Otology & Neurotology
39:931–939 © 2018, Otology & Neurotology, Inc.

Augmented Reality of the Middle Ear Combining Otoendoscopy and Temporal Bone Computed Tomography

*Roberto Marroquin, *Alain Lalande, *Raabid Hussain, †Caroline Guigou, and *†Alexis Bozorg Grayeli

**Le2i Laboratory, University of Burgundy-Franche Comté; and †Otolaryngology-Head and Neck Surgery Department, University Hospital of Dijon, Dijon, France*

Hypothesis: Augmented reality (AR) may enhance otologic procedures by providing sub-millimetric accuracy and allowing the unification of information in a single screen.
Background: Several issues related to otologic procedures can be addressed through an AR system by providing sub-millimetric precision, supplying a global view of the middle ear cleft, and advantageously unifying the information in a single screen. The AR system is obtained by combining otoendoscopy with temporal bone computer tomography (CT).
Methods: Four human temporal bone specimens were explored by high-resolution CT-scan and dynamic otoendoscopy with video recordings. The initialization of the system consisted of a semi-automatic registration between the otoendoscopic video and the 3D CT-scan reconstruction of the middle ear. Endoscope movements were estimated by several computer vision techniques (feature detectors/descriptors and optical flow) and used to warp the CT-scan to keep the correspondence with the otoendoscopic video.
Results: The system maintained synchronization between the CT-scan image and the otoendoscopic video in all experiments during slow and rapid (5–10 mm/s) endoscope movements. Among tested algorithms, two feature-based methods, scale-invariant feature transform (SIFT); and speeded up robust features (SURF), provided sub-millimeter mean tracking errors (0.38 ± 0.53 mm and 0.20 ± 0.16 mm, respectively) and an adequate image refresh rate (11 and 17 frames per second, respectively) after 2 minutes of procedure with continuous endoscope movements.
Conclusion: A precise augmented reality combining video and 3D CT-scan data can be applied to otoendoscopy without the use of conventional neuronavigation tracking thanks to computer vision algorithms. **Key Words:** Augmented reality—Labyrinthine windows—Middle ear—Minimally invasive surgery—Otoendoscopy—Transtympanic surgery.

Otol Neurotol **39:**931–939, 2018.

Augmented reality (AR) is a technology that enriches the sensorial perception by adding virtual contents directly on the user's environment (1,2). In contrast to virtual reality, AR enhances the reality by supplementing it rather than replacing it. The main components needed for an AR system are: a video camera that captures the real image of the surrounding environment, a tracking module that measures the relative position and orientation of the camera, a graphics processing module that processes the virtual objects according to the estimated position of the camera and finally a display that merges and renders the information from the video camera and the virtual objects (1).

Transtympanic procedures have been designed to access target structures in the middle ear cleft through a minimal opening of the tympanic membrane for diagnostic and therapeutic purposes (3–8). These procedures have many theoretical advantages such as reduced bleeding, preserved tympanic membrane position, faster procedure, and simpler postoperative course. Indeed, in an endoscopic transtympanic procedure, the tympanomeatal flap is not elevated and the tympanic membrane remains at its position during the whole procedure. This technique has been applied to in situ drug administration (3,7), labyrinthine fistula diagnosis (4) and ossicular chain repair (5,6). However, the anatomy of the round window is highly variable and drug availability inside the inner ear after intratympanic administration is significantly reduced due to this variability (9). In many cases, transillumination does not allow the visualization of the round window (thick and fibrotic tympanic membrane). Transtympanic procedures are a particular case of transcanal procedures.

The use of an otoendoscope in combination with other surgical instruments reduces the available space in the ear

Address correspondence and reprint requests to Roberto Marroquin, M.Sc., Laboratoire Le2i, Université de Bourgogne Franche-Comté, UFR Sciences et Techniques, allée Alain Savary, 21000 Dijon, France; E-mail: roberto-enrique.marroquin-cortez@u-bourgogne.fr

The financial support is from the Oticon Medical, Société ORL de Bourgogne, CNRS, and Collin Medical SA.
The authors disclose no conflicts of interest.
Supplemental digital content is available in the text.
DOI: 10.1097/MAO.0000000000001922

canal and limits movement. Moreover, the otoendoscope should be kept immobile since movements may tear the tympanic membrane. This immobility results in a reduced field of view (6,10). Several issues related to the transcanal approaches can be addressed through AR, by providing a global view of the middle ear cleft and by advantageously supplementing the otoendoscope, and thus potentially broadening the field of application of transcanal procedures.

AR can, of course, also be used with a standard tympanic membrane flap. In this case, AR would be useful to visualize unexposed middle ear cleft regions. Although ossicular chain repair through a myringotomy by human hand and with current ossicular prostheses is challenging and justifiably uncommon, AR can open the way to keyhole robot-based procedures since robots can perform very stable and yet complex movements through a puncture hole (11). This keyhole robot-based surgery, currently under development (12), is only possible if the surgeon has a clear vision of the target structures and this is exactly what AR provides. In addition, the surgical planning conducted on preoperative imaging can be improved by AR, intraoperatively, for complex middle ear lesions.

Vision-based AR is an application of computer vision which is the branch of computer science that deals with acquiring, processing, and analyzing images and videos. AR, combining CT-scan and operative field images, has been successfully developed for different types of surgeries such as neurosurgery (13), urinary tract endoscopy (14), and oral surgery (13,15). Different computer vision techniques (e.g., image registration, motion tracking, scene reconstruction, segmentation and object classification) have been used to create AR for bronchoscopy (16–18), colonoscopy (19–21), and rhinoscopy (22,23). The theoretical advantage of these techniques is that the correspondence between video and CT-scan is based on video image analysis and not on a conventional tracking system. Vision-based AR approach to surgery potentially reduces installation/set-up costs by not requiring extensive equipment.

In otologic procedures, the challenges are to obtain high precision and enhanced ergonomy (24). We hypothesized that AR based on computer vision techniques can provide sufficient precision with enhanced ergonomy by allowing the unification of information in a single screen, in contrast to conventional navigation systems where the endoscopic feed and the information from imaging modalities such as CT and MRI are available on separate screens. To our knowledge, AR has not yet been implemented in otoendoscopy.

In this work, we aimed at developing a stand-alone AR system for transcanal procedures by combining the routine otoendoscopy video images with 3D images of the middle ear cleft obtained through a high-resolution CT-scan.

MATERIALS AND METHODS

The workflow was divided into three sections: input and experimental setup, registration, and tracking (Fig. 1) (see Figure, Supplemental Digital Content 1, http://links.lww.com/MAO/A651, which shows a detailed schematic of the system workflow).

Experimental Setup

To verify the method in different anatomical configurations, four human temporal bone specimens were included in this study. Two types of experiments were performed: with and without the tympanic membrane on the same temporal bone (Fig. 2). All specimens underwent high-resolution CT-scan (Light Speed, 64 detector rows, General Electric Medical Systems, Buc, France). Axial, coronal, and sagittal views were obtained (field of view 60 mm; slice thickness 0.6 mm; overlapping slice interval 0.3 mm). Each voxel measured $0.6 \times 0.6 \times 0.3\,\text{mm}^3$ (see Figure, Supplemental Digital Content 2, http://links.lww.com/MAO/A652, which depicts the experimental setup).

3D reconstruction of the middle ear cleft, based on DICOM (National Electrical Manufacturers Association, Rosslyn, VA) data, was conducted using the Osirix (Osirix V.5.6, Pixmeo, Geneva, Switzerland) virtual endoscope function. The 3D reconstruction was obtained by placing the virtual endoscope in the external auditory canal facing the umbo, 10 mm lateral to the tympanic membrane. This image of the middle ear cleft structures was used as the reference to warp around the otoendoscopic video. In parallel, otoendoscopy was performed for all temporal bone specimens with a 1.9 mm 0 degree otoendoscope (Hopkins endoscope, Storz, Tuttlingen, Germany) connected to a high definition (HD) recording device (Telepack X, with Telecam mono CCD camera, Storz, Tuttlingen, Germany). A set of otoendoscopy videos were recorded at 25 frames per second (FPS) for 2 minutes each, during which the otoendoscope was manually moved inside the auditory canal with estimated slow (<5 mm/s) or rapid (5–10 mm/s) translations, rotations, and pitches (Table 1).

The proposed system was developed on a desktop computer (iMac, 2.9 GHz Intel Core i5 processor, 8 GB 1600 MHz DDR3 RAM, NVIDIA GeForce GT 750 M 1024 MB graphic card, OS X Yosemite 10.10.1 operating system). An OsiriX plugin was developed, to facilitate the acquisition of the virtual reconstructed CT image, using C++ and Objective-C languages with OpenCV (Open source Computer Vision) which is a library of programming functions aimed at real-time computer vision.

AR was implemented by combining the recorded video images of the external auditory canal and the tympanic membrane with the 3D CT-scan reconstruction of the middle ear cleft.

Registration

The system had two inputs: The 3D CT-scan reconstruction of the middle ear comprising the entire sulcus tympani, and the video recording of the otoendoscopy of the corresponding specimen. The registration between these two inputs began by a manual selection of corresponding points between these two inputs, i.e., the reconstructed CT-scan image and the first otoendoscopic image. These points will be referred to as reference points p_j and p'_j, representing points in the otoendoscopic image and the CT-scan image, respectively. Given N is the number of pair of reference points, $j = \{1,2,\ldots,N\}$.

The tympanic membrane borders were used for the selection of the reference points. To relate the reference points, a registration homography matrix H_R could be calculated such that,

$$p_j \approx H_R p'_j, \tag{1}$$

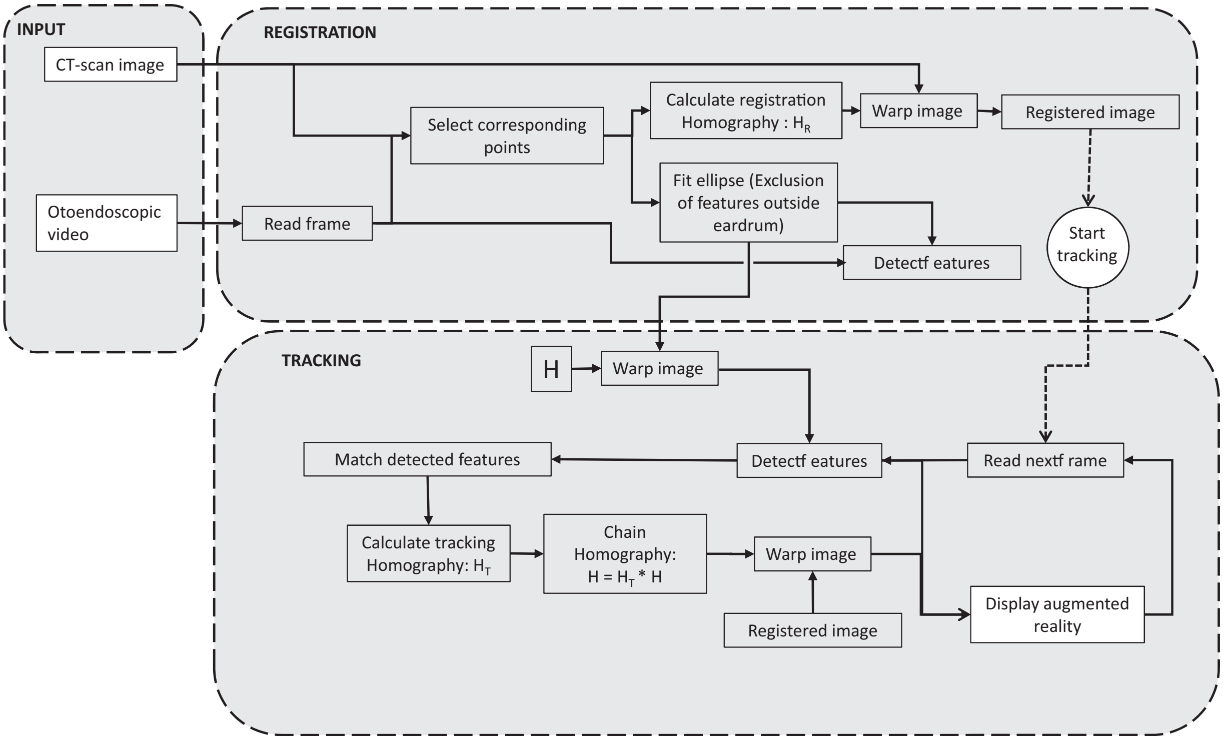

FIG. 1. The workflow. The methodology was divided into three parts: firstly, the input, i.e., the otoendoscopic video stream and the CT-scan image; secondly, the registration of a frame of the otoendoscopic video with the CT-scan image, performed once at the beginning of the process; and thirdly the tracking which consists of estimating the motion of the otoendoscope and of warping the registered CT-scan image to keep the synchronization between the CT-scan image and the otoendoscopic video. This part is repeated every frame. CT indicates computer tomography.

where p_j and p'_j were given in homogeneous coordinates, i.e., $p_j = (x_j, y_j, 1)^T$, $p'_j = \left(x'_j, y'_j, 1\right)^T$ with $(\cdot)^T$ denoting transpose. The registration homography matrix was defined as,

$$H_R = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix}, \tag{2}$$

where the entry $h_{22} = 1$ as H_R can only be defined up to a scale (25) and the rest of the entries were unknowns, i.e., H_R was considered as a projective homography. Theoretically, at least four pairs of reference points ($N \geq 4$) are needed to compute a homography matrix (25). In preliminary trials, an optimum number of correspondence points was not found for $N \geq 6$, thus it was decided to use six pairs during the experiments. After the selection of the reference points, the matrix H_R could be computed by minimizing the back-projection error function as follows:

$$mim \sum_{j=1}^{N} \left(x_j - \frac{h_{00}x'_j + h_{01}y'_j + h_{02}}{h_{20}x'_j + h_{21}y'_j + 1} \right)^2 + \left(y_j - \frac{h_{10}x'_j + h_{11}y'_j + h_{12}}{h_{20}x'_j + h_{21}y'_j + 1} \right)^2, \tag{3}$$

Since we had more constraints than unknowns ($N = 6$), the minimization process (i.e., finding the optimal homography transformation) was conducted according to a least-squares method and the Levenberg–Marquardt algorithm (26).

The computed registration homography (H_R) was then used to warp the CT-scan image to obtain the registered image.

Tracking

Several computer vision techniques were evaluated with the aim of estimating the endoscope movement to maintain correspondence between the video and the warped CT-scan image during the procedure:

1) The optical-flow-based method consists of calculating a pattern of motion between two consecutive frames caused by the movement of an object or the camera. The pattern, represented by a 2D displacement vector, reflects the movement of the selected points from one frame to the next. In this study, the Lucas-Kanade Optical-Flow algorithm (OF) (27) combined with feature from accelerated segment test (FAST) (28) was evaluated.
2) The feature-based method consists of finding image features (e.g., corners, blobs, edges, lines, or other well-localized structures in two dimensions) and tracking them as they move between two consecutive frames. This method involves feature description, detection, and

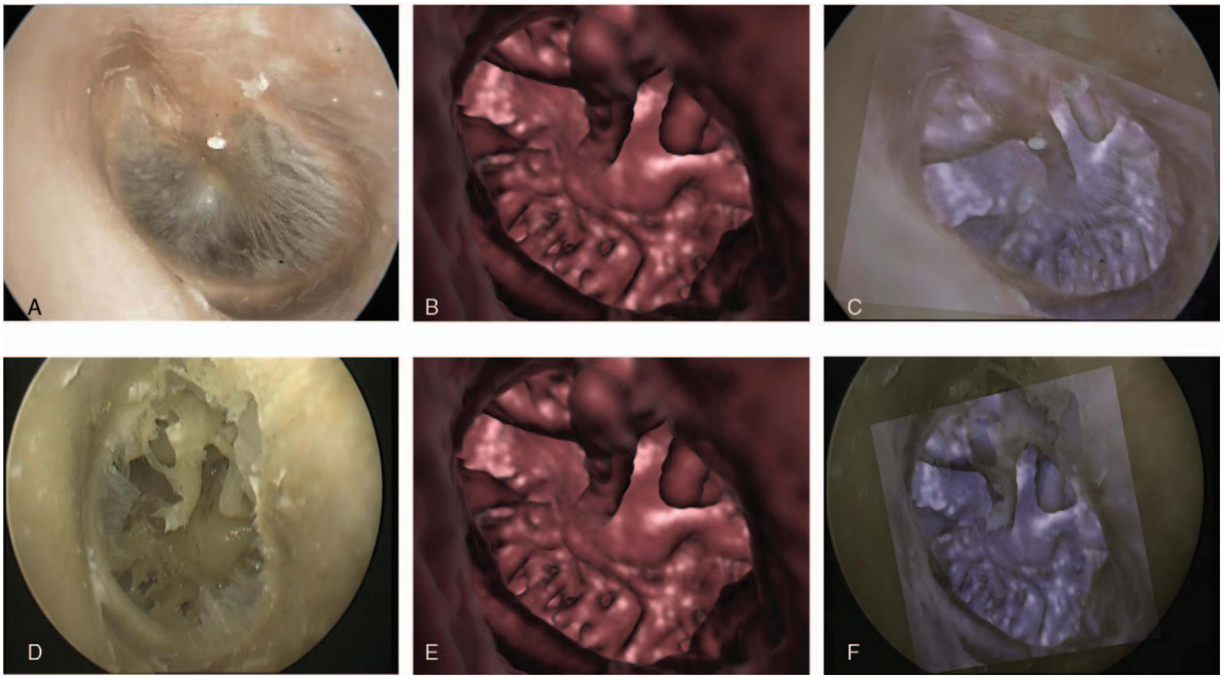

FIG. 2. Otoendoscopic image combined with the CT-scan image. Two different scenarios were considered from the same human temporal bone; with (*A–C*) and without tympanic membrane (*D–F*). *A*, Otoendoscopic image showing the tympanic membrane. *B*, 3D CT-scan image. *C*, Registration between images (*A*) and (*B*) displayed in an AR form. *D*, Otoendoscopic image where the tympanic membrane was removed. *E*, Is the same as (*B*). (*F*) Registration between images (*D*) and (*E*) in an AR form. The images (*C*) and (*F*) were converted from RGB (red-green-blue) color space to BGR (blue-green-red) to better differentiate the otoendoscopic and CT-scan images. AR indicates augmented reality; CT, computer tomography.

matching. The feature-based algorithms evaluated in our study were scale-invariant feature transform (SIFT) (29), speeded up robust features (SURF) (30), and oriented FAST and rotated BRIEF (binary robust independent elementary features, ORB) (31).

A comparison between OF, SIFT, SURF, and ORB methods was conducted to determine the optimum method for our project. Multiple features were detected and described on the otoendoscopic image by each of the above methods. For example, Figure 3A shows the features detected using the SURF method. There was no predefined number of detected features. The number depended on the method and on the image itself. To obtain an accurate homography matrix, coplanar points are essential (25). Since not all the features were in the same plane (tympanic membrane, ear canal), the features outside the tympanic membrane were filtered out using an ellipse representing the tympanic membrane (Fig. 3B and C). The ellipse was obtained by computing a conic fitting algorithm with the reference points on the otoendoscopic image (32). The impact of this filtering method on the accuracy was evaluated. If the ellipse was not used, all detected features inside and outside the tympanic membrane were considered.

A robust feature matching step is then conducted to determine the corresponding feature matches between the current and previous frames. To make the feature matching robust, ambiguous cases, in which more than one feature point could be a possible match, were discarded. Subsequently, the tracking homography matrix (H_T) was estimated by using the previously

TABLE 1. *Experimental conditions*

	Experiments					
Characteristics	1	2	3	4	5	6
Temporal bone no.	1	2	3	4	3	2
X, Y translations	Yes	Yes	Yes	Yes	Yes	Yes
Z translation	No	No	No	Yes	Yes	Yes
Rotation	No	Yes	Yes	Yes	Yes	Yes
Tympanic membrane	Yes	Yes	Yes	Yes	No	No
Speed	Low	Low	High	High	High	High

The experiments were performed during 2 minutes and using six initial reference points. Rotation: rotation around *z* axis. Otoendoscope speed: low <5 mm/s, High ≥5 to 10 mm/s.

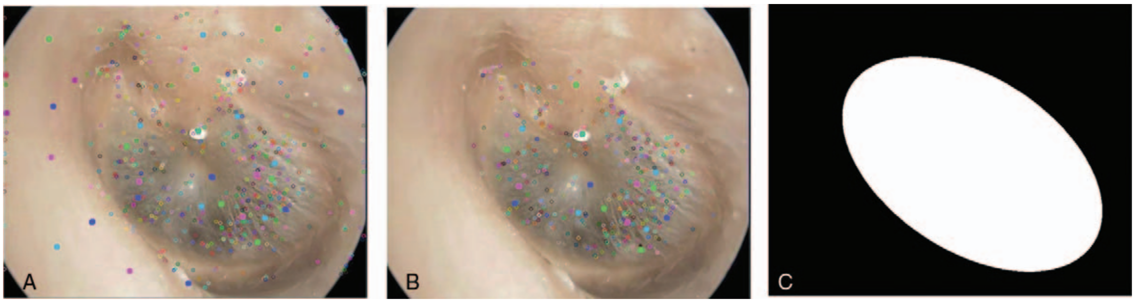

FIG. 3. Features detected with and without the ellipse. *A*, Five hundred fifty three features detected using SURF method without the ellipse. *B*, Three hundred ninety two features detected using SURF method with the ellipse. *C*, Ellipse used to filter out the features outside the tympanic membrane. SURF indicates speeded up robust features.

obtained features matches. This matrix represented the movement between the two frames and its calculation was coupled to a robust process to discard outlier matches (random sample consensus method, RANSAC (33)).

To avoid accumulation of interpolation errors with time in the case of H_T estimation at each frame, we chained the homography H_T as a cumulative homography (see Figure, Supplemental Digital Content 3, http://links.lww.com/MAO/A653, which depicts the homography chaining and the accumulation of interpolation errors):

$$H = H_T * H, \tag{4}$$

where H was initialized as the identity matrix.

Finally, to create the AR view, a linear blend operator (also known as a weighted sum) was used between the otoendoscopic image and the warped registered image:

$$g(x) = \alpha f_0(x) + (1 - \alpha) f_1(x), \tag{5}$$

where $g(x)$ is the AR image, $f_0(x)$ is the otoendoscopic image, $f_1(x)$ is the warped registered image and α is the blending factor which can vary between 0 and 1.

Error Estimation

From Eq. (1) it can be observed that warping a point with H_R yields some error. The mean registration error ($error_R$) can be defined by firstly reconsidering Eq. (1) and letting

$$\hat{p}_j = H_R p'_j, \tag{6}$$

$$\hat{p}_j \approx p_j, \tag{7}$$

where $\hat{p}_j$ is the position of the point p'_j in the otoendoscopic image. From the Euclidean distance d_j between the points p_j and their estimated positions $\hat{p}_j$. The mean registration error for the N pairs of reference points will be computed as:

$$error_R = \sqrt{\frac{\sum_{j=1}^{N} d_j^2}{N}}, \tag{8}$$

In the same manner, the mean tracking error ($error_T$) was calculated by measuring the distance between the real positions of the reference points (manually selected) and their estimated positions by the methodology at different times. Even if the manual selection introduced some error in the evaluation, it provided a simple, robust, and quantitative evaluation of the method at various time-points. First, the manually selected points were defined as p'_j which ideally represents the positions of the reference points p_j at a given time. Then, the positions of the points p_j estimated by the methodology at a given time were defined as:

$$\breve{p}_j = H p'_j, \tag{9}$$

where H was the cumulative matrix. From the Euclidean distance d'_j between the points p'_j and their estimated positions $\breve{p}_j$. The mean tracking error for the N pairs of reference points at a time point was computed as:

$$error_T = \sqrt{\frac{\sum_{j=1}^{N} d'_j 2}{N}} - error_R, \tag{10}$$

To convert the error in pixels into millimeters (mm), the diameter (in mm) of the tympanic membrane was determined across the malleus bone on native CT views. The measuring tool in the 3D multiplanar reconstruction (MPR) function of the OsiriX software was used for this purpose. Then, the same diameter was measured on the video in pixels. The physical error was then computed as follows:

$$error_{mm} = \frac{diameter_{mm}}{diameter_{pixel}} \times error_{pixel}, \tag{11}$$

where $error_{pixel}$ can represent $error_R$ or $error_T$.

As a second series of experiments, the tympanic membrane was removed and two target points were selected on both CT-scan image and the otoendoscopy image: the highest point of the round window niche and the lowest point of the incus. The endoscope was translated, rotated, and displaced along its axis slowly during 2 minutes in all cases. The initial mean registration error and the drift were measured for each target point at 30, 60 and 120 seconds.

Statistics

Data were analyzed using Prism (v. 5, Graphpad Inc., La Jolla, CA). Values were expressed as the mean ± standard deviation. Mean tracking error was compared between methods

TABLE 2. *Comparison of tracking error of methods with and without ellipse*

		Mean Tracking Error (mm)				
Method	FPS	30 seconds	60 seconds	90 seconds	120 seconds	Overall Error (mm)
SURF	17	0.05 ± 0.10	0.18 ± 0.23	0.17 ± 0.15	0.20 ± 0.16	0.15 ± 0.15
SURF-E	18	0.06 ± 0.09	0.11 ± 0.12	0.14 ± 0.19	0.21 ± 0.33	0.13 ± 0.19
SURF versus SURF-E	–	−0.01	0.07	0.03	−0.01	–
SIFT	11	0.06 ± 0.06	0.13 ± 0.10	0.15 ± 0.18	0.38 ± 0.53	0.28 ± 0.38
SIFT-E	10	0.07 ± 0.07	0.12 ± 0.13	0.38 ± 0.43	0.26 ± 0.24	0.21 ± 0.26
SIFT versus SIFT-E	–	−0.01	0.01	−0.23	0.12	–
ORB*	27	0.22 ± 0.34	0.55 ± 0.52	0.20 ± 0.09	0.59 ± 0.66	0.51 ± 0.48
ORB-E**	27	0.24 ± 0.40	0.99 ± 1.24	1.00 ± 1.12	0.90 ± 0.80	0.78 ± 0.91
ORB versus ORB-E***	–	−0.02	−0.44	−0.80	−0.31	–
OF*	30	0.16 ± 0.21	1.47 ± 1.04	1.90 ± 1.47	6.91 ± 11.15	2.66 ± 4.73
OF-E**	28	0.21 ± 0.23	6.88 ± 11.86	64.66 ± 85.90	8.62 ± 12.89	20.09 ± 47.48
OF vs OF-E***	–	−0.05	−5.41	−62.76	−1.71	–

Mean tracking errors and standard deviations for all the experiments for each method at different times. The overall errors were computed as the mean and standard deviation errors per method (each experiment was dealt separately). For the comparison of methods with and without ellipse, the methods without the ellipse were taken as the reference. FPS: maximum number of frames processed per second. SURF indicates speeded up robust features, SIFT, scale-invariant feature transform, ORB, oriented FAST (feature from accelerated segment test), and rotated BRIEF (binary robust independent elementary features), OF, Lucas-Kanade optical-flow algorithm with FAST. The methods using an ellipse have been marked with ''-E''.

*$p < 0.05$ from the ANOVA test between the different methods without an ellipse (SURF, SIFT, ORB, OF).

**$p < 0.05$ from the ANOVA test between the different methods using an ellipse (SURF-E, SIFT-E, ORB-E, OF-E).

***$p < 0.05$ from the paired sample t test between the different methods using an ellipse (SURF versus SURF-E, SIFT versus SIFT-E, ORB versus ORB-E, OF versus OF-E).

with the ellipse (SURF-E, SIFT-E, ORB-E, OF-E) using a one-way ANOVA test followed by Bonferroni correction. The same evaluation was done between methods without the ellipse (SURF, SIFT, ORB, OF). Also, the impact of the ellipse was analyzed using a paired sample t test between the methods with and without the ellipse (SURF versus SURF-E, SIFT versus SIFT-E, ORB versus ORB-E, OF versus OF-E). A p-value <0.05 was considered as significant for all tests.

RESULTS

The system maintained synchronization between the CT-scan image and the otoendoscopic video in all experiments during slow and fast (5–10 mm/s) endoscope movements (see Video, Supplemental Digital Content 4, http://links.lww.com/MAO/A654, which shows the AR system in operation). However, the mean tracking error increased with time in all methods (Table 2). In experimental conditions 5 and 6 without the tympanic membrane, an exact synchronization was observed between the video and the CT-scan reconstruction of the visible ossicles (malleus and incus, Fig. 2D–F) (see Video, Supplemental Digital Content 4, http://links.lww.com/MAO/A654, which demonstrates the correspondence between the visible ossicles during an experiment without tympanic membrane).

Each computer vision method (i.e., SIFT, SIFT-E, SURF, SURF-E, ORB, ORB-E, OF, and OF-E) presents different results owing to their distinct characteristics. OFs and ORBs methods accumulated errors above 1 mm, whereas the precision of SURFs and SIFTs methods was always better than below 1 mm during the entire procedure (Table 2). Surprisingly, the addition of the ellipse seemed to deteriorate the performance of OF and ORB ($p < 0.01$, Table 2), although it did not seem to influence the performance of SIFT and SURF ($p = 0.24$ and $p = 0.53$, respectively). This might be because the generated ellipse (using OF and ORB) excludes some suitable features and includes unsuitable ones.

The processing speeds of each method were also compared with their image refresh rates. The processing speeds were computed by counting the number of frames processed per second (FPS). The time to process one frame includes the time to generate the AR display window and, more importantly, the time to track the endoscope. Moreover, the time to track the endoscope depends essentially on the computer vision method used, and not on the video-recorded FPS, that is why it can be higher or lower than 25 FPS. The OF and OF-E methods were the fastest with 30 and 28 FPS, respectively. They were followed by ORB and ORB-E (both 27 FPS), SURF-E (18 FPS), and SURF (17 FPS). Finally, SIFT and SIFT-E appeared to be the slowest methods with 11 and 10 FPS, respectively.

Table 3 displays the mean tracking error at two particularly useful targets in surgery (the round window niche and the incus) obtained with the SURF-E method in four temporal bones after the removal of the tympanic membrane. For both targets, the mean tracking error remained below 0.15 mm and the maximum error was less than 0.35 mm (obtained for the round window niche). Figure 4 shows an example of this tracking.

TABLE 3. *Mean tracking error at the level of significant targets in the middle ear*

		Drift (mm)			
	Initial MRE (mm)	30 second	60 second	90 second	120 second
Round window niche	0.25 ± 0.16	0.04 ± 0.03	0.13 ± 0.13	0.07 ± 0.04	0.14 ± 0.17
Incus	0.15 ± 0.32	0.04 ± 0.03	0.04 ± 0.04	0.10 ± 0.07	0.09 ± 0.08

Mean tracking error and standard deviation obtained with the SURF-E (speed up robust features with ellipse) method for four experiments on different temporal bones (without tympanic membrane) at the round window niche (highest point of its rim) and the incus (lowest point). Initial MRE: mean registration error (before the tracking). The mean tracking error is independent to the MRE.

DISCUSSION

We developed a stand-alone system that generates AR by combining CT-scan data to otoendoscopy for a minimally invasive approach to middle ear cleft. We compared several image-based algorithms to estimate the otoendoscopic movements and to adapt the virtual CT-based endoscopic images to the video. We showed that SIFTs and SURFs methods are suitable for high precision otological procedures as they provided a precision better than 1 mm which held throughout the procedure, with SURF-E exhibiting highest FPS compared with SURF, SIFT, and SIFT-E. Although providing high image refresh rates, OFs and ORBs methods demonstrated high cumulative tracking errors (close to/higher than 1 mm) making them incompatible for surgeries where sub-millimetric precision is required.

The use of the ellipse had no significant influence on the accuracy of SIFT and SURF methods, however, it deteriorated the performance of OF and ORB, resulting in a higher tracking error. We expected that the use of the ellipse would enhance the results by eliminating the points in the external auditory canal and by taking into account only the coplanar points on the tympanic membrane. In this way, the homography matrix could be more precise. However, the coplanarity of the points on the tympanic membrane was also an approximate assumption and eliminating the points on the external auditory canal did not increase the precision of the matrix. Furthermore, the use of the ellipse was expected to provide faster processing (higher FPS), since a great number of features were filtered out. However, no notable increase in processing speed was observed owing to the fact that the time saved during the feature matching step (as there are lesser features to match) was lost during the ellipse feature filtration process.

Until today, navigation in otology has been considered as an accessory tool which in selected complex cases may guide the surgeon beyond the middle ear limits (24,34,35). The ergonomy of the CT-scan navigation is generally poor since the surgeon is forced to go back and forth between the navigation screen and the microscope view (24). This obstacle is even more important when operators combine a drill to the navigation system (36). Researchers have attempted to improve the ergonomy by a navigation-assisted and co-manipulated robotic arm (37) and more recently the AR in neurosurgery (38) and sinus surgery (13). Indeed, AR promisingly enhances the ergonomy by allowing the surgeons to visualize the operative scene and its corresponding view in different imaging modalities in real time on a single screen (13). Generally, AR systems are combined to a conventional navigation system using optical tracking systems (13,39). Recently, a stand-alone AR system based on fiducial markers and image tracking has been reported for oral surgery (15). The theoretical advantage of such a system

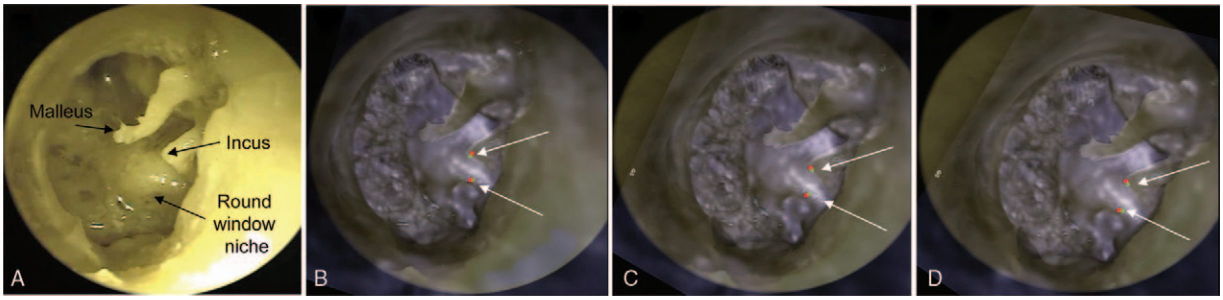

FIG. 4. Target location in the middle ear. After registration of the endoscopic image by using six points on the tympanic rim and the removal of the tympanic membrane the highest point of the round window niche and the lowest point of the incus were selected as targets and tracked on both video endoscopy and 3D CT-scan reconstruction. *A*, Initial otoendoscopic image without the tympanic membrane. *B*, AR before the tracking (initial registration). *C*, AR after 60 seconds. *D*, AR after 120 seconds. On *B*, *C*, and *D* images, the *arrows* indicate the tracking errors at the level of the round window niche and of the incus (reference points in green and estimated points in red). AR indicates augmented reality, CT, computer tomography.

is to reduce errors and inaccuracy due to numerous factors (i.e., patient, imaging, navigation system, and AR system) and to simplify installation and setup.

In otology, the challenge is even greater than in other AR applications since a sub-millimeter precision is necessary. Based on this study, we expect that this level of precision can be reached in laboratory conditions. Although, the AR approach with a method such as SURF-E provides a precise projection of the middle ear elements, it must be extended to incorporate depth information. Moreover, during a transcanal procedure, once the instrument (such as a Rosen needle) goes behind the tympanic membrane, its tip becomes invisible on the video and the reconstructed CT-scan views. Therefore, a virtual reconstruction of the instrument will be necessary to inform the surgeon about the ongoing procedure (to enable the operator to direct the instrument toward the target). This reconstruction could be based on the position information provided by a robot holding the instrument, or even image-based methods (and the depth information could be obtained if three collinear points are identified). This step seems indispensable before the application of this method in the operating room.

Another limitation of this method is the manual setting of the reference points which is a potential source of inaccuracy. The automatization of the initial registration procedure needs to be developed. Positioning several radio-opaque and visible markers inside the external auditory canal before the CT-scan may allow this automatization and increase the precision.

Transtympanic procedures (particular cases of transcanal procedures) have been reported by several other teams as a keyhole approach to the middle ear cleft through the tympanic membrane (5–8). The size of the myringotomy depends on the procedure and the tools available. We hypothesize that an opening as small as the one necessary for a grommet has a very high chance of spontaneous healing. We have already designed and developed an assistant robot capable of performing precise and stable movements inside the middle ear through the inner canal (12). This robot has a remote center of motion (RCM) corresponding to the tool tip enabling the robot to move the tool while the tip is stable. However, the robot can be programmed to move the RCM to the level of the tympanic membrane. In this condition, the tool tip can move inside the middle ear cleft while the tool does not move at the level of the tympanic membrane avoiding rupture of the tympanic membrane during the procedure. Consequently, the combination of AR to robot-based surgery will potentially widen the field of applications.

In conclusion, we have shown that AR can be applied to otoendoscopy with high precision even with rapid endoscope movements. This technique could open important insights in transcanal minimally invasive robot-based procedures.

Acknowledgments: Authors acknowledge the financial support of Oticon Medical, Société ORL de Bourgogne, CNRS, and Collin Medical SA.

REFERENCES

1. Daponte P, De Vito L, Picariello F, et al. State of the art and future developments of the augmented reality for measurement applications. *Measurement* 2014;57:53–70.
2. Van Krevelen DW, Poelman R. A survey of augmented reality technologies, applications and limitations. *Int J Virtual Real* 2010;9:1.
3. Plontke SR, Plinkert PK, Plinkert B, et al. Transtympanic endoscopy for drug delivery to the inner ear using a new microendoscope. *Adv Otorhinolaryngol* 2002;59:149–55.
4. Selmani Z, Pyykkö I, Ishizaki H, et al. Role of transtympanic endoscopy of the middle ear in the diagnosis of perilymphatic fistula in patients with sensorineural hearing loss or vertigo. *ORL J Otorhinolaryngol* 2002;64:301–6.
5. Kakehata S, Futai K, Sasaki A, et al. Endoscopic transtympanic tympanoplasty in the treatment of conductive hearing loss: early results. *Otol Neurotol* 2006;27:14–9.
6. Kakehata S. Transtympanic endoscopy for diagnosis of middle ear pathology. *Otolaryngol Clin North Am* 2013;46:227–32.
7. Mood ZA, Daniel SJ. Use of a microendoscope for transtympanic drug delivery to the round window membrane in chinchillas. *Otol Neurotol* 2012;33:1292–6.
8. Dean M, Chao WC, Poe D. Eustachian tube dilation via a transtympanic approach in 6 cadaver heads. A feasibility study. *Otolaryngol Head Neck Surg* 2016;155:654–6.
9. Alzamil KS, Linthicum FH Jr. Extraneous round window membranes and plugs: possible effect on intratympanic therapy. *Ann Otol Rhinol Laryngol* 2000;109:30–2.
10. Bozzato A, Bozzato V, Al Kadah B, et al. A novel multipurpose modular mini-endoscope for otology. *Eur Arch of Otorhinolaryngol* 2014;271:3341–8.
11. Comparetti MD, Vaccarella A, Dyagilev I, Shoham M, Ferrigno G, De Momi E. Accurate multi-robot targeting for keyhole neurosurgery based on external sensor monitoring. *Proc Inst Mech Eng H* 2012;226:347–59.
12. Nguyen Y, Miroir M, Kazmitcheff G, Ferrary E, Sterkers O, Grayeli AB. From conception to application of a tele-operated assistance robot for middle ear surgery. *Surg Innov* 2012;19:241–51.
13. Li L, Yang J, Chu Y, et al. A novel augmented reality navigation system for endoscopic sinus and skull base surgery: a feasibility study. *PLoS One* 2016;11:e0146996.
14. Nakamoto M, Ukimura O, Faber K, et al. Current progress on augmented reality visualization in endoscopic surgery. *Curr Opin Urol* 2012;22:121–6.
15. Zhu M, Liu F, Chai G, et al. A novel augmented reality system for displaying inferior alveolar nerve bundles in maxillofacial surgery. *Sci Rep* 2017;7:42365.
16. Helferty JP, Higgins WE. Combined endoscopic video tracking and virtual 3D CT registration for surgical guidance. *Proc Int Conf Image Proc* 2002;2:II-II.
17. Merritt SA, Khare R, Bascom R, et al. Interactive CT-video registration for the continuous guidance of bronchoscopy. *IEEE Trans Med Imaging* 2013;32:1376–96.
18. Rai L, Merritt SA, Higgins WE. Real-time image-based guidance method for lung-cancer assessment. *Conf Comput Vis Pattern Recognit Workshops* 2006;2:2437–44.
19. Liu J, Subramanian K, Yoo T, et al. A stable optic-flow based method for tracking colonoscopy images. *Conf Comput Vis Pattern Recognit Workshops* 2008;1–8.
20. Liu J, Subramanian K, Yoo T. Region flow: a multi-stage method for colonoscopy tracking. *Med Image Comput Comput Assist Interv* 2010;13:505–13.
21. Liu J, Subramanian K, Yoo T. An optical flow approach to tracking colonoscopy video. *Comput Med Imaging Graph* 2013;37:207–23.
22. Burschka D, Li M, Ishii M, et al. Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. *Med Image Anal* 2005;9:413–26.
23. Mirota D, Taylor RH, Ishii M, et al. Direct endoscopic video registration for sinus surgery. *Proc SPIE Med Imaging* 2009; 7261:91–9.

24. Kohan D, Jethanamest D. Image-guided surgical navigation in otology. *Laryngoscope* 2012;122:2291–9.
25. Hartley R, Zisserman A. *Multiple View Geometry in Computer Vision*. Cambridge, MA: Cambridge University Press; 2003.
26. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. *J Siam Soc* 1963;11:431–41.
27. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. *IJCAI* 1981;81:674–9.
28. Rosten E, Drummond T. Machine learning for high-speed corner detection. *Comput Vis ECCV* 2006;3951:430–43.
29. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 2004;60:91–110.
30. Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. *Comput Vis ECCV* 2006;3951:404–17.
31. Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF. *Int J Comput Vis* 2011; 2564–71.
32. Fitzgibbon AW, Fisher RB. A buyer's guide to conic fitting. *Proc British Machine Vision Conf* 1995;513–22.
33. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 1981;24:381–95.
34. Selesnick SH, Kacker A. Image-guided surgical navigation in otology and neurotology. *Otol Neurotol* 1999;20:688–97.
35. Gunkel AR, Vogele M, Martin A, et al. Computer-aided surgery in the petrous bone. *Laryngoscope* 1999;109:1793–9.
36. Wenger T, Nowatschin S, Wittmann W, et al. Design and accuracy evaluation of a new navigated drill system for computer assisted ENT-surgery. *Conf Proc IEEE Eng Med Biol Soc* 2011;1233–6.
37. Carney AS, Patel N, Baldwin DL, et al. Intra-operative image guidance in otolaryngology–the use of the ISG viewing wand. *J Laryngol Otol* 1996;110:322–7.
38. Inoue D, Cho B, Mori M, et al. Preliminary study on the clinical application of augmented reality neuronavigation. *J Neurol Surg A Cent Eur Neurosurg* 2013;74:71–6.
39. Liu WP, Richmon JD, Sorger JM, et al. Augmented reality and cone beam CT guidance for transoral robotic surgery. *J Robot Surg* 2015;9:223–33.