Endoscopic Image Clustering with Temporal Ordering Information Based on Dynamic Programming*

Shota Harada[1], Hideaki Hayashi[2], Ryoma Bise[2,3], Kiyohito Tanaka[4], Qier Meng[3], and Seiichi Uchida[2,3]

***Abstract*— In this paper, we propose a clustering method with temporal ordering information for endoscopic image sequences. It is difficult to collect a sufficient amount of endoscopic image datasets to train machine learning techniques by manual labeling. The clustering of endoscopic images leads to group-based labeling, which is useful for reducing the cost of dataset construction. Therefore, in this paper, we propose a clustering method where the property of endoscopic image sequences is fully utilized. For the proposed method, a deep neural network was used to extract features from endoscopic images, and clustering with temporal ordering information was solved by dynamic programming. In the experiments, we clustered the esophagogastroduodenoscopy images. From the results, we confirmed that the performance was improved by using the sequential property.**

I. Introduction

Endoscopic imaging is an important modality for the discovery of various diseases such as stomach cancer, tumors, and ulcerative colitis. An endoscope is used to perform the surgery, for example, for lung cancer and pneumothorax. A common property of endoscopes is that they take a sequence of images according to their camera movement. For example, one observation procedure with esophagogastroduodenoscopy (EGD) takes approximately 30 images. Since the camera movement of an EGD is almost smooth, consecutive images often capture neighboring parts of the same digestive organs.

Presently, it is natural to use some machine learning technique for supporting endoscopic diagnosis by computers. For example, if we can train a deep neural network (DNN), especially a convolutional neural network (CNN) with a sufficient number of training images, we can expect that CNN can estimate the location of the camera position in digestive organs accurately. In addition, we can also expect that CNN can detect some diseases with the estimated location information. To the authors' best knowledge, the collection of a large number of endoscopic images for the training of CNNs is still an open problem. Considering that there is a great variety of camera locations (i.e., observation points inside organs) and diseases, a considerable number of training images must be collected *with an accurate label (, or ground-truth)*. Moreover, this number should be increased significantly by considering that the appearance of an organ will vary for different persons.

The collection of labeled images for a sufficient training set, however, is almost intractable because a significant effort is required by endoscopists to attach a reliable label to each image. Although we can resort to crowdsourcing services (such as Amazon MechanicalTurk) for the attachment of image labels for general object image recognition, we cannot use them for endoscopic images due to the necessity of expert knowledge. We thus need to develop a system that reduces the labeling effort of endoscopists.

In this paper, we propose a new unsupervised learning method for clustering (unlabeled) endoscopic images. The technical highlight of our clustering method is that it utilizes the sequential property of endoscopic images. As mentioned previously, consecutive images may capture neighboring parts; thus, there is a higher probability that the same label is attached. We can utilize this property for the clustering process. Specifically, we can formulate the clustering problem as an unsupervised segmentation problem of an image sequence. In this paper, clustering is performed individually for each image sequence.

The main contributions of this work are as follows:

- We proposed an image clustering algorithm with temporal ordering information. The algorithm can guarantee the global optimality of its solution.
- We showed the effectiveness of using temporal ordering information for the endoscopic image clustering.

II. Related Work

A possible strategy to collect labeled images for a sufficient training set is group-based labeling [1]. If it is possible to gather images, similar in appearance, as a cluster, we can accelerate the labeling process drastically. In the ideal case that an endoscopist observes all the images in a cluster (by listing them on a screen) and finds that all can be labeled the same, they can attach the same label to all the images simultaneously. Even in a more general case where a cluster contains several outliers, it is easier to find and remove them by observing all the images of the cluster at once. This is because outliers will slightly differ in appearance from inliers (i.e., the majority of the cluster members), and this slight difference causes a significant *visual saliency*.

Some studies have been reported the classification of endoscopic images using DNN [2], [3]. In [2], Takiyama

*This work was partially supported by AMED Grant Number JP18lk1010028.

[1]S. Harada is with Graduate School of Systems Life Sciences Kyushu University, 744 Motooka Nishi-ku Fukuoka, 819-0395, Japan (email: shota.harada@human.ait.kyushu-u.ac.jp)

[2]H. Hayashi, R. Bise, and S. Uchida are with Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka Nishi-ku Fukuoka, 819-0395, Japan

[3]B. Ryoma, Q. Meng, and S. Uchida are with Research Center for Medical Bigdata, National Institute of Informatics, Japan

[4]K. Tanaka is with Department of Gastroenterology, Kyoto Second Red Cross Hospital, Japan

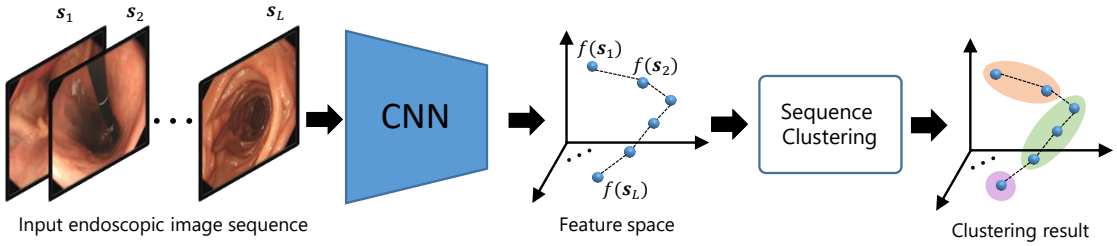

Fig. 1. Overview of the proposed method. The proposed method comprises feature extraction and clustering. The left side indicates feature extraction using a DNN. The right side indicates clustering of the feature vector based on DP. The dashed lines between the features in the feature space indicate that they are temporally adjacent.

et al., EGD images were classified using the CNN trained by a supervised learning method. Moreover, in [3], Jamil *et al.* proposed a classification method for disease classification from endoscopic images. This method was established by a support vector machine using the compact image representation feature vectors transformed from the feature map of a pre-trained CNN.

Owing to the contribution of many studies, various clustering methodologies have been proposed (e.g., [4], [5], [6]). For example, Mai *et al.* proposed a clustering framework that incorporates instance-level and temporal smoothness constraints for coping with a large temporal data [4]. Moreover, in [5], a method of face clustering as a video was proposed. This method achieves constrained clustering by providing constraints of cannot-link and must-link between images using the information of a face image series detected from video. By applying constraints, these methods performed better than some existing methods. From the above, it is effective to introduce domain knowledge, such as must-link and temporal smoothness constraints, into the clustering method to improve performance.

III. Endoscopic Image Sequence Clustering

Fig. 1 provides an overview of the proposed method. The proposed method comprises two parts: feature extraction and clustering. The former extracts a feature vector from each frame of the input endoscopic image sequence using a DNN. The latter conducts clustering of the feature vectors using temporal ordering information.

A. Feature Extraction from Endoscopic Images

A feature vector was extracted from each endoscopic image to enhance the clustering performance. The reason for conducting feature extraction is twofold. One is to reduce the effects of appearance changes. The appearance of an endoscopic image can considerably change by a slight difference of shooting angle and lighting condition, even in the same part of the same organ. The other is the curse of dimensionality. In general, the original image space is high-dimensional. Clustering in the high-dimensional space is inappropriate because the distance metrics become meaningless. Therefore, generally, important features are extracted from the original images before clustering is performed.

For the feature extractor in the proposed method, we used a DNN with CNN layers (DCNN). In [7], it have demonstrated that a CNN trained on a large dataset also learned the ability to extract important information from samples outside the original dataset. We can deal with the problems of appearance change and high dimensionality by using the DCNN. In the proposed method, we used a DCNN pre-trained with a general machine learning task, such as classification and segmentation, where a large number of labeled images were obtained.

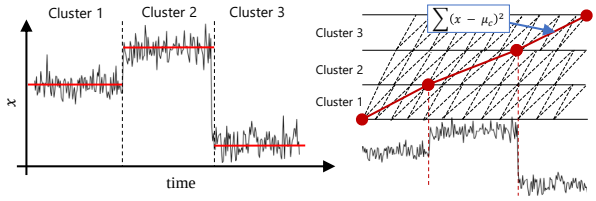

Fig. 2. Outline of our sequence clustering. (Left) An example of the sequence clustering. The vertical lines are cluster boundaries. The horizontal red lines indicate the average values as the representative values of individual clusters. This figure implies that our sequence clustering can be seen as temporal segmentation. (Right) The algorithm to have the globally optimal cluster boundaries. The sequence clustering problem is formulated as an optimal path finding problem and thus can be solved by DP.

B. Clustering with Temporal Ordering Information

Fig. 2 outlines the clustering part. The clustering step of the proposed method was performed by optimizing the minimization of the variance in each cluster as an objective function. Let $\boldsymbol{X} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots; \boldsymbol{x}_L)$ be a sequence of length L, where $\boldsymbol{x}_i$ is a feature vector extracted by DCNN. A k-clustering C of a sequence $\boldsymbol{X}$ is a partition of $1; 2; \cdots L$ into k contiguous subsequences (clusters), $C = fc_1; c_2; \cdots; c_k g$. Each cluster c_i consists of $jc_i j$ images and has an average vector $\boldsymbol{\mu}_c$ as the representative vector. The objective function is defined as

$$C_{\text{opt}}(\boldsymbol{X}; k) = \arg\min_{C \in \mathcal{C}_{L,k}} \sum_{c \in C} \frac{1}{jcj} \sum_{\boldsymbol{x} \in c} (\boldsymbol{x} \quad \boldsymbol{\mu}_c)^2; \tag{1}$$

where C_{opt} is the optimal cluster, and $C_{L,k}$ is the set of all combinations k-clustering of sequence of length L.

To solve (1), we employed dynamic programming (DP) [8]. While DP has been used for optimal nonlinear matching between sequences (i.e., dynamic time warping) and optimal pathfinding, it can solve the temporal segmentation problem, as in Fig. 2. The merit of DP is that it gives a globally optimal solution of (1) within a polynomial time. This is

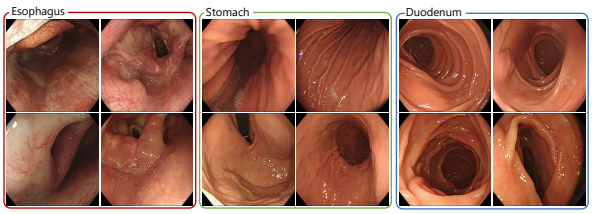

Fig. 3. Examples from the EGD images dataset. This dataset contains the endoscopic images of the esophagus, duodenum, and stomach.

a significant advantage over typical clustering methods that cannot guarantee the quality of their solution.

IV. Experiment

We conducted an experiment to evaluate the capability of the proposed method for endoscopic image clustering. In this experiment, the clustering of organs from an EGD image sequence was performed by unsupervised learning.

A. Experimental Setup

Fig. 3 shows examples from the EGD image dataset. In Fig. 3, the left, middle, and right parts are the esophagus, stomach, and duodenum, respectively. This dataset comprises 3,273 samples of an endoscopic image sequence. Each image sequence includes the esophagus, duodenum, and stomach, and the sequences have variable-length. As the ground truth, each image was manually categorized into one of three organs by a medical expert. Each image has three channels: red, green, and blue. The size of each image is 224 224 pixels, and the maximum and minimum sequence length of the image sequences are 91 and 11, respectively.

For the feature extractor of the proposed method, we used densely connected convolutional networks (DenseNet) [9]. DenseNet consists of DenseBlock that is constructed by a skip connection and a CNN, and a transition layer that is constructed by a CNN and an average pooling layer. In this experiment, we used the parameters of DenseNet trained with ImageNet beforehand.

As a performance measure, normalized mutual information (NMI) was computed between the ground truth and the clustering result. NMI is widely used to evaluate the quality of clustering. The range of NMI is 0 to 1, and a large value indicates a good clustering result.

We compared the results using the proposed method with those using two existing clustering methods. The first method is k-means using a feature vector extracted by DenseNet from the EGD images. The other is k-means using raw EGD images as feature vectors.

The number of clusters in the proposed method was determined for each image sequence based on Akaike's information criterion (AIC) because there is a possibility of being returned to the previous part and shot again when shooting a image sequence. In contrast, the number of clusters of the comparative methods was set as 3 according to the true number of classes.

TABLE I

Mean and standard deviation of NMI.

	mean	std	paired t-test
Ours	0.42	0.09	
CNN with k-means	0.26	0.13	***
Raw with k-means	0.13	0.08	***

***: $p < 0.001$

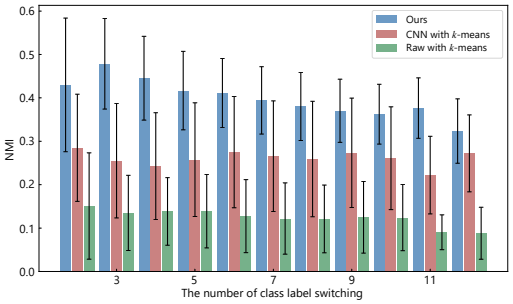

Fig. 4. NMI for the number of class switching.

B. Experimental Results

Table I shows the evaluation result of each clustering method. The average value and standard deviation of the NMI are shown in each row. Furthermore, the significance of the proposed method is indicated by a paired t-test, which was corrected based on Bonferroni correction. Fig. 4 shows the evaluation result of each clustering method with respect to the number of class switching in a sequence. When the number of switching of class labels increases, since the performance of the proposed method may be degraded, we confirm the performance of clustering on the image sequences for each number of class label switching. In Fig. 4, the result of the number of class switching from 2 to 12 is shown.

Fig. 5 shows the result obtained by the proposed method. The clustering results of the comparative methods are also shown for comparison. Figs. 5 (a) show the best and worst case for our clustering result, respectively. Moreover, Fig. 5 (c) shows the best result of clustering obtained by k-means using the CNN feature vectors. The number of clusters of the proposed method determined based on AIC in Figs. 5 (a), (b), and (c) is 5, 3, and 5, respectively.

C. Discussion

Table I confirms the effectiveness of the proposed method for clustering of the endoscopic image sequence. Since the standard deviation of our result was smaller than that of the comparative methods, the proposed method is robust to differences in characteristics among samples. In addition, Fig. 4 shows that our results perform best in the data separated by the number of switching class labels. The clustering performance of the proposed method decreased according to the number of switching class labels when the number of class labels was over 3. However, the clustering performance of our method exceeded the performance of the comparison methods for any number of switching class labels.