IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 8, 2021, accepted November 1, 2021, date of publication November 9, 2021,
date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126835

Extraction of Key-Frames From Endoscopic Videos by Using Depth Information

PRADIPTA SASMAL[1,*], AVINASH PAUL[1,*], M. K. BHUYAN[1], (Senior Member, IEEE), YUJI IWAHORI[1,2], (Member, IEEE), AND KUNIO KASUGAI[3]

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India
[2]Department of Computer Science, Chubu University, Kasugai 487-8501, Japan
[3]Department of Gastroenterology, Aichi Medical University, Nagakute 480-1195, Japan

Corresponding authors: Pradipta Sasmal (s.pradipta@iitg.ac.in) and M. K. Bhuyan (mkb@iitg.ac.in)

The work of Yuji Iwahori was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid Scientific Research (C) under Grant 20K11873, and in part by the Chubu University Grant.

*Pradipta Sasmal and Avinash Paul contributed equally to this work.

ABSTRACT Early detection of colorectal cancer (CRC) can reduce the risk of death. Polyps are the precursor to such cancer. Analyzing the polyps from the most significant frames out of thousands of endoscopy frames is vital for diagnosing and understanding disease. In this article, a deep learning-based monocular depth estimation (MDE) technique is proposed to select the most informative frames (key-frames) of an endoscopic video. In most cases, ground truth depth maps of polyps are not readily available, and that is why the transfer learning approach is adopted in our method. An endoscopic modality generally captures thousands of frames. In this scenario, it is quite essential to discard low-quality and clinically irrelevant frames of an endoscopic video while the most informative frames should be retained for clinical diagnosis. In this view, a key-frame selection strategy is proposed by utilizing the depth information of polyps. In our method, image moment, edge magnitude, and key points are considered for adaptively selecting the key-frames. One important application of our proposed method could be the 3D reconstruction of polyps with the help of extracted key-frames. It gives a surgeon a real-time 3D view of the polyp surface for resection which involves detaching the polyp from its mucosa layer. Also, polyps are localized with the help of extracted depth maps.

INDEX TERMS Key-frames, colorectal cancer (CRC), monocular depth, polyps, 3D reconstruction.

I. INTRODUCTION

Endoscopy is a minimally invasive state-of-the-art medical modality to investigate the gastrointestinal (GI) tract. During endoscopy, an endoscopist looks to find a tumor in the mucosa. The tumor-like growth is called polyps and, if not treated early, may lead to cancer [1]. These polyps are generally found in the colon region and turn into cancerous cells at their advanced stage. Colonoscopy is a medical procedure adopted to detect such anomalies in the colon regions. Colorectal cancer (CRC) is the most occurring cancer, and a significant reason of deaths worldwide [2]. Wireless Capsule Endoscopy (WCE) is an invasive modality to monitor the conditions of the internal viscera of a human body. WCE moves along the gastrointestinal (GI) tract to capture images. It is extensively used to detect polyps in colon regions, which become cancerous if left untreated. Colorectal cancer is the third most prevalent cancer today [3]. During the colonoscopy, doctors comprehensively analyze the detected polyp regions to find the dysplasia in them. Depending on the condition of the polyp nature, they may opt for laparoscopic surgery. However, the number of frames captured during the entire colonoscopy process is so humongous that it challenges the surgeon to infer useful clinical information. Therefore, video summarization techniques are adopted which only retain the clinically informative frames. During WCE, the capsule moves under the peristalsis movement, and it is challenging to control the motion and orientation of the camera. Thus, redundant and clinically non-significant frames are generally obtained in a video sequence. WCE takes nearly 8 hours, capturing close to 50000 frames. A large part of the data is clinically not significant and needs to be removed [4].

Several methods have been proposed for detection, localization and classification of polyps in endoscopy frame [5]–[8]. A recent work focussing on video

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

summarization instead of anomalies detection like bleeding or ulceration is proposed by Li *et al.* [9]. Iakovidis *et al.* [10] used clustering-based methods for video summarization. Similar work based on clustering technique was proposed by Avila *et al.* [11]. However, clustering-based methods are not suitable in noise environments. Endoscopy frames are generally susceptible to noise. Also, redundant frames are captured during the endoscopy, which makes clustering methods perform poorly. Researchers are working on visual attention models, like saliency maps for finding key-frames of videos [12]. Another visual saliency-based attention model was proposed by Ezaj *et al.* [13]. They used motion, color, and texture features for hysteroscopy video summarization. A color histogram comparison-based method was adopted by Mendi *et al.* [14]. They compared the color histogram of successive frames in a video sequence, and key-frames were selected using k-means and PCA whenever a significant change in content was observed. However, this model does not fit into endoscopic videos as most of the frames have similar color information. Recently, dictionary learning-based approaches have been proposed for video summarization [15]. In [16], a gastroscopic video summarization technique based on a dictionary learning approach is proposed. Key-frames are very important and help in better prognosis and clinical management of the disease. Therefore, colonoscopy frames that need immediate medical attention are considered for this study. Malignant polyps usually have a convex shape and are more textured compared to benign polyps. Seitz *et al.*, [17] proposed that polyp size is correlated to the degree of dysplasia. A large and convex type polyp is associated with more severity of dysplasia. Getting a 3D view of the polyp surface can significantly help in resection [18]. A good 3D reconstruction of an object in an image entails dense depth estimation. The 3D view gives shape and size information of a polyp. Depth estimation of endoscopic images is a challenging task as the endoscopic images are monocular.

Attempts have been made to solve it as a per-pixel regression problem, however, supervised learning methods require a lot of training data. It isn't easy to acquire depth data without using stereo cameras or expensive depth sensors, as with endoscopy videos. Thus unsupervised methods are being given more importance. Depth estimation in endoscopic video frames imparts clinical relevance to a physician. 3D reconstruction of the monocular images helps in diagnosis and surgical planning. Recently, depth estimation, especially monocular depth estimation (MDE) has gained high research interest. This is due to its application in scene understanding, robotics, autonomous driving, and Augmented Reality (AR). Finding depth from a single image is an unconstrained problem since many real-world scenes can give the same 2D image, resulting in the same depth maps. Humans perceive depth from cues such as perspective, prior knowledge of sizes of objects, or occlusion. In the literature, both supervised and unsupervised-based methods have been employed for estimating depth.

Eigen *et al.*, [19] introduced a multi-scale information approach that takes care of both global scene structure and local neighboring pixel information. A scale-invariant loss is used for MDE. Similarly, Xu *et al.* [20] formulated MDE as a continuous random field problem (CRF). They fused the multi-scale estimation computed from the inner semantic layers of a CNN with a CRF framework. Instead of finding continuous depth maps, Fu *et al.* [21] estimated depth using an ordinal regression approach. A space-increasing discretization method is introduced by allowing objects at larger depths to have a lesser influence on the depth maps than the objects nearer to the observer.

Depth is generally obtained using sensors like LIDAR, Kinect, or by using stereo cameras. Sensors are expensive, and stereo cameras are not generally used in endoscopy due to several restrictions. Obtaining ground-truth training data for depth estimation is very difficult in endoscopic imaging, so supervised methods are not feasible for endoscopic image reconstruction. Finding correspondence between two images for 3D reconstruction is also difficult in endoscopy videos. It isn't easy to find corresponding features across the frames.

Hence, unsupervised and semi-supervised methods are employed for MDE. Garg *et al.* [22] used binocular stereo image pairs for the training of CNNs and then minimized a loss function formed by the wrapping of the left view image into its right of the stereo pair. Godard *et al.* [23] improved this method by using the left-right consistency criterion. They trained CNNs on stereo images but used a single image for inference. They introduced a new CNN architecture that computes end-to-end MDE. The network was trained with an efficient reconstruction loss function. The state-of-the-art unsupervised MDE method, i.e., Monodepth [23] model has limited application in in-vivo images like endoscopic images. This is because most models leverage outdoor scenes [24] and a few indoor scenes [25] for training, and they use high-end sensors or stereo cameras, while the WCE method only captures monocular images. Hence, it is important to devise a strategy to perform MDE in medical imaging datasets that generally do not have ground truth depth information. That is why, a transfer learning approach is adopted in our method for estimating depth. Transfer learning refers to a learning method where what has been learned in one setting is exploited to improve generalization in another setting [26]. Zero-shot learning is the extreme case of transfer learning where no labeled examples are present. In our method, a zero-shot learning approach for MDE [27] is employed.

The proposed method consists of two main steps. The first step focuses on depth estimation, and the second step extracts key-frames. As mentioned above, a zero-shot learning approach is adopted for depth estimation in endoscopic videos. We propose a framework to select the most informative frames of an endoscopic video sequence. Our method employs a three-criteria approach to identify the key-frames. Subsequently, these key-frames can be used for 3D reconstruction. Our method is unique in the sense that it considers depth information to find key-frames. Finally, any of the

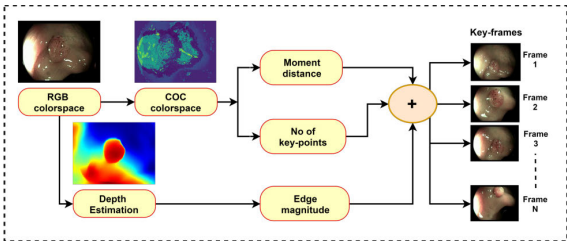

FIGURE 1. Proposed method of finding key-frames.

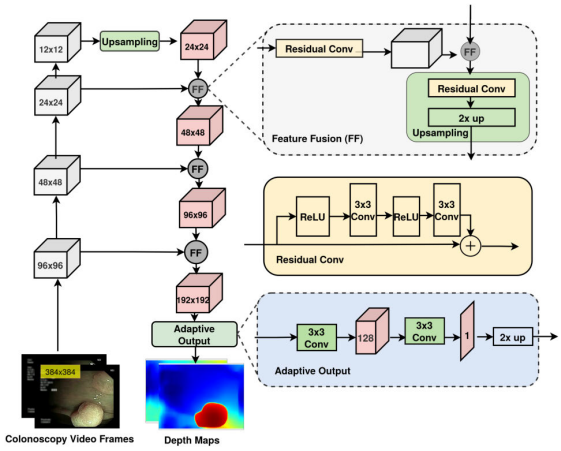

FIGURE 2. Network architecture for depth estimation from colonoscopy video frames; The model is based on a feedforward ResNet architecture [28].

selected key-frames can then be used for 3D reconstruction using a GUI. Experimental results clearly demonstrate the effectiveness of our method in choosing the key-frames and subsequent polyp visualization. The proposed method is elucidated in section II. Experimental results and conclusions are discussed in section III and section IV, respectively.

II. PROPOSED METHOD

A. DEPTH ESTIMATION

Due to the unavailability of ground truth depth data in endoscopy video datasets, a transfer learning approach is adopted for MDE in our proposed method. Lasinger *et al.* [27] proposed a zero-shot learning for depth estimation. The work of Lasinger *et al.* inspires our proposed work for depth estimation as a zero-shot approach.

This section explains how we use monocular images to learn relative depth. As demonstrated in Figure 2, we model monocular relative depth perception as a regression problem. In an end-to-end method to regress pixel-wise relative depth given a batch of input images I, we create a non-linear function $y = f(I, \delta)$ parameterized by δ. The network is built on a feedforward ResNet architecture that generates multi-scale feature mappings [28]. To improve predictions, a progressive refinement technique is used to combine multi-scale variables.

The model was trained for depth maps obtained in three different ways. First, the dataset contains depth maps obtained using LIDAR sensors. This method gives depth maps of high quality. Second, the Structure from Motion (SfM) approach is employed to estimate the depth. The third method of getting depth information from stereo images of the 3D movies dataset. It uses optical flow to find motion vectors from each of the stereo images. Then, the left-right image disparity is used to find a depth map. The dataset contains images that have varying aspect ratios. Sometimes, black bars on frame borders appear in estimated depth maps. So, all the images are cropped to extract only the center portion of the frame. This ensures the framework can handle images of varying aspect ratios. Moreover, the method focuses more on the central part of the image frame. Using the distance of an object from the camera to predict depth leads to sparse 3D reconstructions. This is because depth is estimated by tracking the corresponding features over a series of frames. Then, the induced parallax is used for triangulation and depth estimation. However, the resultant parallax will be small for distant features (like the sky) and won't allow proper reconstruction. Thus, distant objects like the sky are not considered while estimating depth. This addresses the issue of finding correspondences for distant objects.

The disparity map is found by using stereo matching using optical flow. Optical flow successfully handles moderate displacements. The horizontal component of the flow vectors is used as a reference for finding a disparity map. Optical flow is estimated taking either the left or right image as a reference and finding flow from the other. Next, the consistency between both left and right is calculated to discard the pixels with more than one-pixel disparity.

The datasets on which the model is trained are unique because they contain both positive and negative disparities. However, training on ground truth data from different sources has some constraints: 1) The dataset contains images that have only depth (from LIDAR sensors) or disparity images; 2) Data obtained from the SfM technique gives depth images for which scale is not known; 3) The 3D movies dataset gives a ground truth depth which has an unknown shift.

1) LOSS FUNCTION

A shift and scale invariant loss function is chosen to address the problems pertaining to training on three different datasets. Let $\mathbf{d} \in \mathbb{R}^N$ be the computed inverse depth and $\mathbf{d}' \in \mathbb{R}^N$ be ground truth inverse depth, where N is the number of pixels in a frame. Here s and t represent scale and shift, respectively and they are positive real numbers. This can be represented in a vector form by taking $\vec{\mathbf{d}}_i = (\mathbf{d}_i, 1)^\intercal$ and $\mathbf{p} = (s, t)^\intercal$ and thus the loss function becomes:

$$\mathcal{L} = \arg\min_{s,t} \frac{1}{2N} \sum_{i=1}^{N} (s\mathbf{d}_i + t - \mathbf{d}'_i)^2 \tag{1}$$

$$\mathcal{L}(\mathbf{d}_i, \mathbf{d}'_i) = \arg\min_{\mathbf{p}} \frac{1}{2N} \sum_{i=1}^{N} (\vec{\mathbf{d}}_i^\intercal \mathbf{p} - \mathbf{d}'_i)^2 \tag{2}$$

The closed-form solution is given as:

$$\mathbf{p}^{opt} = (\sum_{i=1}^{N} \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^\intercal)^{-1} (\sum_{i=1}^{N} \vec{\mathbf{d}}_i \mathbf{d}'_i) \tag{3}$$

Substituting $\mathbf{p}^{opt}$ into (2) we get:

$$\mathcal{L}(\mathbf{d}_i, \mathbf{d'}_i) = \arg\min_{\mathbf{p}} \frac{1}{2N} \sum_{i=1}^{N} (\vec{\mathbf{d}}_i^{\intercal} \mathbf{p}^{opt} - \mathbf{d'}_i)^2 \tag{4}$$

2) REGULARIZATION TERM

A multi-scale scale-invariant regularization term is used, which does gradient matching to the depth inverse space. This biases discontinuities to be sharp and coincide with ground truth discontinuities. The regularization term can be defined as,

$$\mathcal{L}_r(\mathbf{d}_i, \mathbf{d'}_i) = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N} (|\Delta_x {Q_i}^k| + |\Delta_y {Q_i}^k|) \tag{5}$$

where,

$$Q_i = \vec{\mathbf{d}}_i^{\intercal} \mathbf{p}^{opt} - \mathbf{d'}_i \tag{6}$$

Here Q^k gives the difference of inverse depth maps at a scale k. We use $k = 4$ scale levels, halving the image resolution at each level. Also, the scale is applied before finding x and y gradients.

3) MODIFIED LOSS FUNCTION

The final loss function for a training set of size M, taking into consideration of the regularization term, becomes:

$$\mathcal{L}_{final} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(\mathbf{d}^i, (\mathbf{d'})^i) + \alpha \mathcal{L}_r(\mathbf{d}^i, (\mathbf{d'})^i) \tag{7}$$

Here α is taken as 0.5.

B. SELECTION OF KEY-FRAMES

During the colonoscopy, not all the captured frames are clinically significant. Most of the frames may have redundant information, or may not be useful from a diagnostic perspective. Such frames need to be discarded and the clinically informative frames need to be retained. It is also strenuous and computationally intensive for a physician to investigate each frame of a video sequence. Thus, we propose a key-frame selection technique. Subsequently, 3D reconstruction is done to perform further analysis of the polyps. The key-frame selection method is given in Fig. 1.

1) COLOUR SPACE CONVERSION

Our dataset contains images which are in RGB color space. Taking cues from the human visual system which works on saliency, we changed the color space from RGB to COC which gives a better perception in the medical imaging [29].

The image is subsequently used to find key-frames. A frame should satisfy three criteria before being selected as a key-frame: 1) It should be significantly different from neighboring frames. 2) The key-frame should give significant depth information of a polyp. 3) The polyp should not be occluded in the key-frame. We ensured that the above requirements were met, and they are formulated as follows:

2) IMAGE MOMENT

Image moments give the information of the shape of a region along with its boundaries and texture. Hu moments [30] are considered as they are invariant to affine transformation, and moment distances of consecutive frames are used to identify the redundant frames of a video. Subsequently, the moment difference between consecutive frames are calculated. The frames with a higher moment distance will be considered as the key frames. The moment distance d between two images is calculated as:

$$d = \sum_{i=1}^{i=7} (I_i - I'_i)^2 \tag{8}$$

where, i represents each of a total of 7 moments.

3) EDGE DENSITY

In our proposed method, the key-frames which have significant depth information are only considered for the 3D reconstruction of a polyp. It is observed that the polyp images having more edges have more depth information. The edge information can be obtained with the help of the gradient magnitude of an image. Before finding the gradients, images were smoothed using a Gaussian kernel.

Horizontal and vertical gradients are obtained using Sobel operators S_x and S_y and then the gradient magnitude ΔS is calculated as follows:

$$\Delta S = \sqrt{(S_x)^2 + (S_y)^2} \tag{9}$$

4) KEY-POINT DETECTION

The proposed moment-based key-frame detection method may capture some occluded frames. So, the objective is to select non-occluded key-frames from a group of key-frames that were extracted by our proposed image moment and edge density-based criteria. For this, a key-point detection-based technique is used.

For key-point detection and extraction, we used ORB (Oriented FAST and Rotated BRIEF). ORB is computationally faster and robust to noises in endoscopic images. The frames containing a lesser number of ORB points correspond to occluded polyps.

5) ADAPTIVE KEY-FRAME SELECTION

After finding the moment distance (d), edge magnitude (s), and the number of ORB points (p), we normalize these scores using min-max normalization. This is done so that each of the three scores is reduced to the range of 0 to 1 with both values inclusive. Instead of adding the three scores directly, we use dynamic weights to capture the changes in a video. The variable having more significant variance is given more weightage. Here, w_i is the weight of the normalized score. To consider intra-variable changes, we used the sum of the magnitude of difference between consecutive frame scores as a measure to find weights. We then normalized this score to be used as weights for finding a fused score. The weights are

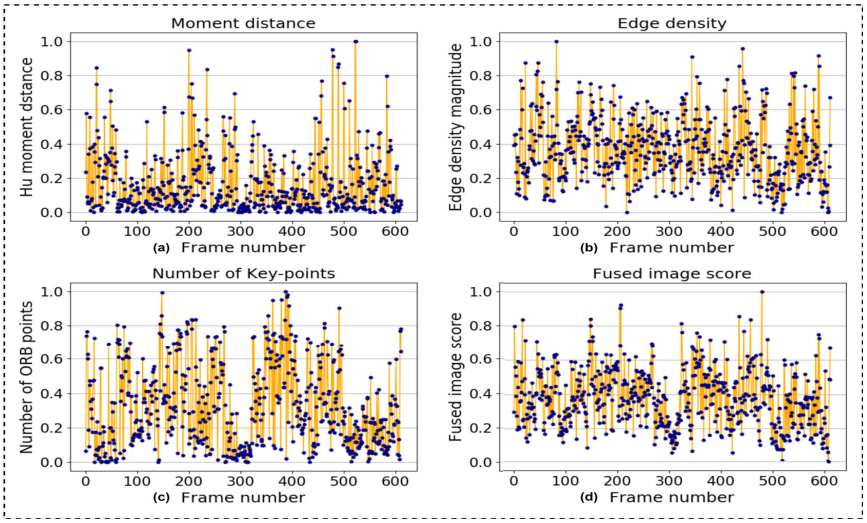

FIGURE 3. Plot of moment distance, edge density, number of key-points and the total fused score vs frame number of a colonoscopy video sequence.

given by:

$$d_1 = \sum_{i=1}^{n} |d_i - d_i'|, \quad s_1 = \sum_{i=1}^{n} |s_i - s_i'|,$$
$$p_1 = \sum_{i=1}^{n} |p_i - p_i'| \tag{10}$$
$$w_1 = \frac{d_1}{d_1 + s_1 + p_1}, \quad w_2 = \frac{s_1}{d_1 + s_1 + p_1},$$
$$w_3 = \frac{p_1}{d_1 + s_1 + p_1} \tag{11}$$
$$f = w_1 d_1 + w_2 s_1 + w_3 p_1 \tag{12}$$

Here, d_1, s_1, p_1 are the sum of magnitudes of difference between consecutive frame scores and f is the fused score obtained by adaptively weighting the three frame scores. The frames with the highest fused scores are selected according to a threshold value which was set as 0.5. The variance of each criterion with frame number is shown in Fig. 3.

III. EXPERIMENTAL RESULTS

The proposed method is evaluated on the publicly available dataset. This dataset contains colonoscopic video sequences from three classes, namely adenoma, serrated and hyperplasic. The adenoma class contains 40 sequences, serrated class contains 15, while the hyperplasic class contains 21 video sequences [32]. In this work, we consider only the frames from the adenoma (malignant) class because this class needs the maximum attention of the physician. The dataset used in this work is publicly available in the url: *http* : *//www.depeca.uah.es/colonoscopy_dataset/*.

For this work, we considered only narrowband images (NBI) as they require less preprocessing and are generally used for polyp classification. The adenoma class contains 40 video sequences of different patients. It contains both patchy and convex polyp sequences. In this work, the frames which have convex polyps are taken for estimating the depth. A few

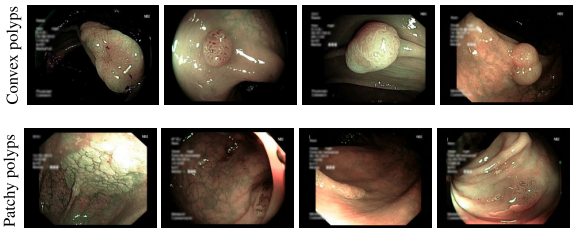

FIGURE 4. Some images of colonoscopy dataset: the first row are the examples of convex polyps and the second row are the examples of patchy polyps.

TABLE 1. Key frame selection and segmentation performance using our method on some of the sequences of CVC-Clinic Database (Sequences with only the elevated polyps are considered).

Sequence	#Key frames selected	mIoU
26-50	5	0.501
104-126	7	0.546
127-151	11	0.721
298-317	2	0.723
343-363	7	0.654
384-408	13	0.723
409-428	8	0.663
479-503	20	0.793
504-528	6	0.695
572-591	4	0.698
592-612	5	0.747

convex and patchy polyp images of the dataset are shown in Fig. 4. We used a pre-trained model trained on diverse datasets by Lasinger *et al.* [27] in our work. A ResNet-based multiscale architecture as proposed by Xian *et al.* [33] is used for depth estimation. Adam optimizer is used with a learning rate of 10^{-4} for layers that are randomly initiated and 10^{-5} for layers initialized with pre-trained weights. Decay rates for the optimizer are set at $\beta_1 = .9$ and $\beta_2 = .999$, training uses a batch size of 8. Due to different image aspect ratios, images

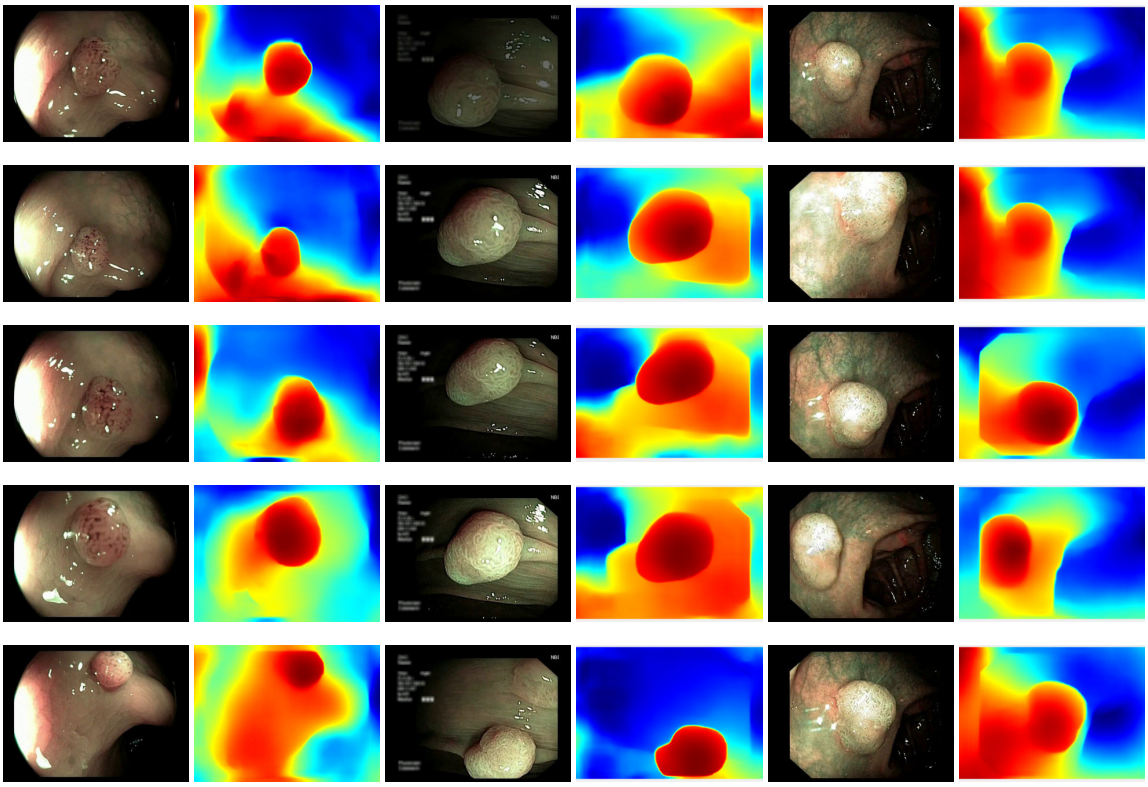

FIGURE 5. Key-frames obtained by our method and their corresponding depth maps. The polyp is visible from different viewing angles in these selected frames.

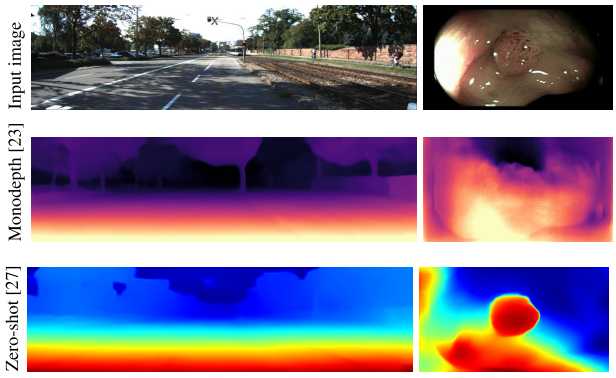

FIGURE 6. Comparison of MDE on two input images, one outdoor and the other one is an endoscopy image. The depth map by Monodepth [23] performs well for outdoor environment while giving unsatisfactory results for the endoscopy image. However, the zero-shot learning method [27] clearly performs well for medical images but cannot accurately estimate the depth in outdoor scenes.

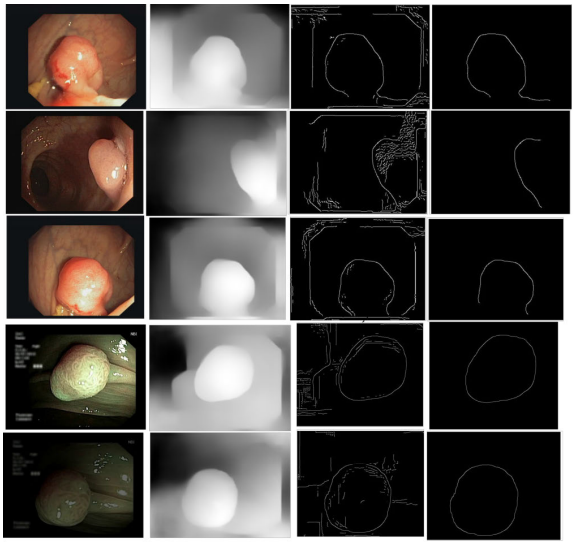

FIGURE 7. Polyp boundary detection using depth map; Column 1: Original endoscopic image, Column 2: Generated depth maps, Column 3: Detected polyp boundary using canny edge detection algorithm, Column 4: Edge refinement using connected component analysis. First three rows of image samples are taken from CVC-Clinic Database [31], the last two rows of images are frames taken from a video sequence of the publicly available dataset [32].

are cropped and augmented for training. The input size of the frames is taken as 384×384.

Our method performs better than the state-of-the-art MDE methods. The depth estimation results are shown in Fig. 6, where the first row represents the input images, while the second and the third row show the comparative results between monodepth model [23] and zero-shot cross-dataset transfer pre-trained model [27]. This clearly shows that monodepth performs well in outdoor environments than our method. However, the Zero-shot learning method is more accurate in predicting depth in endoscopic images.

Our method is the first-of-its-kind in which key-frames are extracted from an endoscopic video using depth maps. Also, it is robust to occlusions. As redundant frames are discarded in our method, it is more convenient for physicians to analyze essential frames of a video sequence. As explained earlier, the moment distance criterion between consecutive frames is

used to ensure that redundant frames are identified and then discarded. The edge magnitude criterion leverages the depth images data to select the best frames. Frames with fewer ORB points have occluded polyps, and these frames are redundant. Adaptive thresholding is used to apply three criteria to obtain essential frames for 3D reconstruction.

The selected key-frames are finally used to reconstruct the 3D surface of the polyp. We have used Facebook's 3D image GUI to view the reconstructed polyp surface; the link to the video is shown here: *https : //youtu.be/PJKfk0Mqu2I*. 3D visualization of a polyp helps in surgeries involving the removal of the polyp from its root. This gives better visualization of polyps for diagnosis. Fig. 5 shows some of the results of key-frame extraction and the corresponding depth maps. No publicly available datasets or methods using them that predict depth maps from endoscopic frames exist. Thus, a comparison between different methods for predicting depth from endoscopic images couldn't be performed.

Another application of our proposed method could be automatic segmentation of polyps in endoscopic images. The depth maps generated by our proposed method can further be used for polyp localization. The canny edge detector is used over the depth maps, and subsequently, polyp boundary is determined by using connected component analysis. Fig. 7 shows localized polyps in some of the endoscopic image samples. The segmentation performance on some of the sequences of the CVC-Clinic Database [31] is shown in Table 1. This dataset contains 25 colonoscopy video sequences. Each sequence contains an average of 25 frames. We defined mIoU as the mean intersection over the union of the segmented polyp masks to the ground truth masks. In polyp segmentation, an IoU score of $\geq$0.5 is generally considered good [34].

IV. CONCLUSION

Our proposed method can determine depth maps using a zero-shot learning approach. The zero-shot learning method performs well on previously unseen classes like endoscopic images. Through this, we extended MDE to in-vivo images, which would be helpful to analyze medical images. The essential frames are picked out from WCE videos with the help of depth information and the proposed three criteria selection strategy. The selection of a threshold value for the final fused score must be empirically set to extract the key-frames. Experimental results show the efficacy of the proposed method in selecting key-frames from endoscopic videos and subsequent segmentation of detected polyps in the key-frames with the help of extracted depth maps. Also, the 3D model could be used in clinical diagnosis and surgeries. One possible extension of this work could be the visualization of polyps in detected key-frames in an augmented reality framework.

ACKNOWLEDGEMENT

The work of Yuji Iwahori was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid Scientific Research (C) under Grant 20K11873, and in part by the Chubu University Grant.

REFERENCES

[1] H. Messmann, *Atlas of Colonoscopy: Techniques, Diagnosis, Interventional Procedures*. Stuttgart, Germany: Thieme, 2006.
[2] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, Apr. 2017.
[3] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. Meester, A. Barzi, and A. Jemal, "Colorectal cancer statistics, 2017," *Cancer J. Clin.*, vol. 67, no. 3, pp. 177–193, 2017.
[4] H.-G. Lee, M.-K. Choi, B.-S. Shin, and S.-C. Lee, "Reducing redundancy in wireless capsule endoscopy videos," *Comput. Biol. Med.*, vol. 43, no. 6, pp. 670–682, 2013.
[5] B.-P. Li and M. Q.-H. Meng, "Comparison of several texture features for tumor detection in CE images," *J. Med. Syst.*, vol. 36, no. 4, pp. 2463–2469, Aug. 2012.
[6] M. P. Tjoa, S. M. Krishnan, and R. Doraiswami, "Automated diagnosis for segmentation of colonoscopic images using chromatic features," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. Conf. (CCECE)*, May 2002, pp. 1177–1180.
[7] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021.
[8] P. Sasmal, M. K. Bhuyan, Y. Iwahori, and K. Kasugai, "Colonoscopic polyp classification using local shape and texture features," *IEEE Access*, vol. 9, pp. 92629–92639, 2021.
[9] B. Li, M. Q.-H. Meng, and Q. Zhao, "Wireless capsule endoscopy video summary," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2010, pp. 454–459.
[10] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Med. Imag. Graph.*, vol. 34, no. 6, pp. 471–478, Sep. 2010.
[11] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, Jr., and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, Jan. 2011.
[12] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.
[13] N. Ejaz, I. Mehmood, and S. W. Baik, "MRT letter: Visual attention driven framework for hysteroscopy video abstraction," *Microsc. Res. Technique*, vol. 76, no. 6, pp. 559–563, Jun. 2013.
[14] E. Mendi, C. Bayrak, S. Cecen, and E. Ermisoglu, "Content-based management service for medical videos," *Telemed. e-Health*, vol. 19, no. 1, pp. 36–41, Jan. 2013.
[15] M. Ma, S. Mei, S. Wan, Z. Wang, and D. Feng, "Video summarization via nonlinear sparse dictionary selection," *IEEE Access*, vol. 7, pp. 11763–11774, 2019.
[16] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, and H. Yu, "Scalable gastroscopic video summarization via similar-inhibition dictionary selection," *Artif. Intell. Med.*, vol. 66, pp. 1–13, Jan. 2016.
[17] U. Seitz, T. L. Ang, F. Dy, T. Sookpaisal, J. Sadikin, I. Marki, F. Thonke, S. Seewald, S. Bohnacker, A. De Weerth, and N. Soehendra, "Colonic polyps and malignant potential—Does size matter?" *Gastrointestinal Endoscopy*, vol. 61, no. 5, Apr. 2005, Art. no. AB264.
[18] R. Law, A. Das, D. Gregory, S. Komanduri, R. Muthusamy, A. Rastogi, J. Vargo, M. B. Wallace, G. S. Raju, R. Mounzer, J. Klapman, J. Shah, R. Watson, R. Wilson, S. A. Edmundowicz, and S. Wani, "Endoscopic resection is cost-effective compared with laparoscopic resection in the management of complex colon polyps: An economic analysis," *Gastrointestinal Endoscopy*, vol. 83, no. 6, pp. 1248–1257, Jun. 2016.
[19] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.
[20] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1426–1440, Jun. 2019.

[21] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
[22] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 740–756.
[23] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.
[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving," in *Proc. CVPR*, Jul. 2011, pp. 3354–3361.
[25] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-D scene structure from a single still image," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[27] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2019, *arXiv:1907.01341*.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, p. 68, 1997.
[30] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[31] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilari no, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.
[32] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.
[33] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 311–320.
[34] M. Yamada, Y. Saito, H. Imaoka, M. Saiko, S. Yamada, H. Kondo, H. Takamaru, T. Sakamoto, J. Sese, A. Kuchiba, T. Shibata, and R. Hamamoto, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

PRADIPTA SASMAL received the B.Tech. degree from the Silicon Institute of Technology, Bhubaneswar, India, in 2011, and the master's (M.Tech.) degree in communication and signal processing specialization from the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India, in 2015. He is currently pursuing the Ph.D. degree with the Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati, India. He has a special interest in the endoscopic image and video processing for detection/segmentation/classification of the polyps with the help of computer vision and artificial intelligence. His research interests include biomedical image/video analysis, machine and deep learning, computer vision, and application of artificial intelligence.

AVINASH PAUL received the B.Tech. degree from the National Institute of Technology, Rourkela, India, in 2018, and the master's (M.Tech.) degree in signal processing specialization from the Indian Institute of Technology (IIT) Guwahati, India, in 2020. He is currently working as a Graphics Validation Engineer at Intel India. His research interests include computer vision, computer graphics, and medical imaging.

M. K. BHUYAN (Senior Member, IEEE) received the Ph.D. degree in electronics and communication engineering from the Indian Institute of Technology (IIT) Guwahati, India, in 2006. He was with the School of Information Technology and Electrical Engineering, University of Queensland, St. Lucia, QLD, Australia, where he was involved in postdoctoral research. Subsequently, he was a Researcher with the SAFE Sensor Research Group, NICTA, Brisbane, QLD. He was an Assistant Professor with the Department of Electrical Engineering, IIT Roorkee, India, and Jorhat Engineering College, Assam, India. He also worked in Indian engineering services. He is currently a Professor with the Department of Electronics and Electrical Engineering, IIT Guwahati, and the Associate Dean of Infrastructure, Planning and Management, IIT Guwahati. In 2014, he was a Visiting Professor with Indiana University and Purdue University, Indiana, USA. He is also working as a Visiting Professor with the Department of Computer Science, Chubu University, Japan. He has almost 25 years of industry, teaching, and research experience. His current research interests include image/video processing, computer vision, machine and deep learning, human computer interactions (HCI), virtual reality and augmented reality, and biomedical signal processing. He was a recipient of the National Award for Best Applied Research/Technological Innovation, which was presented by the Honorable President of India, in 2012, the Prestigious Fullbright-Nehru Academic and Professional Excellence Fellowship, and the BOYSCAST Fellowship.

YUJI IWAHORI (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Electronics, Tokyo Institute of Technology, in 1988. He joined the Educational Centre for Information Processing, Nagoya Institute of Technology, as a Research Associate, in 1988, and he became a Professor with the Centre for Information and Media Studies, Nagoya Institute of Technology, in 2002. Since 2004, he has been a Professor with Chubu University, Japan. He acted as the Department Head of Computer Science, the Head of Graduate Course of Computer Science, the Vice-Dean of the College of Engineering, Chubu University. In the meanwhile, he has been a Visiting Researcher of UBC Computer Science, Canada, since 1991. He has also been a Research Collaborator with IIT Guwahati, since 2010, and with the Department of Computer Engineering, Chulalongkorn University, since 2014. He has become an Honorary Faculty at IIT Guwahati, since 2020. His research interests include computer vision, biomedical image processing, deep learning, and application of artificial intelligence. He received the KES 2008 Best Paper Award and the KES 2013 Best Paper Award from KES International.

KUNIO KASUGAI received the M.D. and the Ph.D. degrees in bioregulation research from Nagoya City University Medical School, Nagoya, Japan. He also worked as a Research Fellow of internal medicine with the University of Michigan. He is currently a Professor of internal medicine with the Division of Gastroenterology and the Vice President of the School of Medicine, Aichi Medical University. He is also the Executive Vice President and Director of the Endoscopy Center, Aichi Medical University Hospital. He holds the society membership of the Japanese Society of Internal Medicine, Japanese Society of Gastroenterology, Japan Gastroenterological Endoscopy Society, and American Gastroenterological Society.