Pacific Graphics 2008
T. Igarashi, N. Max, and F. Sillion
(Guest Editors)

Volume 27 (2008), Number 7

Reconstructing head models from photographs for individualized 3D-audio processing

M. Dellepiane[1], N. Pietroni[1], N. Tsingos[2], M. Asselot[2] and R. Scopigno[1]

[1]Visual Computing Lab, ISTI-CNR, Pisa, Italy
[2]REVES-INRIA, Sophia Antipolis, France

Abstract

Visual fidelity and interactivity are the main goals in Computer Graphics research, but recently also audio is assuming an important role. Binaural rendering can provide extremely pleasing and realistic three-dimensional sound, but to achieve best results it's necessary either to measure or to estimate individual Head Related Transfer Function (HRTF). This function is strictly related to the peculiar features of ears and face of the listener. Recent sound scattering simulation techniques can calculate HRTF starting from an accurate 3D model of a human head. Hence, the use of binaural rendering on large scale (i.e. video games, entertainment) could depend on the possibility to produce a sufficiently accurate 3D model of a human head, starting from the smallest possible input. In this paper we present a completely automatic system, which produces a 3D model of a head starting from simple input data (five photos and some key-points indicated by user). The geometry is generated by extracting information from images and accordingly deforming a 3D dummy to reproduce user head features. The system proves to be fast, automatic, robust and reliable: geometric validation and preliminary assessments show that it can be accurate enough for HRTF calculation.

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computational Geometry and Object Modeling]: Geometric algorithms, languages, and systems

1. Introduction

Three-dimensional computer graphics have reached, in the last few years, impressive levels in the degree of realism and possible personalization. Several different applications (research, videogames, entertainment, medical support) rely on the possibility to enhance the experience of the users by adapting the environment to their peculiar characteristics.

Audio has been always less considered as a mean to provide realism, but techniques like binaural rendering can greatly enhance perception. Binaural rendering can be fully exploited only by using individualized HRTF filters. Otherwise it can produce localization artifacts (please refer to [Beg94] for a comprehensive overview on 3D sound for Virtual Reality and Multimedia).

The head-related transfer function (HRTF) describes how a given sound wave input (parameterized as frequency and source location) is filtered by the diffraction and reflection properties of the head and pinna, before the sound reaches the eardrum and inner ear. Biologically, the source-location-specific prefiltering effects of the head and external ear aid in the neural determination of source location (see [Bla97] for a comprehensive overview about psychophysics of spatial hearing). HRTFs are complicated functions of frequency and the three spatial variables; their shape can noticeably vary between subjects and it's closely related to the features of the head. In particular, the primary role in determining sound perception seems to be associated with the peculiar features of ears, while a secondary contribution is assigned to the size of the main features of the head (nose, chin, head width and height...).

This makes the modeling of accurate individual HRTFs a central issue in the context of audio rendering techniques. Among the several proposed solutions, a very promising direction is the simulation of the measurements made on a 3D head model. Unfortunately, the quality of the final result is strongly related to the accuracy of the geometric model. Laser scanning is a reliable but still expensive technique, so it cannot be considered for the application on the wide public, but only as a reference for validation.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

Is there a way to build a geometrically precise 3D model of an head starting from a very easy-to-provide input? Is it possible to do it automatically?

The system described in the paper is an automatic way to build a 3D model of a human head, starting from a few photos of the subject and some key-points indicated on them. The needed input can be produced in a few minutes, even by naive users, and the rest of the procedure is fast and completely automatic. It involves mechanisms to extract information from the photos provided by the user, and to accordingly deform a starting 3D dummy in order to reproduce the ears and face features of the subject.

The final structure of the system proves to be:

- Usable for large scale application, thanks to the very low amount of input needed: 5 photos, some key-points indicated on them and a single physical measurement.
- Completely automatic. Once the input data are collected, the 3D model reconstruction and the HRTF calculation are performed with no further intervention.
- Fast and reliable: the system is structured to produce realistic results even in the case of low quality or incongruent input data.
- Designed for accurately reproduce the ears shape, but also able to recreate the main face features.
- Geometrically accurate: validation shows that the accuracy in reconstruction is adequate to calculate a satisfactory HRTF.

The structure of the paper is the following: an overview of related work in Section 2 focuses on the issues related to binaural rendering, and briefly presents several methods for 3D face and head reconstruction. Section 3 provides a description of the system, and a detailed overview of its main components. In Section 4 particular attention is devoted to the geometrical validation of obtained model, and to a preliminary analysis of the measurement data obtained from the 3D heads. Finally, conclusions and future work are presented in Section 5.

2. Related Work

Individualized HRTF modeling: Listening to 3D audio over headphones using non-individualized HRTF filters can lead to localization artifacts, like notably increased "inside-the-head" perception and front-back confusions [Bla97, WAKW93]. To address this issue, several approaches have been proposed to model individualized HRTFs for 3D audio processing [Gar05]. They can be classified into four major groups.

The first approach consists in directly measuring HRTFs by placing microphones inside the ears of the subject [KB07]. While leading to improved 3D reproduction, measurement-based approaches usually require dedicated hardware and facilities. Moreover, the process can take up to several hours and is extremely uncomfortable for the user.

Perceptually-oriented techniques use a series of short successive listening tests. They present the subject with sounds located at several target positions using various HRTF sets and guide him in the choice of a composite set that best matches the targets [MMO00]. This procedure might take up to 20 minutes. While ensuring that the results best fit to the user perception, accounting for the entire reproduction pipeline, it remains a challenge to design efficient protocols to quickly select the most appropriate matches.

A third group of approaches attempts to directly model an individualized filter starting from a generic HRTF set (e.g., a dummy head) [Lar01] or by constructing analytical models of the various subcomponents of the HRTFs (e.g., pinna/ head shadowing, shoulder reflection) [DAA99, ADMT01]. These models can be driven by direct measurements of morphological parameters on the subjects or by using photographs [Lar01]. However, linking morphological parameters to changes in the features of the HRTF filters relies on correlation analysis between significant morphological features and spectral features of the HRTFs. This requires a large database and might not be fully reliable [Lar01].

Alternatively, it is possible to simulate virtual measurements of HRTFs filters by running finite element simulations [Kat01, KN06] or ray-casting [ARM06] on a 3D head model. This approach can be compared to virtual gonioreflectometry, used to model BRDFs in computer graphics. Aside from a prohibitive computation time, finite element solutions also require a 3D input mesh of the subject, typically using 3D scanning techniques which limit their practical use. Approximate finite element simulations [TDLD07] can significantly diminish computation time, while providing efficiently estimated individualized HRTFs.

Reconstruction of Head Models: The shape of the ears plays a key role for the reconstruction system. Ear biometrics try to individuate a model for external ear features: [Ian89] is often cited as a very convincing one. Unfortunately, these biometry measures are not well-defined, so it's very hard to use them in automatic methods. Ears are mainly considered in the field of recognition and security by analyzing both 2D [JM06] and 3D [CB05, CB07] ear data. Unfortunately the information extracted and used for recognition is usually not directly linked to geometric features (like curves or size of the ear). Hence, only the part related to image processing solutions can contribute to a 3D reconstruction system.

Aside from trivial texture mapping on low quality geometry, several works in the field of 3D reconstruction focus on the reconstruction of 3D faces from images. Very realistic results can be achieved [HA04, D'A01], but the produced geometry is usually not accurate enough. An approach which is more related to our goal is the morphing of face model to fit images [BV99, Bla06, JHY*05] and 3D scans [BSS07]. These methods are very accurate, but they don't take into account the whole head, and especially the ears.

Regarding 3D head reconstruction from 2D data, some methods [LMT98, LLY04] obtain low resolution complete

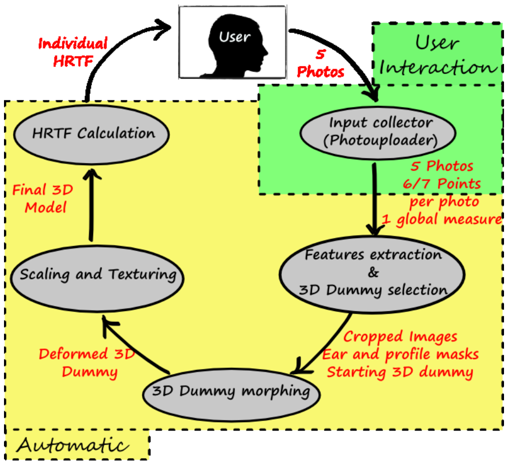


Figure 1: *A scheme of the whole HRTF calculation system*

and textured 3D head models. Also in these cases, geometry is not accurate enough for our aim. [FOT04] can generate accurate models, but a very complex acquisition apparatus (28 digital cameras and two projectors) is needed. Moreover, most of the works don't take into account the scale of the model, that is a key issue for scattering calculation.

In conclusion, a numerical comparison between the cited methods and our system is difficult due to the differences in goals (geometry accuracy vs. visual resemblance, 3D faces vs. 3D heads). Hence, we will consider laser scanning as a reference, since it is the most reliable technique for geometry acquisition.

3. Overview of the system

In order to be able to reach the goals listed in the introduction, the whole system was designed as a custom solution which employs different techniques at their best for this peculiar application. A scheme of the whole system is shown in Figure 1. We can individuate four main components which intervene in the generation of the 3D model. The whole process can be roughly divided in two parts: the first one deals with the treatment of the input given by the user, and the extraction of feature-based data from photographs, in order to select the best 3D dummy to be morphed. In the second part, work is performed on the selected 3D dummy mesh, in order to morph and scale it to resemble the geometric features of the head of the user.

Though, for sake of clarity, the system has been subdivided in several parts, only the first component (Input Collector) needs an intervention by the user. The rest of the system has been designed to be completely automatic.

3.1. Collection of input data from the user

The first issue to be addressed is: what is the lowest possible amount of work we can ask to the user, if we want a robust, automatic, precise and fast system? Provided that digital imaging is the simplest way to encode needed information, several issues arise: how many photos? From which position? Are those data really sufficient?

One of the goals of the design phase was to ensure simplicity: no custom acquisition setup, no need of high quality cameras. The final choice, in order to try to cover the whole head, and to be able to infer 3D information, was to ask for five photos of the head, possibly taken from the same distance: one frontal, two profiles and two three-quarters portraits (see Figure 2 right). A small guide helps the user in taking the photos, by giving some simple hints, like: keep the same expression between the photos, remove carefully hair from ears, use flash if present etc.

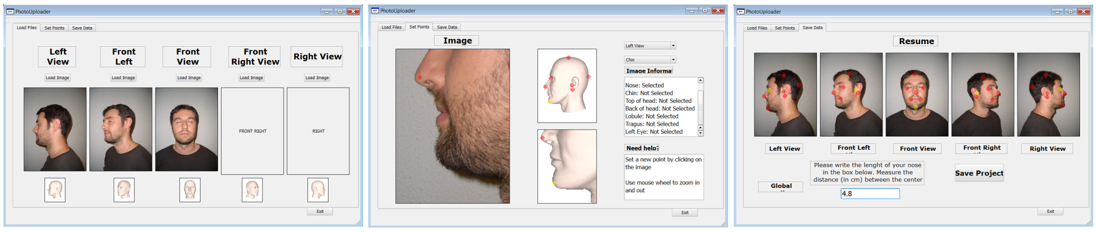
Techniques to analyze human face photos can automatically extract a lot of features (eyes position, nose, profile...). Yet the aim of our work was to build a robust method, able to produce a good result even when particular face features (long beard, tatoos, peculiar haircuts...) are present, or when low quality images are uploaded. To do this, a small further contribution is requested to the user. In order to collect data in a user-friendly way, a simple tool called Photouploader was created. Photouploader was designed as web application, where the user can upload the images and is guided in the data provision.

Photouploader is divided in three tabs, shown in Figure 2. Once the photos are taken, the first tab guides the user in uploading the five photos. The second tab asks the user to indicate some key-points (6 or 7) on each photo. Dummy head images show where are the needed points in a very intuitive way. When all the required points have been selected, in the third tab the user is finally asked to insert a global scale measure (we chose the length of the nose), and save the project into an xml file.

3.2. Features extraction from images and starting 3D dummy selection

Once that input data are collected, the automatic model production process starts. The goal of the first element of the system is the selection of the best starting dummy from a library of 3D heads. The library of 3D heads is composed by ten models obtained via 3D scanning. These model were preferred to *artificial* ones especially because the peculiar ear features were present, and the acquired people were different in age and gender. These models (250k triangles) represent both head and shoulders: they were used also as a reference for the validation of geometric accuracy and HRTF calculation. Moreover, as shown in Figure 3, each 3D model has differently colored parts, which will undergo different morphing policies.

As already stated, although face features are important and



Nose: Selected
Chin: Not Selected
Top of head: Not Selected
Back of head: Not Selected
Lobule: Not Selected
Tragus: Not Selected
Left Eye: Not Selected

Set a new point by clicking on the image

Use mouse wheel to zoom in and out

Please write the lenght of your nose in the box below. Measure the distance (in cm) between the center

4.8

Figure 2: *Screenshots of Photouploader: left, first tab for images upload; center, second tab for key-points picking; right, third tab for data resume and global scaling value insertion*

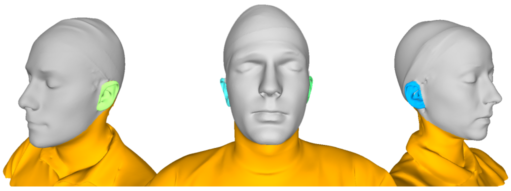


Figure 3: *Three elements of the 3D dummy library*

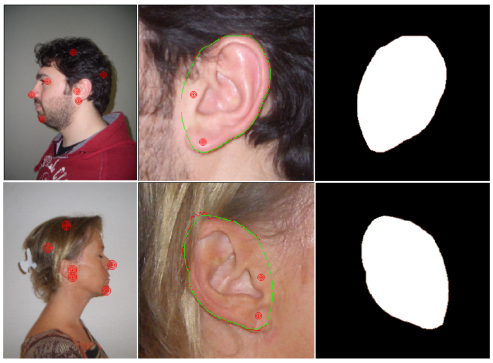


Figure 4: *Two examples of ear external border extraction*

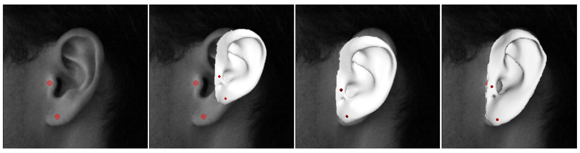


Figure 5: *An example of ear selection: starting image, dummy ear camera position before and after alignment, ear shape after low accuracy morphing*

can't be ignored, the shape of the ears plays a key role in the final HRTF profile. So the dummy which best matches the ear features (extracted from images) will be selected for morphing. Hence the first goal is to extract the external borders of both ears from images. The image is automatically cropped using two of the key-points provided by the user (lobule and tragus) as a reference. Cropped image is then scaled to 256x256 in order to remove high frequency details. Both edge intensity and orientation are calculated using the method proposed in [JM06]. The ear shape is extracted by finding a seed point near to the lobule key-point provided by the user, and then following the ear edge taking into account both intensity and direction of edges.

Several other controls (since the shape of ears is generally similar between subjects, it's possible to apply constraints on the final shape and continuity of extracted line) are added in order to deal with cases where the edges are hard to find or non present (i.e. hair over some part of the ear, earrings). Two examples of external mask extraction are shown in Figure 4. In the second one, even though part of the ear is covered by hair, extracted mask is very similar to ear shape. External masks are used both for dummy selection and as an input for morphing (see next Section). The dummy selection is performed by analyzing each of the ears of the library models (Figure 5). A model of perspective camera is used and a rigid alignment is performed, by modifying extrinsic camera parameters so that the external borders of the 3D ear are best aligned to the extracted mask. Hence, a low accuracy morphing (see next Section) is applied on each aligned 3D ear model, so that it is slightly deformed to fit both the external mask and the internal features. After this operation, a similarity measure between a rendering of the 3D ear and the extracted image (based on the position of feature edges) is calculated, and the most similar dummy is selected. Camera parameters associated to the alignment are stored as well.

A similar approach to the one used for ears is applied to profile and frontal photos to provide best camera data for head deformation (Figure 6). The image is cropped and scaled to 256x256, using the the key-points provided by the user as a reference. Then, in a similar way to the ear extraction, the external mask of the face is created, by using the line extracted after having analyzed the features of the image. Where no information can be found on the image, the key-points, provided by user, help to define a line which resembles head shape. Once a mask is extracted, a rigid alignment which takes into account both extracted mask and user-provided key-points determines the best camera parameters for head morphing.

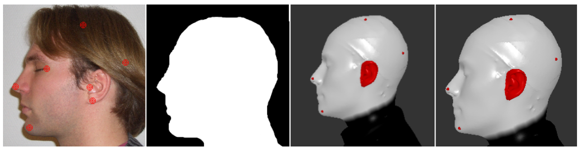


Figure 6: *An example of head alignment: starting image, extracted mask, dummy head camera position before and after alignment*

The final output of this stage includes: ten extracted images (four for the ears, six for the face) and an xml file containing all the necessary information needed for the following steps (new coordinates of the points set by the user, name of selected dummy, camera parameters associated to each view, global scaling value).

3.3. Dummy morphing

The 3D Morpher is the core of all the system: it applies a peculiar deformation to the dummy 3D mesh. Using the set of cameras which defines the alignment of the dummy model with respect to each image, a set of viewport-dependent 2D-to-3D model deformation is calculated. The set of deformations is then combined to morph the dummy model to its final shape. The entire morphing process can be subdivided into the following steps:

Single View Head Deformation Three energy-driven deformations, one for each photo (right and left profile, frontal) are calculated. Each one tries to match the geometry with respect to a single point of view (*view dependent deformation*).

Global Head Deformation View dependent deformations are merged (according to camera positions and orientations) to a global smooth deformation.

Ear deformation Ears, which have been preserved in previous steps, undergo an accurate deformation using close-up ear cameras and images.

Morphing the geometry to match an input image requires the computation of a mapping between the photo and a rendered image of the dummy head (taken from the associated rigid-aligned camera position). This mapping operation is usually referred as *warping* in literature.

3.3.1. Warping Computation

Our warping function is an extension of [WM05] and [POB*07]. It defines a 2D image deformation, that is then applied to the 3D model. It can be summarized as follows : feature images are processed with an edge detector, a stack of feature images is created by downsampling; finally automatic multilevel feature matching defines image deformation (details in [POB*07]).
The original energy function (which has to be minimized in order to compute best deformation) was modified by adding a term $Kp = \sum_{P_i \in keypoints} |(P_i(photo) - P_i(model)|$ that measures the sum of distances between the user-defined key-points on the input photo and the relative key-points on the dummy head (transformed to screen-space coordinates). Energy function can be schematized as follows:

$$E = L2 + \alpha * J + \beta * Kp \qquad (1)$$

where $L2$ is the *per-pixel error* scalar feature strength and J is the *Jacobian term* which controls the smoothness of the warp field. In our setting, the warping calculation starts from an 8x8 triangulated grid up to 64x64 (see [WM05] for details). The values of α and β are set respectively to 400 and 10000.
Once we calculate the warping function, the displacement for each 3D dummy vertex is calculated by projecting the vertex on the associated camera plane, evaluating its warped position, and un-projecting it back to world space (without changing z-value).
The size of images used for morphing is 256x256, as indicated in previous Section. This value represents a good compromise between detail preservation and processing time: higher resolution images could be used, but the gain in detail would not justify the longer time necessary for computation.

3.3.2. Single View Head Deformation

In this phase we apply the warping between a rendered image of the dummy head and the input photograph, using associated camera parameters. In the original method, both external and internal features would be taken into account for deformation. But in a real scenario head photographs could reveal strong sharp features that are difficult to be represented geometrically (such as beard, eyebrows..), furthermore peculiar lighting environments could lead to incorrect edges warping. So the deformation is applied using only the binary masks which define the external profile of the head.
The internal features of the face are then deformed by fitting a group of key-points associated to those indicated by user. Figure 7.1 shows the key-points involved in head warping: for frontal deformation we use 5 key-points: two for the eyes, one for the nose and two for the mouth; while for lateral deformation we chose to use one side eye constraint plus a set of four points around the ear. These points define the bounding box of the ear, so that it is preserved for a latter deformation (see later). Sequence 7.2 shows the deformation process involving the dummy mesh, using one lateral image.
Moreover, the frontal warping is controlled via *symmetrization*. Because of possible non-symmetric head contours extracted from frontal photo or non perfect input image (i.e. tilted or slightly rotated head), simple warping can produce asymmetric head shapes (see Figure 8.1). To overcome this problem we symmetrize the warping as follows:

- We establish a symmetrization line on the rendered image, so that the rendered image is divided into two subspaces (Figure 8.3). The symmetrization line is defined as the line

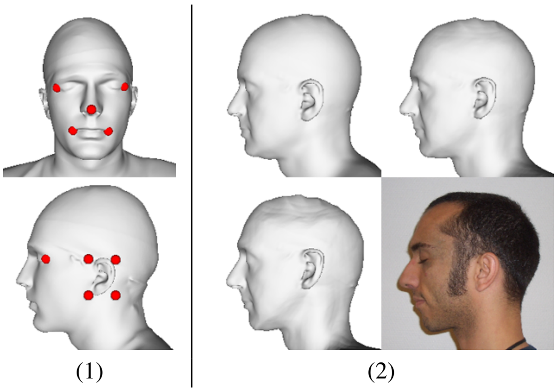


Figure 7: *(1)Keypoints used for frontal and lateral head deformation. (2)Lateral head deformation sequence.*

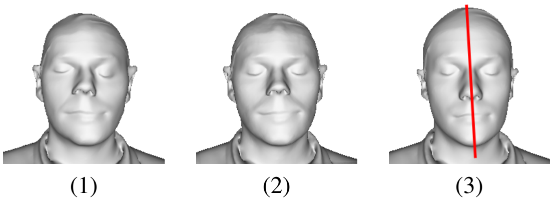


Figure 8: *(1 and 2) Non-symmetrized versus symmetrized ; (3) Symmetrization line of undeformed model.*

passing through the nose key-point and the point in the middle of the eyes key-points.

- We obtain a mapping $Mirr(x,y)$ between the two regions of the rendered dummy, by mirroring over the symmetrization line.
- We finally average the warping of mirrored points:

$$Warp(x,y) = \frac{Warp(x,y) + Mirr^{-1}(Warp(Mirr(x,y)))}{2} \tag{2}$$

Figure 8 .2 shows the effect of warping symmetrization.

3.3.3. Global Head Deformation

At this morphing stage, each vertex can be translated in three different ways (one for each viewpoint). These three camera plane warpings are unified to a single smooth deformation as follows:

- Lateral deformed positions are unified through a weighted sum (weights decrease proportionally respect to ear distances)
- Unified lateral and frontal deformations are summed by assuming they are perpendicular, so that displacements in x- and z-axis are independent: the final displacement in the common direction (y-axis) is a weighted sum of the two contributions.

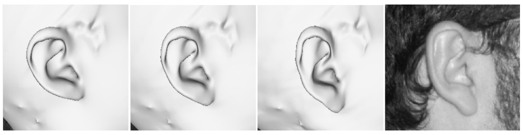


Figure 9: *Example of ear morphing sequence.*

3.3.4. Ear deformation

Accurate ear deformation is key for the final quality of the results: in this case both internal and external features extracted from images can be used to compute the deformation (Figure 9). The morphing sequence is organized as follows: 3D ear rendering is morphed to fit the external ear silhouette, then an additional warping (using the feature stack) matches internal features.

The colors of input ear images are previously modified in order to match the histogram of frequency spectra of the rendering of the dummy model (see [WM05] for details): this further improves final deformation.

3.4. Global scaling and texturing

Scaling is one of the key issues about the accuracy of reconstruction. If the size of the model is incorrect, the computed HRTF will be wrong. The scaling operation is performed using the measure provided by user with the Photouploader (see Section 3.1), which is the nose length. It's sufficient to scale the model by the ratio between this value and the nose length on the model (which can be calculated automatically by setting key-points which are translated accordingly to morphing). This leads back the geometry to its real measures. After a final translation to set the model ears position on the X axis (so that all the produced models are aligned in the same way), HRTF can be automatically calculated, and the process is complete.

A further feature of the system is the possibility to texture the obtained head model: the input photos are deformed to match more precisely the geometry (essentially by applying the inverse warping explained in section 3.3). Shown colored models are obtained by projecting warped images using [CCCS08]. But, for clarity sake, textures are not needed for HRTF calculation.

4. Results

Two results are shown in Figure 10. The entire system proves to be robust and quite fast: the overall time needed to produce the final 3D model, from input collection to model saving, it's less than ten minutes.

Although the visual resemblance of the obtained model is usually satisfactory, the main goal of the entire system was to be able to guarantee sufficient geometric accuracy. For this reason we performed some tests to compare eight models obtained from photos to their corresponding laser scanned

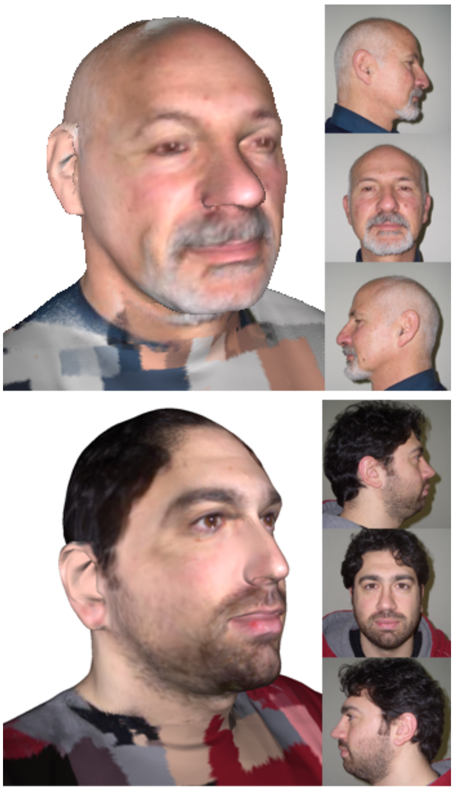


Figure 10: *Two results of processed heads.*

models. Moreover, preliminary HRTF calculation tests were performed on the 3D head couples, in order to obtain a comparison between the resulting simulations.

4.1. Geometric validation

A sub-millimetric precision in geometry reconstruction from photos is a results which is possible only under very particular and controlled conditions. In our case, since the input is provided by the user, and the starting dummy can be very different from the final result, the main goal is to be able to reproduce head features as much as possible. Hence, instead of using purely geometric comparison tools like [CRS98], we compared the results by taking into account two sets of measures, which could represent head features and their influence on the HRTF profile. The first set was the position in space of several key-points, picked on both model. Results of comparison are shown in Table 1. The average error is usually less than 1 cm and the variance of data is not big. These values can be considered as satisfying, especially considering that the input data is solely two-dimensional and the scale factor can introduce inaccuracies.

	Average	Maximum	Variance
Nose	11.3	22.1	0.89
Chin	10.8	21.2	0.03
Left Eye	9.8	16.6	1.22
Right Eye	8.1	11.4	0.46
Left Mouth	11.4	22.2	2.83
Right Mouth	9.8	21.5	1.69
Left Lobule	8.4	13.5	1.89
Left Tragus	6.7	11.5	1.04
Right Lobule	8.3	14.0	1.30
Right Tragus	7.3	10.8	1.83

Table 1: *Distance in mm between key-points of scanned and reconstructed model*

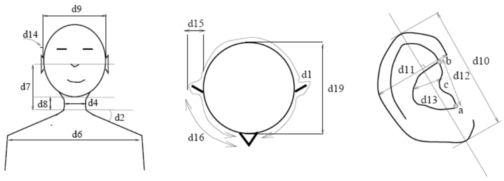


Figure 11: *Some of the measures proposed by [Lar01]*

The second set of measures was extracted from [Lar01], where several ear and head measures were statistically analyzed in order to find which ones were more related to the changes in HRTF profile (Figure 11). We compared the set of 3D models using six measures (three for the head, three for the ears) which are indicated as very important in the conclusions of this work. Results are shown in Table 2. The difference between distances is often less than 5 mm, in particular the internal features (concha size) seem to be preserved accurately. An overall analysis of the data shows that, even if accuracy for some features (like ear size and tragus position) should be further improved, the error bound describes a very reliable system.

4.2. Preliminary assessment of HRTF simulations

A preliminary validation of HRTF simulation on 3D models was performed on couples of laser scanned and reconstructed 3D heads of the same subject.

To simulate HRTFs corresponding to the reconstructed

	Average	Maximum	Variance
Neck (d4)	5.6	15	0.10
Head Size (d9)	5.9	14.5	1.30
Head Size (d19)	3.2	6.7	0.47
Ear size (R) (d10)	3.7	6.3	0.50
Concha size (R) (d12)	1.8	3.4	0.04
Concha size (R)(d13)	0.8	1.3	0.005
Ear size (L) (d10)	3.8	8.2	0.58
Concha size (L) (d12)	2.1	3.8	0.33
Concha size (L) (d13)	1.1	2.5	0.24

Table 2: *Difference in mm between distances indicated in [Lar01]*

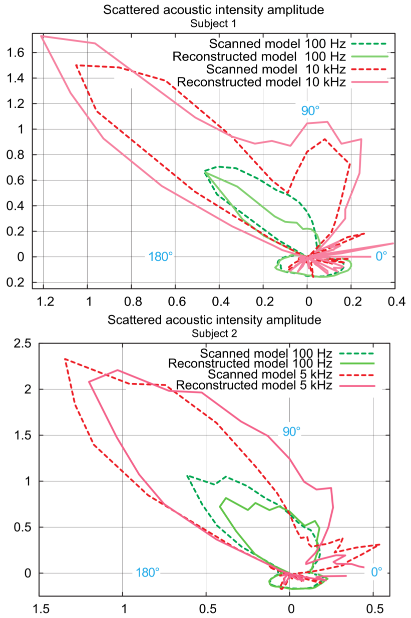


Figure 12: *Two examples of polar plots for measurements on couples of scanned-reconstructed 3D heads.*

geometry, we used a simplified boundary element approach leveraging the Kirchoff approximation. The Kirchoff approximation allows for efficiently computing first order scattering off the reconstructed mesh and can be efficiently implemented using programmable graphics hardware. Please refer to [TDLD07] for details.

We used this approach to computed the scattered field captured by two virtual microphones placed at the entrance of the left and right ear canal (90°), less than 5 millimeters away from the surface of the mesh. The field was produced by a source wave rotating around the head in the horizontal plane, at a 2 meter distance from the center. We simulated the field measurement for 8192 frequencies from D.C. and 22kHz. Our hardware accelerated approach computes an HRTF pair for a given source position in about 200 sec. on a GForce 7600GT.

Computed data are only a subset of a complete individual HRTF, but they provide enough information for a preliminary comparison. Two polar plots of left ear intensity-amplitude for our reconstruction approach compared to a laser scanned model are shown in Figure 12. The intensity-amplitude variation features are aligned for the reconstructed and scanned models. In particular the maxima are reached at 125°, and rear features match between 180° and 360°.

We can hereby state that obtained data from preliminary calculations prove to be encouraging for a future use of 3D reconstructed models for HRTF calculation. A further stage of validation between reconstructed and measured in anechoic chamber HRTFs will provide more information. Moreover, it will be possible to further investigate the importance of the head features (ears vs. face, possible contribution by shoulders) to improve results and possibly further simplify the system.

5. Conclusions and future work

In this paper we presented a system to automatically create 3D head models from a very low amount of input (five photos and some key-points indicated on them). The system integrates image processing techniques with a novel application of 3D morphing, based on a combination of several 2D deformations calculated in different camera spaces. The main current application on the 3D models is the calculation of HRTF. The system proves to be fast, robust and reliable, and both geometric and preliminary HRTF validation are encouraging. Hence, it can be proposed as a new solution to produce accurate 3D head models with a very low amount of input. Future work to further improve the method includes:

- The implementation of more effective methods for face deformation (i.e. implementing part of the contribution of [Bla06]). This would probably lead to a better visual resemblance of the model, widening the application field of the proposed method to geometrically accurate avatar generation. In this case, hair extraction and visualization (which is not considered as a key issue for HRTF calculation) should be taken into account.
- The use of three-quarter images to further refine the mesh geometry.
- The implementation on GPU to make the model generation almost real-time: this could bring the computation time from minutes to seconds.
- The improvement of dummy library: it is currently formed by ten models, but the best solution could be to select a subset of 3D models from a wider set of 3D scanned heads. Moreover, since ears play a key role in all the process, another idea could be to create 3D models with the same face features but different ear shapes.

In conclusion, the proposed system can be considered as a very promising method not only for individualized 3D-audio processing, but also for other applications which need accurate head geometry, produced from a few photographs.

Acknowledgments. This research was funded by the EU FET project CROSSMOD (014891-2, website http://www.crossmod.org). We thank all the people whose photos (and 3D scans) helped in creating and testing the system. We also thank Marco Callieri and Fabio Ganovelli for their support for the development of color projection and morphing components.

References

[ADMT01] Algazi V. R., Duda R. O., Morrison R. P., Thompson D. M.: Structural composition and de-composition of hrtfs. In *Proc. IEEE Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA'01), New Paltz, USA* (2001).

[ARM06] ANDRES S., RÖBER N., MASUCH M.: *HRTF simaulations through acoustic raytracing*. Tech. rep., Otto v. Guericke Un. Magdeburg, Germany, 2006.

[Beg94] BEGAULT D. R.: *3D Sound for Virtual Reality and Multimedia*. Academic Press Professional, 1994.

[Bla97] BLAUERT J.: *Spatial Hearing : The Psychophysics of Human Sound Localization*. M.I.T. Press, Cambridge, MA, 1997.

[Bla06] BLANZ V.: Face recognition based on a 3d morphable model. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 617–624.

[BSS07] BLANZ V., SCHERBAUM K., SEIDEL H.-P.: Fitting a morphable model to 3d scans of faces. In *IEEE ICCV 2007* (2007), pp. 1–8.

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3D faces. In *Siggraph 1999, Computer Graphics Proceedings* (Los Angeles, 1999), Rockwood A., (Ed.), Addison Wesley Longman, pp. 187–194.

[CB05] CHEN H., BHANU B.: Contour matching for 3d ear recognition. In *WACV-MOTION '05: Volume 1* (Washington, DC, USA, 2005), pp. 123–128.

[CB07] CHEN H., BHANU B.: Human ear recognition in 3d. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*, 4 (April 2007), 718–737.

[CCCS08] CALLIERI M., CIGNONI P., CORSINI M., SCOPIGNO R.: Masked photo blending: mapping dense photographic dataset on high-resolution sampled 3d models. *Computers and Graphics* (2008), (under publication).

[CRS98] CIGNONI P., ROCCHINI C., SCOPIGNO R.: Metro: measuring error on simplified surfaces. *Computer Graphics Forum 17*, 2 (June 1998), 167–174.

[D'A01] D'APUZZO N.: Human face modeling frommulti images. In *Proc. of 3rd Int. Image Sensing Seminar on New Dev. in Digital Photogrammetry, Gifu, Japan* (2001), pp. 28–29.

[DAA99] DUDA R. O., AVENDANO C., ALGAZI V. R.: An adaptable ellipsoidal head model for the interaural time difference. In *Proc. IEEE (ICASSP)* (1999), pp. II:965–968.

[FOT04] FUJIMURA K., OUE Y., TERAUCHI T.: Improved 3d head reconstruction system based on combining shape-from-silhouette with two-stage stereo algorithm. In *ICPR '04: Volume 3* (Washington, DC, USA, 2004), pp. 127–130.

[Gar05] GARDNER W.: Spatial audio reproduction: Towards individualized binaural sound. *National Academy of Engineering* (2005).

[HA04] HASSANPOUR R., ATALAY V.: Delaunay triangulation based 3d human face modeling from uncalibrated images. *Computer Vision and Pattern Rec. Workshop* (2004), 75–75.

[Ian89] IANNARELLI A.: Ear identification. *Paramount Publishing Company, Freemont, California* (1989).

[JHY*05] JIANG D., HU Y., YAN S., ZHANG L., ZHANG H., GAO W.: Efficient 3d reconstruction for face recognition. *J. of Pattern Recogn. 38*, 6 (June 2005), 787–798.

[JM06] JEGES E., MATE L.: Model-based human ear identification. *World Automation Congress, 2006. WAC '06* (24-26 July 2006), 1–6.

[Kat01] KATZ B.: Boundary element method calculation of individual head-related transfer function. part I: Rigid model calculation. *Journal Acoustical Soc. Am. 110*, 5 (2001), 2440–2448.

[KB07] KATZ B., BEGAULT D.: Round robin comparison of HRTF measurement systems: preliminary results. In *Proc. 19th Intl. Congress on Acoustics (ICA2007), Madrid, Spain* (2007).

[KN06] KAHANA Y., NELSON P.: Numerical modelling of the spatial acoustic response of the human pinna. *Journal of Sound and Vibration 292*, 1-2 (2006), 148–178.

[Lar01] LARCHER V.: *Techniques de spatialisation des sons pour la réalité virtuelle*. Thèse de doctorat, Université Paris 6 (Pierre et Marie Curie), Paris, 2001.

[LLY04] LEE T.-Y., LIN P.-H., YANG T.-H.: Photo-realistic 3d head modeling using multi-view images. In *ICCSA (2)* (2004), pp. 713–720.

[LMT98] LEE W.-S., MAGNENAT-THALMANN N.: Head modeling from pictures and morphing in 3d with image metamorphosis based on triangulation. In *CAPTECH* (1998), pp. 254–267.

[MMO00] MIDDLEBROOKS J., MACPHERSON E., ONSAN Z.: Psychophysical customization of directional transfer functions for virtual sound localization. *Journal Acoustical Soc. Am. 108*, 6 (2000), 3088–3091.

[POB*07] PIETRONI N., OTADUY M. A., BICKEL B., GANOVELLI F., GROSS M.: Texturing internal surfaces from a few cross sections. *Comp. Graph. Forum 26*, 3 (2007).

[TDLD07] TSINGOS N., DACHSBACHER C., LEFEBVRE S., DELLEPIANE M.: Instant sound scattering. In *Proc. of the Eurographics Symposium on Rendering* (2007).

[WAKW93] WENZEL E., ARRUDA M., KISTLER D., WIGHTMAN F.: Localization using non-individualized head-related transfer functions. *J. Acoustical Soc. Am. 94*, 1 (1993), 111–123.

[WM05] W. MATUSIK M. ZWICKER F. D.: Texture design using a simplicial complex of morphable textures. *SIGGRAPH (ACM Transactions on Graphics)* (2005).