# Image-Based Tracking of the Teeth for Orthodontic Augmented Reality

André Aichert[1,2], Wolfgang Wein[1,2], Alexander Ladikos[1], Tobias Reichl[2], and Nassir Navab[2]

[1] White Lion Technologies AG, Munich, Germany
[2] Chair for Computer Aided Medical Procedures (CAMP), Technische Universität München, Munich, Germany
aichert@cs.tum.edu

**Abstract.** We present image-based methods for tracking teeth in a video image with respect to a CT scan of the jaw, in order to enable a novel light-weight augmented reality (AR) system in orthodontistry. Its purpose is guided bracket placement in orthodontic correction. In this context, our goal is to determine the position of the patient maxilla and mandible in a video image solely based on a CT scan. This is suitable for image guidance through an overlay of the video image with the planned position of brackets in a monocular AR system. Our tracking algorithm addresses the contradicting requirements of robustness, accuracy and performance in two problem-specific formulations. First, we exploit a distance-based modulation of two iso-surfaces from the CT image to approximate the appearance of the gum line. Second, back-projection of previous video frames to an iso-surface is used to account for recently placed brackets. In combination, this novel algorithm allowed us to track several sequences of three patient videos of real procedures, despite difficult lighting conditions. Paired with a systematic evaluation, we were able to show practical feasibility of such a system.

## 1 Introduction

This paper suggests a novel solution for guidance in orthodontic applications with a light-weight monocular video see-through Augmented Reality (AR) system. It targets the guided placement of brackets onto individual teeth in order to improve efficacy and reduce chair time of bracket placement and re-adjustments for dental braces in orthodontic correction, therefore allowing to incorporate pre-procedure simulation and planning. The state of the art for this procedure relies solely on the experience of the orthodontist for both placement of the brackets and the choice of the wire tension between the brackets. In dentistry, low-dose cone-beam CT reconstructions of the jaw are typically obtained with modern digital volume tomography (DVT) devices, with an acceptable dose limit even for teenagers. Related research [1] has developed simulations based on finite element methods from such CT data of teeth and bone. Those simulations could be used in a pre-procedural planning of the optimal bracket placement and

N. Ayache et al. (Eds.): MICCAI 2012, Part II, LNCS 7511, pp. 601–608, 2012.

wire tension, such that patient teeth move in an ideal manner while minimizing rotation. While much less accuracy is needed than for example in implant placement, a realization of the pre-procedure plan requires guided placement of the brackets. Augmentation of a patient video showing a superimposition of the newly placed bracket with its planned position would suffice. Potential benefits include higher efficacy due to a reduction in chair time with fewer follow-up visits for corrections, as well as reduced probability of relapse. Therefore, this routine procedure could be improved with a light-weight monocular augmented reality system, while avoiding the cost and complexity of a full-scale medical AR solution.

## 2 Related Work

One of the central choices which had to be made in realizing this system, is the choice of tracking algorithm [7]. There are several approaches commonly employed for optical tracking, such as marker-based tracking, template-based tracking [2], feature-based tracking [8] and edge-based tracking [5] as well as combinations of these methods. However, due to the nature of our tracking target not all methods can be used. Marker-based tracking is not relevant, because we do not want to augment the scene. When using external tracking systems, a disadvantage besides their high cost is the challenge to keep overall tracking error low, particularly in scenarios such as dental implant placement [12]. This is because the overall system accuracy is limited by the accumulated errors from the tracking system itself, patient registration, hand-eye calibration, synchronization, etc. In order to avoid such an accumulation of errors (as well as additional, expensive equipment) we employ solely image-based tracking methods, which track the patient jaw directly in the video used for the overlay. Template-based tracking by itself is too unstable due to illumination variations, occlusions and the need for an initially textured model of the scene, which we don't have. Feature-based tracking is also not feasible since our scene is mostly textureless and feature-point extractors will not find enough reliable features. This only leaves edge-based tracking methods for use in our system. As can be seen in Fig. 1 edges are a very dominant and a stable feature in the input image. We therefore chose to use edges as our primary tracking modality which is augmented by template-based tracking methods for increased robustness.

## 3 Methods

In our chosen scenario, a DVT volume of the jaw is always available, since it is the basis of the numerical simulation for planning. Therefore, we investigate methods to automatically align a DVT volume with a video image feed, which relates our problem to medical 2D-3D registration [9] and tracking [11]. For an impression of the scenario see figure Fig. 1. We attempt to use all information available from these two modalities and we present a method consisting of two complementary steps. One step provides a good overall alignment, the other step ensures robust

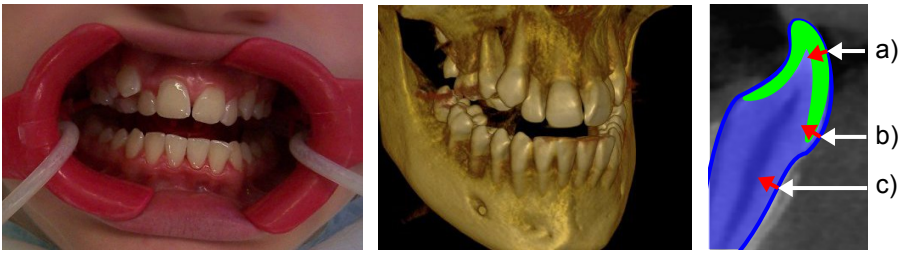


**Fig. 1.** Photograph of a prepared patient (left), a high-quality volume rendering of the jaw from a DVT volume (center) and schematic of distance-based modulation of two iso-surfaces (right)

tracking even in light of changing conditions. The underlying 2D-3D registration problem is solved through an iterative optimization of a similarity metric over a 6 DOF pose.

### 3.1 Dual Iso-Surfaces

In order to compute image similarity, we need a fast method of generating a 2D image from the DVT volume for comparison with the 2D video image. While direct volume rendering is able to create close to photo-realistic images, overly complex methods are too slow for real-time registration. Simpler methods, such as non-polygonal iso-surfaces can be computed extremely efficiently, since they represent only one intensity threshold in the data set. In the patient videos, the shape of the teeth and the gum line contain the most reliable geometric information. Unfortunately, in the DVT volume the gum line itself is an interface between two intensities, i.e. enamel and gum. It can therefore not be retrieved by single iso-surface rendering. We suggest using a modulation based on normal distances of two iso-surfaces for the visualization of the gum line. Related work in "Focus and Context" visualization [6] addressed a similar problem for context modulation, where one iso-surface is shown transparent when close to a second one.

While the teeth are easily visualized as the highest CT intensities (i.e. X-Ray attenuation or Hounsfield units), different types of tissue, including the tongue, cheeks, lip and gum all have almost identical attenuation values. Thus, the lip folding over the gum cannot be separated, and therefore the actual gum line cannot be visualized reliably. As an alternative solution, we visualize the interface between enamel and bone or dentin (i.e. the roots of the teeth). This line approximately follows the course of the gum line and we could verify in experiments that this approximation is bias-free with respect to the registration. Please see Fig. 3 (right) for a direct comparison of the dual iso-surface and the textured iso-surface, in particular the course of the gum line.

We efficiently implement this in a single pass of a GPU ray-caster, with a speed close to single iso-surface rendering. As a ray from the camera center into the scene hits the outer surface representing gum (or dentin instead), its direction is changed to the normal of that surface. The ray then is followed for

only a few millimeters more, possibly hitting the second surface for enamel. See Fig. 1 for three example points on the outer surface (blue) and their relation to the inner surface for enamel (green) in a cross section. If the enamel surface is not found (case (c) in Fig. 1 (right)), the gum surface is fully opaque. If the enamel surface is right next to the outer surface (Fig. 1 (right, a)), the outer surface is fully transparent. Otherwise, the surfaces are blended based on their distance (Fig. 1 (right, b)).

### 3.2 Dissimilarity Metric

Once the DVT volume is rendered, we compare it to the video image. We apply an edge-based similarity metric between the video image and the dual iso-surface rendering. This is based on weighting the distance to the closest edge in the video image with the gradient magnitude of the dual iso-surface rendering. Lower results of this metric indicate that the edges are well-aligned. For efficiency, we compute a distance map from the output of a Canny edge filter [3] of the video frame, since the video image stays constant during the pose optimization. The measure then becomes

$$d = \frac{\sum_{x,y} \left( d(x,y) \cdot g(x,y) \right)}{\sum_{x,y} g(x,y)}$$

where $d(x,y)$ is the distance map of the video edges and $g(x,y) = \|\nabla_{(x,y)} J\|^2$ is the squared image gradient magnitude of the dual iso-surface rendering $J$ at pixel coordinates $(x,y)$. We use an exponent of two, since we put more emphasis on regions with large gradients (i.e. edges), while reducing the weight of areas with only small variations in gradient caused by noise. Fig. 2 shows a plot of the dissimilarity metric against $x$ and $y$ translations parallel to the image plane. Millimeters are measured at the depth of the jaw. Notice a clear minimum at the center of the plot, which represents perfect visual alignment such as seen in Figure 3. Dissimilarity increases more quickly in $y$-direction, since the dominant occlusion edges of the teeth are in $x$-direction in the image. We use an elliptical region of interest, defined by two focal points as the projections of the canines.

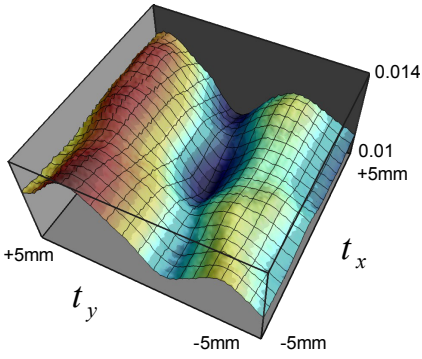


**Fig. 2.** Plot of the dissimilarity metric $d$ for translations parallel to the image plane with a clear minimum for perfect visual alignment at the center

### 3.3 Initial Alignment

In order to increase capture range and speed up the process, we propose another step complementing this edge-based approach. Note that there is a variety of

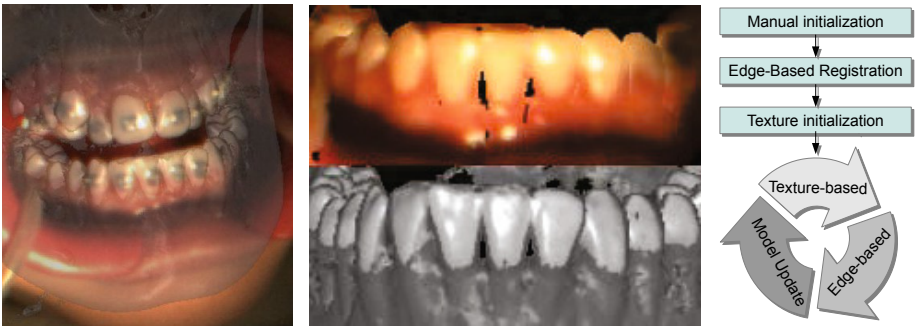


**Fig. 3.** Video image overlaid with an aligned iso-surface (left) and examples of the rendering methods suggested in Sections 3.3 and 3.1 (center) and a flowchart of the prototype system (right)

established tracking techniques that can be used to achieve such a pre-alignment, some may not even need knowledge of the CT.

Since in our scenario the CT is available, we can reproject video pixels into arbitrary views based on the surface shape from the CT and correct alignment on any one video image. If we choose a texture wrapping of a 2D texture for the iso-surface, we can also project this color information directly to the surface. With a textured model of the jaw, we obtain a much simpler mono-modal registration problem, which can be solved using a simple mono-modal similarity metric. We use a simple frontal linear projection of the texture image to the surface. This is acceptable for this application, since both patient motion and changes in view direction are small, as the orthodontist remains on the same side of the patient.

The system can be initialized manually, by asking the orthodontist to move their head and roughly align a transparent view of the CT with his own vision. The system uses edge based registration to get an initial alignment, which in turn is used to initialize the model texture. In the course of the procedure brackets will be placed onto individual teeth. As the difference between the iso-surface model and the reality increases, tracking will become less reliable. To account for these changes, we suggest updating and progressively refining a textured model of the jaw. This however is susceptible to the template update problem (i.e. drift) [10]. In combination with the edge-based approach we are able to exploit the accuracy of the edge-based registration with a quick but reliable pre-alignment using the textured model. The result is a robust alignment with a large capture range, despite challenging image data.

## 4 Experiments

For the evaluation of the proposed system, we created a prototype to study practical feasibility. In cooperation with orthodontic partners we acquired data for the real procedure of three teenage patients. The three videos each are about 20 minutes long and show the whole procedure from the perspective of a camera

mounted to the orthodontist's head. In each case DVT image data is available. In the following, we examine both the registration for single frames and the behavior for short video sequences.

### 4.1 Random Studies

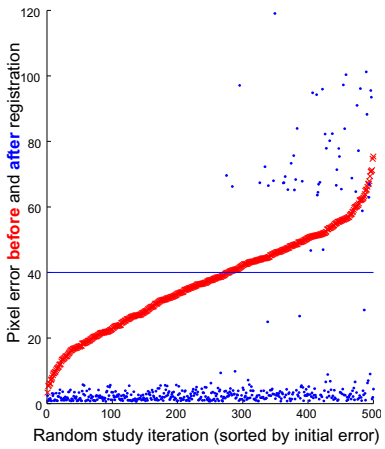


**Fig. 4.** Pixel error of a random study sorted by initial offsets

In a scenario where tracking is performed with the goal of a scene overlay, errors in image pixels are more relevant than in transformation parameter space. We therefore choose interest points on the surface of the teeth and compute the average projection errors. While 2D-3D point correspondences at significant edges and corners between the teeth were defined, they generally resulted in poor visual alignment due to the limited accuracy of placing those landmarks (i.e. target localization error TLE of over 5 pixels). Therefore, a quasi ground truth registration was defined based on the optimal visual alignment by the expert. In several random studies, we perturbed this ground truth alignment for all 6 pose parameters. The parameters were chosen, such that the $x$ and $y$ axis are parallel to the image plane and $z$ is facing the camera with the origin at the center of the jaw. In Fig. 4 (left) we present a random study of 500 iterations as a typical representative. Translation was randomized in a range of $\pm 20$ mm in $x$ and $y$ direction and $\pm 10$ mm in $z$, while rotation in all three axes was randomized in a range of $\pm 10$ degrees, enough to observe a failure of the algorithm in some instances. In the specific view of of the patient, similar to the one shown in Fig. 3 (left), this corresponds to an average pixel offset of 37.2 pixels on the $640 \times 480$ video image, which is well beyond expected inter-frame motion. After removal of 10% outliers, we were able to recover from an average error of 35.4 pixels to just 2.7 pixels, or an average of 2.1 mm and 4 degrees. In Fig. 4 (left) you can see the results of the random study including outliers, sorted by initial pixel error (red). Observe that the algorithm was successful in each case with offsets of less than 40 pixels, which is marked with a blue line in the plot, with the error after registration (blue) well below the red line. Even beyond that threshold, correct alignment is recovered in about 75% of the cases.

### 4.2 Image Sequences

We successfully tracked several sequences of all three patient videos. Visual alignment appeared accurate and reliable, especially around the incisors. Despite the dental prop, patients are moving their jaw during the procedure, which forced us to track upper and lower halves of the jaw independently. We believe that this

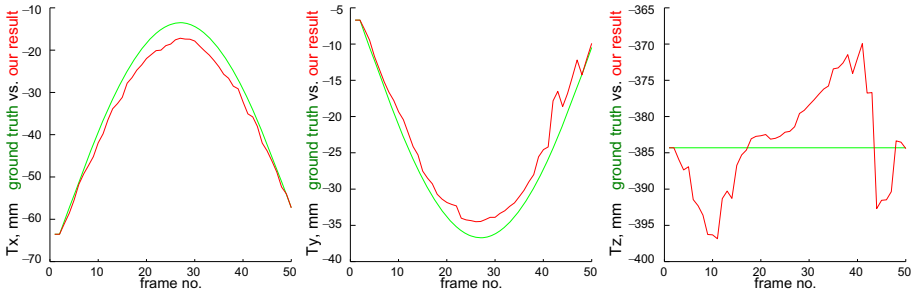


**Fig. 5.** Comparison of ground truth and our tracking results for a synthetic sequence

motion can be modeled with few parameters and included in one optimization, ultimately making alignment more stable, especially at the molars. Separation of mandible and maxilla can be performed automatically by fitting a plane.

In order to quantify behavior over a sequence of frames, we created a synthetic sequence using a textured model. While this experiment is simpler than real patient data, it allows us to compare the computed poses to ground truth. We chose a linear in-plane translation of $x$ and $y$, as well as rotation about these axes and a motion returning to the starting position in 50 frames as a test case. See Fig. 5 (three plots on the left) for plots of the translation components of the resulting poses. Although the in-plane translation was recovered up to about 5 mm, there is an error in the $z$-translation by as much as 15 mm, which is expected for 2D-3D registration. As the $z$-axis is facing the viewer, translations in that direction have little effect on the image (and hence also on the final superimposition). The error in target points was 6.7 pixels average over the whole sequence.

## 5 Conclusion

We presented a tracking solution and novel guidance system for orthodontic correction. We focused on the feasibility of a tracking system based on a CT volume and the patient color video sequence. A multi-step algorithm was devised to use several aspects of the data. The proposed approach includes a dual iso-surface rendering method with distance based modulation to produce fast high-quality images of the gum line, paired with a textured model based pre-alignment and update step. In extensive random studies we could show correct registration of single images; more importantly, several sequences of real procedures were successfully tracked. In conclusion, we enabled a novel application of augmented reality in an orthodontics routine procedure. Future work could focus on a parameterization of jaw movement for concurrent tracking of both halves, as well as better handling of occlusion by the orthodontist's tool. While this work focused on recursive tracking for high accuracy, detection of the prop or the teeth could complement the current method (e.g. using an advanced approach such as

[4]). In the future our prototype has to be integrated with simulation and planning capabilities in order to create a fully practical solution, and a systematic quantitative evaluation of tracking accuracy performed.

**Acknowledgements.** We would like to thank Dr. V. Rummel, Dortmund, Germany, for his support during data acquisition and his valuable medical feedback.

## References

1. Ammar, H.H., Ngan, P., Crout, R.J., Mucino, V.H., Mukdadi, O.M.: Three-dimensional modeling and finite element analysis in treatment planning for orthodontic tooth movement. American Journal of Orthodontics and Dentofacial Orthopedics 139, 59–71 (2011)
2. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision 56(3), 221–255 (2004)
3. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698 (1986)
4. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of texture-less objects. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
5. Klein, G., Murray, D.: Full-3D edge tracking with a particle filter. In: Proc. British Machine Vision Conference (BMVC 2006), vol. 3, pp. 1119–1128. BMVA, Edinburgh (September 2006)
6. Krüger, J., Schneider, J., Westermann, R.: ClearView: An interactive context preserving hotspot visualization technique. IEEE Transactions on Visualization and Computer Graphics 12(5) (September-October 2006)
7. Lepetit, V., Fua, P.: Monocular model-based 3D tracking of rigid objects: A survey. In: Foundations and Trends in Computer Graphics and Vision, pp. 1–89 (2005)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
9. Markelj, P., Tomazevic, D., Likar, B., Pernus, F.: A review of 3D/2D registration methods for image-guided interventions. Medical Image Analysis (2010)
10. Matthews, I., Ishikawa, T., Baker, S.: The template update problem. In: Proceedings of the British Machine Vision Conference (September 2003)
11. Mori, K., Deguchi, D., Sugiyama, J., Suenaga, Y., Toriwaki, J., Maurer Jr., C.R., Takabatake, H., Natori, H.: Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. Medical Image Analysis 6(3), 321–336 (2002)
12. Wanschitz, F., Birkfellner, W., Figl, M., Patruta, S., Wagner, A., Watzinger, F., Yerit, K., Schicho, K., Hanel, R., Kainberger, F., Imhof, H., Bergmann, H., Ewers, R.: Computer-enhanced stereoscopic vision in a head-mounted display for oral implant surgery. Clinical Oral Implants Research 13, 610–616 (2002)