2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Wireless Capsule Endoscopy Video Summarization: A Learning Approach Based on Siamese Neural Network and Support Vector Machine

Jin Chen, Yuexian Zou*, Yi Wang
ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen 518055, China
*{zouyx@pkusz.edu.cn}

***Abstract*—Wireless capsule endoscopy video summarization (WCE-VS) is highly demanded for eliminating redundant frames with high similarity. Conventional WCE-VS methods extract various hand-crafted features as image representations. Researches show that such features only reflect the low-level characteristics of single frame and essentially are not effective to capture the semantic similarity between WCE frames. Motivated by the salient property of Siamese neural network (SNN) in mapping similar image pairs closer while mapping dissimilar image pairs apart in the feature space, a novel learning-based WCE-VS method is proposed in this paper. Specifically, with the availability of labelled similar and dissimilar pairs of WCE frames, SNN is trained with a contrastive loss function to extract high level semantic features. Furthermore, for similarity judgment, to avoid the challenge of manually setting optimal threshold in conventional methods, we creatively cast it into a supervised classification problem implemented by a linear SVM. Extensive experiments validate the effectiveness and efficiency of our proposed method.**

Keywords—Wireless capsule endoscopy; video summarization; Siamese neural network; linear SVM classifier; similarity judgment

I. INTRODUCTION

Wireless Capsule Endoscopy (WCE) is a novel medical equipment for non-invasive gastrointestinal disease detection. After WCE is swallowed by the patient, one whole WCE examination process will capture 30000-80000 frames. However, as shown in Fig. 1(a), there are a large number of redundant frames with high similarity in the WCE video [1]. The major obstruction associated with this advanced technology is that such huge amount of frames will cause long view time for clinicians. Hence, it is of great significance to develop a video summarization algorithm used for eliminating these redundant WCE frames.

Literature reviews of the general video summarization technique (VST) show that the majority of VSTs contain following steps as shown in Fig. 1(b). Firstly, for each frame, the extraction of effective features is conducted. Secondly, the shot segmentation by judging the similarity between adjacent frames sequentially is carried out. As a result, the video could be segmented into different shots, in each of which the frames are with certain similarity. Finally, the key frames are extracted in each shot.

For WCE video summarization (WCE-VS), there has already existed numerous researches following the general VST. Jia Sen Huo [1] extracts each WCE frame's features based on HSV color histogram and a block edge directivity descriptor method. Through

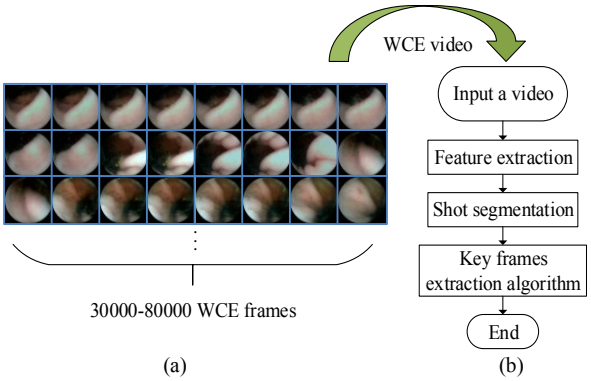

Fig. 1. Video summarization for WCE videos: (a) is a WCE video with a large amount of redundant frames, (b) is the general video summarization framework.

experiments, a fixed threshold is selected to cut the WCE video into different shots. Lastly, key frames are extracted using a relation matrix rank method. Yixuan Yuan [2] fuses HSV color features, LBP texture features and HOG shape features based on the information entropy. By comparing the distance of information entropies between two consecutive frames with an automatic threshold, the original WCE video is segmented into several shots. For each shot, AP clustering method is adopted to select key frames. Jin Chen [3] takes both HSV color histogram-based color features and GLCM-based texture features into consideration. A *W*-parametric mean value threshold is set adaptively to judge the similarity between adjacent WCE frames sequentially, for making shot segmentation. Eventually, in each shot, the key frames are extracted by an adaptive K-means clustering algorithm.

In general, all these methods discussed above are fairly effective to meet their own demands. However, considering the endless of the need for accuracy and efficiency in medical diagnosis, we are encouraged to develop more effective WCE-VS method. According to our analysis on their methodologies, it can be concluded that feature representation of frames and the similarity judgment between frames are essential to these methods. Careful evaluation shows that the existing WCE-VS methods employed the commonly used feature descriptors, such as HSV histogram, GLCM feature, HOG feature and so on. Obviously, these features are considered as hand-crafted image representations in computer vision [4], which reflects the low-level characteristics (color, texture, shape, etc.) of single frame. Analysis and experimental results show that such features are not able to capture the high-level semantic similarity between WCE frames, which would result in unsatisfactory summaries because of the

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

semantic gap [5]. Meanwhile, the fusion of hand-crafted features in [3] commonly takes high computational cost. Besides, the existing WCE-VS methods normally face a practical and challenging problem, which is manually setting a suitable threshold for similarity judgment in making the shot segmentation. These observations motivate us to develop a learning-based approach to automatically learn high-level semantic features, and discriminate the similar or dissimilar WCE frames without setting any threshold.

According to recent research outcomes [6-8], we note that Siamese neural network (SNN) is able to learn high-level semantic features suitable for discriminating pairs of similar and dissimilar images. Specifically, a contrastive loss function of SNN is optimized so that the Euclidean distance of similar pairs in feature space is small while that of dissimilar pairs is large. Inspired by the salient property of SNN, a learning-based approach is developed for WCE-VS, where SNN is employed to learn the high level semantic features with similar and dissimilar pairs of WCE frames as training dataset. Then, in order to avoid manually setting a threshold for similarity judgments, the Euclidean distance corresponding to similar and dissimilar pair of WCE frames is taken as input to train a linear SVM classifier for similarity judgment without any threshold. After that, through judging the similarity of adjacent WCE frames, the WCE video is segmented into different shots. Lastly, as gradient varying characteristic exists in each shot, we employ the adaptive K-means clustering algorithm [3] to preserve key frames and remove redundant frames for each shot. According to the essential working principle in developing WCE-VS, our proposed method is termed as WCE-VS-SNN-SVM for clear and concise presentation in this paper.

The rest of the paper is organized as follows. Section II describes the feature extraction based on Siamese neural network. Section III presents the shot segmentation by a linear SVM classifier. Section IV describes key frame extraction procedure via an adaptive K-means clustering algorithm. Then, experiments and performance analysis are presented in Section V. Section VI concludes the paper.

II. Feature Extraction Based on Siamese Neural Network

Our designed Siamese neural network (SNN) is similar to the one used in [9]. It consists of two deep convolutional neural networks (CNN) and a loss layer [10] and its diagram is shown in Fig. 2. In SNN, the two CNNs share the same parameters $\boldsymbol{W}$ and they are taken as feature extractors (CNNFE) [11]. According to the outputs of the two CNNFEs, the loss value is calculated in the loss layer for training the whole Siamese neural network. Our goal is to get the well-learned CNNFE being able to extract high-level semantic features of each WCE frame, after the training procedure of SNN.

For Fig. 2, the inputs are the training pairs of WCE frames and the binary label of the pairs, denoted as $\{\boldsymbol{X}_i, \boldsymbol{X}_i', Y_i\}(i=1,2,\ldots,n)$. Specifically, $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ are the i-th pair of WCE frames. Y_i is the label of the i-th pair of WCE frames. $Y_i = 0$ if the WCE frames $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ are deemed similar and $Y_i = 1$ if they are deemed dissimilar. n means the number of training pairs of WCE frames.

The CNNFE contains 7 layers, including two convolution layers, two max pooling layers, two fully-connected layers, and a "feat" layer which is actually a fully-connected layer. The layer parameters of the CNNFE are shown in Table I. We use Rectified Linear Units (ReLu) as the non-linear activation function in convolution layers and fully-connected layers. By the forward pass of CNN, the output of the CNNFE can be denoted as $G(\boldsymbol{X}_i, \boldsymbol{W})$ which represents the extracted feature vector of WCE frame $\boldsymbol{X}_i$, where $G(.)$ is the non-linear mapping function of CNNFE and $\boldsymbol{W}$ is the shared parameters of the two CNNFEs.

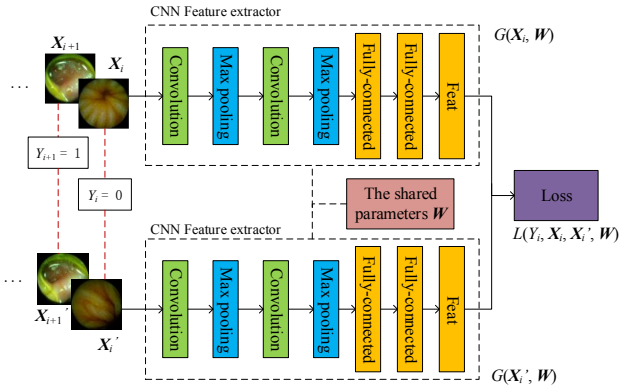

Fig. 2. The diagram of the training process of our designed SNN.

TABLE I. The Architecture Parameters of CNNFE.

Layer	Type	Number of maps or neurons	Kernel size	Stride
1	Convolution	20	5×5	1
2	Max pooling	20	2×2	2
3	Convolution	50	5×5	1
4	Max pooling	50	2×2	2
5	Fully-connected	500	--	--
6	Fully-connected	300	--	--
7	Feat	100	--	--

In the loss layer, the contrastive loss function [9] is adopted to measure the loss value for training the Siamese neural network. Firstly, the parameterized Euclidean distance between $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ is calculated as:

$$D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W}) = \left\| G(\boldsymbol{X}_i, \boldsymbol{W}) - G(\boldsymbol{X}_i', \boldsymbol{W}) \right\|_2 \tag{1}$$

where $D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ represents the similarity degree of the i-th pair of WCE frames. The lower $D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ is, the more similar $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ will be. As a result, the contrastive loss function for the i-th pair of WCE frames $L(Y_i, \boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ can be formulated as:

$$\begin{aligned} L(Y_i, \boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W}) = (1 - Y_i) \times L_S\left(D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})\right) \\ + Y_i \times L_D\left(D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})\right) \end{aligned} \tag{2}$$

$$L_S(D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})) = \frac{1}{2} \times \left(D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})\right)^2 \tag{3}$$

$$L_D\left(D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})\right) = \frac{1}{2} \times \left\{\max\left(0, m - D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})\right)\right\}^2 \tag{4}$$

where $L_S(.)$ in (3) is a loss function for labeled pairs of similar WCE frames. $L_D(.)$ in (4) is a loss function for labeled pairs of dissimilar WCE frames, and $m > 0$ is a margin ($m = 1$ in this paper). It is noticeable that $L_S(.)$ is a monotone increasing loss function and $L_D(.)$ is a monotone decreasing loss function. During the training procedure to minimize $L(Y_i, \boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ in (2), it would result in low values of $D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ for similar pairs and high values of $D(\boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W})$ for dissimilar pairs [12]. Finally, the objective function of our designed SNN is given by:

$$\arg\min_{\boldsymbol{W}} \sum_{i=1}^{n} L\left(Y_i, \boldsymbol{X}_i, \boldsymbol{X}_i', \boldsymbol{W}\right) + \frac{\lambda}{2} \times \|\boldsymbol{W}\|_2^2 \tag{5}$$

where λ is a regularization parameter. In our study, we employ the stochastic gradient descent (SGD) method to obtain the optimal

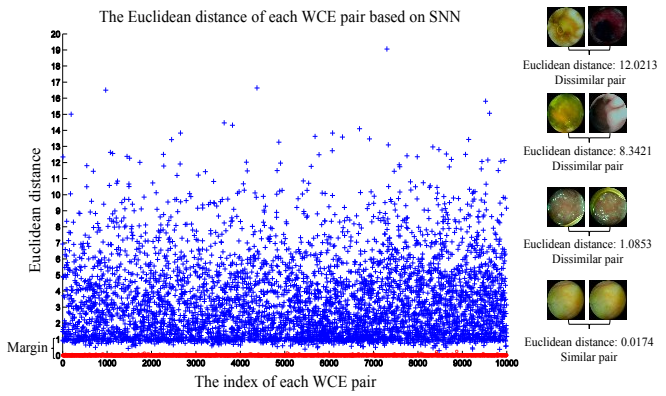

Fig. 3. The Euclidean distances of 10000 simialr and dissimilar WCE pairs in our learning-based feature space. Red circle (o) indicates the Euclidean distance generated from similar pair while blue cross (+) denotes that from dissimilar pair.

solution $\boldsymbol{W}^*$ iteratively [13]. The gradient can be calculated by back-propagation algorithm through the loss, and the total gradient is the sum of the contributions from two CNNFEs [9]. The optimal value of the loss function would converge to a small and stable value, indicating the Siamese neural network has been trained properly. After the training process, the optimal parameters $\boldsymbol{W}^*$ of the CNNFE is learned. Then, for any input $\boldsymbol{X}_i$ to the CNNFE, its corresponding feature is generated as:

$$\boldsymbol{Fea}_i = G\left(\boldsymbol{X}_i, \boldsymbol{W}^*\right) \tag{6}$$

where $\boldsymbol{Fea}_i \in \mathcal{R}^{p_7}$ is the generated feature of the WCE frame $\boldsymbol{X}_i$, p_7=100 is the number of neurons in the "feat" layer of CNN feature extractor.

For visualizing the effectiveness of our learned features by SNN, we conduct a simple experiment: the Euclidean distances of 10000 pairs of WCE frames are computed in the feature space and shown in Fig. 3, where red circle (o) indicates the Euclidean distance generated from similar pair while blue cross (+) denotes that from dissimilar pair. Carefully analyzing the results shown in Fig. 3, we observe: 1) there is a distinguishable margin between the red circles and blue crosses; 2) the Euclidean distances of different similar pairs tend to zero while that of different dissimilar pairs scatter. This phenomena is intuitive since different dissimilar pair has different dissimilar degree. To verify this opinion, we choose 4 image pairs shown in the right column of Fig. 3. From bottom to top, it is clear to see that the Euclidean distances are listed from small to large (0.0174 to 12.0213). Correspondingly, we can see the similarity degree of the image pairs decreases. Actually, only the bottom pair is labeled as similar pair and other three pairs are labeled as dissimilar pairs. Be worth mentioning the training dataset is labeled by gastroenterologist. From above theoretical derivation and experimental results, we are confident that the trained SNN is able to capture the higher level semantic feature embedded in the training dataset.

III. Shot Segmentation by a Linear SVM Classifier

In video summarization, video shots are the basic structural building blocks of a video sequence [14]. By analyzing the WCE videos, there just exists sudden shot transition, without any gradual transitions which often appear in public videos. Therefore, by judging whether the adjacent WCE frames are similar or not sequentially, the shot boundaries are detected in the positions where dissimilar adjacent WCE frames occur.

According to the observation of Fig. 3, it is clear that the similar and dissimilar WCE pairs can be classified by setting a threshold. However, the manual threshold often makes the algorithm sensitive to unseen data, so the selection of a suitable threshold is a challenging work. Besides, we also observe some outliers in Fig. 3, which may affect the selection of the threshold. To avoid this problem, it triggers us to choose the linear SVM classifier to automatically obtain the optimal margin by the availability of the pair labels. Hence, in this section, we propose a novel method to realize shot segmentation of WCE videos, where a linear SVM classifier is trained to distinguish the dissimilar adjacent WCE frames without setting any threshold.

Specifically, the training dataset for the linear SVM classifier is the same with the dataset used for training the Siamese neural network, which contains the similar and dissimilar pairs of WCE frames. As described in Section II, high-level semantic features are generated by well-learned CNNFE of SNN. From Fig. 3, the Euclidean distances calculated in such feature space are distinguishable for similar and dissimilar pairs. Therefore, it is straightforward to consider the Euclidean distances corresponding to a pair WCE frames in feature space as a new feature for similarity judgment, which is calculated as:

$$Dis_i = \left\| \boldsymbol{Fea}_i - \boldsymbol{Fea}_i' \right\|_2 \tag{7}$$

where Dis_i is the Euclidean distance of the i-th pair of WCE frames, $\boldsymbol{Fea}_i$ and $\boldsymbol{Fea}_i'$ are the extracted features for $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$, respectively. It is obvious that when Dis_i is small/large, ($\boldsymbol{X}_i$, $\boldsymbol{X}_i'$) should be a similar/dissimilar pair. Essentially, this is a binary classification problem and a linear SVM classifier is a good candidate for its efficiency and good performance [16]. The label of ($\boldsymbol{X}_i$, $\boldsymbol{X}_i'$) is denoted as Y_i^{SVM}, where Y_i^{SVM} = +1 when $\boldsymbol{X}_i$ is similar to $\boldsymbol{X}_i'$ and Y_i^{SVM} = -1 when they are dissimilar. With these descriptions, the classification problem can be achieved by the following optimization problem [15]:

$$\begin{aligned} \arg\min_{\mathbf{w}_{\mathrm{SVM}}} \ & \frac{1}{2} \times \mathbf{w}_{\mathrm{SVM}}{}^T \mathbf{w}_{\mathrm{SVM}} + \\ & C \times \sum_{i=1}^{n} \left[\max\left(1 - Y_i^{\mathrm{SVM}} \mathbf{w}_{\mathrm{SVM}}{}^T Dis_i, 0\right) \right]^2 \end{aligned} \tag{8}$$

where $\mathbf{w}_{\mathrm{SVM}}$ is the parameters of linear SVM classifier, C > 0 is a penalty parameter. In our study, the linear SVM classifier is implemented by a well-known open source library LibSVM [16]. For a test pair, its label Y_{test} can be predicted by:

$$Y_{\mathrm{test}} = \begin{cases} +1 & \mathbf{w}_{\mathrm{SVM}} Dis_{\mathrm{test}} \geq 0 \\ -1 & \mathbf{w}_{\mathrm{SVM}} Dis_{\mathrm{test}} < 0 \end{cases} \tag{9}$$

where Dis_{test} is the Euclidean distance of the test WCE pair.

For presentation clarity, the proposed shot segmentation algorithm is illustrated in Fig. 4. It is easy to understand that the Euclidean distance between adjacent WCE frames need to be calculated firstly. Accordingly, the adjacent WCE frames can be classified as similar or not based on the output of the linear SVM classifier. When the output of the linear SVM classifier turns to be -1, the position of the shot boundary is determined accordingly. In the end, the WCE video can be segmented into different shots, in which WCE frames bear temporal correlation and spatial similarity.

IV. Key Frame Extraction By An Adaptive K-means Clustering Algorithm

Research [3] has shown that one of the differences between WCE video and public video is the gradient changing characteristic in one shot. One example is shown in Fig. 5. In one WCE video shot, the adjacent WCE frames are of high similarity, but it is also noted that

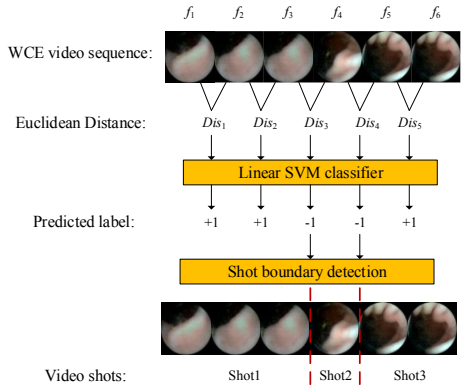

Fig. 4. Shot segmentation based on linear SVM classifier.

the first frame may be deemed dissimilar to the last frame by medical professionals since they may carry on different medical information. However, for the public video shot (from the Open Video Project [20]), the frames in the same shot usually represent the same scene, so the first frame is similar to the last one in one shot. This observation from WCE videos warns us we cannot just adopt the conventional key frame extraction methods which just select the first frame or the last frame as the key frame. A new key frame extraction method should be developed for WCE-VS to guarantee less information loss. In the following context, to extract key frames for each shot, a K-means clustering algorithm [3] is taken. The proposed key frame extraction method for each WCE shot is summarized as follows.

Algorithm 1: K-means clustering method for WCE-VS

Input: a WCE video shot: $\{f_1, f_2, f_3, \ldots, f_N\}$,where N is the number of WCE frames in the shot.

Output: K clusters $\{Class_1, Class_2, \ldots, Class_K\}$, K key frames $\{Key_{Class_1}, Key_{Class_2}, \ldots, Key_{Class_K}\}$

Step1: Initialization: the number of classes $K = 1$;$Class_1 \leftarrow f_1$; the centroid of $Class_1$ (denoted as $Center_{Class_1}$) $\leftarrow f_1$;

Step2: Extract feature vectors for each WCE frame by the CNNFE, denoted as $\{\boldsymbol{Fea}_1, \boldsymbol{Fea}_2, \ldots, \boldsymbol{Fea}_N\}$

Step3: Get the next frame f_i. If the frame is out of the WCE video shot, goto Step6; Else continue.

Step4: Calculate the Euclidean distance between f_i and $Center_{Class_n}$ (n=1, 2,…, K), denoted as $Dis(f_i, Center_{Class_n})$.

Step5: Find the cluster $Class_t$ which is closest to f_i:

$$Dis(f_i, Center_{Class_t}) = \min_{n=1,2,\ldots,K} (Dis(f_i, Center_{Class_n})) \tag{10}$$

Step6: Judge the similarity between f_i and $Center_{Class_t}$ by the linear SVM classifier, and the predicted label is denoted as Y_i'.

Step7: *Classification rules*:

$$\text{If} \quad Y_i' = -1 \Rightarrow \begin{cases} K = K + 1 \\ Class_K \leftarrow f_i \\ Center_{Class_K} \leftarrow f_i \end{cases} ; \tag{11}$$

$$\text{Else} \Rightarrow \begin{cases} Class_t \leftarrow f_i \\ \text{update } Center_{Class_t} \end{cases} ; Center_{Class_t} = \frac{1}{M} \times \sum_{m=1}^{M} \boldsymbol{Fea}_m^{Class_t}$$

where M is the number of frames in $Class_t$, $\boldsymbol{Fea}_m^{Class_t}$ is the m-th frame's feature vector in $Class_t$. Go to Step3.

Step8: Extract the key frames in the shot. The key frame is decided by selecting the frame which is closest to the centroid of each cluster. Therefore, the key frames of the shot are denoted as $\{Key_{Class_1}, Key_{Class_2}, \ldots, Key_{Class_K}\}$.

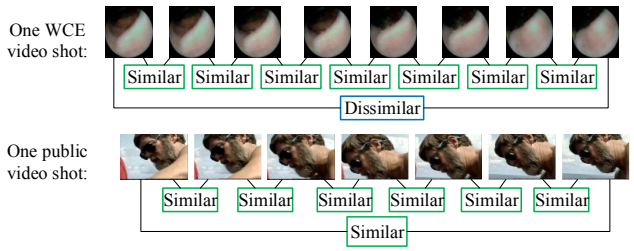

Fig. 5. The differences between WCE video shots and public video shots.

V. Experiments and Discussions

The experiments presented in this section aim to evaluate the performance of our proposed WCE-VS-SNN-SVM method. Firstly, we describe the learning process of the Siamese neural network and the linear SVM classifier in Section V-A. Then, the performance of WCE-VS-SNN-SVM is analyzed on WCE videos in Section V-B. Lastly, in order to evaluate the extensibility of our proposed method, we test it on a public dataset in Section V-C.

A. Learning Process of Siamese Neural Network and Linear SVM Classifier

The dataset for training SNN and the linear SVM classifier needs to be constructed firstly. There are 50 WCE videos which respectively have 30000-80000 WCE frames. These videos are captured from 50 patients. Under the guidance of gastroenterologist, 7875 pairs of similar WCE frames and 7986 pairs of dissimilar WCE frames are built from 3 WCE videos which are randomly selected from the 50 WCE videos. In order to make three-fold cross-validation, we construct 3 datasets, termed as DataSet1, DataSet2 and DataSet3, each of which contains 5250 similar pairs and 5324 dissimilar pairs as training dataset, 2625 similar pairs and 2662 dissimilar pairs as testing dataset.

For training SNN described in Section III, the input WCE frames are all set to the size of 96 × 96. Especially, the SNN is implemented by a well-known deep learning open source library Caffe [17]. The learning rate is one of the most important super-parameters to influence the learning ability of SNN. Fig. 6 shows the learning curves of SNN with different learning rates on the training dataset of DataSet1. It is clear to see that while the learning rate is set to 0.01, the training loss could converge to a stable and small value on about 20000 epochs. The same conclusions can be drawn on the training dataset of DataSet2 and DataSet3.

For the linear SVM classifier, the training dataset is the same as that of SNN, and the linear SVM classifier is implemented by a well-known open source library LibSVM [16]. From Table II, we find the testing accuracy rate can achieve 94.83% on average when using the linear SVM classifier to judge the similarity of two WCE frames.

B. Video Summarization Performance on WCE Videos

After getting the well-learned SNN and the linear SVM classifier, we can use them to make video summarization for WCE videos. For evaluating the performance of our proposed WCE-VS-SNN-SVM method, three general evaluation indicators including F-measure, compressed ratio (CR) and computation time T_{CT} are taken into account. F-measure is calculated by the two evaluation indicators, recall and precision. The F-measure value close to one indicates both high recall and precision, and vice versa [5]. Additionally, practical application demand higher F-measure, larger compression ratio and less computation time.

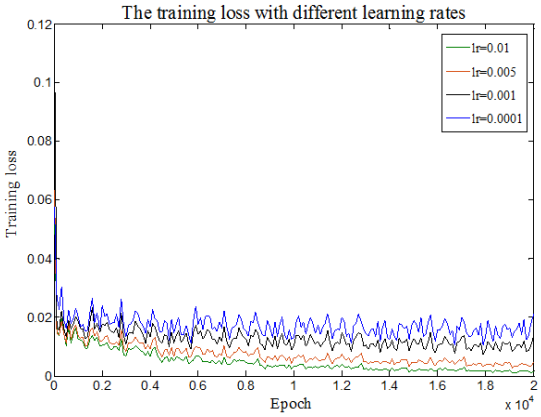

Fig. 6. The training loss with different learning rates.

TABLE II. THE ACCURACY OF SIMILARITY JUDGMENT BY THE LINEAR SVM CLASSIFIER.

Cross validation	DataSet1	DataSet2	DataSet3	Average
Training accuracy rate	100.00%	99.99%	99.99%	99.99%
Testing accuracy rate	95.18%	94.63%	94.69%	**94.83%**

$$Recall = TP/(TP+FN);\ Precision = TP/(TP+FP) \quad (12)$$

$$F\text{-measure}=2(Recall \times Precision)/(Recall + Precision) \quad (13)$$

$$CR = N_{redundant} / N_{total} \quad (14)$$

$$T_{CT} = T_{train} + T_{VS} \quad (15)$$

where true positive (*TP*) is the quantity of correct matches between the ground truth and the VS method. False negative (*FN*) is the number of frames which are in the ground truth but not in the result by the VS method. In contrast, False positive (*FP*) is the number of frames which are in the video summarization results but not in the ground truth [3]. $N_{\text{redundant}}$ is the amount of redundant WCE frames VS method yields. N_{total} denotes frames quantity in the WCE video. T_{CT} is computation time of the method, and it contains the time of training the model T_{train} and the time of video summarization T_{VS}.

The comparative methods are CCT-MRFE-RMR [1] and WCE-RIE [3], aiming at making video summarization on WCE videos. Both of the two methods use the hand-crafted features and threshold-based similarity judgment method. But CCT-MRFE-RMR set a fixed threshold through a large number of experiments, and WCE-RIE proposes a data-driven method to set threshold adaptively. The performance of these WCE video summarization methods are described as follows.

Comparison of *F*-measure: We construct 5 WCE video clips, and there are 500 consecutive frames in each clip. These frames don't exist in the training dataset. What's more, medical professionals have labeled the ground truth key frames for the 5 WCE video clips. The results of *F*-measure are shown in Table III. Comparing with CCT-MRFE-RMR and WCE-RIE, our proposed method WCE-VS-SNN-SVM achieves the highest average F-measure of 84.75%. Besides, the variance of *F*-measure in WCE-VS-SNN-SVM achieves 0.12% which is the lowest among all. It can be concluded that WCE-VS-SNN-SVM performs more effectively and robustly over different WCE data sets.

Comparison of *CR*: We randomly select 3 WCE videos (Video1, Video2 and Video3) from the 50 WCE videos excluding the 3 WCE videos for training. The results of *CR* are shown in Table IV. The

TABLE III. THE F-MEASURE RESULT COMPARISON.

Clips	CCT-MRFE-RMR	WCE-RIE	WCE-VS-SNN-SVM
	F-measure	***F-measure***	***F-measure***
Clip1	62.82%	73.40%	84.30%
Clip2	83.26%	90.12%	85.33%
Clip3	70.19%	83.57%	90.39%
Clip4	70.66%	79.61%	82.14%
Clip5	64.45%	83.02%	81.59%
Average	70.28%	81.94%	**84.75%**
Variance	0.65%	0.37%	**0.12%**

TABLE IV. THE *CR* RESULT COMPARISON.

Videos	Total number of frames	CCT-MRFE-RMR	WCE-RIE	WCE-VS-SNN-SVM
		CR	***CR***	***CR***
Video1	38421	88.31%	79.17%	86.13%
Video2	35403	83.57%	81.37%	84.72%
Video3	44173	87.09%	79.94%	84.25%
Average		86.32%	80.16%	**85.03%**

TABLE V. THE COMPUTATION TIME RESULT COMPARISON.

Method	Training time		The time of VS			
			Video1	***Video2***	***Video3***	***Speed***
CCT-MRFE-RMR	---		2002s	1872s	2758s	0.056s/frame
WCE-RIE	---		81070s	72511s	91494s	2.077s/frame
WCE-VS-SNN-SVM	SNN	SVM	568s	533s	651s	**0.015**s/frame
	150m	245s				

average *CR* of our proposed method achieves 85.03%, indicating it has an acceptable capacity to eliminate redundancy frames. Combining with the results of *F*-measure, our proposed method WCE-VS-SNN-SVM performs more effectively than WCE-RIE with both *F*-measure and *CR*. CCT-MRFE-RMR achieves the highest value of *CR*, the lowest average value of *F*-measure and the highest variance value of *F*-measure, which means CCT-MRFE-RMR can extract the fewest key frames but loss more informative WCE frames. In conclusion, WCE-VS-SNN-SVM can preserve few key frames while reserving much useful information with high effectiveness.

Comparison of computation time: We use the same 3 WCE videos (Video1, Video2 and Video3) to record the computation time by different VS methods. The results are obtained on a PC with Intel Core i5-3470 3.2GHz CPU, 8GB RAM and NVIDIA GeForce GTX 970 GPU. To avoid out of memory, in the process of video summarization, 500 WCE consecutive frames are processed each time until to the end of each WCE video. As shown in Table V, though our proposed method contains the time for training the SNN and SVM, the training process is off-line. The time for video summarization by our proposed method is much less than that by CCT-MRFE-RMR and WCE-RIE. Actually, the less computation time of video summarization is of great importance for practical application.

C. *Video Summarization Performance on a Publice dataset*

For evaluating the extensibility of WCE-VS-SNN-SVM, a public dataset containing 50 videos from Open Video Project [20] is adopted in this experiment. The compared methods are DT [19], OVP [20] and VSUMM [21] which are the baseline methods for video summarization on this public dataset. As described in Section IV, for the public video shot, the frames in the same shot usually represent the

same scene. Therefore, instead of the adaptive K-means clustering method in WCE-VS-SNN-SVM for key frame extraction, we simply select the frame nearest to the centroid of each shot as key frame. However, as described in Section III, the shot segmentation algorithm in our proposed method is specially designed for WCE videos where just sudden shot transitions exist. Thus WCE-VS-SNN-SVM has limits on the videos which have gradual shot transitions. For the public videos where just sudden shot transitions exist, WCE-VS-SNN-SVM still performs effectively. As shown in Fig. 7, the results of our proposed method WCE-VS-SNN-SVM contains more informative frames than that of DT and VSUMM, and less redundant frames than that of OVP.

VI. Conclusions

In this paper, we propose a learning-based WCE-VS method based on SNN and a linear SVM classifier. By using a large amount labeled pairs of similar and dissimilar WCE frames as training dataset, Siamese neural network is trained to learn high-level semantic features which are suitable for discriminating the similarity. According to the Euclidean distances of such labeled pairs in feature space, a linear SVM-based binary classifier is trained for similarity judgment without any threshold, which is used to make shot segmentation. Lastly, for removing redundant frames in each shot, an adaptive K-means clustering algorithm is employed to preserve key frames, based on the well-learned SNN and SVM. Extensive experiments prove the effectiveness and efficiency of our method on WCE videos and such public videos where just sudden shot transitions exist.

In the future work, we will try to consider more effective shot detection methods to improve the extensibility of our proposed method on public videos.

Acknowledgment

This work was partially supported by the Shenzhen Science & Technology Fundamental Research Program (No: JCYJ20150430162332418 and No: JCYJ20160330095814461).

References

[1] H. Jia Sen, Z. Yue Xian, and L. Lei, "An advanced WCE video summary using relation matrix rank," in *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, 2012, pp. 675-678.

[2] Y. Yixuan and M. Q. H. Meng, "Hierarchical key frames extraction for WCE video," in *Mechatronics and Automation (ICMA), 2013 IEEE International Conference on*, 2013, pp. 225-229.

[3] J. Chen, Y. Wang, and Y. X. Zou, "An adaptive redundant image elimination for Wireless Capsule Endoscopy review based on temporal correlation and color-texture feature similarity," in *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, 2015, pp. 735-739.

[4] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. Hand-Crafted Features for Pedestrian Gender Recognition," presented at the Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 2015.

[5] I. Mehmood, M. Sajjad, and S. Baik, "Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure," *Journal of Medical Systems,* vol. 38, pp. 1-9, 2014/07/19 2014.

[6] C. Liu, "Probabilistic Siamese Network for Learning Representations," University of Toronto, 2013.

[7] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3279-3286.

Fig. 7. The results of different VS methods on one public video where just sudden shot transitions exist (informative frames are marked on green box, redundant frames are marked on red box).

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 539-546 vol. 1.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1735-1742.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, pp. 2278-2324, 1998.

[11] X.-X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," *Pattern Recognition,* vol. 45, pp. 1318-1325, 2012.

[12] N. Carlevaris-Bianco and R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, 2014, pp. 2769-2776.

[13] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, ed: Springer, 2012, pp. 421-436.

[14] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 16, pp. 82-91, 2006.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *J. Mach. Learn. Res.,* vol. 9, pp. 1871-1874, 2008.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.,* vol. 2, pp. 1-27, 2011.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick*, et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," presented at the Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, 2014.

[18] F. Yanan, L. Haiying, C. Yu, Y. Tingfang, L. Teng, and M. Q. Meng, "Key-frame selection in WCE video based on shot detection," in *Intelligent Control and Automation (WCICA), 2012 10th World Congress on*, 2012, pp. 5030-5034.

[19] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," *International Journal on Digital Libraries,* vol. 6, pp. 219-232, 2006/04/01 2006.

[20] Open Video Project, "http://www.open-video.org/"

[21] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters,* vol. 32, pp. 56-68, 2011.