

3D Surgical Overlay with Markerless Image Registration Using a Single Camera

Junchen Wang¹(✉), Hideyuki Suenaga², Liangjing Yang¹, Hongen Liao³, Takehiro Ando¹, Etsuko Kobayashi¹, and Ichiro Sakuma¹

¹ The University of Tokyo, Tokyo, Japan
wangjunchen@gmail.com

² The University of Tokyo Hospital, Tokyo, Japan

³ Tsinghua University, Beijing, China

Abstract. Minimum invasive surgery can benefit from surgical visualization, which is achieved by either virtual reality or augmented reality. We previously proposed an integrated 3D image overlay based surgical visualization solution including 3D image rendering, distortion correction, and spatial projection. For correct spatial projection of the 3D image, image registration is necessary. In this paper we present a 3D image overlay based augmented reality surgical navigation system with markerless image registration using a single camera. The innovation compared with our previous work lies in the single camera based image registration method for 3D image overlay. The 3D mesh model of patient's teeth which is created from the preoperative CT data is matched with the intraoperative image captured by a single optical camera to determine the six-degree-of-freedom pose of the model with respect to the camera. The obtained pose is used to superimpose the 3D image of critical hidden tissues on patient's body directly via a translucent mirror for surgical visualization. The image registration performs automatically within approximate 0.2s, which enables real-time update to tackle patient's movement. Experimental results show that the registration accuracy is about 1 mm and confirm the feasibility of the 3D surgical overlay system.

1 Introduction

Three dimensional (3D) image overlay based augmented reality (AR) surgical navigation has been introduced many times in the literature [3–5, 8]. The 3D image with both horizontal and vertical parallax is displayed by a lens array monitor (3D display) which can be observed without wearing glasses. In our recent work [12], we proposed an integrated 3D image overlay based surgical visualization solution including 3D image rendering, distortion correction, and spatial projection. Given an anatomical mesh model derived from CT data, the 3D image of the model, which has the same geometric dimensions as the original organ, can be rendered in real time with the help of a graphics processing unit (GPU). To superimpose the 3D image on patient's body with correct overlay, the pose of the anatomical model (i.e., the digitalized patient organ in the image

space) with respect to the 3D display has to be determined intraoperatively. 3D measurement systems (e.g., Polaris Tracking System) are usually employed for this purpose. The pose determination is composed of two steps. One is to calculate the transformation from the 3D display to the 3D measurement system, which is called display-camera calibration. The other is to calculate the transformation from the 3D measurement system to the patient (i.e., the pose of the anatomical model with respect to the 3D measurement system), which is called image registration. The device calibration is performed offline only once because the 3D display and the 3D measurement system are relatively fixed during surgery. However the image registration suffers from patient's movement, which requires the image registration process to be performed automatically in real time.

In the previous work [3–5, 8], an optical tracking system was employed for image registration using a manual marker-based registration method. The involvement of markers will hamper common surgical workflow, and the attachment of markers is either invasive or infeasible in many cases. Automatic markerless image registration is preferable in surgical navigation. In our previous work [10, 13], a low-priced stereo camera was employed replacing the Polaris tracking system for the image registration task in oral and maxillofacial surgery. A teeth contour tracking method was proposed to calculate the pose of patient's teeth with respect to the stereo camera without manual intervention. However, this method is a 3D-3D matching method requiring that the organ to be registered have sharp 3D contour features; and these features should be easily reconstructed by the stereo camera three-dimensionally. Such conditions are quite strict, hence limit the applicable scope of the 3D contour tracking method. Furthermore, in the previous method, only silhouette features in the captured image pair are used while other useful visual clues such as image gradients on non-silhouette edges are ignored. The incorporation of these information could improve the accuracy and robustness of image registration.

In this study, we further simplify the hardware for image registration by using a single camera. We present a 3D surgical overlay system with automatic markerless image registration. The image registration is achieved by matching the 2D shape of teeth's mesh model with the intraoperative 2D image captured by the camera. The display-camera calibration is performed by solving a perspective-n-point (PnP) problem (i.e., the estimation of camera's extrinsic parameters given its intrinsic parameters and n-point 3D-2D correspondences). The proposed system enables real-time correct 3D image overlay on patient's head and neck area for surgical visualization in oral and maxillofacial surgery.

2 System Overview

The proposed system consists of a 3D display, a translucent mirror (AR window), a monochrome camera and a workstation for information processing, as shown in Fig. 1(a). The 3D display is composed of a high pixel per inch (ppi) liquid crystal display (LCD) and a hexagonal lens array which is placed in front of the LCD.

The 3D image of a mesh model (e.g., in the form of an STL file) is created using computer generated integral imaging (CGII) techniques and can be projected to a specified location and orientation with respect to the 3D display [12]. The 3D image is further overlaid on the patient by the translucent mirror. Surgeons will see a superimposed 3D image through the AR window to acquire visualized anatomical information. In this study, the goal of the system is to realize intraoperative augmented reality visualization in oral and maxillofacial surgery.

Figure 1(b) shows the involved coordinate systems in the overlay system. We denote by T_C , T_D , T_M the camera, 3D display, and model coordinate systems, respectively. T_D is the world coordinate system in which the 3D image is rendered as described in our CGII rendering algorithm [12]. T_M actually represents the image space where the patient is digitalized (e.g., by CT scanning). To overlay the 3D image correctly on the patient, the transformation from T_D to T_M denoted by \mathbf{T}_D^M should be determined. Because we have $\mathbf{T}_D^M = \mathbf{T}_D^C \mathbf{T}_C^M$, this raises two problems: display-camera calibration and image registration.

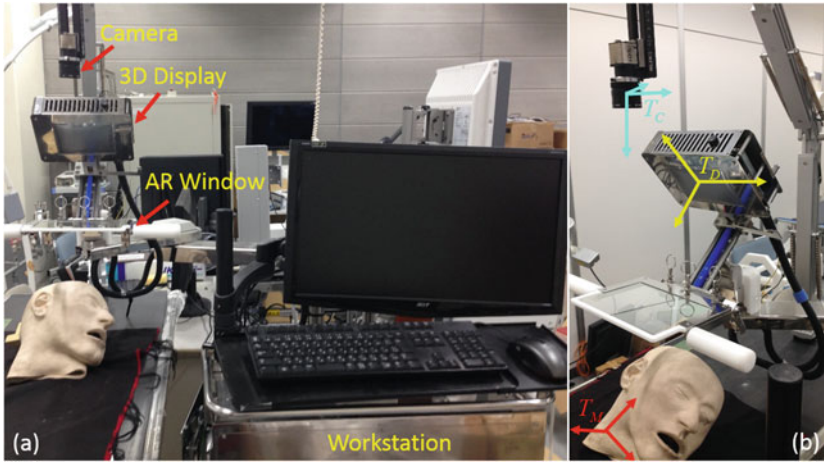


Fig. 1. (a) System overview. (b) Coordinate systems.

3 Display-Camera Calibration

Display-camera calibration is to determine the transformation \mathbf{T}_D^C , which can be formulated as a camera extrinsic calibration problem. The 3D image of a known geometry in T_D is projected by the 3D display, and the 2D image of the 3D image is captured by the camera through the AR window. In the captured image, the 2D-3D correspondences are established. The pose $(\mathbf{R}_C^D, \mathbf{t}_C^D)$ of the 3D display¹ with respect to the camera can be estimated by solving

$$\min_{\langle \mathbf{R}_C^D, \mathbf{t}_C^D \rangle} \sum_{i=1}^N \|\mathbf{K} \left(\mathbf{R}_C^D, \mathbf{t}_C^D \right) \mathbf{X}_i - \mathbf{x}_i\|^2 \quad (1)$$

¹ $\det(\mathbf{R}_C^D) = -1$ due to the reflection of the AR window.

where $\mathbf{X}_i \leftrightarrow \mathbf{x}_i, i = 1 \cdots N$ are 3D-2D correspondences in the form of homogeneous coordinates; \mathbf{K} is the intrinsic matrix of the camera; $\|\mathbf{X}_i - \mathbf{x}_i\|$ denotes the underlying 2D distance between \mathbf{X}_i and \mathbf{x}_i .

Figure 2(a) shows the calibration model which is a 5×5 planar ball array. The captured 2D image through the AR window is shown in Fig. 2(b). The centers of the projected balls are automatically detected using a simple threshold followed by an ellipse fitting. Figure 2(c) shows an example of ball center detection. Equation (1) is well known as a PnP problem which can be easily solved using a nonlinear least squares technique [1]. \mathbf{T}_D^C is the inverse of the pose $(\mathbf{R}_C^D, \mathbf{t}_C^D)$.

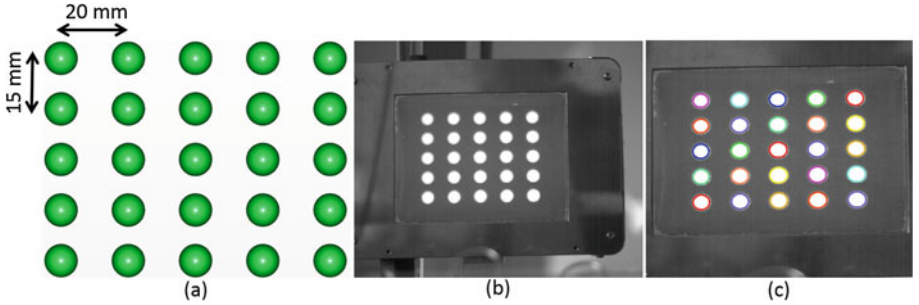


Fig. 2. (a) Calibration model. (b) Captured image of (a). (c) Automatic ball center detection.

4 Image Registration

Image registration is to determine the transformation \mathbf{T}_C^M . Unlike the fixed spatial relationship during surgery in the device calibration, \mathbf{T}_C^M suffers from patient movement and varies when patient’s pose is changed. This requires the image registration to be performed automatically in real time. We propose a 3D-2D registration method by matching patient’s 3D teeth model (created from pre-operative CT data) with the intraoperative 2D camera image based on Ulrich’s method [9]. Because teeth are rigid and can be easily exposed to an optical camera, they can serve as “natural markers” for registering anatomical information in the head and neck area.

4.1 Problem Formulation

Our problem is to find the best pose (\mathbf{R}, \mathbf{t}) so that the 2D projection of the 3D model using the projection matrix $\mathbf{K}(\mathbf{R}, \mathbf{t})$ is most consistent with the 2D image. 2D projection views of the 3D model can be rendered using computer graphics API, such as OpenGL. To measure the consistency between the projected model shape E_{2D} and the image $I(x, y)$, we use the following similarity proposed by Steger [7]

$$s(E_{2D}, I) = \frac{1}{N} \sum_{i=1}^N \frac{|\langle \nabla I(x_i, y_i), \mathbf{d}_i \rangle|}{\|\nabla I(x_i, y_i)\| \cdot \|\mathbf{d}_i\|} \quad (2)$$

where $E_{2D} \stackrel{\text{def}}{=} \{x_i, y_i, \mathbf{d}_i\}_{i=1}^N$ is a set of edge points (x_i, y_i) with associated direction vectors \mathbf{d}_i representing the normal direction; $\nabla I(x_i, y_i)$ is the image gradient at (x_i, y_i) ; $\langle \cdot, \cdot \rangle$ denotes dot product. The absolute value $|\cdot|$ on the numerator will ignore local contrast polarity change. $s(E_{2D}, I)$ ranges from $[0, 1]$.

Given (\mathbf{R}, \mathbf{t}) , the projected 2D shape E_{2D} of the 3D model is extracted as follows. First, the intrinsic matrix \mathbf{K} is used to set the view frustum of the virtual camera so that the virtual camera has the same projection geometry as the real camera. Then, the 3D model is rendered into a 3-channel RGB image whose RGB values represent the normal vector on the corresponding surface of the 3D model. Next, the image tensor (2×2 matrix) at each pixel of the RGB image is calculated, whose largest eigen value represents the edge strength of the pixel [6]. The edge strength corresponds to the face angle of the corresponding 3D edge of the model. Subsequently, a threshold is applied to the edge strength to suppress the pixel whose corresponding face angle is below a certain value (e.g., 30°). Finally, non-maximum suppression is performed for edge thinning and the remaining edge pixels with their gradient vectors constitute the 2D shape E_{2D} . Figure 3 shows the extracted 2D shape of a left molar model with the suppression angle of 35° . A straightforward idea is to find optimal (\mathbf{R}, \mathbf{t}) so that (2) is maximized.

4.2 Aspect Graph Based Matching

It is impossible to directly optimize (2) unless the start point is quite near to the true pose. However, we do not have the prior knowledge about the pose of the model. We instead adopt a view-based approach. An aspect graph-based matching method proposed by Ulrich [9] is used for fast 3D-2D pose estimation.

Offline Aspect Graph Building. Views are generated by specifying view-points (virtual camera positions) in a spherical coordinate system (SCS) whose origin is set to be the center of 3D model's bounding box. The viewpoint range is specified by $[r_{\min}, r_{\max}]$, $[\varphi_{\min}, \varphi_{\max}]$, and $[\theta_{\min}, \theta_{\max}]$, which is a spherical

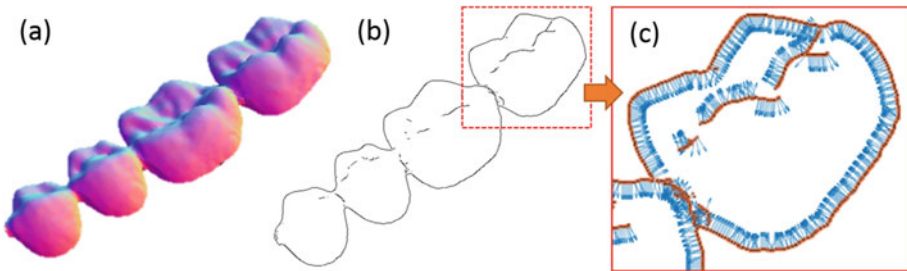


Fig. 3. (a) Rendered RGB image of a molar model. (b) Projected 2D shape of (a). (c) Associated direction vectors.

quadrilateral. r , φ , and θ represent altitude, longitude, and latitude, respectively. The generated views are clustered into aspects according to their mutual similarities calculated using (2) on the overlapped pixels. The aspect in this context means a cluster of views (can be one view) whose mutual similarities are higher than a threshold t_c . A complete-linkage clustering method is used for view clustering. After the clustering is finished, the aspect is downsampled to the next higher image pyramid level and the clustering process is repeated. As a result, we can obtain a hierarchical aspect graph spanning different image levels.

Online 2D Matching. After the hierarchical aspect graph has been built, it is ready to use in the online search phase for pose estimation. We assume the proposed method is used for real-time image registration, in which case the input data is a video stream. A tracking-matching-optimization strategy is proposed for robust and fast pose estimation. The output of the tracking is used to confine the search space at the top image level to a tight bounding box encompassing the object. The tracking-learning-detection (TLD) framework [2] is incorporated for tracking the 2D appearance of the 3D model over a video stream at a higher image level whose resolution is close to 512×512 . Let B_{target}^n be the tracked bounding box at the top image level n , denoted by $I^n(x, y)$. All aspects at the top level of the hierarchical aspect graph are examined within B_{target}^n in $I^n(x, y)$. An aspect is represented by its shape features $E_{2D} = \{x_i, y_i, \mathbf{d}_i\}_{i=1}^N$. To search for a match of an aspect, E_{2D} is scaled, rotated, and translated by discrete steps as follows

$$x'_i = x_i \sigma \cos \gamma - y_i \sigma \sin \gamma + t_x \quad (3)$$

$$y'_i = x_i \sigma \sin \gamma + y_i \sigma \cos \gamma + t_y \quad (4)$$

where σ is scaling factor; γ is rotation angle; (t_x, t_y) is translation. The similarity between the transformed E_{2D} and the image $I^n(x, y)$ is calculated using (2). Those 2D poses $(\sigma, \gamma, t_x, t_y)$ with resulting similarity exceeding a threshold t_s are stored in a candidate list. These candidates are either refined in the child aspects by searching close neighboring poses, or discarded due to lower similarity than t_s . All candidates are tracked down along the hierarchical level until reaching the bottom image level. The best candidate is considered as the candidate with the highest similarity score. The 3D pose can be recovered from the 2D pose of the best match at the bottom level and the pose of its associated aspect.

4.3 3D Pose Refinement

The accuracy of the obtained 3D pose from matching is usually insufficient due to the discrete step widths when searching for matches. Iterative optimization using an iterative closest point (ICP) algorithm is performed to refine the 3D pose by alternately identifying feature correspondences and estimating the pose. The corresponding 3D point (X_i, Y_i, Z_i) of the shape feature (x_i, y_i) can be recovered

using the z buffer value of OpenGL. The sub-pixel edge point (x'_i, y'_i) near (x_i, y_i) is localized along the direction of \mathbf{d}_i using the method proposed in [11]. The refined pose can be calculated by solving the PnP problem $(X_i, Y_i, Z_i) \leftrightarrow (x'_i, y'_i)$. The above procedure is repeated until convergence. Usually, several iterations will lead to satisfactory convergence.

5 Experiments and Results

5.1 Experimental Setting

Figure 4(a) shows the experimental scene. The LCD (6.4 inch) of the 3D display has a resolution of 1024×768 pixels with a pixel pitch of 0.13 mm (200 ppi). The micro lens array has lens pitches of 0.89 mm and 1.02 mm in the vertical and horizontal directions, respectively.

A mandibular phantom was created using a 3D printer from real patient's CT data as shown in Fig. 4(b). Fiducial points were made in the front teeth area and molar area of the phantom with known positions in the image (model) space, for accuracy evaluation. The front teeth model and the molar model shown in Fig. 4(c) are used for image registration. Which model should be used depends on the exposed area (front teeth area or molar area).

The camera (UI-3370CP-M-GL, IDS Imaging Development Systems GmbH, Germany) employed in the experiments has a resolution of 2048×2048 pixels with maximum frame rate of 80 frames per second (fps). Camera calibration was performed in advance to obtain the intrinsic matrix and remove lens distortion. The computer used in the experiments has an Intel® Core™ i7-4820K CPU (3.7GHz) and a NVIDIA® GeForce GTX TITAN Black GPU. The GPU is used to accelerate the aspect graph building process and the online matching process by parallel computing.

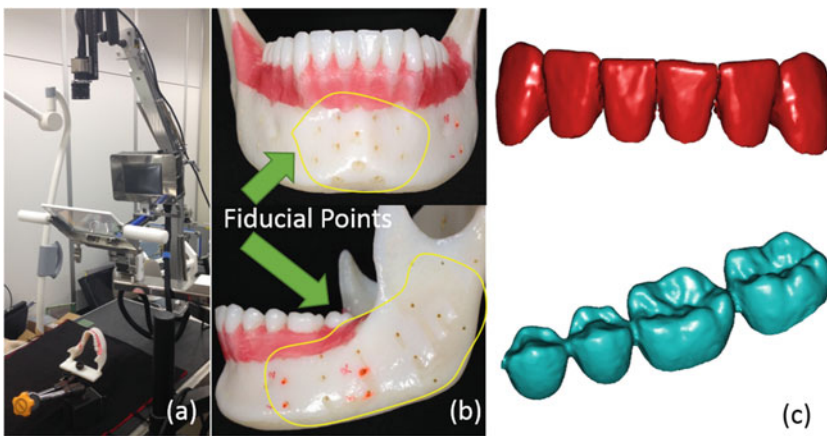


Fig. 4. (a) Experimental scene. (b) Mandibular phantom with fiducial points. (c) Teeth models for image registration.

5.2 Display-Camera Calibration

Figure 5 shows the automatic processing flow of the display-camera calibration. The original image was smoothed using a Gaussian filter. White spots were segmented out using a threshold followed by classifications according to their roundness and areas. The contours of the extracted spots were approximated by ellipses whose centers were used for PnP estimation. The PnP estimation yielded a geometric estimation error of 1.9 pixels.

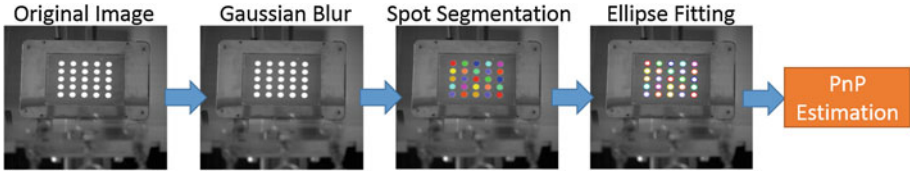


Fig. 5. Processing flow of display-camera calibration.

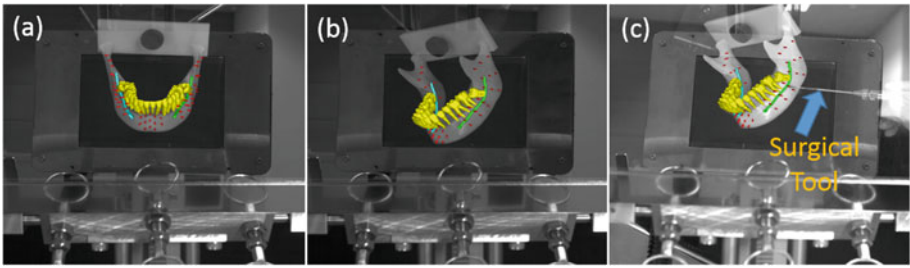


Fig. 6. Image registration results using (a) front teeth model (b) left molar model. (c) Target registration evaluation.

5.3 Image Registration Evaluation

Image registration was performed by matching the model with the video stream captured through the AR window. The distance between the camera and the phantom was approximately 710 mm. The registration process took approximately 0.2 s yielding an update frame rate of 5 fps. Figure 6(a) and (b) show the image registration results using the front teeth model and the left molar model (see Fig. 4(c)), respectively, with critical structures (tooth roots and nerve channels) and fiducial points (in red) overlaid on camera's view.

Target registration errors (TREs) on the fiducial points (see Fig. 4(b)) were calculated to evaluate the registration accuracy as follows. A surgical tool (dental drill) was used to approach individual fiducial points on the phantom under the guidance of the virtually overlaid fiducial points on camera's view. The physical distance between the indicated position and the real position of a fiducial point

on the phantom was measured as the error on that fiducial point. The average error distance in an evaluation area was calculated as the TRE in that area. For TRE calculation in the front teeth (left molar) area, the front teeth (left molar) model was used for the image registration. Figure 6(c) shows the accuracy evaluation process. The accuracy evaluation results yielded TREs of 0.8 mm in the front teeth area (15 points) and 1.1 mm in the molar area (18 points).

5.4 3D Surgical Overlay

After display-camera calibration and image registration, the necessary spatial information for 3D display has become available. Figure 7 shows the 3D overlay of the tooth roots and nerve channels, observed through the AR window. The visualized information could be used to guide surgical operation.

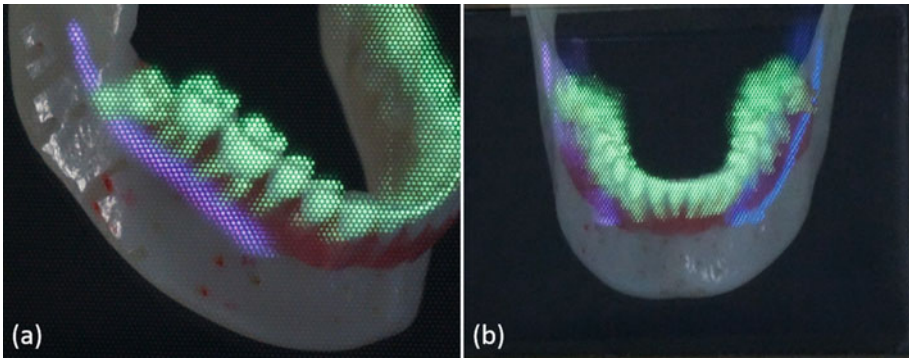


Fig. 7. 3D surgical overlay by (a) molar matching and (b) front teeth matching.

6 Conclusion

This paper presents a 3D surgical overlay system with automatic markerless image registration using a single camera. Teeth are rigid and easy to be exposed to a camera, making it possible to match a teeth model with an intraoperative camera image for the image registration task. The registered 3D images representing anatomical structures are superimposed on the patient via a translucent mirror for augmented reality surgical visualization. In this study, the application in dental surgery was demonstrated using our proposed system. Given the fact that the maxillary teeth are fixed with the skull, the proposed method may also be used for surgical navigation in the craniofacial region. In that case, the error compensation in the area far from the registration features can be a challenging work.

Acknowledgment. This work was supported in part by JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Multidisciplinary Computational Anatomy) JSPS KAKENHI Grant Number 26108008, and a Translational Research Network Program grant from the Ministry of Education, Culture, Sports, Science and Technology, Japan. H. Liao was supported in part by National Natural Science Foundation of China (Grant No. 81427803, 61361160417, 81271735).

References

1. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, Cambridge (2004). ISBN: 0521540518
2. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(7), 1409–1422 (2012)
3. Liao, H., Hata, N., Nakajima, S., Iwahara, M., Sakuma, I., Dohi, T.: Surgical navigation by autostereoscopic image overlay of integral videography. *IEEE Trans. Inf. Technol. Biomed.* **8**(2), 114–121 (2004)
4. Liao, H., Inomata, T., Sakuma, I., Dohi, T.: 3-D augmented reality for MRI-guided surgery using integral videography autostereoscopic image overlay. *IEEE Trans. Biomed. Eng.* **57**(6), 1476–1486 (2010)
5. Liao, H., Ishihara, H., Tran, H.H., Masamune, K., Sakuma, I., Dohi, T.: Precision-guided surgical navigation system using laser guidance and 3D autostereoscopic image overlay. *Comput. Med. Imaging Graph.* **34**(1), 46–54 (2010)
6. Sebastian, T., Klein, P., Kimia, B.: On aligning curves. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(1), 116–125 (2003)
7. Steger, C.: Occlusion, clutter, and illumination invariant object recognition (2002)
8. Tran, H., Suenaga, H., Kuwana, K., Masamune, K., Dohi, T., Nakajima, S., Liao, H.: Augmented reality system for oral surgery using 3D auto stereoscopic visualization. In: Fichtinger, G., Martel, A., Peters, T. (eds.) *MICCAI 2011, Part I*. LNCS, vol. 6891, pp. 81–88. Springer, Heidelberg (2011)
9. Ulrich, M., Wiedemann, C., Steger, C.: Combining scale-space and similarity-based aspect graphs for fast 3D object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(10), 1902–1914 (2012)
10. Wang, J., Suenaga, H., Hoshi, K., Yang, L., Kobayashi, E., Sakuma, I., Liao, H.: Augmented reality navigation with automatic marker-free image registration using 3D image overlay for dental surgery. *IEEE Trans. Biomed. Eng.* **61**(4), 1295–1304 (2014)
11. Wang, J., Kobayashi, E., Sakuma, I.: Coarse-to-fine dot array marker detection with accurate edge localization for stereo visual tracking. *Biomed. Sig. Process.* **15**, 49–59 (2015)
12. Wang, J., Suenaga, H., Liao, H., Hoshi, K., Yang, L., Kobayashi, E., Sakuma, I.: Real-time computer-generated integral imaging and 3D image calibration for augmented reality surgical navigation. *Comput. Med. Imaging Graph.* **40**, 147–159 (2015)
13. Wang, J., Suenaga, H., Yang, L., Liao, H., Kobayashi, E., Takato, T., Sakuma, I.: Real-time marker-free patient registration and image-based navigation using stereovision for dental surgery. In: Liao, H., Linte, C.A., Masamune, K., Peters, T.M., Zheng, G. (eds.) *MIAR 2013 and AE-CAI 2013*. LNCS, vol. 8090, pp. 9–18. Springer, Heidelberg (2013)