2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Depth Estimation for Oral Cavity by Shape from Shading with Endoscope

1st Xi Wu
School of Aerospace Engineering
Tsinghua University
Beijing, China
xiwu21@mails.tsinghua.edu.cn

2nd Gangtie Zheng
School of Aerospace Engineering
Tsinghua University
Beijing, China
gtzheng@mail.tsinghua.edu.cn

***Abstract*—Abstract—Tracheal intubation for patients with respiratory infectious diseases requires doctors to wear a full set of protective clothing, which takes a certain time. How to protect doctors from infection when facing an emergency operation has become an important issue. The intubation robot may solve this contradiction. To provide visual information for real-time path planning for robotic intubation, this study recovers depth information about the oral environment using the low-cost and widely used endoscopic. Since the oral cavity is small and has less texture, the Shape from Shading (SFS) method may be a good choice for oral depth estimation. This paper proposes the "oral elbow" hypothesis, filters outliers caused by saliva, calculates the 3-D contour map, and highlights the contour map features from different views. Oral images are obtained from a healthy person and a silicon dummy. This work expands the application scenarios of depth estimation to the oral environment; provides depth information for the visual navigation of the intubation surgical robot.**

***Index Terms*—Tracheal Intubation, Oral Cavity, Monocular Depth Estimation, Shape from Shading**

I. Introduction

When operating endotracheal intubation for patients with infectious respiratory diseases, anesthesiologists always face a high risk of infection. They need to wear protective clothing and masks to prevent infection via droplets suspended in the air and physical contact with the patient [1]. On one hand, the wearing usually takes more than ten minutes, which may delay the first aid treatment for patients. On the other hand, improper protection increases the infection risk of doctors. If reliable mechanical intubation is applied instead of manual intubation, the efficiency of the first aid treatment and the safety of the doctor are both guaranteed.

The rapid development of Computer-Assisted Surgery (CAS) in recent years has promoted the development of machine intubation. After Hemmerling et al. (2012) designed the first robotic tracheal intubation system in 2012 [2], similar designs of robotic intubation systems have emerged [3], [4]. Visual studies on the oral cavity mainly focused on teeth that are easy to extract features. However, few studies involve depth estimation of the oral passage due to the intrinsic limitation of oral scenes. Without texture or regular shape on the oral mucosa, it isn't easy to extract features [5]. The non-rigid characteristics and the deformation caused by physiological activities bring great difficulties to image registration [5]. And the narrow oral cavity is hard to accommodate devices such as RGB-D cameras which can directly obtain depth information. These above factors lead to limitations in depth estimation for oral scenes relative to other scenes.

Estimating depth from a 2-D image is key for 3-D reconstruction [6], having promising application value in the medical imaging field. The traditional techniques are summarized as shapes from X, for example,

- Shape from stereo method [7] simulates the human eyes observing from two viewpoints and calculates the disparity map to get the depth map. But the equipment size limits its application in the narrow oral cavity.
- Shape from Motion (SFM) method [8] estimates the depth using a 2-D image sequence, but it is not suitable for a non-rigid environment such as the oral scene.
- Shape from texture method [9] uses the perspective and texture information but is still not suitable for the texture-less and irregularly shaped scene of the oral cavity.
- Shape from Shading (SFS) method [10] uses the optical information of the image. The method may be chosen based on the simple thought that the further the object, the darker the corresponding image region.

Besides, there are many methods applied to depth estimation. For example, deep neural networks can estimate dense depth maps from a single image in an end-to-end manner [11]. However, the neural network method requires large amounts of data and faces generalization difficulty.

Based on the above analysis of the oral cavity characteristics and depth estimation methods, we choose to obtain the monocular 2-D oral images through the tiny, low cost and widely used medical endoscope and then recover the oral surface depth based on the SFS method. In this study, a novel hypothesis is proposed. The human oral and larynx passages are approximately regarded as an elbow with an elliptical cross-section. The hypothesis is based on the observation that when the light illuminates the section of the tube, the place with a larger incident angle and further distance turns out darker. The SFS method recovers the surface shape through image brightness and light incident angle, suitable for application to the homogeneous and texture-less oral mucosa in the oral cavity. This study samples oral and larynx

979-8-3503-2365-8/23/$31.00 ©2023 IEEE

images from a healthy person and a silicon dummy, adopting the SFS method and recovering the relative depth shown in a 3-D contour map. We find that the top view of the 3-D contour map well presents the oral contour, and the side view intuitively shows the bending direction of the "oral elbow" which indicates the next moving direction for the endoscope. Therefore, depth information may contribute to the route plan of robotic tracheal intubation. Under the assumption of "oral elbow", it is possible to distinguish the envelope of the edge of the oral cavity based on the contour map. Overall, this study expands the depth estimation application scenarios to the oral mucosa which has less texture and is difficult to extract features; and hopes to provide depth information for the visual navigation of the tracheal intubation robot.

II. Related Work

Shape from Shading (SFS) recovers shape from the shading variation in the image [12]. Shading can be quantified with image grayscale value, which in the SFS method depends on the angle between the light source direction and the surface normal. Since SFS method was first developed by Horn [10] in 1970, many different solutions [13], [14], [15], [16], [17] have emerged to solve the nonlinear first-order partial differential equation that recovers depth from image brightness. The SFS problem is known to be ill-posed [18], usually illustrated with "the crater illusion" [19] and "Bas-relief Ambiguity" [18]. It could relieve the problem if assuming that all parameters of the light source, surface reflectivity, and camera are known [20]. Some researchers apply SFS in medical images, like in colonoscopy images [21], stomach endoscopic images [20], et al.

Oral depth estimation is limited due to its narrow space and less-textured surface. Some work involves SFS method for human internal surface similar to oral mucosa [20], [21], [22], but only presents qualitative illustration. Qiu et al (2018) extend monocular depth estimation in the oral cavity and introduce laser light markers to generate more features [5]. However, the detected area is limited to view outside the mouth, far from deep enough for tracheal intubation.

III. Methods

For the less-texture oral surface, this paper recovers depth information based on the intensity of 2-D grayscale images, adopting the SFS method. The SFS method is based on the Lambertian hypothesis that describes an ideal diffuse reflection surface. Our work sets the following three approximations for the oral environment:

- Since most of the oral cavity is composed of the oral mucosa, it's assumed to be a Lambertian surface with approximately constant reflectivity.
- Consider the oral and larynx passages as an elbow with an elliptical cross-section which is the top view. This approximation is inspired by the light and dark variations on the tube wall when the light illuminates. It mainly expresses the rough shape and reflective properties of oral and larynx passages.
- Since the front end of the endoscope is roughly perpendicular to the "elbow" cross-section during the intubation process, it can be considered that the light source is approximately perpendicular to the incident.

The SFS method we adopted is based on Lambert's cosine theorem, that is, for Lambertian surface, the intensity of the reflected light on the object surface is proportional to the cosine of the light incident angle. The Lambert cosine law is expressed as:

$$E = I\rho\cos\varphi \tag{1}$$

where E is the image brightness, I is the light source intensity, and φ is the surface reflectivity. When the light source is vertically incident, the incident angle of a point is equal to the surface slant φ at the same place. Therefore, the surface slant φ can be recovered through the reflected intensity which is represented by the image brightness. Here, we assume that the oral surface conforms to the Lambertian hypothesis with constant reflectivity.

Before operating the SFS method, we first convert the RGB image to the grayscale image first and then calculate the intensity gradient field:

$$\begin{aligned} E_x &= \partial E/\partial x = E(x+1,y) - E(x,y) \\ E_y &= \partial E/\partial y = E(x,y+1) - E(x,y) \end{aligned} \tag{2}$$

The recovered shape is expressed as $z(x,y)$, and the recovering process involves the surface gradient p, q, surface slant φ, and tilt θ.

According to Lambert's Cosine Law, we get the maximum intensity when $\cos\varphi = 1$. And the tilt θ is calculated by the intensity gradients in the x and y directions. For any point,

$$\begin{aligned} \cos\varphi &= E/E_{max} \\ \tan\theta &= E_y/E_x \end{aligned} \tag{3}$$

The slant and tilt are related to the direction with the largest gradient of the point on the surface.

$$\begin{aligned} \delta z &= \delta l \times \sin\varphi \\ \delta x &= \delta l \times \cos\varphi\cos\theta \\ \delta y &= \delta l \times \cos\varphi\sin\theta \end{aligned} \tag{4}$$

where l is the magnitude of the largest gradient.

Therefore, we get the surface gradient:

$$\begin{aligned} p &= \frac{\sin\varphi}{\cos\varphi\cos\theta} = \frac{E}{\sqrt{E_{max}^2 - E^2}} \times \frac{\sqrt{L_x^2 + L_y^2}}{L_x} \\ q &= \frac{\sin\varphi}{\cos\varphi\sin\theta} = \frac{E}{\sqrt{E_{max}^2 - E^2}} \times \frac{\sqrt{L_x^2 + L_y^2}}{L_y} \end{aligned} \tag{5}$$

Now we achieve the conversion from the intensity gradient field to the surface gradient field $G(x,y) = p(x,y)\cdot\boldsymbol{i} + q(x,y)\cdot\boldsymbol{j}$. Next, we operate depth estimation from the gradient field. Given the surface depth $z(x,y)$, the surface gradient is represented as $\nabla z(x,y)$.

Our goal is to find a $z(x, y)$ that minimizes the least square error between the calculated gradient p, q, and the theoretical gradient $\partial z/\partial x, \partial z/\partial y$. The objective function is below:

$$\iint F(\nabla z, G)dxdy \tag{6}$$

$$F(\nabla z, G) = \|\nabla z - G\|^2 = (\partial z/\partial x - p)^2 + (\partial z/\partial y - q)^2 \tag{7}$$

Its Euler Lagrange equation is:

$$2(\partial^2 z/\partial x^2 - \partial p/\partial x) + 2(\partial^2 z/\partial y^2 - \partial q/\partial y) = 0 \tag{8}$$

It has the form of a Poisson equation after simplification:

$$\nabla^2 z = \nabla G \tag{9}$$

For 2-D images, pixel points are discrete. So the above partial differential equation in continuous space can be rewritten as the partial differential equation in discrete space. And then we solve the equations by Discrete Fourier Transform (DFT).

IV. Results and Discussion

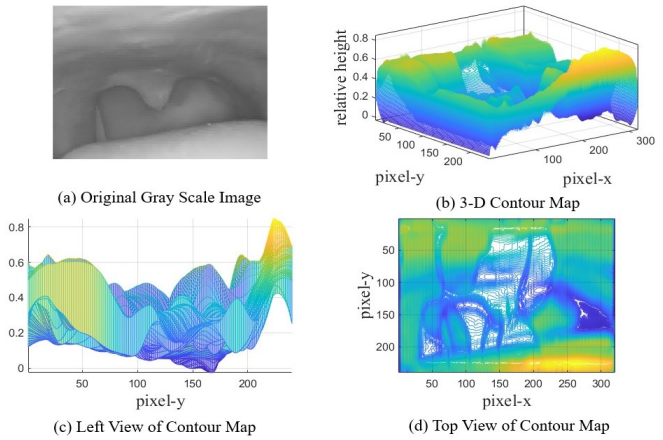
For a single image, the depth estimation process is:

- Converting the RGB image to the grayscale image. For a grayscale image, skip this step.
- Removing factors that clearly do not meet the Lambertian assumption, like teeth and saliva.
- Operating SFS algorithm and presenting depth information in the 3-D contour map.
- Illustrating the left view and the top view of the 3-D contour map to better understand the depth information.
- Integrating the original grayscale image, 3-D contour map, the left view, and the top view for this imaging case.

A. Silicon Dummy

This research operates endoscope intubation for a silicon dummy from outside the mouth to the front of the trachea. In clinical endotracheal intubation surgery, doctors first insert the laryngoscope to the front of the trachea, where our endoscope stops. Our endoscope captures the RGB image series during its intubation. Then we choose two images of representative places in the mouth and larynx. Operating the depth estimation process for the oral image and the larynx image, we get Fig. 1 and Fig. 2 separately. Here the x-y plane is consistent with the x-y plane of the pixel coordinate system. And the depth z is expressed as the height from the oral surface to the x-y plane. Since the monocular depth estimation can only estimate the relative depth but not the absolute depth, the depth z is normalized and falls within [0,1].

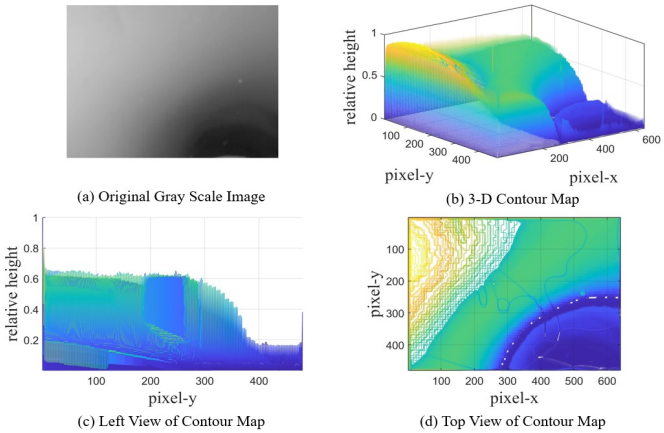
The chosen oral image excludes teeth, to avoid the effect of teeth with apparently different reflectivity from oral mucosa. We get a discrete relative depth map for the oral image in Fig. 1. Creatively, we acquire valuable depth information inspiring for endoscope navigation via the left view and the top view of the 3-D contour map. The left view is perpendicular



(a) Original Gray Scale Image (b) 3-D Contour Map (c) Left View of Contour Map (d) Top View of Contour Map

Fig. 1. The combined figure of the oral image for the silicon dummy. (a) the original grayscale image; (b)the 3-D contour map; (c) the left view of the 3-D contour map; (d) the top view of the 3-D contour map. The x-y plane is consistent with the x-y plane of the pixel coordinate system. The depth z is the relative distance from the oral surface to the imaging plane. The blue color means small z and deep distance. The image contains no teeth.

to the oral section and shows the moving direction of the endoscope. And the field of the top view illustrates the endoscope position on the oral section plane. Accordingly, the two movement directions of an endoscope for the next moment are available from the left and the top view.



(a) Original Gray Scale Image (b) 3-D Contour Map (c) Left View of Contour Map (d) Top View of Contour Map

Fig. 2. The combined figure of the larynx image for the silicon dummy. (a) the original grayscale image; (b) the 3-D contour map; (c) the left view of the 3-D contour map; (d) the top view of the 3-D contour map. The larynx part on the image is more in line with the elbow shape.

The chosen larynx image is more in line with the elbow shape. It's purely about the light and dark variations. Since the endoscope captures one side of the "larynx elbow", the 3-D depth map is like a slop. The left view clearly shows the depth descending to one side. And it's obvious to see the dark passage of the larynx on the top view.

Overall, the depth estimation process based on the SFS method proves to be effective for silicon dummy. More information for endoscopic movement navigation, such as the endoscope position in the oral section and its next step direction, is indicated from the left view and top view of the

3-D contour map. Next, we expand the depth estimation for the healthy person.

B. Healthy Person

Our research recruits a healthy volunteer and operates endoscope insertion slowly from outside the mouth to the base of the tongue, stopping before causing a retching reaction. The endoscope does not get deep into the larynx of the healthy person. So the larynx case is only analyzed on a silicon dummy. The endoscope captures the RGB image series, then we choose three representative places: outside the mouth, above the middle of the tongue, and above the base of the tongue. Operating the depth estimation process for the images of three places, we get Fig. 3, Fig. 4, and Fig. 5 separately.

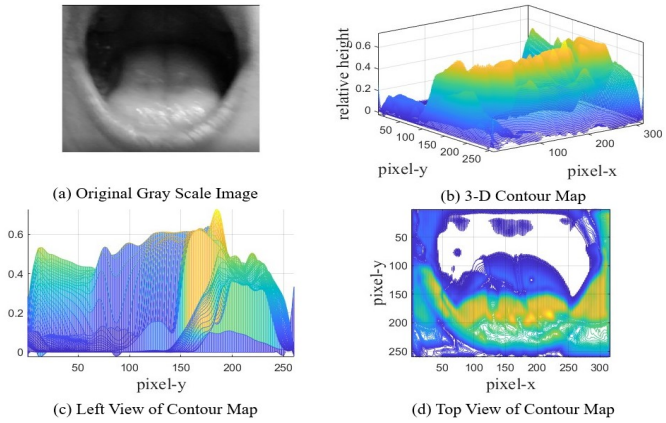


Fig. 3. The combined figure of the human oral image outside the mouth. (a) the original grayscale image; (b) the 3-D contour map; (c) the left view of the 3-D contour map; (d) the top view of the 3-D contour map. The x-y plane is consistent with the x-y plane of the pixel coordinate system. The depth z is the relative distance from the oral surface to the imaging plane. The blue color means small z and deep distance.

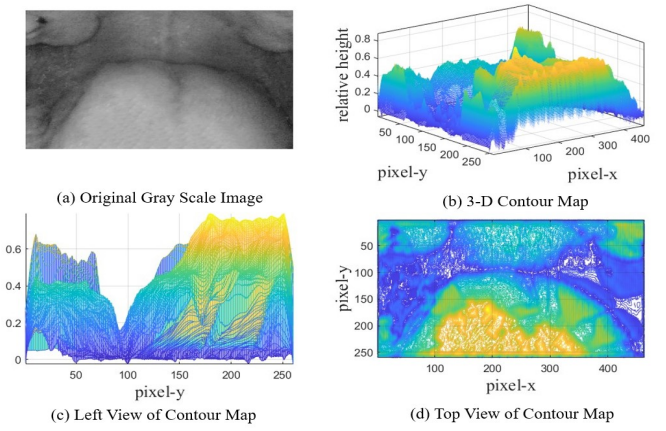


Fig. 4. The combined figure of the human oral image above the middle of the tongue. (a) the original grayscale image; (b)the 3-D contour map; (c) the left view of the 3-D contour map; (d) the top view of the 3-D contour map. The "U" shape of depth variation on the left view is consistent with the vertical section of the elbow. The grayscale image is already performed outlier detection and replacement before applying the SFS algorithm.

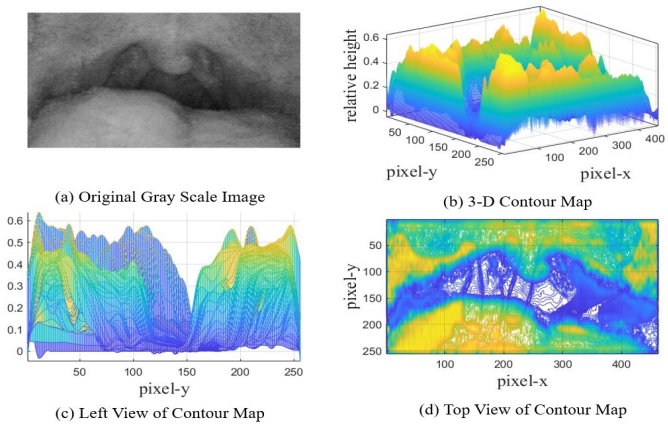


Fig. 5. The combined figure of the human oral image above the base of the tongue. (a) the original grayscale image; (b) the 3-D contour map; (c) the left view of the 3-D contour map; (d) the top view of the 3-D contour map. The "U" shape on the left view is a little different from that of Fig. 4 with the movement of the endoscope. The grayscale image is already performed outlier detection and replacement before applying the SFS algorithm.

When the endoscope is outside the mouth, there is almost no light illumination in the deep part of the mouth, and the pixel value in that area is almost zero. The dark area is processed as blank.

Inside the oral cavity, we see several discrete reflective points on the surface, and the brightness value is significantly higher than that of the surrounding area. That's the specular reflection caused by the small amount of saliva on the oral mucosa of a healthy person. To relieve the local reflection problem, we detect and replace the outliers before performing the SFS method. If an element differs from the area median by more than three times the absolute deviation of the median, we consider it an outlier and replace it with its nearest non-outlier.

The left view and top view of the 3-D contour map still highlight the 3-D information of a certain aspect that is valuable for endoscope navigation in future work. Outside the oral cavity, the contour in the top view indicates the lip contour, which may help to determine the insertion position. Inside the oral cavity, the top view highlights the shape of the oral cavity cross-section. Similar to the dummy case, we can adjust the endoscope position on the section plane until it stays in the center of the top view. The side view highlights the bend angle for the "oral elbow". Quantitatively the curve's descending slope is calculated from the side view. Accordingly, we can plan the next moving direction which is perpendicular to the section plane.

V. Conclusion and Prospect

In this study, we fully considered the characteristics of the oral environments and the oral cavity and larynx passage, proposing the "oral elbow" hypothesis, and recovering the oral surface depth by getting a gradient field from the intensity gradient field based on the SFS method. We get the 3-D contour map, and highlight the features of oral contour and bending direction separately with top and side

views. That features further provide the position on the oral section plane and the next moving direction of the tip of the endoscope for endoscope navigation. However, depth is coarse, relative values rather than precise, absolute values due to the limitations of the monocular depth estimation and the oral scene. That calls for future study with multiple methods and our work may be one of the inspirations. Our work shows the thoughts of the "oral elbow" hypothesis and the decomposition of moving direction to be indicated in the left view and the top view. Most importantly, our study is fully aware of the importance of image brightness for depth estimation in the oral environment and recovers the oral depth information.

For future applications in tracheal intubation robot systems, route design can utilize the curve descending slope of the left view of the 3-D contour map and the endoscope position from the oral section center according to the field of the top view. The slope remains consistent either relative depth or absolute depth. In this paper, we decompose endoscope motion into two directions of perpendicular to the oral section and on the oral section plane, which is separately indicated from the left view and the top view.

Expensive equipment like the RGB-D camera can directly generate the depth map. The equipment miniaturization and integration can promote its application in the oral scene, providing more information for depth estimation, which is crucial for the development of autonomous intubation robots in the future.

REFERENCES

[1] M. Zuo, Y. Huang, W. Ma, Z. Xue, J. Zhang, Y. Gong, and L. Che, "Expert recommendations for tracheal intubation in critically ill patients with noval coronavirus disease 2019," *Chinese Medical Sciences Journal*, vol. 35, no. 2, pp. 105–109, 2020.

[2] T. Hemmerling, R. Taddei, M. Wehbe, C. Zaouter, S. Cyr, and J. Morse, "First robotic tracheal intubations in humans using the kepler intubation system," *British journal of anaesthesia*, vol. 108, no. 6, pp. 1011–1016, 2012.

[3] X. Wang, Y. Tao, X. Tao, J. Chen, Y. Jin, Z. Shan, J. Tan, Q. Cao, and T. Pan, "An original design of remote robot-assisted intubation system," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[4] X. Cheng, G. Jiang, K. Lee, and Y. N. Laker, "Intubot: Design and prototyping of a robotic intubation device," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1482–1487.

[5] L. Qiu and H. Ren, "Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2197–2204.

[6] A. Bhoi, "Monocular depth estimation: A survey," *arXiv preprint arXiv:1901.09402*, 2019.

[7] S. C. De Vries, A. M. Kappers, and J. J. Koenderink, "Shape from stereo: A systematic approach using quadratic surfaces," *Perception & Psychophysics*, vol. 53, no. 1, pp. 71–80, 1993.

[8] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International journal of computer vision*, vol. 9, no. 2, pp. 137–154, 1992.

[9] J. Aloimonos, "Shape from texture," *Biological cybernetics*, vol. 58, no. 5, pp. 345–360, 1988.

[10] B. K. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," 1970.

[11] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612–1627, 2020.

[12] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[13] K. Ikeuchi and B. K. Horn, "Numerical shape from shading and occluding boundaries," *Artificial intelligence*, vol. 17, no. 1-3, pp. 141–184, 1981.

[14] E. Rouy and A. Tourin, "A viscosity solutions approach to shape-from-shading," *SIAM Journal on Numerical Analysis*, vol. 29, no. 3, pp. 867–884, 1992.

[15] J. Oliensis, "Shape from shading as a partially well-constrained problem," *CVGIP: Image Understanding*, vol. 54, no. 2, pp. 163–183, 1991.

[16] A. Pentland, "Shape information from shading: a theory about human perception," in *[1988 Proceedings] Second International Conference on Computer Vision*. IEEE, 1988, pp. 404–413.

[17] T. Ping-Sing and M. Shah, "Shape from shading using linear approximation," *Image and Vision computing*, vol. 12, no. 8, pp. 487–498, 1994.

[18] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," *International journal of computer vision*, vol. 35, no. 1, pp. 33–44, 1999.

[19] A. P. Pentland, "Local shading analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 170–187, 1984.

[20] E. Prados and O. Faugeras, "Shape from shading. handbook of mathematical models in computer vision 375–388," 2006.

[21] B. L. Craine, E. R. Craine, C. J. O'Toole, and Q. Ji, "Digital imaging colposcopy: corrected area measurements using shape-from-shading," *IEEE Transactions on medical imaging*, vol. 17, no. 6, pp. 1003–1010, 1998.

[22] M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Metric depth recovery from monocular images using shape-from-shading and specularities," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 25–28.