An Endoscopic 3D Scanner based on Structured Light

Christoph Schmalz[a,b,*], Frank Forster[b], Anton Schick[b], Elli Angelopoulou[a]

[a] *University of Erlangen-Nuremberg, Pattern Recognition Lab, Martensstrasse 3, 91058 Erlangen, Germany*
[b] *Siemens AG, CT T DE HW2, Otto-Hahn-Ring 6, 81739 Munich, Germany*

Abstract

We present a new endoscopic 3D scanning system based on Single-Shot Structured Light. The proposed design makes it possible to build an extremely small scanner. The sensor head contains a catadioptric camera and a pattern projection unit. The paper describes the working principle and calibration procedure of the sensor. The prototype sensor head has a diameter of only 3.6mm and a length of 14mm. It is mounted on a flexible shaft. The scanner is designed for tubular cavities and has a cylindrical working volume of about 30mm length and 30mm diameter. It acquires 3D video at 30 frames per second and typically generates approximately 5000 3D points per frame. By design, the resolution varies over the working volume, but is generally better than 200µm. A prototype scanner has been built and is evaluated in experiments with phantoms and biological samples. The recorded average error on a known test object was 92µm.

Keywords: Endoscopy, 3D Scanning, Catadioptric camera, Single-Shot Structured Light

1. Introduction

1.1. Motivation

Endoscopes are an important tool in medicine. They are a key component in minimally invasive surgery, but many procedures involving endoscopes can be challenging for the physician. This is due to the impaired depth perception, as most existing endoscopes acquire and display monocular images only. One way to overcome this limitation would be an endoscope capable of acquiring the precise three-dimensional geometry of its field of view. Such an 3D-endoscope would permit synthesizing wide baseline stereoscopic images for surgeons, providing them with an intuitive 3D visualization rather than with flat images. It also has the potential to assist in robotic navigation. Moreover, it would make it easier to perform absolute 3D measurements, such as the area and volume of a pathological structure. Finally, it might simplify solving advanced tasks such as coverage analysis (i.e. checking if 100% of a surface has been seen in the course of an inspection) or registration of endoscopic images with data generated in a preoperative CT or MR scan.

We present a novel flexible endoscope, based on Single Shot Structured Light, that can perform accurate 3D imaging at 30Hz while still having a fairly small diameter of 3.6mm. The main contributions of the paper are the calibration algorithm for the sensor and the experimental evaluation of the measurement results on phantom models and ex vivo tissue samples. We reconstruct the surface of our test objects from image sequences acquired during controlled sensor motion. For the future it is planned to recover the sensor motion automatically from a second camera, which is not yet functional in the current prototype. However, it is already possible to create high-quality cavity reconstructions with a mean error of below 100µm.

The paper is organized as follows: the next section discusses the state of the art regarding 3D endoscopy. Section 3 contains a short overview of the measurement principle behind Single Shot Structured Light. Section 4 describes the hardware setup, which consists of an axial configuration of a pattern projector and a catadioptric camera. Sections 5 and 6 discuss two important aspects of the algorithms used, namely calibration of the camera and the projector, and data processing. Experimental results with a phantom model and biological samples are presented in section 7. Finally, the conclusion and discussion are given in section 8.

2. 3D Endoscopy

Solutions for 3D endoscopy that use a standard endoscope and obtain 3D data with only computer vision algorithms and no additional hardware are very convenient. An ideal solution would convert a single color image into a 3D data set. A well-known technique for this is Shape-from-Shading (SfS). Okatani and Deguchi (1997) propose to use it for 3D endoscopy. However, without any of the assumptions conventionally used, SfS becomes a very complex task. In medical endoscopy there is neither a remote light source with approximately parallel rays of illumination, nor a camera that can be described via the orthographic projection model, nor a Lambertian surface.

*Corresponding Author: christoph.schmalz@informatik.uni-erlangen.de

Okatani and Deguchi overcome these issues to some extent by modelling the light source as an imaginary single point source at the projection center and by extracting equal distance contours from the image. However, their technique still has limitations such as inherent topological ambiguities and issues due to interreflections and non-uniform reflectance. Wu et al. (2010) present a more stable SfS reconstruction. They merge several shape data sets into a complete 3D model of the scene where the respective positions and orientations of the endoscope are obtained by an additional tracking system. However, they report a computation time of several minutes for a complete model of a bone.

Another monocular solution for generating 3D data is Shape-from-Motion (SfM, also known as Structure-from-Motion) (Thormahlen et al., 2002; O. et al., 2011; Wang et al., 2008; Zhou et al., 2010; Hu et al., 2010). It requires a sequence of images taken by a camera that moves relative to the scene (or vice versa). Features are tracked over at least two consecutive frames. Given enough features, the 3D position of the tracked points can be estimated up to a scale factor using projective geometry. The distribution and quantity of trackable features in the scene determines the density of the resulting point cloud. SfM implementations also often have difficulties providing live feedback. Typically, a whole image sequence has to be processed before 3D data can be computed. The lag described in the literature varies widely; Hu et al. (2010) report a processing time of several minutes while Grasa et al. (2009); O. et al. (2011) demonstrate an SfM system running at 25 Hz. Additional challenges are posed by non-rigid scenes, which may prevent a reconstruction or - in the worst case - cause artifacts. It is also important to mention that both SfS and SfM generate 3D data only up to scale; it cannot be used for absolute measurement, but it is suitable for stereo view synthesis. Hu et al. report an RMS reprojection error of their tracked features of around 1 pixel. This corresponds to a mean residual error of 1.68mm between their reconstruction and a ground truth surface; however it is unclear how the scale of the metric reconstruction was determined.

A natural way to overcome the lack of depth perception is to image the scene from two distinct viewpoints separated by a known baseline (interpupillary distance). Such a setup is typically realized using two imaging sensors with two distinct lenses (Durrani and Preminger, 1995). There are, however, also alternative set-ups such as a single lens behind two pupil openings combined with a lenticular array on a single sensor chip. Such a design permits a small endoscope diameter (Tabaee et al., 2009; D. et al., 2010), but results in a 'weak' 3D effect. With stereo algorithms the quality of the 3D data depends on the optical structure of the scene; featureless areas or viewpoint-dependent glares tend to cause problems.

A widely-used alternative to stereo vision is Structured Light where one of the stereo cameras is replaced by a projector or, more generally, a light source. Armbruster and Scheffler (1998) describe a rather large endoscope (targeted at industrial applications) based on the well-known phase-shifting approach for the illumination pattern (Creath, 1986). In Kolenovic et al. (2003), the authors present a conceptually similar miniaturized holographic interferometer that can acquire data at a rate of 5Hz. The capsule has a diameter of 10mm and three protruding arms to provide three different illumination directions. They employ temporal phase shifting to get rid of the disturbing zero order and the complex-conjugate image arising in digital holography. The reported quality is impressive, but endocopes using phase shifting tend to be unsuitable for moving objects. Hayashibe et al. (2006) propose a 3D endoscope based on triangulation with a scanning laser line. The necessary triangulation angle is created by using two endoscopes, one for illumination and one for observation, resulting in a rather large and impractical setup. In general, many Structured Light systems described in the literature have an overall diameter of the endoscope significantly greater than 10 mm and are consequently too large for many medical applications. An exception is Clancy et al. (2011), who present a fiber-optic addon for the instrument channel of a rigid endoscope. They project dots of different wavelengths onto the scene and try to identify the wavelength in the camera image. However, given the color filters in typical cameras, this is a difficult task and the resulting density of 3D points is very low.

Time-of-Flight (ToF) methods have also been considered for endoscopic imaging (Penne et al., 2009). The advantage of ToF is that it does not suffer from occlusion, unlike triangulation-based methods such as Stereo or Structured Light. At the same time, it is very challenging to build a small endoscope with an integrated ToF-sensor. Penne et al. (2009) did not miniaturize the hardware to the required level, but rather used a rigid endoscope with fiber optics for illumination and observation. The authors report an average error of 0.89mm for measurements of a plastic cube with a side length of 15mm at a standoff distance of 30mm. Surface texture and volume scattering in biological tissue (a significant effect for the infrared illumination typically used in ToF sensors) pose problems.

Detailed surveys of 3D reconstruction techniques for endoscopic applications can be found in Mountney et al. (2010) and Mirota et al. (2011).

3. Structured Light

Structured Light is a method for 3D reconstruction of surfaces based on triangulation. One or multiple illumination patterns are projected onto a scene and observed by a camera. The basic principle is illustrated in figure 1. The projected light pattern defines a set of planes (or other suitable surfaces) in space. The intersection of the camera ray CX with the corresponding light plane give the 3D coordinates of the point X. The task of finding the right plane for a given camera ray is known as the correspondence problem. However, the camera rays and the

light planes can only be calculated if both the camera and the projector are calibrated (see section 5).

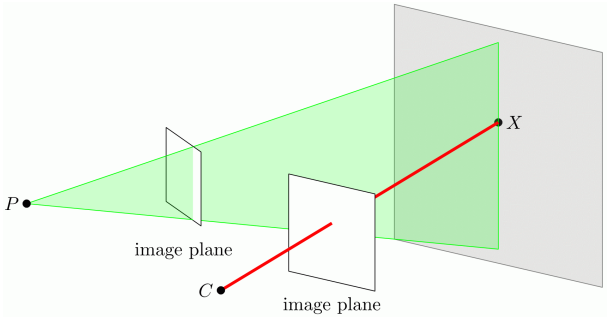

Figure 1: Active Triangulation principle. A light plane from projector P falls on the object and is observed by camera C. In a calibrated system the 3D coordinates of the object point X can be reconstructed.

There are numerous variants (Salvi et al., 2010) of Structured Light methods. Single Shot techniques need only one image of the scene to reconstruct 3D data. This has two main advantages. Firstly, it allows to measure moving scenes, which is critical for eventual applications on live patients. Secondly, it simplifies the miniaturization of the projection hardware that is necessary for endoscopic applications. Only a single static pattern has to be projected and thus no moving parts are required.

The proposed system uses a single-shot color ring pattern (see figures 2 and 5). This type of pattern can be robustly detected in the camera image and yields relatively many 3D points per frame. The sequence of the color rings that make up the pattern is based on pseudorandom arrays. Their defining property is that subarrays with a minimal length L occur at most once. Therefore, observing a subarray of sufficient length (a sequence of stripes) in the camera image allows one to deduce its index in the projected pattern. Thus the correspondence problem is solved. In the example pattern shown in figure 2, we interpret the color changes between subsequent stripes as array elements (also termed symbols) rather than the absolute colors. Single symbols like G-B+ are not unique, but longer sequences are guaranteed to occur only once. If such a characteristic sequence of color changes is detected in the camera image, the corresponding stripe indices in the projection pattern can be recovered and triangulation can be performed. Why are color changes used instead of the directly observed colors? The projected colors may be distorted by the object texture and are therefore hard to detect reliably in the camera image. The color changes between neighboring rings can be recognized more easily (Schmalz and Angelopoulou, 2010).

4. Structured Light Endoscope

The Structured Light Endoscope consists of a catadioptric camera and a slide projector. The design is illustrated in figure 3. Views of the working prototype system are

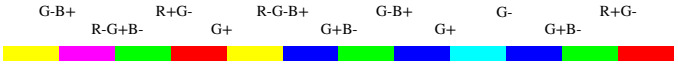

Figure 2: Example color stripe pattern. The letters symbolize the color channels. A plus sign denotes channel rising at the edge, a minus sign a falling channel. Constant channels are omitted.

shown in figures 4 and 5. The endoscope head has a diameter of 3.6mm and contains a complete Structured Light triangulation system in an unusual axially-aligned configuration. This results in a relatively long triangulation base while keeping the diameter small. For stability the hardware is housed inside a cylindrical glass tube. The light source is external and connected via a glass fiber in the flexible shaft. This means almost no heat needs to be dissipated from the head. Still, the pattern slide is exposed to very high light intensities, which causes bleaching in organic dyes. Therefore the pattern is realized as a multilayer interference filter on a glass substrate. The camera is catadioptric, that is it uses a mirror in addition to lenses (Swaminathan et al., 2006). In our design the mirror is spherical, which makes for a wide field of view at the cost of a strong fisheye distortion towards the rim of the mirror. To compensate the varying camera resolution, the outer rings of the pattern are wider, so that they appear approximately equally spaced in the camera image. At the moment the Structured Light pattern consists of 15 rings with distinct colors (figure 5).

The wide-angle catadioptric measurement camera mainly observes the space to the sides of the endoscope. It has a blind spot directly ahead as the camera chip sees only itself in that direction (see figure 6). Because of its small size of $1.2mm \times 1.2mm$, the camera chip has a resolution of only 400×400 pixels. Furthermore, a large area of the chip cannot be used for measurements due to the blind spot. In future hardware revisions this aspect will be improved.

A second camera with a regular narrow-angle lens observes the area in front of the endoscope and gives additional guidance for the operator. The necessary illumination is provided by two white micro LEDs located at the front of the scanner. It also allows tracking of feature points on the object surface for more robust registration of the Structured Light data. The optical design of the sensor is described in more detail in Schick et al. (2011). The measurement volume of the current prototype is a cylinder with diameter and length of approximately 30mm.

5. Calibration

The scanning system needs to be calibrated in order to generate metric 3D data. The calibration of the proposed system consists of two parts. The first is the camera calibration, which is needed to calculate the rays of view for each camera pixel. The second part is the projector calibration, which yields a set of cones corresponding to the edges between the projected color rings. Range data can

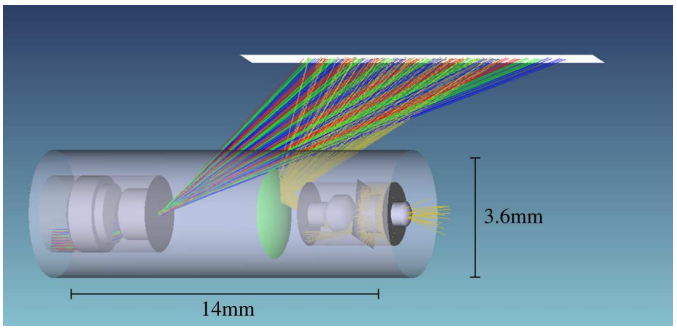

Figure 3: Schematic hardware setup. The light for the projector is supplied through a glass fiber (not shown). A ring-shaped slide projects colored cones to the side (only rays to one side are shown). They are observed by the camera via a curved mirror (shown in green). A second camera provides a front view (yellow rays to the right). In this diagram, the wall thickness of the glass housing is zero and so no refraction of the rays occurs.

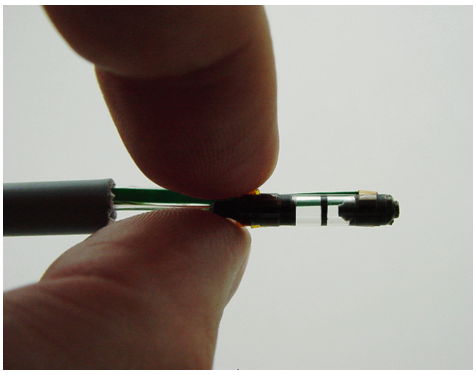

Figure 4: The physical realization of the prototype design. It has a diameter of 3.6mm and a length of 14mm.

then be computed by intersecting the rays of view with the appropriate cones.

We provide a short derivation of the most important formulas used for calibration and measurements with the endoscopic Structured Light 3D scanning system. A more complete treatment of these raytracing fundamentals can be found in Glassner (1989). All rays are expressed in camera coordinates. This coordinate system has the x and y axes in the image plane, the z axis along the optical axis and the origin in the optical center of the camera.

5.1. Camera Calibration

As the camera is a catadioptric system with a spherical mirror, it falls in the catagory of non-single-viewpoint cameras. Two approaches for calibration were investigated. One is a standard model of a pinhole camera with radial and tangential distortion (Zhang, 2000). This model has 8 intrinsic parameters and 6 extrinsic parameters per pose of the calibration target. It assumes a single viewpoint, so it is not a perfect model for our setup. However, we found that it can still be applied. This is because the relatively low image resolution of 400×400 pixels, in conjunction with the imaging geometry, results in an object space resolution of between 100µm and 200µm at a typical distance

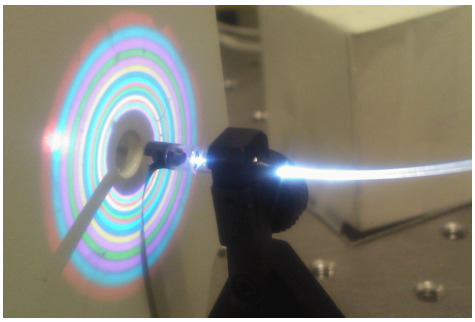

Figure 5: Prototype system in operation. The connection cable for the camera causes a small shadow. In a future version transparent wires will be used. The inner rings are less wide to compensate camera distortion.

of 10mm. In the peripheral area the resolution is even lower because of the large amount of distortion in the image. This mostly masks the errors caused by the varying viewpoint.

5.1.1. Pinhole Model

In the camera coordinate system, the ray of view for image coordinates (x_i, y_i) of a pinhole camera is the set of all points X with

$$X = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} (x_i - c_x)d_x \\ (y_i - c_y)d_y \\ f \end{pmatrix} = O_p + \lambda T_p \quad (1)$$

with the pixel pitch (d_x, d_y), the principal point (c_x, c_y), the focal length f and the free parameter $\lambda > 0$. If the camera exhibits image distortion, that has to be corrected first, for example using Zhang's model (Zhang, 2000).

5.1.2. Pinhole-Mirror-Tube Model

The second calibration method augments the pinhole camera with a reflective sphere and an encasing glass tube. The reflection and refraction processes are explicitly modeled. The augmented model has 8 additional intrinsic parameters (see Table 1).

A sphere with the center C and radius r is the set of points X with

$$(X - C_m)^2 = r_m^2 \quad (2)$$

Plugging eq. 1 into eq. 2 and simplifying, the intersection point of a ray with the sphere must fulfill

$$a\lambda^2 + b\lambda + c = 0 \quad (3)$$

with $a = T_p^2$, $b = 2(O_p - C_m) \cdot T_p$ and $c = (O_p - C_m)^2 - r_m^2$. Ignoring degenerate cases, the desired intersection point with the mirror is

$$O_m = O_p + \frac{-b - \sqrt{b^2 - 4ac}}{a} T_p \quad (4)$$

The surface normal in this point is

$$N_m = \frac{O_m - C_m}{\|O_m - C_m\|} \quad (5)$$

parameter type	number of parameters
camera pose	6 per pose
focal length	1
principal point	2
radial distortion	3
tangential distortion	2
mirror position	3
mirror radius	1
tube origin	2
tube rotation	2

Table 1: Camera calibration parameters. The first five types are used in the pinhole model with radial and tangential distortion. The augmented model with mirror and glass tube has additional parameters, listed in the last four rows. Because of spherical symmetry, the mirror rotation is irrelevant. Also, the rotation of the tube around its axis is degenerate. The origin of the tube is defined as the intersection point of its center line with the arbitrary plane z=0. Thus, it needs only two parameters. The inner and outer tube radius, as well as its index of refraction, are known and therefore not part of the optimization at all.

And the reflected ray

$$X = O_m + \lambda \left(T - 2N_m \left(N_m \cdot T\right)\right) = O_m + \lambda T_m \quad (6)$$

The glass housing is modeled as a pair of co-axial cylinders. A cylinder with the point C_c on its centerline, the axis direction A_c and the radius r_c is the set of points X with

$$\left((C_c - X) \times A_c\right)^2 = r_c^2 \quad (7)$$

Plugging eq. 1 into eq. 7 and simplifying, we obtain another quadratic equation

$$a\lambda^2 + b\lambda + c = 0 \quad (8)$$

with $a = \left(T_m \times A_c\right)^2$, $b = 2\left(\left(O_m - C_c\right) \times A_c\right) \cdot \left(T_m \times A_c\right)$ and $c = \left(\left(O_m - C_c\right) \times A_c\right)^2 - r_c^2 A_c^2$. Again ignoring degenerate cases, the intersection point of the ray with the cylinder is

$$O_r = O_m + \frac{-b + \sqrt{b^2 - 4ac}}{a} T_m \quad (9)$$

The normal N_r in the intersection point is

$$N_r = \frac{O_r - C_r}{\|O_r - C_r\|} \quad (10)$$

where C_r is the closest point to O_r on the cylinder center line, $C_r = C_c + \frac{(O_r - C_c)A_c}{A_c^2} A_c$.

Applying Snell's law with the refraction indices n_1 and n_2 we get the refracted ray direction T_r from the incident ray direction T_m as

$$T_r = \frac{n_1}{n_2} T_m - \left(\frac{n_1}{n_2} cos\gamma_i + \sqrt{1 - sin^2\gamma_t}\right) N_r \quad (11)$$

where $sin^2\gamma_t = \frac{n_1}{n_2}\left(1 - cos^2\gamma_i\right)$ and γ_i is the angle of the incident ray with the surface normal. A second refraction is computed with the outer cylinder of the glass tube to obtain the final ray of view in the augmented camera model.

The reverse problem of finding the image coordinates for a given point P in space, taking into account refraction and reflection, is more complex. It can be solved by optimizing the image coordinates with respect to the object space error.

5.1.3. Results

Both models are initialized with the results of the standard calibration algorithm by Zhang for the poses, focal length and distortion parameters. For the initialization of the additional parameters in the pinhole+mirror+tube model, we use the default design values of the mirror and the tube. The parameters of both models are optimized with the non-linear Levenberg-Marquardt algorithm (More, 1978). The quantity being optimized is the object space error, which can be defined as follows.

The distance d of a point P to a ray is simply

$$d = \frac{\|(O - P) \times T\|}{\|T\|} \quad (12)$$

with the cross product $\times$.

Each calibration point has known 3D coordinates in the world coordinate system and known image coordinates in the camera, which can be used to calculate a ray of view with the equations given above. For the pinhole model the 'plain' rays are used, while for the pinhole+mirror+tube model the rays are additionally reflected and refracted. The sum of the ray-point distances over all calibration points k in all poses l is the object space error e_{cam}.

$$e_{cam} = \sum_l \sum_k d(k, l) \quad (13)$$

Note that many other camera calibration algorithms minimize the distorted image plane error. This is not a good metric for our highly distorted images. The border of the image is "compressed", and points in the center have a relatively higher weight in the optimization. Therefore we evaluated the performance of both the undistorted image plane error and the object space error as error metrics. The object space error gave slightly better results. Since it is also faster to compute and its optimization is more stable, it is the metric that is currently being used.

We also tested two different ways of generating the calibration points. One was a classic dot grid target from Edmund Optics. The dots were localized with ellipse fits to their contours. Unfortunately, this is error-prone, especially in the peripheral area where the image distortion is high. The alternative method used an active target in the form of the display of a Fuijitsu UH900 mini-notebook. The calibration marks on this target were generated with the algorithm outlined in Schmalz et al. (2011). This

method displays a series of coded patterns on the display to uniquely identify each pixel. Virtual calibration marks can then be defined with subpixel precision. With help of the known pixel size (only 94µm in the case of the UH900) pixel coordinates can be converted into metric coordinates.

The results of the camera calibration can be seen in Table 2. The active target approach gives better results than the dot grid target. Also, the extended model with mirror and tube gives better results than the simple pinhole model. Of course, using more parameters naturally reduces the residual error. Therefore we first evaluated the error of the pinhole-mirror-tube model with default design values for the 8 additional parameters (optimizing only the parameters already used in the pure pinhole model). This resulted in a marked improvement in the RMS errors and proved that the extended model is sound. In a third run all parameters were optimized, yielding a further improvement of the residual errors. Unfortunately, there are no ground truth calibration parameters to compare our results against, but the optimized parameters for the mirror and the glass tube were within the manufacturing tolerances of their design values.

For both the classic and the active calibration six different camera poses were used. We took care to use a comparable set of poses for both calibrations. The active target yielded 1504 marks, the dot grid 811 marks. Compared to the spatial camera resolution of between 100µm and 200µm, the absolute values of the resulting object space errors are small.

RMS error [mm]	classic	active
pinhole	0.0864	0.0674
pmt default	0.0802	0.0546
pmt optimized	0.0707	0.0501

Table 2: Endoscopic camera calibration results after optimization of the object space error. The active target gives better results than the classic target. The full model with pinhole camera, mirror and glass tube (abbreviated pmt) is better than the simple pinhole model, even when the additional parameters are set to their default design values. Optimizing the full parameter set results in a further improvement.

5.2. Projector Calibration

For the projector calibration we model the projected rings as light cones. A planar dot grid calibration target is acquired in several poses while the cones are being projected (figure 6). In each pose, the surface of the target defines a plane in space. This plane can be calculated from the locations of the dots as observed by a calibrated camera. The camera simultaneously observes the color rings, which are the intersections of the target surface plane with the set of projected light cones. Thus each pose yields a set of points on the surface of each projected cone. The cones to which the points belong are identified using the algorithm described in section 6.

An acute cone with the axis A, the vertex V and the angle $\theta < \frac{\pi}{2}$ is the set of all points X with

$$A \cdot \left(\frac{X - V}{\|X - V\|} \right) = cos\theta \tag{14}$$

The distance of a point P to the cone is

$$d = \|P - V\| \cdot sin\left(min(\delta, \frac{\pi}{2})\right)$$

with $\delta = acos\left(\frac{P-V}{\|P-V\|} \cdot A\right) - \theta$. Eq. 14 can also be written as

$$(X - V)^T \mathbf{M} (X - V) = 0 \tag{15}$$

with

$$\mathbf{M} = AA^T - \mathbf{1} \cdot cos^2\theta \tag{16}$$

and

$$A \cdot (X - V) > 0 \tag{17}$$

Finally, the parameters of the cones are fitted, again using the Levenberg-Marquardt algorithm. The error metric in this case is the distance between the calibration points and the intersection points of the corresponding camera rays with the cones (also see figure 8).

What is the intersection point of a ray and a cone? Plugging eq. 1 into eq. 15 and simplifying we obtain a quadratic equation

$$a\lambda^2 + b\lambda + c = 0 \tag{18}$$

where $a = T^T\mathbf{M}T$, $b = T^T\mathbf{M}(O - V)$ and $c = (O - V)^T\mathbf{M}(O - V)$. Excluding all degenerate cases, the sought after intersection point is

$$X_c = O + \frac{-b - \sqrt{b^2 - ac}}{a}T \tag{19}$$

To determine the parameters of the light cones in the projector calibration, the objective function

$$e_{proj} = \sum_m \sum_l \sum_k \|P(k, l, m) - X_c(k, l, m)\| \tag{20}$$

is minimized. Here m iterates over all cones, l over all target poses and k over all calibration points P visible in the current pose.

An alternative metric to optimize would be the sum of the orthogonal distances between the calibration points and the cones, without calculating the intersection points with the rays of view. However, eq. 20 gave better results, as it mirrors the way depth data is computed in the calibrated sensor (which is also based on eq. 19).

For reasons of numerical stability, we do not optimize each cone separately but enforce a common vertex and axis for all the cones. However, the refraction of the light rays in the glass tube causes a shift in the apparent z-position

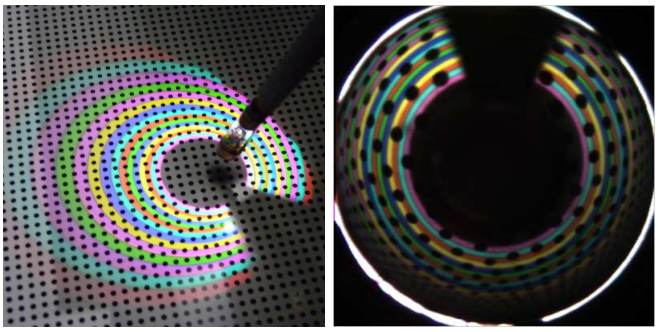

Figure 6: Calibration target for the projector calibration. Left: Outside view. Right: Camera view. The dots are used to determine the pose of the target relative to the calibrated camera. The projected rings yield the calibration points for the light cones. In the right image, the black area at the top is caused by the camera cable. The black area in the center is caused by the camera chip looking at itself in the mirror.

of the vertex, which depends on the opening angle of the cone. This z-offset is determined separately for each cone in a second optimization step.

An example input image is shown in figure 6. The number of projector calibration points recovered from six poses of the target is illustrated in figure 7. The outermost and innermost cones do not have enough points for a reliable calibration. The calibration points for cone 9 and the optimization result can be seen in figure 8. The different cones are nested with a varying opening angle, sharing the same vertex (except for the small z-offset due to refraction). The RMS residuals of the cone fittings for the different camera calibrations are shown in figure 9. The residuals range from 100µm to 300µm. Again, this is mainly due to the limited resolution of the camera, especially in the peripheral area. The quality of the projector calibration is approximately equal for all four camera calibrations. This suggests that the main error source is the localization of the projected rings in the camera image. Futhermore, since a dot grid target is mandatory for the projector calibration (the projected rings cannot be seen on the glossy surface of the digital display), the advantage of the active target in the camera calibration (see Table 2) is neutralized.

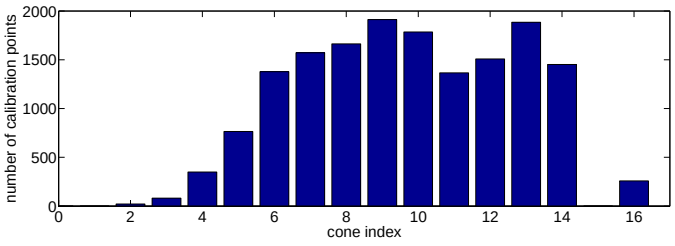

Figure 7: Number of calibration points per cone. While the projection slide has only 15 rings, two additional rings can be defined by counting the black area before the first and after the last proper ring. However, because of image quality limitations, the outermost and innermost rings do not yield enough points for a stable calibration.

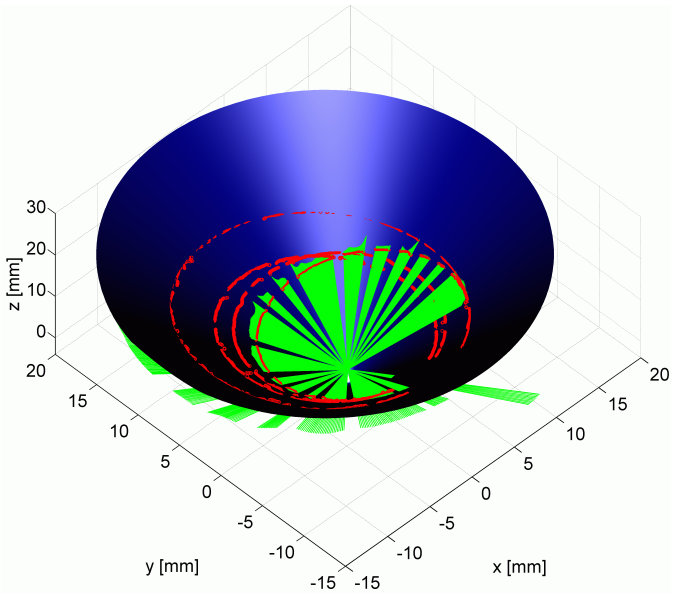

Figure 8: Light cone calibration example for a single cone. The blue surface model is fitted to the red calibration points. Each red circle corresponds to one view of the calibration target. The green rays indicate the errors for one set of calibration points. We minimize the distance along the ray of view, not the orthogonal distance of the calibration points to the cone.

6. Structured Light Decoding

The images seen by the endoscopic camera typically have a limited quality with low contrast and a high degree of noise. We therefore applied the robust decoding algorithm outlined in Schmalz and Angelopoulou (2010). This method is able to cope with colored and textured objects as well as with challenging image quality. We give a short summary of the method here. First, a watershed transform of the input image is performed. From the superpixels in the resulting oversegmented image a region adjacency graph is built. Each superpixel becomes a vertex in the graph. The color assigned to the vertex is the median of the original image pixels belonging to that superpixel. The edges of the region adjacency graph describe the color changes between neighboring vertices. They are scored according to how well they fit possible color changes in the projected pattern. To find correspondences between the projected pattern and the camera image, unique sequences of edges representing the color changes in the projected pattern have to be found in the region adjacency graph. If a matching sequence is found, the correspondence information is propagated to the neighboring vertices in a best-first-search. Once all possible regions have been mapped to projected stripes, the stripe edge locations in the original image are localized to subpixel precision. The triangulation between the projected light planes (or cones in our case) and the viewing rays from the camera can then be performed to obtain 3D data. The advantages of this graph-based decoding method are the robust color assignments of the superpixels, the absence of fixed thresholds

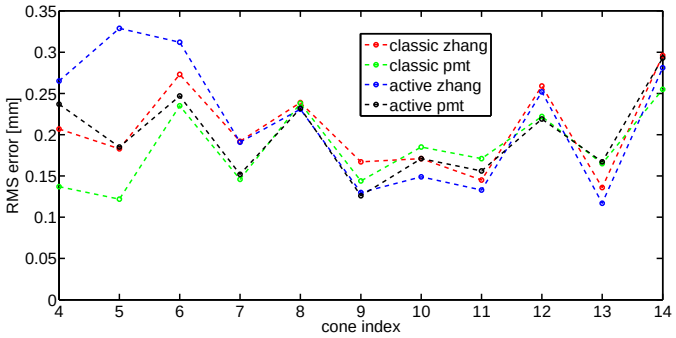

Figure 9: Cone parameter optimization results with four different camera calibrations (see table 2). 'Zhang' refers to to the simple pinhole model with distortion. The acronym pmt denotes the 'pinhole-mirror-tube' model. Cones 0 to 3 could not be fitted because of a lack of calibration points. The largest differences are found for the outermost cones (4 to 6) as the image distortion is higher towards the image border.

for color changes and the ability to sidestep any disruptions of the pattern in the camera image by finding alternative paths in the region adjacency graph. Furthermore, the implementation is fast and can be run in real-time on current hardware.

An example input image with the projected ring pattern is shown in figure 10. The associated decoding result can be seen in figure 11. The final depth data is shown in figure 12.

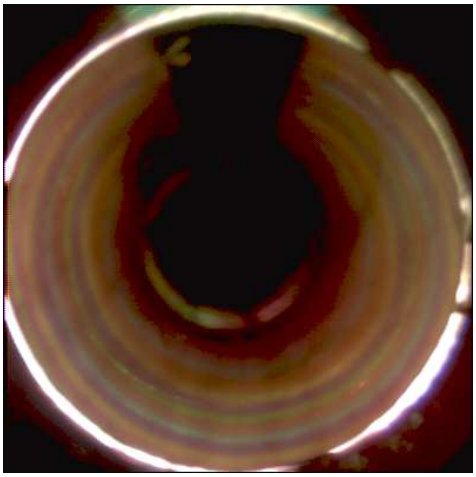

Figure 10: Example input image, gamma adjusted for better visibility. The dark spot in the center is the mirror image of the camera chip. The dark area at the top is the shadow of the connection cable. The contrast of the color rings is relatively low. The bright white ring is an artifact caused by stray light and should be eliminated in future hardware revisions.

With the current camera resolution and projector geometry, a single input image yields about 5000 data points. However, the user is typically interested in the reconstruction of a complete 3D model of the scene. Therefore, the endoscope has to be moved through the cavity. The individual point clouds recovered from successive frames have to be registered to each other and merged. The overlap between successive images is very large if the sensor is moving slowly compared to the frame rate of 30Hz. Registration algorithms like Iterative Closest Point (LM-ICP) (Fitzgibbon, 2003) could be used, but there may be degenerate cases, like constant-diameter cylindrical cavities, where this algorithm can fail. Therefore, we propose to guide the registration process by motion estimation with help of the second camera. The main measurement camera cannot be used for this purpose because the projected pattern moves with the sensor head and masks the underlying scene motion. The auxiliary front camera does not see the pattern and feature tracking or optical flow can be used to derive an initial estimate of the camera translation and rotation between two frames (Raudies and Neumann, 2009).

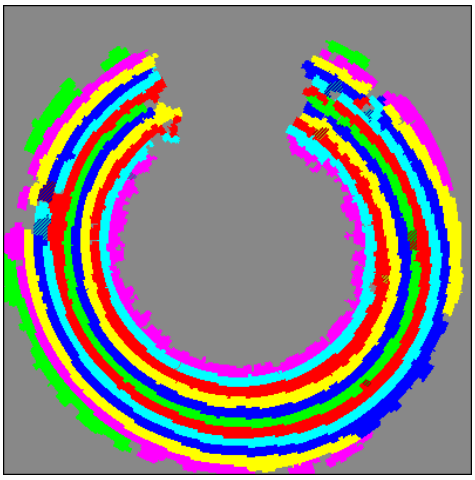

Figure 11: Decoded result for the input image from figure 10 with recovered ideal projected ring colors. Hatched areas could not be identified uniquely. In the gray areas no color stripe information could be recovered or was available.

7. Evaluation

The prototype scanner was evaluated in four distinct experimental setups. In a first test we measured a simple planar test object in various poses. In another experiment, an artificial cavity in a block of plastic was used. For this experiment ground truth CAD data is available for comparison. In a third experiment we used a colon phantom. Finally, we used the windpipe of a lamb to check the performance on biological tissue.

Unfortunately, because of hardware limitations the front camera currently cannot be used. Therefore, the scanner was moved in a controlled fashion using a manual x-y translation stage. The registration of the individual scans was then performed with the help of the known fixed offsets between the datasets. As the offsets are only used for initialization of the registration, they do not have to be perfectly correct. In fact, it must be expected that the off-

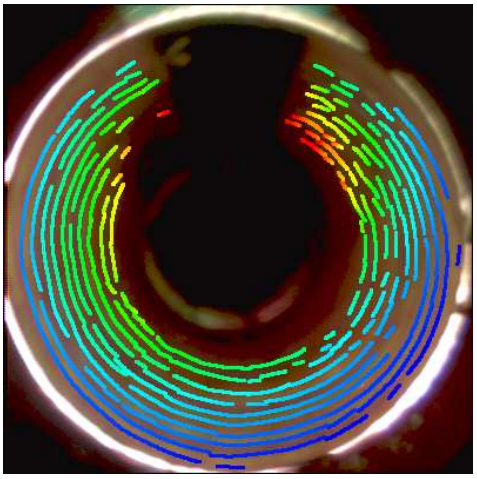

Figure 12: Decoding result for the input image in figure 10 superimposed on the input image. Depth data can only be computed at the edges between two rings. The colors indicate the z-coordinate of the recovered points. The range in this example is 4mm (blue) to 18mm (red).

sets that can be estimated with optical flow techniques will be subject to error as well. The workflow for the different reconstructions was as follows:

1. Compute a 3D point cloud from every single frame.
2. Perform an initial registration of the individual point clouds by applying the known translation between consecutive frames.
3. Optimize the coordinate tranformations between the individual point clouds using ICP.
4. Merge the points clouds into one and merge points closer than 0.3mm.
5. Smooth the resulting point cloud using the method of Vollmer et al. (1999) with a radius of 1.5mm.
6. Perform a Poisson surface reconstruction (Kazhdan et al., 2006) and remove large faces from the resulting mesh (optional).

7.1. Planar target

In a simple test a planar object was measured in two different poses relative to the sensor. The sensor calibration was based on a camera calibration with a classic dot grid target and the pinhole+mirror+tube model. In the first pose, the test plane was positioned approximately 10mm in front of the sensor, with its normal parallel to the camera z-axis. In this pose almost all rings are completely visible in the camera image (figure 13). However, due to the axial setup of the scanner, the triangulation angle is relatively low for points in the forward direction. The angle varies considerably across the working space (mainly in z direction) and typically lies between 10 and 2 degrees. In the frontal pose, the resulting standard deviation of the depth values from the plane was 153µm. In the second pose the test plane was positioned approximately 9mm to the side of the sensor head, its normal perpendicular to the camera z-axis. Here, only part of each ring can be observed in the camera image (figure 14), but the triangulation angle is larger. Consequently, the standard deviation from the plane was only 88µm. Both figure 13 and figure 14 reveal a systematic component of the error that depends on the ring number. This means that the cone parameters that were computed with the procedure outlined in section 5.2 are slightly erroneous. Possibilities for correcting this effect, e.g. better input data or better models, are a topic for future research.

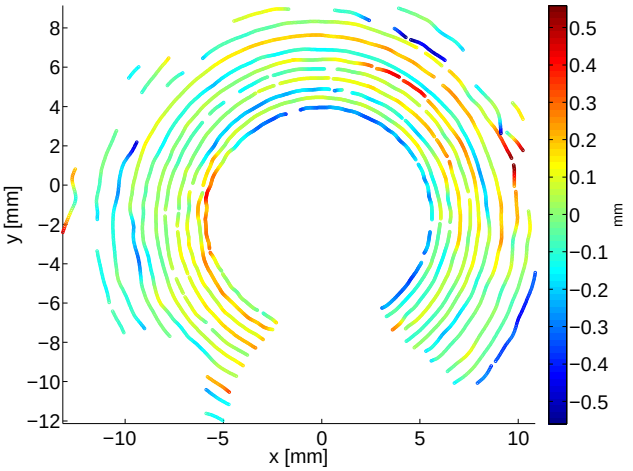

Figure 13: Depth errors of the plane in frontal view for 6794 points. The standard deviation is 153µm. The colors encode the z-error in a local coordinate system whose x and y axes are aligned with the best-fit plane.

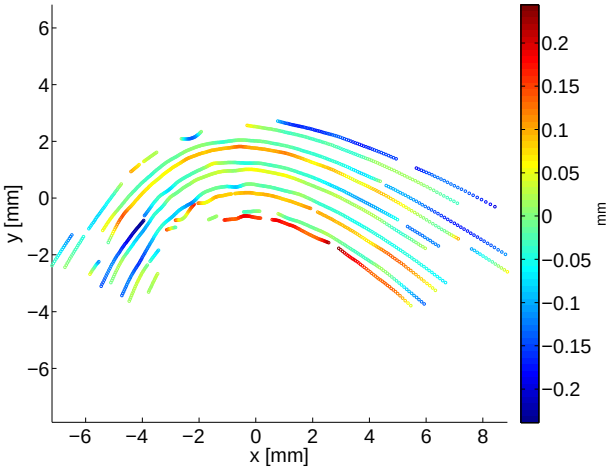

Figure 14: Depth errors of the plane in side view for 2216 points. The standard deviation is 88µm. The colors encode the z-error in an in-plane coordinate system.

7.2. Generic cavity

The result for the hollow plastic block is shown in figure 15. The input data consisted of a sequence of 41 im-

ages. Between successive images the sensor was moved in steps of 0.5 mm along the z-direction of the sensor coordinate system. The initial merged point cloud contained 258610 points; after thinning 11323 remained. This final point cloud was aligned to the ground truth CAD model using LM-ICP (Fitzgibbon, 2003). The average error between the reconstructed points and the original CAD data is 92µm. This result was achieved using a sensor calbration based on the pinhole+mirror+tube model with a classic dot grid target. With the pmt model and an active target the average error was 98µm. Zhang's pinhole camera model with a classic target gave an average error of 108µm. Finally, Zhang's model with the active target resulted in a slightly larger error of 138µm. This behaviour of the error again differs from the camera calibration, where the active pmt model fared best and the pinhole model with a classic target fared worst. Further investigations are necessary here. However, for all calibration methods the results are quite good, considering the size of the reconstructed cavity (approximately 32mm long with a diameter around 13mm) and the low triangulation angle.

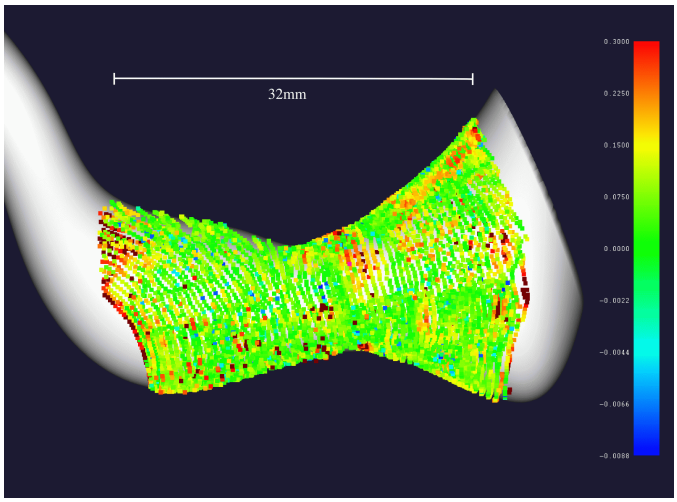

Figure 15: Measurement result for an artificial cavity. The colors encode the error relative to the ground truth CAD data (mm). The average error was 0.092mm.

7.3. Colon Phantom

A rubber replica of a human colon was also measured with the endoscopic sensor. The colon diameter was approximately 40mm and therefore at the upper limit of the current sensor prototype. Nevertheless, good reconstruction results could be obtained. A sequence of 50 frames was recorded with the setup shown in figure 16. From this set of images 267260 points could be recovered. After registration the average point distance was 0.108mm. A thinning step reduced the number of points to 93587. Next, a Poisson surface reconstruction (Kazhdan et al., 2006) was performed, which resulted in a watertight mesh without any holes. From this, the large artificial faces closing the holes were removed, giving a final surface consisting of 39288 vertices. The reconstructed shape clearly shows the folds of the colon (figure 17). Unfortunately each fold also causes a shadow, leading to some holes in the data.

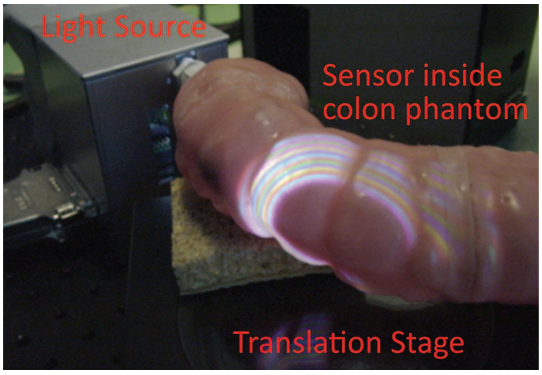

Figure 16: Experimental setup for the colon phantom measurement. The sensor is inside the cavity, which rests on a manual translation stage.

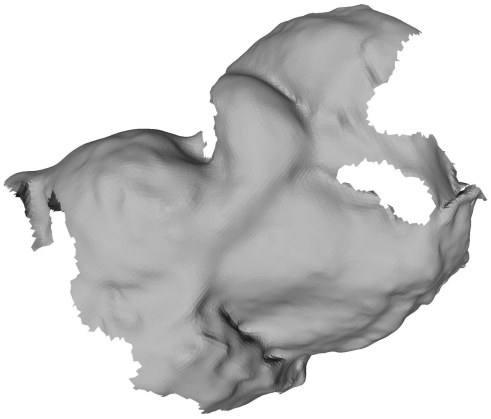

Figure 17: Colon phantom surface reconstructed from 50 frames. The folds are clearly visible, but also cause shadows which result in holes in the recovered surface.

7.4. Windpipe

Figure 19 shows the measured surface of the windpipe, which has a diameter of approximately 14mm. The input data consisted of a sequence of 26 frames. These images yielded 131146 points with an average distance of 0.057mm. The point density is markedly higher here because of the smaller diameter compared to the colon phantom. After thinning, 7417 points remained and were again used for a Poisson surface reconstruction (Kazhdan et al., 2006). Overly large faces were removed from the mesh. The result shows that the sensor works even on biological surfaces, which can be difficult because of volume scattering and highlights. The data quality is very promising. Even the ripples at the "bottom" side could be recovered.

Figure 18: Ex-vivo lamb trachea sample.

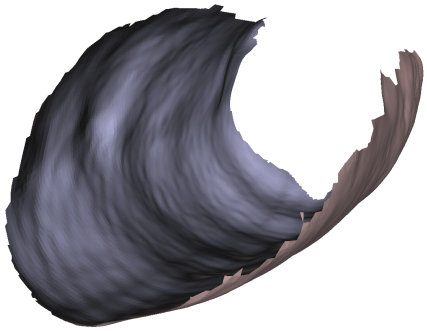

Figure 19: Inner surface of a lamb's windpipe created from 26 images. No additional smoothing was applied. Note the recovered longitudinal ripples at the bottom. The missing area at the top is due to the camera connection cable.

8. Conclusion and Future Work

In this paper we presented a new flexible 3D endoscope with a diameter of 3.6mm. To the best of our knowledge it is the first 3D endoscope based on Single-Shot Structured Light as well as the smallest Structured Light setup presented so far. The endoscope does not contain moving parts and can be built in a robust and cost-efficient way. It acquires 3D data at 30Hz with minimal lag and is not affected by movement. The accuracy of about 0.1mm (including the error due to point cloud alignment and merging) is quite competitive, especially in relation to the small size of the endoscope. Several experiments demonstrate the endoscope's performance with phantoms and biological specimens. Data acquisition works even with tissues that have challenging optical properties, e.g. the color of the tissue proves to be unproblematic despite the use of color rings. So far there is no experience regarding dynamic effects like smoke or bleeding, which may occur during clinical practice.

In the short term we intend to automate the process of registering the individual scans to obtain a complete 3D model. One way this could be achieved is motion estimation via feature tracking or optical flow. The feature tracking module is already implemented, but could not yet be evaluated due to the malfunctioning front camera. An alternative method to obtain the required data is to use the main camera and rapidly switch between the color ring pattern and simple white light illumination. This would also allow sythesis of "natural" stereo images without the overlaid ring pattern, which is potentially distracting. In a future step we plan to perform non-rigid registration between our data and CT or MRI scans.

Non-rigid scenes remain a challenge. Although 3D data from individual frames can be reconstructed with the proposed Single Shot technique, motion in the scene may cause the registration step between data from subsequent frames to fail, even when the sensor motion is known. However, it may be possible to parameterize the admissible types of surface deformation and include those parameters in the optimization process. This is a topic for future research.

Acknowledgements

We would like to thank Sarah Hempel for providing the colon phantom and many ideas for future experiments.

Armbruster, K., Scheffler, M., 1998. Messendes 3D-Endoskop. Horizonte 12, 15–16.

Clancy, N.T., Stoyanov, D., Maier-Hein, L., Groch, A., Yang, G.Z., Elson, D.S., 2011. Spectrally encoded fiber-based structured lighting probe for intraoperative 3d imaging. Biomed. Opt. Express 2, 3119–3128.

Creath, K., 1986. Comparison of phase-measurement algorithms, in: Proceedings of the SPIE, p. 19.

D., S., M., S., P., P., G.Z., Y., 2010. Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: MICCAI, pp. 275–282.

Durrani, A.F., Preminger, G.M., 1995. Three-dimensional video imaging for endoscopic surgery. Computers in biology and medicine 25, 237–247.

Fitzgibbon, A.W., 2003. Robust registration of 2D and 3D point sets. Image and Vision Computing 21, 1145–1153.

Glassner, A.S., 1989. An Introduction to Ray Tracing. Academic Press Inc.

Grasa, O.G., Civera, J., Guemes, A., Muoz, V., Montiel, J.M.M., 2009. Ekf monocular slam 3d modeling, measuring and augmented reality from endoscope image sequences, in: 5th Workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery, held in conjunction with MICCAI2009.

Hayashibe, M., Suzuki, N., Nakamura, Y., 2006. Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management. Medical Image Analysis 10, 509–519.

Hu, M., Penney, G., Figl, M., Edwards, P., Bello, F., Casula, R., Rueckert, D., Hawkes, D., 2010. Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. Medical Image Analysis PMID: 21195656.

Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson surface reconstruction, in: Proceedings of the fourth Eurographics symposium on Geometry processing, p. 61–70.

Kolenovic, E., Osten, W., Klattenhoff, R., Lai, S., von Kopylow, C., Jüptner, W., 2003. Miniaturized digital holography sensor for distal three-dimensional endoscopy. Applied Optics 42, 5167–5172.

Mirota, D.J., Ishii, M., Hager, G.D., 2011. Vision-based navigation in image-guided interventions. Annual Review of Biomedical Engineering 13, 297–319.

More, J., 1978. The Levenberg-Marquardt algorithm: implementation and theory. Numerical Analysis , 105–116.

Mountney, P., Stoyanov, D., G.Z., Y., 2010. Three-dimensional tissue deformation recovery and tracking. Signal Processing Magazine, IEEE 27, 14 –24.

O., G., J., C., J., M., 2011. Ekf monocular slam with relocalization for laparoscopic sequences, in: Robotics and Automation (ICRA), 2011 IEEE International Conference on, pp. 4816 – 4821.

Okatani, T., Deguchi, K., 1997. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. Computer Vision and Image Understanding 66, 119–131.

Penne, J., Höller, K., Stürmer, M., Schrauder, T., Schneider, A., Engelbrecht, R., Feussner, H., Schmauss, B., Hornegger, J., 2009. Time-of-flight 3-D endoscopy. Medical Image Computing and Computer-Assisted Intervention (MICCAI) , 467–474.

Raudies, F., Neumann, H., 2009. An efficient linear method for the estimation of ego-motion from optical flow, in: Proceedings of the 31st DAGM Symposium on Pattern Recognition, Springer-Verlag, Berlin, Heidelberg. pp. 11–20.

Salvi, J., Fernandez, S., Pribanic, T., Llado, X., 2010. A state of the art in structured light patterns for surface profilometry. Pattern Recognition 43, 2666–2680.

Schick, A., Forster, F., Stockmann, M., 2011. 3D measuring in the field of endoscopy, in: Proceedings of the SPIE, p. 808216.

Schmalz, C., Angelopoulou, E., 2010. A graph-based approach for robust single-shot structured light, in: Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference on, p. 80–87.

Schmalz, C., Forster, F., Angelopoulou, E., 2011. Camera calibration with active targets. Optical Engineering , to appear.

Swaminathan, R., Grossberg, M.D., Nayar, S.K., 2006. Non-single viewpoint catadioptric cameras: Geometry and analysis. International Journal of Computer Vision 66, 211–229.

Tabaee, A., Anand, V.K., Fraser, J.F., Brown, S.M., Singh, A., Schwartz, T.H., 2009. Three-dimensional endoscopic pituitary surgery. Neurosurgery 64, 288–295.

Thormahlen, T., Broszio, H., Meier, P., 2002. Three-dimensional endoscopy, in: Medical Imaging in Gastroenterology and Hepatology, p. 199–212.

Vollmer, J., Mencl, R., Mueller, H., 1999. Improved laplacian smoothing of noisy surface meshes, in: Computer Graphics Forum, p. 131–138.

Wang, H., Mirota, D., Hager, G., Ishii, M., 2008. Anatomical reconstruction from endoscopic images: Toward quantitative endoscopy. American journal of rhinology 22, 47.

Wu, C., Narasimhan, S.G., Jaramaz, B., 2010. A multi-image shape-from-shading framework for near-lighting perspective endoscopes. International Journal of Computer Vision 86, 211–228.

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 22.

Zhou, J., Zhang, Q., Li, B., Das, A., 2010. Synthesis of stereoscopic views from monocular endoscopic videos, in: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, p. 55–62.

Christoph Schmalz received his diploma in Physics from the University of Erlangen-Nuremberg, Germany, in 2005. Currently he is working on a PhD thesis about Single Shot Structured Light scanning in a joint project of Siemens AG and the Pattern Recognition Lab in Erlangen. His research interests include image processing, camera calibration and 3D reconstruction.

Dr. Frank Forster is a program manager for 2D and 3D Machine Vision at Siemens Corporate Technology. He received his Master of Science in Computer Science from the State University of New York at Albany in 1998, his diploma in Computer Science in 2000 from the University of Wuerzburg and in 2005 his PhD degree in natural sciences from the Technical University of Munich. Since 2002 he is a member of the Corporate Technology department at Siemens AG in Munich, where he is involved in vision based sensor systems for medical or industrial purposes.

Dr. Anton Schick received his Diploma and PhD degree in Physics from Technische Universität München in 1983 respectively 1988. Currently, he is a Principal Research Scientist at the Corporate Technology Division at Siemens AG in Munich (CT T DE HW2) and holds over thirty granted patents. Previously, he was director of the development unit of Optical Solutions at Siemens I DT EA. His research group developed an extremely fast confocal 3D measurement technique that has since been commercialized by the Siemens EA business unit. He has more than twenty years of experience in the field of optical technologies, and regards the field as multidisciplinary; a view which has allowed deep insights into optical design, electro-optic systems and laser technology.

Dr. Elli Angelopoulou received her PhD in Computer Science from the Johns Hopkins University in 1997. She did her postdoc at the GRASP Laboratory at the University of Pennsylvania. Her research is focused on multispectral imaging, reflectance analysis, image forensics and shape reconstruction. She has served on the program committees of ICCV, CVPR and ECCV and is an associate editor of MVA and JISR. She is a member of the OSA and the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence.