performance is inadequate. Although two different video clips contain the same subject, the detected features in each of the videos have very poor correlation to one another. The conclusion is that the SFM approach, using a sparse point map, provides more accurate relative depth information about the feature points than the SFS, however is quite limited since heavily relies on correlation amongst the feature points. On the other hand, SFM approach, which exploits a dense set of points and therefore provides a more accurate shape, is able to successfully discriminate between individuals using different video clips for training and testing (84 % recognition rate). The authors finally suggest fusing the two approaches to obtain the best from the two.

6.5 Multimodal Systems

Since this chapter is focused on the strategies to perform biometric ear recognition under uncontrolled data acquisition conditions (ideally fully covert ones), we will leave out all those multimodal proposals which would imply an aware user participation, e.g. recognition involving some combination of ear and palmprint ([37] or of ear and fingerprint [2, 39]) or of ear and signature [67]. While these systems improve performance of identification for aware and somehow collaborative users, they are of course not feasible for the kind of settings that we are addressing here. A more complete review of multimodal systems involving ear recognition can be found in [5, 74]. The systems that we will shortly describe rather join ear recognition with gait or face, especially side face, which can be acquired even at a reasonable distance without user participation.

On the other hand, it is also to mention the possibility, as for example with face, to use ear in a wider interpretation of multimodality, i.e. to combine ear+ear in some way. We will first mention these approaches.

A final aspect to anticipate is that, in many cases, the true novelty of presented methods is in the kind of fusion they adopt, rather than in the techniques exploited for the single modalities. Fusion in a multibiometric system may be performed at feature-level, which implies to combine the different templates which must have a compatible structure, and to train a specific system on the obtained combination. This approach is the most expensive one, but fully preserves the original information. A second choice is to perform fusion at matching score or ranking level. In general, this is the preferred solution, because already existing unimodal systems can be exploited and enough information is still present in the result (e.g., in identification this is usually a list of scores or ranks). Finally, fusion can be performed at decision level, which is the cheapest one; however, any useful information to possibly identify problems from one system has been lost.

6.5.1 Matching and Fusion of Different Ear Aspects

Yan and Bowyer analyze in [109] the performance of 2D and 3D ear recognition under different conditions. The images belong to the initial core acquired at the University of Notre Dame. The approaches considered include PCA (widely known as "eigen-ear") with 2D intensity images, achieving 63.8 % rank-one recognition; PCA with range images (often referred as 2.5D images), achieving 55.3 %; Hausdorff matching of edge images extracted from range images, achieving 67.5 %; and ICP matching of the 3D data, that with some improvements passes from 84.1 to 98.7 %. ICP-based matching achieves the best performance, and also shows good scalability with size of datasets.

After this preliminary study, the same authors investigate the effect of fusing the results from the above techniques [110]. They test multi-modal (2D+3D), multi-algorithm (PCA+ICP) and multi-instance (two images for both enrollment and testing) modalities. All provide an improvement over a single biometrics. Multi-modal combinations include 2D PCA with 3D ICP, 2D PCA with 3D PCA, and 2D PCA with 3D edge-based approach. The best results are achieved by multi-modal 2D PCA fused with 3D ICP. To combine 2D PCA-based and 3D ICP-based ear recognition, the authors propose a fusion rule at matching score level based on the interval distribution between rank-1 and rank-2 recognition rates. The rank-1 recognition rate achieves 91.7 % with 302 subjects in the gallery. For the multi-algorithm tests, three different algorithms are used to process 3D data: ICP-based algorithm, PCA-based algorithm and edge-based algorithm. After score normalization, the weighted sum rule is used for combinations. The best performance is achieved when combining ICP and edge-based algorithm on the 3D data (90.2 %). In general, all the approaches perform much better in multi-instance experiments with multiple images used to represent one subject. ICP approach used to match a two-image-per-person probe against a two-image-per-person gallery reaches the highest rank-1 recognition rate of 97 % with two images in the gallery and two used as probe.

Along a similar ear-ear combination line, in [13] a kind of feature-level fusion is proposed for verification. SIFT is used to extract relevant features from ear images at different poses. Features are then merged according to an appropriate fusion rule to produce a single feature template called the Fused Template. The similarity of SIFT features extracted from the live image and Fused Template of the enrolled user is measured by their Euclidean distance. IITK database in a version with 1060 images is used to validate the performance of the method. The images are collected from a distance of 2.5 m. The camera is moved circularly and the subject is at the center. Orientation of the camera facing the ear in frontal pose is considered as 0°. The ear images are captured at −40°, −20°, +0°, +20° and +40° placing the camera tripod at fixed landmark positions. Only images of right ear are collected. Two images per pose (angle) are obtained in a short interval. There are 10 images per subject. The database contains 1060 images from 106 subjects. The images obtained are normalized to 648 × 486. Features are extracted from the ear images at −40°, +0° and +40°. The pose variation is chosen so to reduce the correlation and get a more complete representation of the subject. Illumination is normalized by histogram fitting, taking the image at 0° as reference. The dimension of the final feature vector is reduced by eliminating redundant points between the fusing images, i.e. those found in overlapping regions of the different images. Surviving keypoints are concatenated to form a composite feature template. During matching, the SIFT features of each keypoint p_i of the probe image are matched to those of every keypoint t_j of the template in the database. The pair of matching keypoints (p_k, t_l) with minimum distance is removed. The matching process is continued for the remaining points until no pair is found. The decision of whether the probe template is matched or not depends on the number of matching points. Comparing performance with non-fused and fused templates, FAR decreases from 11.75 % down to 6.51 %, FRR from 11.59 to 2.86 %, and accuracy increases from 88.32 to 95.32 %.

6.5.2 Matching and Fusion of Ear and Face

Victor, Bowyer and Sarkar [90] were perhaps the first to compare the performance of ear and face biometrics. They used PCA for both traits, and the final result of their experiments was that face is better for recognition, despite possible expression and other kinds of variations like make up and beard which are not found in ear. However, they did not study the effect of combining the two biometrics.

It is worth noticing that, in principle, approaches fusing ear and side face are realistically more suitable for uncontrolled recognition, since the two traits can be acquired in one shot (actually, there is only one image including the two traits) and any kind of fusion is often avoided. Fusing frontal face and ear requires two shots instead, and either a controlled setting or a capture setting assuring images from two points of view (frontal and side) on the subject to be recognized (with only one device we should be sure that sooner or later the subject will turn in different poses). Moreover, fusion is required at some level.

6.5.2.1 Ear and Frontal Face

Among the first results of combining face and ear biometrics we find those in the already mentioned work in [26]. Combination experiments exploit a very simple feature-level combination technique. The normalized and masked images of the ear and face are concatenated to form a combined face-plus-ear image. These new images undergo PCA processing. CMC curves for the experiments suggest that the multimodal biometrics offers a significant improvement of overall performance.

A similar approach is used in [118], with the difference that users are chimeric ones, i.e., ears are from USTB II database, while faces with different poses are from ORL database. Feature vectors are extracted from chained images using Full-Space Linear Discriminant Analysis (FSLDA). Rank-one recognition rate of multimodal biometrics is 98.7 % (apparently notwithstanding the number of dimensions) and is higher than performance of single traits. It is worth noticing that ORL faces (visually) present more significant variations than USTB II ears, and this seems to be the reason for having ear performance better than face.

Face and ear are matched separately in [54], and face is captured in frontal pose. For localization of ear region, Triangular Fossa and Antitragus are detected manually on ear image, and then used to apply a complete ear localization technique. Both face and ear are cropped from the respective images. After geometric normalization, histogram equalization is performed. Both face and ear images are convolved with Gabor wavelet filters to extract spatially enhanced Gabor features for both face and ear. The feature vectors extracted from Gabor face and ear responses can be further characterized and described by normal (Gaussian) distributions. Gaussian Mixture Model (GMM) is applied to the high-dimensional Gabor face and Gabor ear responses separately for quantitive measurements, and Expectation Maximization (EM) algorithm is used to estimate density parameters in GMM. This produces two sets of feature vectors which are then fused using Dempster-Shafer theory. The approach is tested on the IIT Kanpur multimodal database. Database of face and ear consists of 2 face and 2 ear images per person for 400 individuals. For the present evaluation, only frontal view faces are used with uniform lighting, and minor changes in facial expression. Face recognition achieves 91.96 % and ear recognition achieves 93.35 %, while fusion reaches 95.53 %. Despite the results, this method cannot be deemed feasible for uncontrolled settings due to the quality required for images. Moreover, the manual intervention by a human operator hinders its use for massive applications.

3D ear and face are fused in [24]. As for 3D ear, the approach presented in [23] is used. For the 2D face recognition, Active Shape Model technique is used to extract a set of facial landmarks. A series of Gabor filters are applied at the locations of facial landmarks, and the responses are calculated. The Gabor features are stored in the database as the face model, and are compared with those extracted from a probe image during testing. The match scores of the ear recognition and face recognition are fused at score level using a weighted sum to fuse the results after normalizing them. Weights are experimentally determined. Experiments are conducted using a gallery set of 402 video clips and a probe of 60 video clips (images). The result achieved for rank-1 identification for the 2D face recognition is 81.67 %, for 3D ear recognition 95 %, and for fusion 100 %. This methodology presents the limitations discussed in Sect. 6.4.2 regarding ear processing.

A multibiometric extension of the approach in [113] described in Sect. 6.3.2.2 is presented in [107]. The same HMAX approach is adopted for both face and ear, using different filters: face images are processed with Gabor filter and ear images with Gaussian filter. Resulting features are classified by K-NN and SVM classifiers, in order to match results. Fusion occurs on the match scores obtained from the last stages. The system is tested on a dataset which has 10 different images of each of 40 distinct chimeric users. The authors randomly pair faces from ORL and ears from USTB to obtain a multimodal dataset for each of these 40 persons. The best results with 40 classes are always obtained with SVM which achieves 88.3 % recognition rate on faces, 81.2 % on ear, and finally 89.5 % on multimodal fusion.

The work presented in [51] especially addresses identification during video conferences using face and ear. Face features include color features, computed in Hue, Saturation and Value (HSV) color space, and 2D wavelet based features, approximated by Generalized Gaussian Density (GGD). For ear, only GGD is used. The Kullback-Leibler Distance (KLD) measure is used to match GGD features from both traits on probe and gallery templates. Fusion exploits min-max normalization and sum rule. The system is tested on a dataset of 45 subsets built for this aim. As expected, performance is improved by fusion.

The work presented in [43] starts from the consideration that multimodal systems may perform even worse than unimodal ones if one or more modalities fall on degenerated data, and the adopted fusion rules do not avoid derived problems to propagate to the multimodal result. The proposed solution is an adaptive feature weighting scheme which is based on the Sparse Coding Error Ratio (SCER) index. Such index is able to measure the different reliability between face and ear images due to illumination, pose, expression (for face), occlusion, and corruption. The authors present two multimodal methods based on Space Representation based Classification (SRC), which integrate the new SCER index: Multimodal SRC (MSRC)and Multimodal RSC (MRSC), where RSC is Robust Sparse Coding. Appearance-based features of face and ear are extracted by separately applying PCA, and are then directly chained. The approach is tested against pixel corruptions, as well as face images with sunglasses or scarf, but it is not clear which are the results achieved with real pose and illumination variations.

6.5.2.2 Ear and Profile Face

Of course, profile face image may contain less discriminant information than frontal view. Nevertheless, it contains the ear, which can be used as well if the system includes this possibility, and also some complementary features with respect to frontal ones (e.g., the profile shape of the nose). Among the works addressing ear and profile face fusion, one of the earliest is presented in [115]. There is no need for a fusion step, like in [26], since the face profile silhouette is already combined with the ear in a single image. Actually, considering also the ear region is a way to overcome the limitation of many recognition techniques based on face profile. Such techniques usually rely on a number of fiducial points and on their relations. However, their reliability decreases with the peculiar appearance of some anatomical elements a concave nose, protruding lips, or a flat chin. Furthermore, the number and position of fiducial points vary with expression changes even for the same person. Full-Space Linear Discriminant Analysis (FSLDA) is therefore applied here to the profile image containing the ear region too. Experiments are performed on USTB III database. Using only 10 features, the recognition rate reaches 86.10 %. As the feature number increases to 100, the recognition rate reaches 96.2 %.

A group of similar approaches are presented in [103–105]. They all use USTB III or subsets, so that ear is contained in face profile images. However, in a series of common initial steps, ear region alone is cropped from the original image in a new one, and both ear and profile face images undergo Wiener filtering to emphasize the features. Then the ear images are resized in a proportional way. Finally, histogram equalization is used to produce an image with equally distributed intensity values. The differences among the different proposals are related to the nonlinear associated feature extraction algorithm, where the kernel matrices are computed separately from the corresponding image data. Further differences are the fusion and classification strategies adopted. In [103] polynomial kernel is used for feature extraction from ear data, and sigmoid kernel for profile face data. Then kernel canonical correlation analysis (KCCA) is used for feature fusion. The minimum-distance classifier is used for classification, and the system achieves a recognition rate of 98.68 %. Gaussian kernel is used in [104] to compute kernel matrices for both ear and profile face feature extraction. Afterwards, the authors fuse the two obtained matrices using product, average and weighted sum, and then apply kernel Fisher discriminant analysis (KFDA) to the fused matrices. In experiments KFDA is applied to single biometrics to compare their behavior with the one achieved by fusion. Ear and profile face alone achieve respectively 91.77 and 93.46 % recognition rate, while fusion results are 96.41 % for Average rule, 96.20 % for Product rule and 96.84 % for Weighted-sum rule. The further different system in [105] uses FSLDA to extract features for both traits. Decision level fusion of ear and profile face is carried out using the combination methods of Product, Sum and Median rules according to the Bayesian theory, and to a modified Vote rule for two classifiers. The latter is adopted to suitably support the setting with two classifiers: if each classifier assigns the given sample to a different class, the votes would appear equal so that the system would not be able to make a decision. With no combination, the ear achieves 94.05 % recognition rate compared with 88.10 % of profile face. As for the combination, recognition rate is 96.43 % for Product rule, 97.62 % for Sum rule, 97.62 % for Median rule, and 96.43 % for the Modified Vote rule.

Differently from the above approaches, the work in [92] adopts a kind of pose normalization strategy before further processing. PCA and KPCA are used to represent subspaces and to extract features of ear and face images. The basic idea underlying pose transformation is that the feature spaces corresponding to ear and face images in some pose are transformed respectively into those of a perfectly frontal ear and a perfectly profile face using pose transformation matrices. Then the generated feature sets of ear and face are fused. The paper adopts three fusion modes, namely serial strategy, parallel strategy (the two feature vectors to fuse play the role of the real and complex part of a vector in the real space), and kernel canonical correlation analysis (KCCA). The nearest neighbor classifier is used, and tests are performed on USTB III database. Experimental results for ear and face unimodal traits highlight better performance with pose transformation and KPCA for both traits, with a minimum achieved at 45° rotation of respectively 60 and 30 % recognition rate. The advantages of pose transformation are more evident with greater rotations. Results on fusion show that serial strategy is the best. It is to say that a clear improvement after fusion is not always evident.

6.5.3 Ear and More

A mixed solution between fully controlled settings and the relaxation of some related constraint is the availability of an equipped “biometric tunnel”. An example is the University of Southampton Multi-Biometric Tunnel presented in [85]. It is a constrained environment designed to address the requirements of airports and other high throughput settings. It can acquire a number of contactless biometrics in a non-intrusive manner and with not much participation by the user, who is aware of the acquisition process anyway. The system uses synchronized cameras to capture gait and additional cameras to capture images from the face and one ear, as an individual walks through the tunnel. The path allowed to the subject is quite narrow and inside a volume where illumination and other conditions are controlled. As for the ear, a digital photograph is taken when a subject passes through a light beam at the end of the tunnel. The camera uses a wide field of view to accomodate a large range of subject heights and walking speeds. A high shutter speed minimizes motion blur. In addition, two flash cameras provide sufficient light for the shutter speed and reduce the presence of possible shadows, since the flash guns are positioned to point from above and below the ear. The system has been tested with the fusion of gait and ear in [83]. Ear recognition follows the same approach of the already cited work in [21]. The rank-1 recognition performance for visible ears is 77 %. This is lower than the performance in previous publications due to the less constrained ear images, which include wider occlusion. Score fusion is used to combine gait and ear recognition results. The distance measures returned by each algorithm are normalised using an estimate of the offset and scale of the measures between different subjects. Rank-1 recognition increases to 86 % fusing ear and gait with sum rule.

Ear, face and gait are exploited in [114]. For each modality, feature extraction step creates a Gabor wavelet representation, which then undergoes principal component analysis (PCA). During matching step, after normalization, fusion is performed at score level, testing both weighted sum and weighted product. Chimeric users are created from three different databases for the three modalities. For face images, ORL face database includes 120 images corresponding to 40 subjects with three images per person, taken at different times, under different illumination conditions and different facial expression. Ear images come from USTB ear database, from where 120 images corresponding to 40 subjects with three images per person are extracted. Gait silhouette images come from CASIA gait database, from which 120 average silhouette images over one gait cycle were extracted from video sequences corresponding to 40 subjects with three average silhouette images per person. The best unimodal performance is achieved for all three modalities by setting the Gabor parameters orientation ad scale respectively at $\mu = 7$ and $\nu = 4$. At False Acceptance Rate (FAR) set at 0.1 %, face achieves 65 % Genuine Acceptance Rate (GAR), ear achieves 82.5 %, and gait achieves 72.5 %. With the same Gabor parameters, multimodal fusion achieves the best performance with z-mean normalization and weighted product reaching 97.5 % GAR at 0.1 % FAR.

In [65], face, ear and iris scores are fused using a fuzzy approach. In the identification phase, the input face and ear images are compared with the gallery images by Fisherfaces and Fisherears, and measuring the Euclidian distance. For iris, they calculate the Hamming distance between the iris codes. In each of the three cases, five identifiers are obtained as output that will be ranked according to their distances. The ranked identities of the first five subjects returned by each of the three lists are then passed to the fuzzy fusion module along with their normalized scores. A further contribution comes from the normalized scores from soft biometrics on the face, namely gender, ethnicity and color of the eyes. During identification, the soft biometric information of input identifiers is fed into the fusion module. Datasets used are CASIA 1.0 for the iris, USTB for the ear and FERET for the face. Fuzzy rules are used during fusion. For a FAR of 1 %, the GAR achieved by fuzzy fusion reaches 98.13 %, while face, ear and iris respectively achieve GAR of 94, 95.2 and 89 %. Not only performance of the multimodal system is higher, but the fusion method is better than other methods tested with the same settings.

6.6 Conclusions

Unconstrained ear processing still poses a number of research problems. Differently from face, ear cannot undergo expression variations. However, due to its strong three-dimensional characterization, its correct detection and recognition can be significantly affected by illumination, pose and occlusion. Moreover, the similar texture and color distribution of the region around the ear, as well as the noise introduced by hair, can further hinder detection. A still open problem is the correct ear detection in uncontrolled settings. Some works in literature like [57] propose special acquisition stands with a hole allowing to capture the ear in a very precise way. However, such approaches seem to rather aim at detaching the recognition problem from the problem of detection, that unavoidably also affects the following processing steps. In real world applications, e.g., video surveillance of a controlled environment, this kind of detection is completely unfeasible.

While illumination seems easier to address, through histogram equalization or other techniques, it may still modify the 3D appearance of the ear, so affecting both detection and recognition. In the same way, occlusion can hinder both a correct localization and a reliable recognition. A possible solution for a correct detection is for example the use of localized region based active contours model [57] and similar approaches where techniques usually applied to the whole image are made local to be more robust to this kind of problems. Even for recognition, the best way to address occlusions seems to be a localized approach, such as HERO [35], or modular neural networks [41].

Pose variations are the source of the most difficult detection and recognition problems. Illumination and occlusion usually affect only limited regions, so that a localized approach can effectively exploit the remaining ones. Ear is divided into regions, with or without anatomical valence, and results from the different regions are fused for better robustness. However, pose variations globally modify the overall appearance of the ear. 3D techniques seem to be the most effective way to solve these problems, since they rely on a 3D model expressing the full structure of the ear in a way independent from rotations along any axis. However, 3D techniques are still very expensive, and most of all cannot still be exploited without user participation. A promising alternative for detection is Gabor Jets, which is exploited in [96], while for recognition a solution seems to be provided by methods exploiting non linear embeddings, like LLE and Improved LLE [101, 102], and NPE [120], or by multiview approaches. In particular, video sequences are exploited in the Structure From

Motion technique proposed in [22], which gives good results also on "hard" datasets. The physical proximity of face and ear, if not the true inclusion in side images, further suggest the use of the biometric fusion of the two modalities. Images can be fused at the feature level [103], or at the score level as in most approaches.

A final note regards the acquisition environment of existing datasets. Images as well as video sequences are (almost) always captured indoor, with a uniform background, and usually at a reasonable distance. In a similar way, (almost) all systems in literature have been tested in these "easier" conditions, where environmental noise is often at a minimum. Ear biometrics applied outdoors and at a distance may pose further problems, mainly for a correct detection which also affects a correct recognition.

References

1. Abate AF, Nappi M, Riccio D, Ricciardi S (2006) Ear recognition by means of a rotation invariant descriptor. In: Proceedings of the 18th international conference on pattern recognition–ICPR 2006, vol 4, pp 437–440
2. Abate AF, Nappi M, Riccio D, De Marsico M (2007) Face, ear and fingerprint: designing multibiometric architectures. In: Proceedings of 14th international conference on image analysis and processing–ICIAP 2007, pp 437–442
3. Abaza A, Hebert C, Harrison MAF (2010) Fast learning ear detection for real-time surveillance. In: Proceedings of the 4th IEEE international conference on biometrics: theory applications and systems–BTAS 2010, pp 1–6
4. Abaza A, Ross A (2010) Towards understanding the symmetry of human ears: a biometric perspective. In: Proceedings of the 4th IEEE international conference on biometrics: theory applications and systems–BTAS 2010, pp 1–7
5. Abaza A, Ross A, Harrison MAF, Nixon MS (2013) A survey on ear biometrics. ACM Comput Surveys 45(2), Article 22
6. Abdel-Mottaleb M, Zhou J (2006) Human ear recognition from face profile images. In: Zhang D, Jain AK (eds) Proceedings of the international conference on biometrics–ICB 2006, LNCS 3832, pp 786 – 792
7. Akkermans AHM, Kevenaar TAM, Schobben DWE (2005) Acoustic ear recognition for person identification. In: Proceedings of the AutoID'05, pp 219–223
8. Alvarez L, Gonzalez E, Mazorra L (2005) Fitting ear contour using an ovoid model. In: Proceedings of the 39th annual international carnahan conference on security technology–CCST '05, pp 145–148
9. Ansari S, Gupta P (2007) Localization of ear using outer helix curve of the ear. In: Proceedings of the international conference on computing: theory and applications–ICCTA 2007, pp 688–692
10. Arbab-Zavar B, Nixon MS, Hurley DJ (2007) On model-based analysis of ear biometrics. In: Proceedings of the 1st IEEE international conference on biometrics: theory, applications, and systems–BTAS 2007, pp 1–5
11. Arbab-Zavar B, Nixon MS (2008) Robust log-gabor filter for ear biometrics. In: Proceedings of the 19th international conference on pattern recognition–ICPR 2008, pp 1–4
12. Bach FR, Jordan MI (2003) Kernel independent component analysis. J Mach Learn Res 3:1–48
13. Badrinath GS, Gupta P (2009) Feature level fused ear biometric system. In: Proceedings of the 7th international conference on advances in pattern recognition–ICAPR '09, pp 197–200
14. Battisti F, Carli M, De Natale FGB, Neri AA (2012) Ear recognition based on edge potential function. In: Proceedings of the SPIE 8295, image processing: algorithms and systems X; and parallel processing for imaging applications II, Feb 9, p 829508. doi:10.1117/12.909082

15. Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features. In: Proceedings of the 9th european conference on computer vision–ECCV 2006, pp 404–417
16. Bhanu B, Chen H (2003) Human ear recognition in 3D. In: Proceedings of multimodal user authentication workshop (MMUA), Santa Barbara, CA, pp 91–98
17. Brown M, Lowe DG (2002) Invariant features from interest point groups. In: Proceedings of the 13th British machine vision conference, pp 253–262
18. Burge M, Burger W (1997) Ear biometrics for machine vision. In: Proceedings of 21 workshop of the Austrian association for pattern recognition
19. Burge M, Burger W (1998) Ear biometrics. In: Jain AK, Bolle R, Pankanti S (eds) Biometrics: personal identification in networked society. Kluwer Academic Publishers, Boston, pp 273–286
20. Bustard JD, Nixon MS (2008) Robust 2D ear registration and recognition based on SIFT point matching. In: 2nd IEEE international conference on biometrics: theory, applications and systems–BTAS 2008, pp 1–6
21. Bustard JD, Nixon MS (2010) Toward unconstrained ear recognition from two-dimensional images. IEEE Trans Syst Man Cyber Part A Syst Human 40(3):486–494
22. Cadavid S, Abdel-Mottaleb M (2007) Human identification based on 3D ear models. In: Proceedings of the 1st IEEE international conference on biometrics: theory, applications, and systems–BTAS 2007, pp 1–6
23. Cadavid S, Abdel-Mottaleb M (2008) 3-D ear modeling and recognition from video sequences using shape from shading. IEEE Trans Info Forensics Security 3(4):709–718
24. Cadavid S, Mahoor MH, Abdel-Mottaleb M (2009) Multi-modal biometric modeling and recognition of the human face and ear. In: Proceedings of the IEEE international workshop on safety, security and rescue robotics–SSRR 2009, pp 1–6
25. Cai J, Goshtasby A (1999) Detecting human faces in color images. Image Vision Comput 18(1):63–75
26. Chang K, Victor B, Bowyer KW, Sarkar S (2003) Comparison and combination of ear and face images in appearance-based biometrics. IEEE Trans Pattern Anal Mach Intell 25(8):1160–1165
27. Chen H, Bhanu B (2004) Human ear detection from side face range images. In: Proceedings of the international conference on pattern recognition (ICPR 2004), vol 3, pp 574–577
28. Chen H, Bhanu B (2005) Shape model-based 3D ear detection from side face range images. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition–workshops (CVPR 2005), p 122
29. Chen H, Bhanu B (2007) Human ear recognition in 3D. IEEE Trans Pattern Anal Mach Intell 29(4):718–737
30. Chen H, Bhanu B (2009) Efficient recognition of highly similar 3D objects in range images. IEEE Trans Pattern Anal Mach Intell 31(1):172–179
31. Choraś M, Choraś RS (2006) Geometrical algorithms of ear contour shape representation and feature extraction. In: Proceedings of the intelligent systems design and application–ISDA 2006, IEEE CS Press, vol II, pp 451–456, Jinan, China
32. Cummings AH, Nixon MS, Carter JN (2010) A novel ray analogy for enrollment of ear biometrics. In: Proceedings of the 4th IEEE international conference on biometrics: theory applications and systems–BTAS 2010, pp 1–6
33. Curless B (1999) From range scans to 3D models. ACM SIGGRAPH, Comput Graph 33(4): 38–41
34. Debevec P (1999) Image-based modeling, rendering and lighting. Comput Graph 33(4):46–50
35. De Marsico M, Michele N, Riccio D (2010) HERO: human ear recognition against occlusions. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops–CVPRW 2010, pp 178–183
36. Dong J, Mu Z (2008) Multi-pose ear recognition based on force field transformation. In: Proceedings of the 2nd international symposium on intelligent information technology application–IITA 2008, vol 3, pp 771–775

37. Faez K, Motamed S, Yaqubi M (2008) Personal verification using ear and palm-print biometrics. In: Proceedings IEEE international conference on systems, man and cybernetics–SMC 2008, pp 3727–3731
38. Fleck M, Forsyth D, Bregler C (1996) Finding naked people. In: Proceedings of the European conference on computer vision–ECCV 1996, vol 2, pp 592–602
39. Gnanasivam P, Muttan S (2011) Ear and fingerprint biometrics for personal identification. In: Proceedings of the international conference on signal processing, communication, computing and networking technologies–ICSCCN 2011, pp 347–352
40. Grabham NJ, Swabey MA, Chambers P, Lutman ME, White NM, Chad JE, Beeby SP (2013) An evaluation of otoacoustic emissions as a biometric. IEEE Trans Info Forensics Security 8(81):174–183
41. Gutierrez L, Patricia M, Lopez M (2010) Modular neural network integrator for human recognition from ear images. In: Proceedings of the 2010 international joint conference on neural networks–IJCNN 2010, pp 1–5
42. Huang C, Lu G, Liu Y (2009) Coordinate direction normalization using point cloud projection density for 3D ear. In: Proceedings of the 4th international conference on computer sciences and convergence information technology–ICCIT 2009, pp 511–515
43. Huang Z, Liu Y, Li C, Yang M, Chen L (2013) A robust face and ear based multimodal biometric system using sparse representation. Pattern Recogn 46(8):2156–2168
44. Hurley DJ, Nixon MS, Carter JN (1999) Force field energy functionals for image feature extraction. In: Proceedings of the British machine vision conference 1999–BMVC99 BMVA, pp 604–613
45. Hurley DJ, Nixon MS, Carter JN (2000) Automatic ear recognition by force field transformations. IEE colloquium on visual biometrics (Ref.No. 2000/018), pp 7/1–7/5
46. Hurley DJ, Nixon MS, Carter JN (2002) Force field energy functionals for image feature extraction. Image Vision Comput 20(5–6):311–317
47. Hurley DJ, Nixon MS, Carter JN (2005) Force field feature extraction for ear biometrics. Comput Vision Image Underst 98:491–512
48. Iannarelli A (1989) Ear identification. In forensic identification series. Paramont Publishing Company, Fremont
49. Islam SMS, Bennamoun M, Davies R (2008) Fast and fully automatic ear detection using cascaded adaBoost. In: Proceedings of the IEEE workshop on applications of computer vision–WACV 2008, pp 1–6
50. Jain A, Bolle R, Pankanti S (eds) (1998) Biometrics: personal identification in networked society. Kluwer Academic Publishers, Boston
51. Javadtalab A, Abbadi L, Omidyeganeh M, Shirmohammadi S, Adams CM, El-Saddik A (2011) Transparent non-intrusive multimodal biometric system for video conference using the fusion of face and ear recognition. In: Proceedings of the 9th annual international conference on privacy security and trust–PST 2011, pp 87–92
52. Jeges E, Mate L (2006) Model-based human ear identification. In: Proceedings of the world automation congress–WAC, pp 1–6
53. Kisku DR, Mehrotra H, Gupta P, Sing JK (2009) SIFT-based ear recognition by fusion of detected keypoints from color similarity slice regions. In: Proceedings of the international conference on advances in computational tools for engineering applications–ACTEA 2009, pp 380–385
54. Kisku DR, Sing JK, Gupta P (2009) Multibiometrics belief fusion. In: Proceedings of the 2nd international conference on machine vision–ICMV 2009, pp 37–40
55. Kisku DR, Gupta S, Gupta P, Sing JK (2010) An efficient ear identification system. In: Proceedings of the 5th international conference on future information technology–futuretech 2010, pp 1–6
56. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK (2012) Face recognition performance: role of demographic information. IEEE Trans Info Forensics Security 7(6):1789–1801
57. Kumar A, Hanmandlu M, Kuldeep M, Gupta HM (2011) Automatic ear detection for online biometric applications. In: Proceedings of the 3rd national conference on computer vision, pattern recognition, image processing and graphics–NCVPRIPG 2011, pp 146–149

58. Kumar A, Wu C (2012) Automated human identification using ear imaging. Pattern Recogn 45:956–968
59. Lades M, Vorbruggen JC, Buhmann J, Lange J, Von Der Malsburg C, Wurtz RP, Konen W (1993) Distortion invariant object recognition in the dynamic link architecture. IEEE Trans Comput 42(3):300–311
60. Liu W, Wang Y, Li SZ, Tan T (2004) Null space-based kernel fisher discriminate analysis for face recognition. In: Proceedings of the 6th IEEE international conference on automatic face and gesture recognition–FG 2004, pp 369–374
61. Liu H (2011) Multi-view ear recognition by patrial least square discrimination. In: Proceedings of the 3rd international conference on computer research and development–ICCRD 2011, vol 4, pp 200–204
62. Liu H, Zhang D (2011) Fast 3D point cloud ear identification by slice curve matching. In: Proceedings of the 3rd international conference on computer research and development–ICCRD 2011, vol 4, pp 224–228
63. Lu L, Zhang X, Zhao Y, Jia Y (2006) Ear recognition based on statistical shape model. In: Proceedings of the IEEE international conference on innovative computing, information and control, vol 3, pp 353–356
64. Messer K, Matas J, Kittler J, Luettin J, Maitre G (1999) XM2VTSDB: the extended M2VTS database. In: Proceedings of the international conference on audio- and video-based person authentication–AVBPA '99, pp 72–77
65. Monwar MM, Gavrilova M, Wang Y (2011) A novel fuzzy multimodal information fusion technology for human biometric traits identification. In: Proceedings of the 10th IEEE international conference on cognitive informatics and cognitive computing–ICCI*CC 2011, pp 112–119
66. Morano RA, Ozturk C, Conn R, Dubin S, Zietz S, Nissanov J (1998) Structured light using pseudorandom codes. IEEE Trans Pattern Anal Mach Intell 20(3), pp 322–327
67. Monwar M, Gavrilova M (2008) FES: a system for combining face, ear and signature biometrics using rank level fusion. In: Proceedings of the 5th international conference on information technology: new generations–ITNG 2008, pp 922–927
68. Moreno B, Sanchez A, Velez JF (1999) On the use of outer ear images for personal identification in security applications. In: Proceedings of the IEEE 33rd annual international carnahan conference on security technology, pp 469–476
69. Nanni L, Lumini A (2007) A multi-matcher for ear authentication. Pattern Recogn Lett 28(16):2219–2226
70. Nanni L, Lumini A (2009) Fusion of color spaces for ear authentication. Pattern Recogn 42(9):1906–1913
71. Nanni L, Lumini A (2009) A supervised method to discriminate between impostors and genuine in biometry. Expert Syst Appl 36(7):10401–10407
72. Nosrati MS, Faez K, Faradji F (2007) Using 2D wavelet and principal component analysis for personal identification based On 2D ear structure. In: Proceedings of the international conference on intelligent and advanced systems–ICIAS 2007, pp 616–620
73. Passalis G, Kakadiaris IA, Theoharis T, Toderici G, Papaioannou T (2007) Towards fast 3D ear recognition for real-life biometric applications. In: Proceedings of the IEEE conference on advanced video and signal based surveillance–AVSS 2007, pp 39–44
74. Pflug A, Busch C (2012) Ear biometrics: a survey of detection, feature extraction and recognition methods. IET Biometrics 1(2):114–129
75. Prakash S, Jayaraman U, Gupta P (2009) A skin-color and template based technique for automatic ear detection. In: Proceedings of the 7th international conference on advances in pattern recognition–ICAPR 2009, pp 213–216
76. Prakash S, Jayaraman U, Gupta P (2009) Connected component based technique for automatic ear detection. In: Proceedings of the 16th IEEE international conference on image processing–ICIP 2009, pp 2741–2744
77. Prakash S, Gupta P (2011) An efficient ear recognition technique invariant to illumination and pose. Telecommun Syst 52(3):1–14 http://dx.doi.org/10.1007/s11235--011-9621-2

78. Prakash S, Gupta P (2012) A rotation and scale invariant technique for ear detection in 3D. Pattern Recogn Lett 33, pp 1924–1931
79. Raposo R, Hoyle E, Peixinho A, Proenca H (2011) UBEAR: A dataset of ear images captured on-the-move in uncontrolled conditions. In: Proceedings of the IEEE workshop on computational intelligence in biometrics and identity management–CIBIM 2011, pp 84–90
80. Reisfeld D, Wolfson H, Yeshurun Y (1995) Context-free attentional operators: the generalized symmetry transform. Int J Comput Vision 14(2):119–130
81. Riccio D, Tortora G, De Marsico M, Wechsler H (2012) EGA-ethnicity, gender and age, a pre-annotated face database. In: Proceedings of the 2012 IEEE workshop on biometric measurements and systems for security and medical applications–BioMS 2012, pp 38–45
82. Said EH, Abaza A, Ammar H (2008) Ear segmentation in color facial images using mathematical morphology. In: Proceedings of the biometrics symposium–BSYM 2008, pp 29–34
83. Samangooei S, Bustard JD, Seeley RD, Nixon MS, Carter JN (2011) Acquisition and analysis of a dataset comprising gait, ear and semantic data. In: Bhanu B, Govindaraju (eds) Multi-biometrics for human identification, pp 277–301. Cambridge University Press, Cambridge
84. Sana A, Gupta P, Purkai R (2007) Ear biometrics: a new approach. In: P Pal (ed) Advances in pattern recognition. World Scientific Publishing, New York, pp 46–50
85. Seely RD, Samangooei S, Lee M, Carter JN, Nixon MS (2008) The University of Southampton multi-biometric tunnel and introducing a novel 3D gait dataset. In: Proceedings of the 2nd IEEE international conference on biometrics: theory, applications and systems–BTAS 2008, pp 1–6
86. Shailaja D, Gupta P (2006) A simple geometric approach for ear recognition. In: Proceedings of the 9th international conference on information technology–ICIT 2006, pp 164–167
87. Shih H-C, Ho CC, Chang H-T, Wu C-S (2009) Ear detection based on arc-masking extraction and AdaBoost polling verification. In: Proceedings of the 5th international conference on intelligent information hiding and multimedia signal processing–IIH-MSP 2009, pp 669–672
88. Sun C, Mu Z-C, Zeng H (2009) Automatic 3D ear reconstruction based on epipolar geometry. In: Proceedings of the 5th international conference on image and graphics–ICIG 2009, pp 496–500
89. Takala V, Ahonen T, Pietikäinen M (2005) Block-based methods for image retrieval using local binary patterns. In: Proceedings of the 14th scandinavian conference–SCIA 2005, LNCS 3540, pp 882–891
90. Victor B, Bowyer K, Sarkar S (2002) An evaluation of face and ear biometrics. In: Proceedings of 16th international conference on pattern recognition, vol 1, pp 429–432
91. Viola P, Jones MM (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
92. Wang Y, Mu Z-C, Liu K, Feng J (2007) Multimodal recognition based on pose transformation of ear and face images. In: Proceedings of the international conference on wavelet analysis and pattern recognition–ICWAPR 2007, vol 3, pp 1350–1355
93. Wang Y, Mu Z-C, Zeng H (2008) Block-based and multi-resolution methods for ear recognition using wavelet transform and uniform local binary patterns. In: Proceedings of the 19th international conference on pattern recognition–ICPR 2008, pp 1–4
94. Wang X, Yuan W (2010) Gabor wavelets and general discriminant analysis for ear recognition. In: Proceedings of the 8th world congress on intelligent control and automation–WCICA 2010, pp 6305–6308
95. Wang Z-Q, Yan X-D (2011) Multi-scale feature extraction algorithm of ear image. In: Proceedings of the international conference on electric information and control engineering–ICEICE 2011, pp 528–531
96. Watabe D, Sai H, Sakai K, Nakamura O (2008) Ear biometrics using jet space similarity. In: Proceedings of the Canadian conference on electrical and computer engineering–CCECE 2008, pp 1259–1264
97. Watabe D, Sai H, Sakai K, Nakamura O (2011) Improving the robustness of single-view ear-based recognition under a rotated in depth perspective. In: Proceedings of the 2011 international conference on biometrics and kansei engineering–ICBAKE, pp 179–184

98. Woodard DL, Faltemier TC, Ping Y, Flynn PJ, Bowyer KW (2006) A comparison of 3D biometric modalities. In: Proceedings of the conference on computer vision and pattern recognition workshop–CVPRW 2006, p 57
99. Wu J, Brubaker SC, Mullin MD, Rehg JM (2008) Fast asymmetric learning for cascade face detection. IEEE Trans Pattern Anal Mach Intell 30(3):369–382
100. Wu H-L, Wang Q, Shen H-J, Hu L-Y (2009) Ear identification based on KICA and SVM. In: Proceedings of the WRI global congress on intelligent systems–GCIS 2009, vol 4, pp 414–417
101. Xie Z-X, Mu Z-C (2007) Improved locally linear embedding and its application on multipose ear recognition. In: Proceedings of the international conference on wavelet analysis and pattern recognition–ICWAPR 2007, vol 3, pp 1367–1371
102. Xie Z-X, Mu Z-C (2008) Ear recognition using LLE and IDLLE algorithm. In: Proceedings of the 19th international conference on patten recognition–ICPR 2008, pp 1–4
103. Xu X-N, Zhichun Mu Z-C (2007) Feature fusion method based on KCCA for ear and profile face based multimodal recognition. In: Proceedings of the IEEE international conference on automation and logistics, pp 620–623
104. Xu X-N, Mu Z-C, Yuan L (2007) Feature-level fusion method based on KFDA for multimodal recognition fusing ear and profile face. In: Proceedings of the international conference on wavelet analysis and pattern recognition–ICWAPR 2007, vol 3, pp 1306–1310
105. Xu X-N, Mu Z-C (2007) Multimodal recognition based on fusion of ear and profile face. In: Proceedings of the 4th international conference on image and graphics–ICIG 2007, pp 598–603
106. Xu H, Mu Z-C (2008) Multi-pose ear recognition based on improved locally linear embedding. In: Proceedings of the congress on image and signal processing–CISP 2008, vol 2, pp 39–43
107. Yaghoubi Z, Faez K, Eliasi M, Eliasi A (2010) Multimodal biometric recognition inspired by visual cortex and support vector machine classifier. In: Proceedings of the international conference on multimedia computing and information technology–MCIT 2010, pp 93–96
108. Yan P, Bowyer KW (2005) ICP-based approaches for 3d ear recognition. In: Proceedings of SPIE 5779:282–291
109. Yan P, Bowyer KW (2005) Empirical evaluation of advanced ear biometrics. In: Proceedings of the IEEE computer vision and pattern recognition workshops–CVPRW 2005, vol 3, pp 41–48
110. Yan P, Bowyer KW (2005) Multi-biometrics 2D and 3D ear recognition. In: Kanade T, Jain A, Ratha NK (eds) Proceedings of the international conference on audio- and video-based person authentication–AVBPA 2005, LNCS 3546, pp 503–512
111. Yan P, Bowyer KW (2007) Biometric recognition using 3D ear shape. IEEE Trans Pattern Anal Mach Intell 29(8):1297–1308
112. Yan P, Bowyer KW (2007) A fast algorithm for ICP-based 3D shape biometrics. Comput Vision Image Underst 107(3):195–202
113. Yaqubi M, Faez K, Motamed S (2008) Ear recognition using features inspired by visual cortex and support vector machine technique. In: Proceedings of the international conference on computer and communication engineering–ICCCE 2008, pp 533–537
114. Yazdanpanah AP, Faez K, Amirfattahi R (2010) Multimodal biometric system using face, ear and gait biometrics. In: Proceedings of the 10th international conference on information science, signal processing and their applications–ISSPA 2010, pp 251–254
115. Yuan L, Zhichun Mu Z, Ying Liu Y (2006) Multimodal recognition using face profile and ear. In: Proceedings of the 1st international symposium on systems and control in aerospace and astronautics–ISSCAA 2006, pp 887–891
116. Yuan L, Mu Z-C (2007) Ear detection based on skin-color and contour information. In: Proceedings of the international conference on machine learning and cybernetics, vol 4, pp 2213–2217
117. Yuan L, Mu ZC (2007) Ear recognition based on 2D images. In: Proceedings of the 1st IEEE international conference on biometrics: theory, applications and systems–BTAS 2007, pp 1–5

118. Yuan L, Mu Z-C, Xu X-N (2007) Multimodal recognition based on face and ear. In: Proceedings of the international conference on wavelet analysis and pattern recognition–ICWAPR 2007, vol 3, pp 1203–1207
119. Yuan L, Zhang F (2009) Ear detection based on improved AdaBoost algorithm. In: Proceedings of the international conference on machine learning and cybernetics, vol 4, pp 2414–2417
120. Yuan L, Mu Z-C (2012) Ear recognition based on local information fusion. Pattern Recogn Lett 33:182–190
121. Zeng H, Mu Z-C, Wang K, Sun C (2009) Automatic 3D ear reconstruction based on binocular stereo vision. In: Proceedings of the IEEE international conference on systems, man and cybernetics–SMC 2009, pp 5205–5208
122. Zeng H, Mu Z-C, Yuan L, Wang S (2009) Ear recognition based on the SIFT descriptor with global context and the projective invariants. In: Proceedings of the 5th international conference on image and graphics–ICIG 2009, pp 973–977
123. Zhang D, Lu G (2002) Shape-based image retrieval using generic Fourier descriptor. Sig Process Image Commun 17(10):825–848
124. Zhang H-J, Mu Z-C, Qu W, Liu L-M, Zhang C-Y (2005) A novel approach for ear recognition based on ICA and RBF network. In: Proceedings of the 2005 international conference on machine learning and cybernetics, vol 7, pp 4511–4515
125. Zhang H, Mu Z (2008) Compound structure classifier system for ear recognition. In: Proceedings of the IEEE international conference on automation and logistics - ICAL 2008, pp 2306–2309
126. Zhang Z, Liu H (2008) Multi-view ear recognition based on B-Spline pose manifold construction. In: Proceedings of the 7th world congress on intelligent control and automation - WCICA 2008, pp 2416–2421
127. Zhou J, Cadavid S, Abdel-Mottaleb M (2010) Histograms of categorized shapes for 3D ear detection. In: Proceedings of the 4th IEEE international conference on biometrics: theory applications and systems - BTAS 2010, pp 1–6