USING COMPUTER VISION TO GENERATE CUSTOMIZED SPATIAL AUDIO

Ankur Mohan, Ramani Duraiswami, Dmitry N. Zotkin, Daniel DeMenthon, Larry S. Davis

Perceptual Interfaces and Reality Laboratory, UMIACS, University of Maryland, College Park 20742
{ankur, ramani, dz, daniel, lsd}@umiacs.umd.edu

ABSTRACT

Creating high quality virtual spatial audio over headphones requires real-time head tracking, personalized head-related transfer functions (HRTFs) and customized room response models. While there are expensive solutions to address these issues based on costly head trackers, measured personalized HRTFs and room responses, these are not suitable for widespread or easy deployment and use. We report on the development of a system that uses computer vision to produce customizable models for both the HRTF and the room response, and to achieve head-tracking. The system uses relatively inexpensive cameras and widely available personal computers. Computer-vision based anthropometric measurements of the head, torso, and the external ears are used for HRTF customization. For low-frequency HRTF customization we employ a simple head-and-torso model developed recently [1]. For high frequency customization we employ measured pinna characteristics as an index into a database of HRTFs [2]. For head tracking we employ an online implementation of the POSIT algorithm [3] along with active markers to compute head pose in real-time. The system provides an enhanced virtual listening experience at low cost.

Partial support of NSF award #0205271 and ONR grant N000140110571 is gratefully acknowledged.

1. INTRODUCTION

The use of the auditory modality to convey information to the user of a human-computer interface is an emerging application area. Different parameters of the audio signal such as repetition rate, pitch, timbre, intensity, spatial position and ambience can be manipulated to represent information, creating a virtual audio scene. Algorithms to render sound with the specific set of the above mentioned parameters are necessary for sonification, along with an agreed mapping of the data to the audio representation space. Another area where a similar task arises is in virtual/augmented reality applications where one seeks to create a "convincing" multimodal scene. For the auditory modality, this amounts to synthesizing an audio scene that sounds "natural", so that to the user the synthetic scene is indistinguishable from the real one. It is sufficient to recreate the stimulus that the person receives from the real scene (the acoustic wave pressure at the eardrum) to achieve this.

One of the primary requirements of these audio user interfaces is the need to reposition the rendered audio scene in response to the user motion. Indeed, our perception of the environment is strongly characterized by our location relative to the objects being perceived; if the user rotates her head, in her coordinate system the direction of the stationary objects changes. Applications in virtual and augmented reality that seek to create convincing experiences require an ability to track the motion of a person and quickly attend to such motion to reposition the virtual environment. In this paper, we describe a set of algorithms to synthesize a virtual audio environment that includes real-time tracking of head position and orientation using an inexpensive setup.

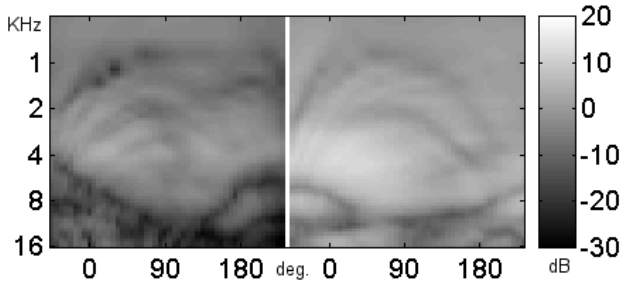

Fig. 1. Variations in the spectral content of the sound along the cone of confusion.

2. AUDIO SPATIALIZATION BASICS

As noted above, it is enough to recreate the stimuli to create the same perception. For the auditory modality, this amounts to putting together binaural and spectral cues encoded in the head-related transfer function (HRTF), environmental cues and dynamics.

2.1. Head-Related Transfer Function

Research in human sound localization dates back to 1907 with Lord Rayleigh's work [4], with the development of interaural time difference and interaural level difference (ITD and ILD) cues for the sound perception. However, ITD and ILD alone can't explain localization in the vertical plane, or, more generally, along the geometrical locations giving rise to the same ITD and ILD values. It was later hypothesized and verified that sound scattering on certain parts of human anatomy, especially outer ear (pinna), also provides cues that are responsible for sound localization [5], [6]. The outer ear has a complicated shape; sound arriving at the ear from different directions enacts in interaction with pinna cavities, ridges and notches, undergoing complex reflections, diffraction and interference. This results in significant changes in *the spectrum of the sound* which reaches the eardrum. Other weaker cues are produced by sound scattering by the head, called head shadowing, and sound reflection off the torso, called shoulder bounce.

The spectral changes are characterized by the so-called head related transfer function (HRTF). If $H_l(r, \varphi, \theta, \omega)$ is the spectrum of the sound at the eardrum (of the left ear) when the source is positioned at the point with the spherical coordinates (r, φ, θ) where ω is the frequency, and $H(\omega)$ is the spectrum of the sound at the center of the head as if the listener is not present, then the HRTF $h_l(r, \varphi, \theta, \omega)$ (for the left ear) is defined as

$$h_l(r, \varphi, \theta, \omega) = H_l(r, \varphi, \theta, \omega)/H(\omega),$$

with the same definition for the right ear. The HRTF is then a function of the frequency, the direction of the sound source and the distance to it. Example of a measured HRTF for the human subject is shown in Figure 1; the plots show the magnitude of HRTF for the contralateral and the ipsilateral ears as source moves around the cone of confusion at 45 degree of azimuth for elevations from -45 to 225 degrees on horizontal axis. Usually, the dependence on distance is weak for distant (more than about 50 cm away) sources and is ignored. When HRTF-filtered sound is presented to the listener through headphones, the perception of sound coming from that direction will arise. However, several problems remain to be solved, described below.

2.2. Acquisition of HRTFs

Because of the individual differences in the anatomy, the HRTF naturally exhibits significant variations, which means that the spectrum of the signal reaching the eardrum for the same source position will be different for different persons (even when the own ear of a person is modified slightly using putty or tape, the localization performance degrades very significantly [7]). The results of using non-personalized HRTF for synthesis vary from acceptable shifts in the perceived and the true position of the sound source to inability to perceive any source motion at all. Obtaining a personalized set of HRTF for an individual is a time-consuming and costly process; it is usually done by direct measurement, where a tiny microphone is inserted into the ear canal of an individual and a loudspeaker is playing predefined signals at all possible positions around the person, covering the measurement sphere in small steps (five or ten degrees of resolution in both azimuth and elevation is common). For all experimental HRTF measurement methods, obtaining reliable data at low frequencies has always been a problem due to the relatively long signal length necessary, so that the reflections off the room walls and measurement equipment start to matter. For increased realism and to compensate for these problems, we use an analytical head-and-torso model [1] to synthesize HRTF for low frequencies and to blend them with measured HRTFs using body parameters detected using computer vision (Section 5).

2.3. Reverberation and dynamic cues

Another challenge is the presence of several other elements of the audio scene besides the HRTF-based cues which alone are not sufficient for "natural" synthesis. Two other important sets of cues are the reverberation cues and the dynamic cues.

Presence of reverberation is more important than possession of the correct HRTFs for perception of externalization (a sense that the source is present in the environment at some distance from the head). Human perceive the distance to the source by essentially comparing the intensity of the direct signal with the intensity of the reverberation; if reverberation is not mixed in, the weird perception that the source is located at the correct azimuth and elevation but is positioned extremely close to the head or on the surface of the head occurs. Reverberation can be added by physics-based simulation of room acoustics as in the classical Allen-Berkeley image method [8], which is fast enough for real-time processing and is sufficient for achieving the perception of an externalized sound source.

Another property of the source that is external to the user is that it stays stable with respect to a moving user. If the source is motionless, the direction of arrival of the sound signal is changing when the user turns her head in the user's coordinate system. If this cue is not present, externalization is hard because the only stationary point during the rotation is the center of the head, and that is where the source is perceived. To stabilize the audio scene, head tracking is necessary. Multiple methods can be used, including electromagnetic, gyroscopic, mechanical and optical trackers. In the application examples described below, we use the optical tracker based on computer vision described below (Section 4).

2.4. Acoustic rendering pipeline

The acoustic rendering pipeline consists of filtering operations repeatedly done on the incoming audio data stream. Given the position of the virtual sound source with respect to the user's coordinate system, we retrieve the HRTF for a source direction and convolve the block of data with the corresponding head-related impulse response (HRIR) (obtained by inverse Fourier transform of the HRTF). Then, we compute the positions of the first few reflections from the room geometry, and process these image sources in the same way (because they are also located at a particular positions in space and should be perceived as such). The reverberation tail is then added by another convolution operation with a long reverberation filter that is fixed for a given room; the rationale behind this is that the perceived reverberation pattern beyond the first few reflections does not depend much on the position of the source and the receiver in the room. (It does depend significantly though on the room size and wall properties). The rendering pipeline is described in detail in [9] to which we refer the interested reader.

3. PARTIAL AUDIO CUSTOMIZATION

The recently released HRTF database [10], available on the Web at [11], contains HRTF measurements of 43 human subjects and two mannequin measurements (KEMAR with large and small pinnae) along with measurements of 10 ear parameters and 17 body parameters. We use eight ear parameters for matching (see [2] for details) and select the HRTF for the person that has the most similar measurements. It is not known yet whether some features of ear are more important for the localization, so we weight all parameters equally in the matching. We take a picture of the ear of the new system user P and identify few key points on the ear image manually. From these key points measurement values are extracted. If the measured value of i^{th} parameter for the user P is $\hat{d}_i$, the value in the database is d_i and the variance of this parameter in the database across all subjects is $Var(d_i)$, then the error for this parameter is $\varepsilon_i = (\hat{d}_i - d_i)/\sqrt{Var(d_i)}$, the total error is a sum of ε_i and the subject minimizing total error is chosen as a best match and the HRTF set of this subject is used to spatialize the audio for the user P. Matching is done separately for left and right ear, which sometimes results in different matching database subjects for left and right ears.

4. REAL-TIME VIDEO-BASED POSE TRACKING

We propose to solve the pose estimation problem using computer vision. We describe a computer vision system to compute head pose that works in real-time with a cheap and widely available web camera and is accurate and robust. Head pose estimation is a well studied field and many techniques have been proposed to solve this problem [12], [13], [14], [15]. Most of these require a detailed head model, are not real time and compute only a limited range of poses. Our method is based on POSIT (Pose with Orthography and Scaling with Iterations) [3], [16]. POSIT needs four or more non-coplanar model points and their corresponding image projections. It first finds an approximate pose assuming that the image formation model is the scaled orthographic projection (SOP). Approximate depths are computed for each feature point and the points are repositioned on the lines of sight at these depths,

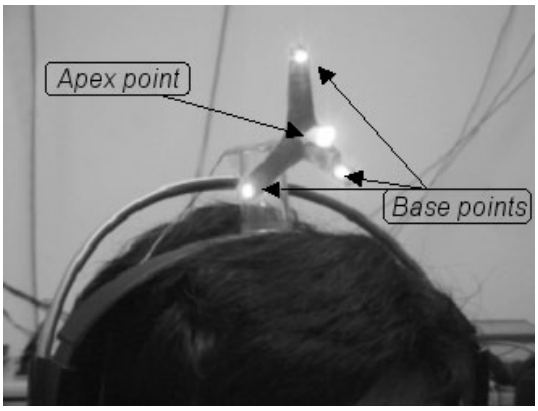

Fig. 2. The tracked object mounted on the system headphones.

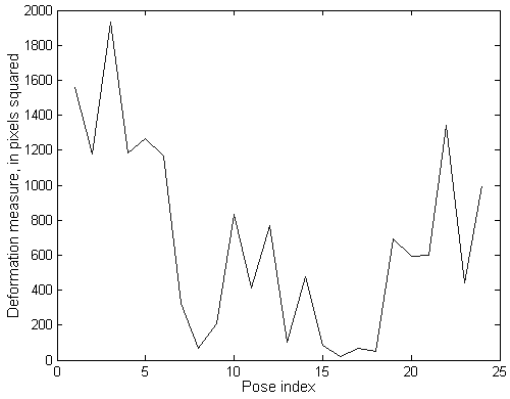

Fig. 3. Deformation error measure versus pose index.

repeating iteratively until convergence to an accurate SOP image and an accurate pose. In our case, the four points necessary for POSIT are four miniature incandescent light bulbs mounted on the structure on the headphones (Figure 2). The camera lens is covered with an infrared filter that blocks off visible light. Only the projections of the lights are visible on the image, and their image coordinates can be obtained by thresholding and then connected component search [17].

We now have to determine the correspondences between the model points and the image projections. For four model points, there are 24 possible matchings between the model and image points. These matchings will result in different poses, out of which only one is the correct pose. POSIT finds the translation vector and transformation matrix that transforms the object onto the camera coordinate system so that the feature points fall on the lines of sight of the image points. If the wrong correspondence has been established, the object must be deformed to adjust it to the line of sight. This deformation can be quantitatively determined by the value of the deformation measure $G = |I \cdot J| + |I \cdot I - J \cdot J|$, where I and J are the first two rows of the transformation matrix found by POSIT. The deformation measure will be zero if the transformation matrix is a scaled rotation matrix, that is if $|I| = |J|$ and $|I \cdot J| = 0$. This deformation measure can be used to solve the correspondence problem completely for a limited number of poses, or to reduce the number of viable poses [3].

The deformation measure is very small for all 24 poses for a symmetrical (tetrahedral) object. For an asymmetrical object, the number of poses for which the deformation measure is small (around four, visible as four local minimas in Figure 3) and doesn't change by increasing the degree of asymmetry. Furthermore, most of the time the apex point is mistaken as one of the base points in the incorrect poses (Figure 2). This suggests that if we select the apex point correctly, we can narrow down to the correct pose most of the times. The apex point is detected by using a bigger light there so that its image projection is larger and can be easily differentiated from the projection of the other model points. Note that perspective enhances this differentiation since in most head positions the apex light is closer to the camera than the other lights.

Once we have the correct correspondence for the apex point, the deformation measure can be computed for the six possible matchings between the base points and their image projections. The pose that results in the smallest value of the deformation measure is selected as the correct pose. However, there are still certain ambiguous positions where an acyclic permutation of the base points also results in a viable pose. This can be dealt with by selecting among the two ambiguous poses the pose for which the user's head is rotated by 90 degrees or less away from the camera.

Note that six pose computations have to be performed for the permutations of the base points for each frame. This computation can be speeded up by noting that we don't need the actual pose for selecting the good one from the bad ones. The six poses will lie far apart in pose space. Therefore, an approximation to the correct pose is enough for differentiating good poses from bad ones. Hence we can use POS (POSIT without the iterations) [3] to compute an initial approximate pose. Once we know the correct correspondence, we can compute the full pose using POSIT. When sufficient information to compute the pose is not available (e.g., the head is occluded or is not entirely in the field of view of the camera), pose computation is suspended and is resumed when all four points are visible again. The system works in real time and is accurate and robust.

5. VIDEO ESTIMATION OF BODY PARAMETERS

Acquisition of a personalized set of HRTFs necessary for audio spatialization is difficult at low frequencies. To compensate for that, the "snowman" head-and-torso model described in [1] is used. For this model, it is necessary to obtain head radius, torso radius and neck height. We have developed a simple computer vision system to obtain these measurements from an image of the subject. We apply background subtraction to separate the subject's body from the background, using a simple color-based scheme which models the background pixels according to the angle subtended by the color vector at the r, g, b axes. This angle also serves as the discrimination measure to separate the foreground from the background. The scheme works only if the background is sufficiently dissimilar to the foreground (the subject's body) in color space. This is however not an important limitation in our case since we have full control over the background. Then, thresholding followed by connected components operation eliminates small spurious regions while morphological operators smooth the contours of the foreground component, and finally the contour detection operator from the OpenCV library extracts the foreground silhouette. The thinnest region of the silhouette is the neck diameter, the broadest region is the torso diameter and the broadest region

above the neck is the head radius; these dimensions are obtained by scanning the image from left to right. To calibrate the mapping from pixel coordinates to world coordinates and to find approximate room dimensions for the reverberation synthesis algorithm, we measure the image projection of known length in the world.

6. PERFORMANCE AND APPLICATIONS

The video tracking and audio rendering subsystems form the basis of several applications described below. They work together in real-time on a dual processor PC. No specialized hardware (such as DSP boards) are used, and multithreaded programming is utilized to efficiently load both CPUs. The video tracking is done at camera frame-rate (30 fps). The audio rendering is done in blocks of 2048 samples which translates to the latency of less than 50 ms at a sampling rate of 44.1 KHz. The audio rendering pipeline has to submit the freshly computed block of data before the block that is currently playing is done; this enforces constraints on how much processing can be done. Generally, as much recomputation of the acoustic geometry and image sources (obtained by reflections of true sources in room walls) to update the rendering filters and account for the source and the receiver motion is done as possible in this limited time frame. In the described hardware configuration, it is possible to compute reflections of up to fifth order in real time with one acoustic source and up to third order with 8 sources.

Subjectively, the system is characterized as extremely convincing by a large group of people (about 150 so far) in the simple test where the sound source is "attached" to a small object that a person can move around in 3D-space herself. Even though the sound is delivered through headphones, people are generally tricked into believing that the sound is coming from the object; externalization achieved is very good, and localization of the sound coincides with the location of the object. In more formal tests, the person had to point in the direction of the sound source presented at a random position in 3D-space. These tests showed the negative impact of using non-individualized HRTF sets; further experiments with semi-customization using algorithms described in Section 3 revealed that the localization performance is improved by 25-30% by our simple method of customization (see [9]). According to subject reports, customization (as well as implementation of low-frequency "snowman" HRTF model from [1]) also improves the subjective quality of the audio scene rendering.

We have created several sample applications using video-based tracking and audio rendering systems described above. The simplest one is the demonstration mentioned above where a sound source is rendered in the 3D-space at the position of a small object that the user can move. Ironically, this simplest demo turned out to be the most impressive, probably due to the strong motor-sensory feedback. Another interesting application that is enjoyed by most participants is the shooting (or listening) game with spatial audio interface. In the game, the participant is immersed in the virtual world with several objects flying around and can move using keyboard or head motion commands. The objects are producing sounds (such as play music, make noise, or broadcast the audio news from the Internet) and spatialized audio algorithms are used to render them in their proper spatial positions to achieve consistency between audio and video modalities. The concept demonstrated by the game is the extension of the limited video field of view by a full sphere audio field; new targets are often located first by the sound they make and only then the player brings them into visual field of attention. The player can follow the object to listen to it, or shoot it and break it if she doesn't like its sound, in which case new targets with different sounds reemerge shortly.

7. CONCLUSIONS

We described the algorithms for using computer vision for real-time user pose tracking and determining body parameters in synthesis of spatial audio; we also briefly discussed the audio spatialization algorithms. Our work is aimed at enabling low-cost spatial audio using typical office PCs and inexpensive cameras. The developed methods are currently used in ongoing research in our lab. This research includes customizable spatial audio user interfaces for low-sighted population and studies on sonification and increasing separability of perception of multiple audio streams, using certain manipulations such as spatialization and pitch and timbre shifts.

8. REFERENCES

[1] V. R. Algazi, R. O. Duda, D. M. Thompson (2002). "The use of head-and-torso models for improved spatial sound synthesis", Proc. AES 113th Convention, Los Angeles, CA.

[2] D. N. Zotkin, R. Duraiswami, L. S. Davis, A. Mohan, V. Raykar (2002). "Virtual audio system customization using visual matching of ear parameters", Proc. IEEE ICPR 2002, Quebec City, Canada, pp. 1003-1006.

[3] D. DeMenthon and L. Davis (1995). "Model-based object pose in 25 lines of code", Intl. J. Comp. Vision, vol. 15, pp. 123-141.

[4] J. W. Strutt (Lord Rayleigh) (1907). "On our perception of sound direction", Philosophical Mag., vol. 13, pp. 214-232.

[5] D. Wright, J. Hebrank, and B. Wilson (1974). "Pinna reflections as cues for localization", J. Acoust. Soc. Am., vol. 56(3), pp. 957-962.

[6] W. M. Hartmann (1999). "How we localize sound", Physics Today, November 1999, pp. 24-29.

[7] M. B. Gardner and R. S. Gardner (1973). "Problem of localization in the median plane: effect of pinna cavity occlusion", J. Acoust. Soc. Am., vol. 53(2), pp. 400-408.

[8] J. Allen and D. Berkeley (1979). "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., vol. 65(5), pp. 943-950.

[9] D. N. Zotkin, R. Duraiswami, and L. S. Davis. "Rendering localized spatial audio in a virtual auditory space", IEEE Trans. on Multimedia, in press.

[10] V. R. Algazi, R. O. Duda, D. P. Thompson, and C. Avendano (2001). "The CIPIC HRTF database", Proc. IEEE WASPAA01, New Paltz, NY, pp. 99-102.

[11] http://interface.cipic.ucdavis.edu/

[12] R. Azuma and G. Bishop (1994). "Improving static and dynamic registration in an optical see-through HMD", Proc. ACM SIGGRAPH 1994, Orlando, FL, pp. 197-204.

[13] T. Darrell, B. Moghaddam, and A. P. Pentland (1996). "Active face tracking and pose estimation in an interactive room", IEEE CVPR 1996, San Francisco, CA, pp. 67-71.

[14] P. Fieguth and D. Terzopoulos (1997). "Color based tracking of heads and other mobile objects at video frame rates", Proc. IEEE CVPR 1997, Puerto Rico, pp. 21-27.

[15] M. L. Cascia, S. Sclaroff and V. Athitsos (2000). "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models", IEEE PAMI, vol. 22(4), pp. 322-336.

[16] P. David, D. DeMenthon, R. Duraiswami, and H. Samet (2002). "SoftPOSIT: Simultaneous pose and correspondence determination", Proc. ECCV 2002, Copenhagen, Denmark, vol. 3, pp. 698-714.

[17] A. Mohan (2002). "Real time head pose estimation for creating virtual auditory spaces", M.S. Thesis, Dept. of Electrical and Comp. Eng., Univ. of Maryland, College Park, MD.