The Endoscopogram: a 3D Model Reconstructed from Endoscopic Video Frames

Qingyu Zhao[1], True Price[1], Stephen Pizer[1,2], Marc Niethammer[1], Ron Alterovitz[1], Julian Rosenman[1,2]

[1] Computer Science,UNC Chapel Hill, NC, United States
[2] Radiation Oncology, UNC Chapel Hill, NC, United States

Abstract. Endoscopy enables high resolution visualization of tissue texture and is a critical step in many clinical workflows, including diagnosis and treatment planning for cancers in the nasopharynx. However, an endoscopic video does not provide 3D spatial information, making it difficult to use in tumor localization, and it is inefficient to review. We introduce a pipeline for automatically reconstructing a textured 3D surface model, which we call an endoscopogram, from multiple 2D endoscopic video frames. Our pipeline first reconstructs a partial 3D surface model for each input individual 2D frame. In the next step (which is the focus of this paper), we generate a single high-quality 3D surface model using a groupwise registration approach that fuses multiple, partially overlapping, incomplete and deformed surface models together. We generate endoscopograms from synthetic, phantom, and patient data and show that our registration approach can account for tissue deformations and reconstruction inconsistency across endoscopic video frames.

1 Introduction

Modern radiation therapy treatment planning relies on imaging modalities like CT for tumor localization. For throat cancer, an additional kind of medical imaging, called endoscopy, is also taken at treatment planning time. Endoscopic videos provide direct optical visualization of the pharyngeal surface and provide information, such as a tumor's texture and superficial (mucosal) spread, that is not available on CT due to CT's relatively low contrast and resolution. However, the use of endoscopy for treatment planning is significantly limited by the fact that (1) the 2D frames from the endoscopic video do not explicitly provide 3D spatial information, such as the tumor's 3D location; (2) reviewing the video is time-consuming and the optical views do not provide the full geometric conformation of the throat.

In this paper, we introduce a pipeline for reconstructing a 3D textured surface model of the throat, which we call an endoscopogram, from 2D video frames. The model provides (1) more complete 3D pharyngeal geometry; (2) efficient visualization; (3) the opportunity for registration with CT, thereby enabling transfer of the tumor location and texture into the CT space.

State-of-the-art endoscopic reconstruction techniques have been mostly applied in applications like colonoscopy inspection [1, 2] and bone reconstruction for orthopedic surgeries [3]. However, such methods can not be directly used for nasopharyngoscopy reconstruction, because they cannot efficiently deal with the following three challenges: (1) non-Lambertian surfaces; (2) poorly known shape priors; (3) non-rigid deformation of tissues across frames. Our proposed pipeline deals with these problems using (1) a Shape-from-Motion-and-Shading (SfMS) method [4] incorporating a new reflectance model for generating single-frame-based partial reconstructions; (2) a novel geometry fusion algorithm for non-rigid fusion of multiple partial reconstructions. Since our pipeline does not assume any prior knowledge on environments and shapes, it can be readily generalized to other types of endoscopic reconstruction applications.

In this paper we focus on the geometry fusion step mentioned above. The challenge here is that all individual reconstructions are only partially overlapping due to the constantly changing camera viewpoint, may have missing data (holes) due to camera occlusion, and may be slightly deformed since the tissue may have deformed between 2D frame acquisitions. Our main contribution in this paper is the design of a novel groupwise surface registration algorithm that can deal with these limitations. An additional contribution is an outlier geometry trimming algorithm based on robust regression. We generate endoscopograms and validate our registration algorithm with data from synthetic CT surface deformations and endoscopic video of a rigid phantom and real patients.

2 Endoscopogram Reconstruction Pipeline

The input to our system (Fig. 1) is a video sequence of hundreds of consecutive frames $\{\mathcal{F}_i | i = 1...N\}$. The output is an endoscopogram, which is a textured 3D surface model derived from the input frames. We first generate for each frame $\mathcal{F}_i$ a reconstruction $\mathcal{R}_i$ by the SfMS method. We then fuse multiple single-frame reconstructions $\{\mathcal{R}_i\}$ into a single geometry $\mathcal{R}$. Finally, we texture $\mathcal{R}$ by pulling color from the original frames $\{\mathcal{F}_i\}$. We will focus on the geometry fusion step in Section 3 and briefly introduce the other techniques in the rest of this section.

Shape from Motion and Shading (SfMS). Our novel reconstruction method [4] has been proven to be efficient in single-camera reconstruction of live endoscopy data. The method leverages sparse geometry information obtained

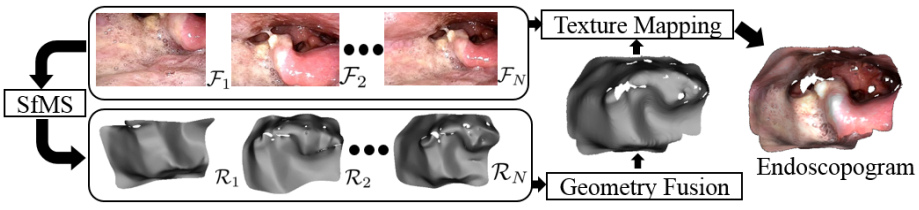


Fig. 1. The endoscopogram reconstruction pipeline.

by Structure-from-Motion (SfM), Shape-from-Shading (SfS) estimation, and a novel reflectance model to characterize non-Lambertian surfaces. In summary, it iteratively estimates the reflectance model parameters and the SfM-regularized SfS reconstruction surface for individual frames. One drawback of this method is that large tissue deformation and lighting changes across frames can induce inconsistent individual SfS reconstructions. Nevertheless, our experiments show that this kind of error can be well compensated in the subsequent geometry fusion step. In the end, for each frame $\mathcal{F}_i$, a reconstruction $\mathcal{R}_i$ is produced as a triangle mesh and transformed into the world space using the camera position parameters estimated from SfM. Mesh faces that are nearly tangent to the camera viewing ray are removed because they correspond to occluded regions. The end result of this is that the reconstructions $\{\mathcal{R}_i\}$ have missing patches and different topology and are only partially overlapping with each other.

Texture Mapping. The goal of texture mapping is to assign a color to each vertex v^k (superscripts refer to vertex index) in the fused geometry $\mathcal{R}$, which is estimated by the geometry fusion (Section 3) of all the registered individual frame surfaces $\{\mathcal{R}'_i\}$. Our idea is to find a corresponding point of v^k in a registered surface $\mathcal{R}'_i$ and to trace back its color in the corresponding frame $\mathcal{F}_i$. Since v^k might have correspondences in multiple registered surfaces, we formulate this procedure as a labeling problem and optimize a Markov Random Field (MRF) energy function. In general, the objective function prefers pulling color from non-boundary nearby points in $\{\mathcal{R}'_i\}$, while encouraging regional label consistency.

3 Geometry Fusion

This section presents the main methodological contributions of this paper: a novel groupwise surface registration algorithm based on N-body interaction, and an outlier-geometry trimming algorithm based on robust regression.

Related Work. Given the set of partial reconstructions $\{\mathcal{R}_i\}$, our goal is to non-rigidly deform them into a consistent geometric configuration, thus unifying the tissue deformation and minimizing reconstruction inconsistency among different frames. Current groupwise surface registration methods often rely on having or iteratively estimating the mean geometry (template) [5]. However, in our situation, the topology change and partially overlapping data renders initial template geometry estimation almost impossible. Missing large patches also pose serious challenges to the currents metric [6] for surface comparison. Template-free methods have been studied for images [7], but it has not been shown that such methods can be generalized to surfaces. The joint spectral graph framework [8] can match a group of surfaces without estimating the mean, but these methods do not explicitly compute deformation fields for geometry fusion.

Zhao et. al. [9] proposed a pairwise surface registration algorithm, Thin Shell Demons, that can handle topology change and missing data. We have extended this algorithm into our groupwise situation.

3.1 Thin Shell Demons

Thin Shell Demons uses geometric virtual forces and a thin shell model to compute physically realistic deformation. The so-called forces $\{f\}$ between two surfaces $\{\mathcal{R}_1, \mathcal{R}_2\}$ are vectors connecting corresponding vertex pairs, i.e. $\{f(v^k) = u^k - v^k \mid v^k \in \mathcal{R}_1, u^k \in \mathcal{R}_2\}$ (with some abuse of notation, we use k here to index correspondences). The algorithm regards the surfaces as elastic thin shells and produces a physically realistic deformation field $\phi : \mathcal{R}_1 \to \mathcal{R}_2$ by iteratively minimizing the energy function $E(\phi) = \sum_{k=1}^{M} c(v^k)(\phi(v^k) - f(v^k))^2 + E_{shell}(\phi)$. The first part penalizes inconsistency between the deformation vector and the force vector applied on a point and uses a confidence score c to weight the penalization. The second part minimizes the thin shell deformation energy, which is defined as the integral of local bending and membrane energy:

$$E_{shell}(\phi) = \int_{\mathcal{R}} \lambda_1 W(\sigma_{mem}(p)) + \lambda_2 W(\sigma_{bend}(p)), \tag{1}$$

$$W(\sigma) = Y/(1-\tau^2)((1-\tau)tr(\sigma^2) + \tau tr(\sigma)^2), \tag{2}$$

where Y and τ are the Young's modulus and Poisson's ratio of the shell. σ_{mem} is the tangential Cauchy-Green strain tensor characterizing local stretching and is computed by $J_\varphi(p)^T J_\varphi(p)$, where $J_\varphi(p)$ is the Jacobian of the tangential transformation $\varphi : T_p \to T_{\phi(p)}$ at local point p. The bending strain tensor σ_{bend} characterizes local curvature change and is computed by the shape operator difference $\varphi^T \widetilde{\Lambda}_{\phi(p)} \varphi - \Lambda_p$.

3.2 N-body Surface Registration

Our main observation is that the virtual force interaction is still valid among N partial shells even without the mean geometry. Thus, we propose a groupwise deformation scenario as an analog to the N-body problem: N surfaces are deformed under the influence of their mutual forces. This groupwise attraction can bypass the need of a target mean and still deform all surfaces into a single geometric configuration. The deformation of a single surface is independent and fully determined by the overall forces exerted on it. With the physical thin shell model, its deformation can be topology-preserving and not influenced by its partial-ness. With this notion in mind, we now have to define (1) mutual forces among N partial surfaces; (2) an evolution strategy to deform the N surfaces.

Mutual Forces. In order to derive mutual forces, correspondences should be credibly computed among N partial surfaces. It has been shown that by using the geometric descriptor proposed in [10], a set of correspondences can be effectively computed between partial surfaces. Additionally, in our application, each surface $\mathcal{R}_i$ has an underlying texture image $\mathcal{F}_i$. Thus, we also compute texture correspondences between two frames by using standard computer vision techniques [11]. To improve matching accuracy, we compute inlier SIFT correspondences only between temporally close frame pairs, i.e. $\{\mathcal{F}_i, \mathcal{F}_j \mid 0 < |i-j| \leq T\}$. Finally, these SIFT matchings can be directly transformed to 3D vertex correspondences via the SfSM reconstruction procedure.

In the end, any given vertex $v_i^k \in \mathcal{R}_i$ will have M_i^k corresponding vertices in other surfaces $\{\mathcal{R}_j | j \neq i\}$, given as vectors $\{f^\beta(v_i^k) = u^\beta - v_i^k, \beta = 1...M_i^k\}$, where u^β is the β^{th} correspondence of v_i^k in some other surface. These correspondences are associated with confidence scores $\{c^\beta(v_i^k)\}$ defined by

$$c^\beta(v_i^k) = \begin{cases} \delta(u^\beta, v_i^k) & \text{if } \langle u^\beta, v_i^k \rangle \text{ is a geometric correspondence,} \\ \bar{c} & \text{if } \langle u^\beta, v_i^k \rangle \text{ is a texture correspondence,} \end{cases} \tag{3}$$

where δ is the geometric feature distance defined in [10]. Since we only consider inlier SIFT matchings using RANSAC [11], the confidence score for texture correspondences is a constant $\bar{c}$. We then define the overall force exerted on v_i^k as the Nadaraya-Watson kernel-weighted average based on confidence scores: $\bar{f}(v_i^k) = \sum_{\beta=1}^{M_i^k} c^\beta(v_i^k) f^\beta(v_i^k) / \sum_{\beta=1}^{M_i^k} c^\beta(v_i^k)$.

Deformation Strategy. With mutual forces defined, we can solve for the group deformation fields $\{\phi_i\}$ by optimizing independently for each surface

$$E(\phi_i) = \sum_{k=1}^{M_i} c(v_i^k)(\phi(v_i^k) - \bar{f}(v_i^k))^2 + E_{shell}(\phi_i), \tag{4}$$

where M_i is the number of vertices that have forces applied. Then, a groupwise deformation scenario is to evolve the N surfaces by iteratively estimating the mutual forces $\{f\}$ and solving for the deformations $\{\phi_i\}$. However, a potential hazard of our algorithm is that without a common target template, the N surfaces could oscillate, especially in the early stage when the force magnitudes are large and tend to overshoot the deformation. To this end, we observe that the thin shell energy regularization weights λ_1, λ_2 control the deformation flexibility. Thus, to avoid oscillation, we design the strategy shown in Algorithm 1.:

Algorithm 1 N-body Groupwise Surface Registration

1: Start with large regularization weights: $\lambda_1(0), \lambda_2(0)$
2: In iteration p, compute $\{f\}$ from the current N surfaces $\{\mathcal{R}_i(p)\}$
3: Optimize Eq. 4 independently for each surface to obtain $\{\mathcal{R}_i(p+1)\}$
4: $\lambda_1(p+1) = \sigma * \lambda_1(p),\ \lambda_2(p+1) = \sigma * \lambda_1(p)$, with $\sigma < 1$
5: Go to step 2 until convergence.

3.3 Outlier Geometry Trimming

The final step of geometry fusion is to estimate a single geometry $\mathcal{R}$ from the registered surfaces $\{\mathcal{R}_i'\}$ [12]. However, this fusion step can be seriously harmed by the outlier geometry created by SfMS. Outlier geometries are local surface parts that are wrongfully estimated by SfMS under bad lighting conditions (insufficient lighting, saturation, or specularity) and are drastically different from all other surfaces (Fig. 2a). The sub-surfaces do not correspond to any part in other surfaces and thereby are carried over by the deformation process to $\{\mathcal{R}_i'\}$.

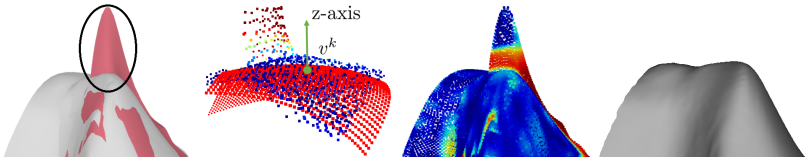


Fig. 2. (a) 5 overlaying registered surfaces, one of which (pink) has a piece of outlier geometry (circled) that does not correspond to anything else. (b) Robust quadratic fitting (red grid) to normalized $\mathcal{N}(v^k)$. The outlier scores are indicated by the color. (c) Color-coded $\mathcal{W}$ on $\mathcal{L}$. (d) Fused surface after outlier geometry removal.

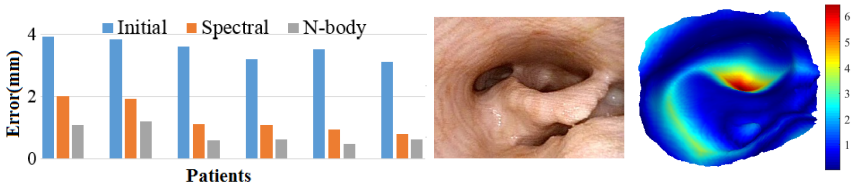
Our observation is that outlier geometry changes a local surface's topology (branching) and violates many differential geometry properties. We know that the local surface around a point in a smooth 2-manifold can be approximately presented by a quadratic Monge Patch $h : U \to \mathbb{R}^3$, where U defines a 2D open set in the tangent plane, and h is a quadratic height function. Our idea is that if we robustly fit a local quadratic surface at a branching place, the surface points on the wrong branch of outlier geometry will be counted as outliers (Fig. 2b).

We define the 3D point cloud $\mathcal{L} = \{v^1, ...v^P\}$ of P points as the ensemble of all vertices in $\{\mathcal{R}'_i\}$, $\mathcal{N}(v^k)$ as the set of points in the neighborhood of v^k and $\mathcal{W}$ as the set of outlier scores of $\mathcal{L}$. For a given v^k, we transform $\mathcal{N}(v^k)$ by taking v^k as the center of origin and the normal direction of v^k as the z-axis. Then, we use Iteratively Reweighted Least Squares [13] to fit a quadratic polynomial to the normalized $\mathcal{N}(v^k)$ (Fig. 2b). The method produces outlier scores for each of the point in $\mathcal{N}(v^k)$, which are then accumulated into $\mathcal{W}$ (Fig. 2c). We repeat this robust regression process for all v^k in $\mathcal{L}$. Finally, we remove the outlier branches by thresholding the accumulative scores $\mathcal{W}$, and the remaining largest point cloud is used to produce the final single geometry $\mathcal{R}$ [12] (Fig. 2d).

4 Results

We validate our groupwise registration algorithm by generating and evaluating endoscopograms from synthetic data, phantom data, and real patient endoscopic videos. We selected algorithm parameters by tuning on a test patient's data (separate from the datasets presented here). We set the thin shell elastic parameters $Y = 2, \tau = 0.05$, the energy weighting parameters $\lambda_1 = \lambda_2 = 1, \sigma = 0.95$, the frame interval $T = 0.5s$, and the texture confidence score $\bar{c} = 1$.

Synthetic Data. We produced synthetic deformations to 6 patients' head-and-neck CT surfaces. Each surface has 3500 vertices and a 2-3cm cross-sectional diameter, covering from the pharynx down to the vocal cords. We created deformations typically seen in real data, such as the stretching of the pharyngeal wall and the bending of the epiglottis. We generated for each patient 20 partial surfaces by taking depth maps from different camera positions in the CT space. Only geometric correspondences were used in this test. We measured the



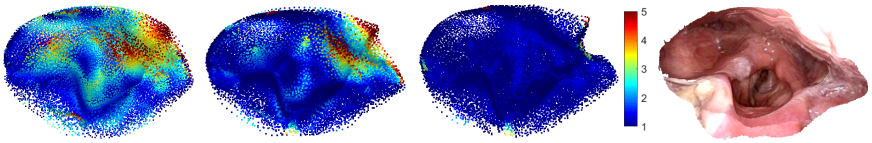
(a) (b)

Fig. 3. (a) Error plot of synthetic data for 6 patients. (b) A phantom endoscopic video frame (left) and the fused geometry (right) with color-coded deviation (in millemeters) from the ground truth CT.

registration error as the average Euclidean distance of all pairs of corresponding vertices after registration (Fig. 3a). Our method significantly reduced error and performed better than a spectral-graph-based method [10], which is another potential framework for matching partial surfaces without estimating the mean.

Phantom Data. To test our method on real-world data in a controlled environment, we 3D-printed a static phantom model (Fig. 3b) from one patient's CT data and then collected endoscopic video and high-resolution CT for the model. We produced SfMS reconstructions for 600 frames in the video, among which 20 reconstructions were uniformly selected for geometry fusion (using more than 20 surfaces for geometry fusion won't further increase accuracy, but will be computationally slower). They were first downsampled to ∼2500 vertices and then rigidly aligned to the CT space. Since the phantom is rigid, the registration plays the role of unifying inconsistent SfMS estimation. No outlier geometry trimming was performed in this test. We define a vertex's deviation as its distance to the nearest point in the CT surface. The average deviation of all vertices is **1.24mm** for the raw reconstructions and is **0.94mm** for the fused geometry, which shows that the registration can help filter out inaccurate SfMS geometry estimation. Fig 3b shows that the fused geometry resembles the ground truth CT surface except in the farther part, where less data was available in the video.

Patient Data. We produced endoscopograms for 8 video sequences (300 frames per sequence) extracted from 4 patient endoscopies. Outlier geometry trimming was used since lighting conditions were often poor. We computed the



(a) (b) (c) (d)

Fig. 4. OD plot on the point cloud of 20 surfaces (a) before registration; (b) after registration; (c) after outlier geometry trimming. (d) The final endoscopogram.

overlap distance (OD) defined in [14], which measures the average surface deviation between all pairs of overlapping regions. The average OD of the 8 cases is **1.6±0.13mm** before registration, **0.58±0.05mm** after registration, and **0.24±0.09mm** after outlier geometry trimming. Fig. 4 shows one of the cases.

5 Conclusion

We have described a pipeline for producing an endoscopogram from a video sequence. We proposed a novel groupwise surface registration algorithm and an outlier-geometry trimming algorithm. We have demonstrated via synthetic and phantom tests that the N-body scenario is robust for registering partially-overlapping surfaces with missing data. Finally, we produced endoscopograms for real patient endsocopic videos. A current limitation is that the video sequence is at most 3-4 seconds long for robust SfM estimation. Future work involves fusing multiple endoscopograms from different video sequences.

Acknowledgements This work was supported by XXXXXXX.

References

1. Armin, M., Visser, H., Chetty, G., Dumas, C., Conlan, D., Grimpen, F., Salvado, O.: Visibility map: A new method in evaluation quality of optical colonoscopy. In: MICCAI. (2015) 396–404
2. Hong, D., Tavanapong, W., Wong, J., Oh, J., de Groen, P.C.: 3D reconstruction of virtual colon structures from colonoscopy images. Computerized Medical Imaging and Graphics **38**(1) (2014) 22–23
3. Wu, C., Narasimhan, S.G., Jaramaz, B.: A multi-image shape-from-shading framework for near-lighting perspective endoscopes. International Journal of Computer Vision **86**(2) (2010) 211–228
4. XXXXX, X.: Xxxxx. In: XXXXX. (??) ??–??
5. Durrleman, S., Prastawa, M., Korenberg, J., Joshi, S., Trouv, A., Gerig, G.: Topology preserving atlas construction from shape data without correspondence using sparse parameters. In: MICCAI. (2012) 223–230
6. Durrleman, S., X. Pennec, A. Trouv, N.A.: Statistical models of sets of curves and surfaces based on currents. Med. Image Anal. **13**(5) (2009) 793–808
7. Balci, S.K., Golland, P., Shenton, M., Wells, W.M.: Free-form b-spline deformation model for groupwise registration. In: MICCAI. (2007) 2330
8. Arslan, S., Parisot, S., Rueckert, D.: Joint spectral decomposition for the parcellation of the human cerebral cortex using resting-state fmri. In: IPMI. (2015) 85–97
9. Zhao, Q., Price, J.T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: Surface registration in the presence of topology changes and missing patches. In: Medical Image Understanding and Analysis. (2015) 8–13
10. Zhao, Q., Pizer, S., Niethammer, M., Rosenman, J.: Geometric-feature-based spectral graph matching in pharyngeal surface registration. In: MICCAI. (2014) 259–266
11. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)

12. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: SIGGRAPH. (1996) 303–312
13. Green, P.: Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society **46**(2) (1984) 149–192
14. Huber, D.F., Hebert, M.: Fully automatic registration of multiple 3d data sets. Image and Vision Computing **21**(7) (2003) 637–650