Markerless Image-to-Face Registration for Untethered Augmented Reality in Head and Neck Surgery

Christina Gsaxner[1,2], Antonio Pepe[1], Jürgen Wallner[2], Dieter Schmalstieg[1], and Jan Egger[1,2]

[1] Institute of Computer Graphics and Vision, Graz University of Technology, Austria
gsaxner@tugraz.at
[2] Department for Oral and Maxillofacial Surgery, Medical University of Graz, Austria

Abstract. In the treatment of head and neck cancer, physicians can benefit from augmented reality in preparing and executing treatment. We present a system allowing a physician wearing an untethered augmented reality headset to see medical visualizations precisely overlaid onto the patient. Our main contribution is a strategy for markerless registration of 3D imaging to the patient's face. We use a neural network to detect the face using the headset's depth sensor and register it to computed tomography data. The face registration is seamlessly combined with the headset's continuous self-localization. We report on registration error and compare our approach to an external, high-precision infrared tracking system.

Keywords: Augmented reality · 3D registration · head and neck cancer

1 Introduction

Medical applications can benefit from augmented reality (AR) interfaces , e.g., by providing a more intuitive mental mapping from 3D imaging data to the patient [1]. In particular, immersive AR systems combine natural 3D interaction with an increased spatial perception of 3D structures [2]. In this contribution, we present a method for immersive medical visualization in the head and neck area using a commercial AR headset with optical see-through display, the Microsoft HoloLens (Microsoft Corporation, Redmond, WA, USA). Our system works in an unprepared environment and achieves registration based on facial surface matching at real-time frame rates by registering 3D imaging data directly to the patient's face as observed by the headset's depth sensor. Moreover, we take advantage of the highly accurate built-in self-localization capabilities of the headset. The combination of facial detection and self-localization enables fully untethered, real-time markerless registration.

Related work Usually, image data is acquired offline, e.g., through magnetic resonance (MRI) or computed tomography (CT) imaging, and must be registered

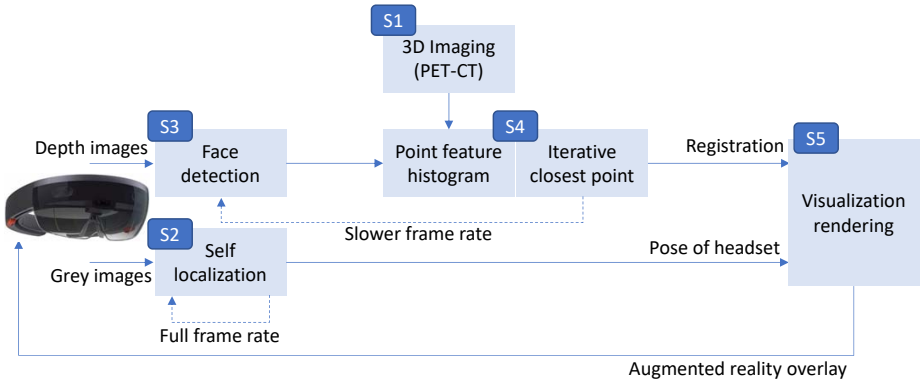

Fig. 1. Overview of the proposed five-step registration pipeline.

to the patient in the physician's view with high accuracy, thus establishing a relationship between physical and virtual space. Several studies for medical AR do not estimate this relationship at all, instead relying on the manual placement of medical content with respect to the virtual world [3, 4]. Others establish this correspondence using outside-in tracking methods based on markers rigidly attached to the patient [5], or external devices, e.g., optical or depth sensors [6, 7]. However, such outside-in approaches require complicated preparation and must be calibrated *in situ*, which disrupts clinical workflow and therefore hinders acceptance. As an alternative, inside-out methods utilizing self-localization techniques have recently been explored in the context of medical AR using intraoperative X-ray images or manually selected landmarks as a registration strategy [8, 9].

Contribution For applications involving the head and face, such as in surgery planning of head and neck cancer, the opportunity arises to use facial features directly for both registration and tracking. Thus, in our contribution, we present a strategy for markerless, inside-out image-to-face registration, which, in combination with the self-localization of the headset, enables untethered real-time AR to aid physicians in the treatment and management of head and neck cancer.

2 Methods

Our goal was to build a system for 3D image registration using only the headset hardware. Our solution combines two sensor pipelines, the headset's self-localization, and facial localization using the headset's depth sensor. The self-localization runs an algorithm for simultaneous localization and mapping (SLAM) on a dedicated hardware accelerator, fed by multiple cameras on the headset, and delivers highly robust and accurate camera poses [10]. The depth sensor provides the ability to detect the patient's face as represented in the 3D image

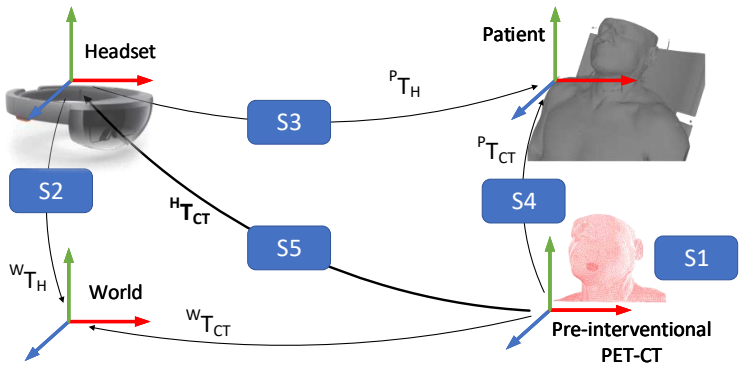

Fig. 2. Coordinate systems and their transformations to be computed during image-to-patient registration. The goal is to find ${}^{H}T_{CT}$, the relative pose of the PET-CT coordinate system CT with respect to the physician, denoted by H.

data. The depth sensor faces forward and is pre-registered with the user's field of view, conveniently allowing direct superposition of computer-generated visuals. We build a pipeline that performs competitive sensor fusion [11] between the self-localization component and a custom pipeline for face registration in five steps, labeled S1-S5, as shown in Figure 1:

- S1 Obtain and preprocess medical image data
- S2 Obtain an update from the self-localization
- S3 Apply a neural network for face detection on incoming depth images, followed by extraction of a point cloud
- S4 Coarsely register the 3D image data to the depth image using point feature histograms; refine the registration using an iterative closest point method
- S5 Render overlay using the registration obtained by combining S2 with S3/S4

Out of these five steps, only S1 is performed offline. All other steps are run online, but S3/S4 can be run at a lower than full frame without affecting overall system performance. To correctly overlay virtual content with the physical world, we need to estimate ${}^{H}T_{CT}(t)$, the rigid 3D transformation which correctly positions content in the coordinates of pre-interventional CT acquisitions CT with respect to the physician wearing the headset H at time t. Consequently, a series of transformations (Figure 2) has to be estimated as follows:

$$ {}^{H}T_{CT}(t) = {}^{W}T_{H}^{-1}(t) \cdot {}^{W}T_{H}(t_0) \cdot {}^{P}T_{H}^{-1}(t_0) \cdot {}^{P}T_{CT}(t_0) \cdot \mathbf{P}_{CT}, \tag{1} $$

with W and P representing the world and patient coordinate system, respectively, and $\mathbf{P}_{CT}$ denoting the point cloud representation of the patients' skin surface recovered from CT, obtained in S1. We describe each of the steps S1-S5 in detail in the following sections.

2.1 Medical imaging data processing

In the pre-interventional, offline step S1, we acquire ^{18}F-fluorodeoxy-D-glucose positron emission tomography-computed tomography (^{18}F-FDG PET-CT) data, which is essential in the diagnosis and evaluation of head and neck carcinomae due to its ability to combine functional information from PET with anatomical information from CT [12]. Volumetric CT image data of the patient is segmented into skin surface and anatomically relevant structures for visualization. Polygonal meshes are extracted using Marching Cubes algorithm; then, a point cloud representation of the skin surface $\mathbf{P}_{CT}$ is created for usage in consecutive registration steps. Similarily, tumor sufaces are extracted from co-registered ^{18}F-FDG PET acquisitions, which exhibit high contrast for metabolically active tumors.

2.2 Self-localization

Step S2 obtains ${}^{W}T_{H}$, the poses of the surgeon's viewpoint with respect to world coordinates, using the headset's SLAM-based self-localization system [13]. We use the camera poses delivered by the headset and associate them with the face model created in S3 using the registration procedure of S4. We do not use the geometric model of the SLAM system, since it is too coarse for our purposes.

2.3 Face detection and extraction

Step S3 denotes the acquisition of a point cloud representation of the patient's face from the depth sensor. A region of interest (ROI) around the patient's head is found automatically and in real-time by using a neural network. It relies on a single-shot-multibox detector (SSD) [14] using a ResNet-10 architecture, trained for face detection. SSD performs object localization in a single forward pass by regressing a bounding box around objects. If detection is successful, the ROI is mapped to the depth image to create a point cloud using an inverse perspective transformation $({}^{P}T_{H})^{-1}$. Given a position in the depth frame $\mathbf{m} = [u, v]$ in pixel units and the depth camera's intrinsic matrix K, the corresponding scene point in camera coordinates $\mathbf{p} = [x, y, z]^T$ can be calculated by

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} d(u,v) = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} d(u,v), \tag{2}$$

where $d(u, v)$ denotes the depth at $[u, v]$. This inverse projection is applied to all pixels within the ROI around the patient's face, resulting in a point cloud $\mathbf{P}_P = \{\mathbf{p}_2, \mathbf{p}_2, ..., \mathbf{p}_N\}$ which represents the face in headset coordinates.

2.4 Aligning pre-interventional data with the patient

For step S4, we take advantage of the distinctive nature of a human's facial features to enable a markerless, automatic, two-stage registration scheme inspired

by the method proposed by Holz et al. [15]. ${}^{P}T_{CT}$ is the transformation that aligns the point cloud from pre-interventional imaging $\mathbf{P}_{CT}$ with the target point cloud representing the patient $\mathbf{P}_P$. To compute an initial alignment, we adopt registration based on fast point feature histograms (FPFH) [16]. FPFH features are computed in both point clouds and reciprocally matched using 1-nearest-neighbor search, resulting in $\kappa_f = \{(\mathbf{f}_P, \mathbf{f}_{CT})\}$, a set of correspondence points found by matching FPFH features of $\mathbf{P}_P$ and $\mathbf{P}_{CT}$. The fast global registration algorithm by Zhou et al. [17] is applied to compute an initial transformation ${}^{P}\hat{T}_{CT}$ such that distances between corresponding points are minimized:

$$E({}^{P}\hat{T}_{CT}) = \sum_{(\mathbf{f}_P, \mathbf{f}_{CT}) \in \kappa_f} \rho(||\mathbf{f}_P - {}^{P}\hat{T}_{CT}\mathbf{f}_{CT}||), \tag{3}$$

where $\rho(.)$ is a scaled German-McClure estimator, a robust penalty for optimization. The initial transformation ${}^{P}\hat{T}_{CT}$ is then refined using point-to-plane ICP [18], resulting in the final alignment ${}^{P}T_{CT}$. We define the correspondence set as the actual 1-nearest-neighboring points $\kappa = \{(\mathbf{p}_P, \mathbf{p}_{CT})\}$ and optimize

$$E({}^{P}T_{CT}) = \sum_{(\mathbf{p}_P, \mathbf{p}_{CT}) \in \kappa} ((\mathbf{p}_P - {}^{P}T_{CT}\mathbf{p}_{CT}) \cdot \mathbf{n}_{pp})^2, \tag{4}$$

where $\mathbf{n}_{pp}$ is the normal of point $\mathbf{p}_P$. This combination of rapid feature matching and ICP refinement allows a robust, accurate and fast computation of ${}^{P}T_{CT}$.

2.5 Visualization using augmented reality

In step S5, we use the transformation obtained in previous steps to render virtual content in a way that it is anchored to world coordinates using ${}^{W}T_{CT}$. The SLAM system is almost entirely free of drift; so, as long as the patient remains stationary, ${}^{W}T_{CT}$ can be computed at a much lower framerate than ${}^{W}T_{H}$. This makes the system comfortable to use, allowing the surgeon to look away from the patent's face and preserve registration of virtual objects when returning to the patient, even without re-detection of the face, simply by receiving an update of ${}^{W}T_{H}(t)$. If the patient moves, re-detection of the face leads to instant re-registration of the overlaid virtual objects, by simply updating ${}^{P}T_{CT}$. Figure 3 shows an example of bones and tumoral masses registered with the patient.

3 Experiments and results

We evaluted end-to-end registration using phantom heads by 3D-printing CT scans from step S1. We chose eight subjects with cancerous tumors in the head/neck area, for which PET-CT imaging was available. Furthermore, we show our application's feasability with a human subject. To avoid unjustifiable radiation exposure, a MRI scan was used to extract the isosurface of the skin.

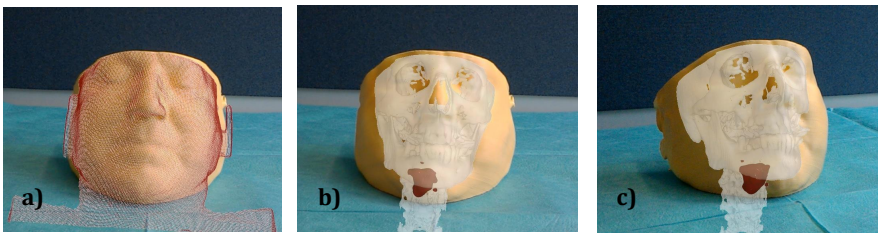

Fig. 3. Example AR visualization on a patient phantom. By registering a point cloud to the patient's face as in a), bones and a tumoral mass can be overlaid as shown in b). Registration persists if the physician changes his viewpoint, seen in c).

3.1 Target registration error

The target registration error (TRE) is computed to evaluate the end-to-end registration accuracy of the proposed system. It is defined by

$$TRE = \frac{1}{N_k} \sum_{k=1}^{N_k} ||{}^{H}T_{CT}\mathbf{m}_k^{CT} - \mathbf{m}_k^{H}||, \tag{5}$$

where N_k is a number of reference points; $\mathbf{m}^{CT}$ and $\mathbf{m}^{H}$ are the reference points in the CT data and the headset's view, respectively. Since data obtained from the clinical routine is used in this study, there are no fiducial markers in pre-interventional data, which could be used as reference points for TRE computation. Therefore, $N_k = 5$ landmarks, namely, the left/right inner canthus, the tip of the nose and the left/right labial commissure of mouth, were labeled manually in CT data and later selected in the operator's view to obtain $\mathbf{m}^{H}$ and $\mathbf{m}^{CT}$.

We repeated TRE measurement 10 times for each patient phantom as well as the human subject, at distances ranging from 50 cm to 90 cm between operator and patient, using slightly changing viewing angles. Table 1 summarizes the TRE for phantoms 1-8 as well as the human subject, averaged over all measurements.

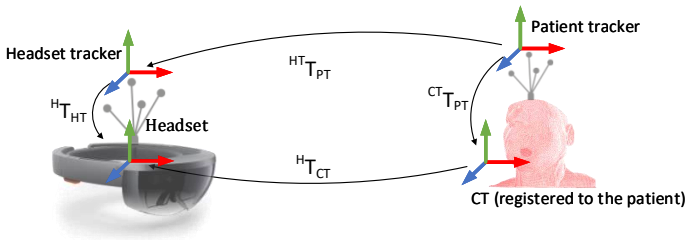

Fig. 4. Coordinate systems for comparing ${}^{H}T_{CT}$ to ${}^{HT}T_{PT}$ from an external infrared tracking system. ${}^{CT}T_{PT}$ and ${}^{H}T_{HT}$ are estimated using hand-eye calibration.

Table 1. Target registration error (TRE) of five reference points, as well as error in translation E_t and rotation E_r between the transformation ${}^{S}T_{CT}$ and the equivalent transformation ${}^{S}T_{CT}^{OT}$ obtained from an external infrared tracking system.

Subject		1	2	3	4	5	6	7	8	Human	Total
TRE (mm)	Mean	9.4	9.4	10.2	6.7	10.5	11.4	7.1	10.2	8.1	9.2
	Sd	2.3	0.8	1.7	1.3	2.2	2.2	0.7	1.4	1.0	1.5
E_t (mm)	Mean	5.4	2.0	3.1	6.7	3.1	2.3	7.0	3.0	4.3	3.9
	Sd	2.1	1.9	3.0	1.8	1.7	1.7	2.3	2.0	1.7	1.8
E_r (°)	Mean	4.9	5.7	5.4	11.0	2.9	2.4	4.4	5.6	9.8	4.9
	Sd	3.7	2.4	3.5	2.5	1.8	1.9	1.9	3.2	1.8	2.4

3.2 Comparison with a high-precision external tracking device

We compare ${}^{H}T_{CT}$ derived from our application with the transformation computed by an external infrared tracking system, consisting of 15 OptiTrack Flex 13 cameras (NaturalPoint, Inc., Corvallis, OR, USA). We rigidly attached a set of non-collinear retro-reflective markers to the headset and our patient phantoms or human subject for the computation of ${}^{HT}T_{PT}$, the relative pose of the patient tracker with respect to the HoloLens tracker from the OptiTrack. To correlate these transformations, ${}^{H}T_{HT}$, which calibrates the headset tracker and the virtual camera of the HoloLens, as well as ${}^{CT}T_{PT}$, the transformation from the patient tracker to the CT coordinate system (already registered to the patient), needed to be estimated by hand-eye calibration [19], as shown in Figure 4. Thus, we can compute ${}^{H}T_{CT}^{OT}$, the reference transformation obtained by the OptiTrack system, as ${}^{H}T_{CT}^{OT} = {}^{H}T_{HT} \cdot {}^{HT}T_{PT} \cdot ({}^{CT}T_{PT})^{-1}$. To quantify the error between transformations, we evaluate the error in distance as well as the angular error separately. The results obtained from all patient phantoms, as well as the results from our experiments with a human subject, are summarized in Table 1.

4 Discussion and conclusion

We presented a novel end-to-end solution to the image-to-patient registration problem in AR using optical see-trough headsets. Our markerless registration scheme works fully automatically in an unprepared environment, by exploiting the distinct characteristics of human faces. It computates the transformation aligning pre-interventional 3D data with the patient in the surgeon's view. We evaluated accuracy with patient phantoms and a human test person, reporting a mean TRE of 9.2 ± 1.5 mm and an average error in comparison to a high-precision optical tracking system of 3.9 ± 1.8 mm in translation and 4.9 ± 2.4 ° in rotation. The accuracy of ${}^{H}T_{CT}$ is subject to several error sources, partly due to hardware restrictions: A residual error remains due to the rather low quality of point cloud representation acquired from the headset's depth sensor. Moreover, inaccuracies and latency of the HoloLens self-localization may affect the overall

precision. Finally, optical see-through display calibration was not considered, as we expect that future hardware will support auto-calibration using eye tracking.While our system does not yet achieve the sub-millimeter precision required for image-guided intervention, it represents a promising all-in-one tool for immersive treatment and intervention planning in the management of head and neck cancer. As others before us [2, 8], we believe that the Microsoft HoloLens has great potential for clinical and educational applications in medicine, especially since the imminent release of the HoloLens 2, which has much improved hardware and software capabilities. As a next step, we plan a clinical evaluation of our system involving a patient study, for which ethics approval has recently been obtained. This study should demonstrate the benefits of AR to physicians in the treatment of head and neck cancer. Other future work includes a more refined visualization and 3D interaction to provide guidance to surgeons during intervention planning.

References

1. Sielhorst, T., Feuerstein, M., Navab, N.: Advanced Medical Displays: A Literature Review of Augmented Reality. J. Disp. Technol. **4**(4), 451–467 (2008)
2. de Oliveira, M.E., Debarba, H.G., Lädermann, A., Chagué, S., Charbonnier, C.: A Hand-Eye Calibration Method for Augmented Reality Applied to Computer-Assisted Orthopedic Surgery. Int. J. Med. Robot **15**(2), e1969 (2019)
3. Pratt, P., et al.: Through the HoloLens Looking Glass: Augmented Reality for Extremity Reconstruction Surgery Using 3D Vascular Models with Perforating Vessels. European Radiology Experimental **2**(1), 2 (2018)
4. Incekara, F., Smits, M., Dirven, C., Vincent, A.: Clinical Feasibility of a Wearable Mixed-Reality Device in Neurosurgery. World Neurosurgery **118**, e422–e427 (2018)
5. Ahn, J., Choi, H., Hong, J., Hong, J.: Tracking Accuracy of a Stereo-camera-based Augmented Reality Navigation System for Orthognathic Surgery. J. Oral Maxillofac. Surg. (2019)
6. Chen, X., et al.: Development of a Surgical Navigation System Based on Augmented Reality Using an Optical See-through Head-mounted Display. J. Biomed. Inform. **55**, 124–131 (2015)
7. Hung Hsieh, C., Der Lee, J., Tsai Wu, C.: A Kinect-based Medical Augmented Reality System for Craniofacial Applications Using Image-to-Patient Registration. Neuropsychiatry **07**(06), 927–939 (2017)
8. Hajek, J., et al.: Closing the Calibration Loop: An Inside-Out-Tracking Paradigm for Augmented Reality in Orthopedic Surgery. In: MICCAI. pp. 299–306 (2018)
9. Mahmoud, N., et al.: On-patient See-Through Augmented Reality Based on Visual SLAM. Int. J. Comput. Assist. Radiol. Surg. **12**(1), 1–11 (2017)
10. Klein, G.: Registration on HoloLens. Keynote talk at ISMAR (2017)
11. Durrant-Whyte, H.F.: Sensor models and multisensor integration. Int. J. of Robot. Res. **7**(6), 97–113 (1988)
12. Castaldi, P., Leccisotti, L., Bussu, F., Micciché, F., Rufini, V.: Role of (18)F-FDG PET-CT in Head and Neck Squamous Cell Carcinoma. Acta Otorhinolaryngologica Italica **33**(1), 1–8 (2013)
13. Klein, G., Murray, D.: Parallel Tracking and Mapping for Small AR Workspaces. In: ISMAR. pp. 1–10 (2007)

14. Liu, W., et al.: SSD: Single Shot MultiBox Detector. In: ECCV. pp. 21–37 (2016)
15. Holz, D., Ichim, A.E., Tombari, F., Rusu, R.B., Behnke, S.: Registration with the Point Cloud Library: A Modular Framework for Aligning in 3-D. IEEE Robot. & Autom. Mag. **22**(4), 110–124 (2015)
16. Rusu, R.B., Blodow, N., Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D Registration. In: ICRA. pp. 3212–3217 (2009)
17. Zhou, Q.Y., Park, J., Koltun, V.: Fast Global Registration. In: ECCV (2016)
18. Chen, Y., Medioni, G.: Object Modelling by Registration of Multiple Range Images. Image Vis. Comp. **10**(3), 145–155 (1992)
19. Tsai, R.Y., Lenz, R.K.: A New Technique for Fully Autonomous and Efficient 3D Robotics Hand/Eye Calibration. IEEE Trans. Robot. Autom. **5**(3), 345–358 (1989)