Spatio-Temporal Residual Networks for Slide Transition Detection in Lecture Videos

Zhijin Liu[1,2], Kai Li[1,2,*], Liquan Shen[1,2], Ran Ma[1,2], and Ping An[1,2]
[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
[e-mail: kailee@shu.edu.cn]
*Corresponding author: Kai Li

Received October 22, 2018; revised February 9, 2019; accepted February 25, 2019; published August 31, 2019

Abstract

In this paper, we present an approach for detecting slide transitions in lecture videos by introducing the spatio-temporal residual networks. Given a lecture video which records the digital slides, the speaker, and the audience by multiple cameras, our goal is to find keyframes where slide content changes. Since temporal dependency among video frames is important for detecting slide changes, 3D Convolutional Networks has been regarded as an efficient approach to learn the spatio-temporal features in videos. However, 3D ConvNet will cost much training time and need lots of memory. Hence, we utilize ResNet to ease the training of network, which is easy to optimize. Consequently, we present a novel ConvNet architecture based on 3D ConvNet and ResNet for slide transition detection in lecture videos. Experimental results show that the proposed novel ConvNet architecture achieves the better accuracy than other slide progression detection approaches.

Keywords: Lecture video, slide transition, 3D ConvNet, ResNet

This work was supported by the Project of National Natural Science Foundation of China(No. 61601278), “Chen Guang” project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (No. 17CG41), and the Program of Shanghai Academic Research Leader (No. 16XD1401200).

http://doi.org/10.3837/tiis.2019.08.011 **ISSN : 1976-7277**

1. Introduction

Nowadays, with the rapid development of Internet, there are many ways for users to acquire knowledge, in which online education is an important method. However, many raw lecture videos, recorded in classes and conference rooms, are time-consuming for learners. According to statistics, the number of uploaded videos on YouTube increases at 400 hours per minute. If these videos are not structured, users will be buried in the large number of videos. It discourages their passion on studying and reduces their study interests. To facilitate browsing, the video provider should display the lecture videos with section-wise annotations or titles, while the process is tedious and the large size of e-learning repositories today makes this approach cumbersome. However, it is quite difficult if a learner wants to quickly scan the contents of a particular lecture among a series of lecture videos. Therefore, automatic summarizing lecture videos is essential for e-learning and other applications.

In this paper, we focus on the slide transition detection, which is significant in lecture video summarization. The lecture videos are recorded by pan-tilt-zoom (PTZ) cameras and we distinguish the type of lecture video by the camera recording method, such as stationary camera, camera motion and camera switch. Because lecture videos may contain projected slides, the speaker, and the audience, some disturbance inevitably happens, e.g., camera motion, camera switch, and people movement. And, a slide transition always occurs in a short time along with the content changes on the projected screen, which is hard to find manually. It is a longstanding and challenge research topic to detect slide transition base on the spatial and temporal features across video frames.

Due to the complex noise interruption, there are few methods to detect slide transition across various types of lecture videos. Some approaches are designed to extract visual features such as color histogram, SIFT, HOG and wavelet to measure the appearance similarity between adjacent frames. However, these algorithms fail to take into consideration that people movement, camera motion and camera switch, such as switching from the computer screen to the speaker, often resulting in a significant appearance change of the video. Some other approaches are targeted for specific type of lecture video, such as a single PTZ camera without camera switch, and a fixed camera. These methods are limited to several specific types of lectures videos without universality.

In this paper, we present a novel ConvNet architecture based on 3D ConvNet and ResNet for detecting slide transition in lecture videos, as shown in **Fig. 1**. It is worth mentioning that Convolutional Neural Network (ConvNet) is playing a huge role in image classification and identification [1], [2]. Traditional ConvNets can extract video frame spatial features, which is important for understanding the image content, but it ignores the temporal evolutional among adjacent video frames, especially for slide change detection. Thus, we apply 3D ConvNet to learn the spatio-temporal features in videos with the convolution kernels extended from 2D to 3D. With the number of stacked layers, 3D ConvNet cost more memory and is difficult to train. To solve this problem, Residual Network (ResNet) is introduced [3]. Therefore, we combine 3D ConvNet and ResNet to propose a new C3D Residual block. The new ConvNet architecture cost less training time and is easier to optimize. In addition to slide transition detection, the saptio-temporal network has also shown its application in shot boundary detection, first-person video summarization, video retrieval and so on.

We divide the lecture videos into six types and compare our approach with other slide progression detection method for evaluation. Experimental results show that the proposed

novel ConvNet architecture is able to handle the six types of lectures videos and achieves the best accuracy than others.

To summarize, our contributions in this paper are: (a) a simple, yet effective C3D Residual block is presented to extract the spatio-temporal features for lecture videos. The new residual block eases the training of networks and can be easily combined with other network blocks; (b) we group several video frames into a frame volume who contains the slide change. Creatively, the slide transition detection in lecture videos turns into a classification problem for the spatio-temporal residual network model; (c) as shown in experiment, the spatio-temporal residual network is effective and achieve best results on six types of lecture videos even with complex camera motion and camera switch.

The rest of this paper is organized as follows. Sec. 2 review the related work of the slide transition detection. Sec. 3 describes the proposed ConvNet architecture based on C3D ConvNet and ResNet. Sec. 4 reports the experimental results. Sec. 5 concludes this paper.

2. Related Work

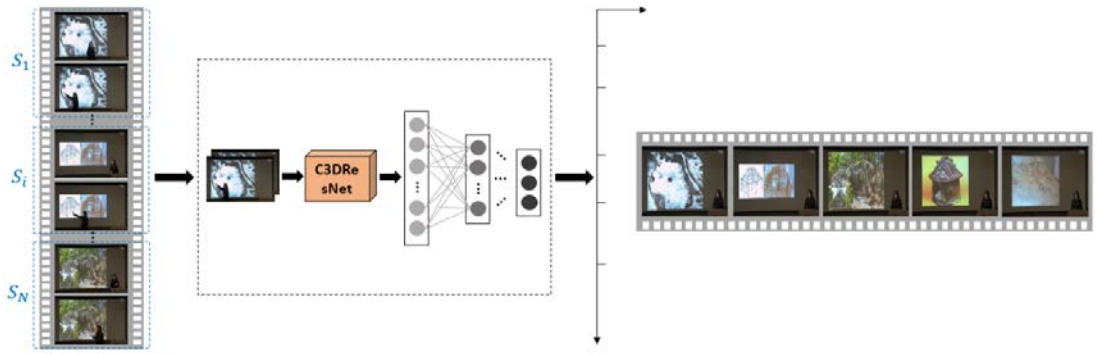
Some proposed methods of detecting slide transition, focus on the low level visual features, such as color histogram, corner points, Gabor filter and edge information, to measure the appearance similarity across frames. For instance, Ma et al. [4] analysed the color-histograms of images to segment a video into different shots. Whereas, the frames appearance will change with the camera motion. Hyun et al. [5] and Jeong et al. [6] extracted SIFT features, which are invariant to image scaling and rotation, and match them between two adjacent frames. The slide transition is detected when the similarity based on SIFT similarity is smaller than a threshold. The detecting accuracy are not reliable as the value of threshold is to be discussed. Li et al. [7] tracked the SIFT features across the whole video and the slide progression happens at the appearance and disappearance of mostly feature trajectories. However, it is limited to a single PTZ camera and camera switches are not allowed. Complementary features are also introduced in [8], [9] for visual similarity measurement. However, if there are camera motions in the lecture video, appearance difference is not always effective and parameter estimation for the similarity threshold is trivial.

To address these problems, shot boundary detection [10], [11], [12] and meta-data, such as audio signal [13], [14], transcripts [15], [16] approaches are proposed to summarize lecture videos. Subudhi et al. [10] compared frames histograms to detect shot transition and calculate edge function based on three contrast features to estimate content and non-content frames. Mohanta et al. [11] utilized local and global features to detect and classify shot boundary and then a multilayer perceptron network was applied to detect no change, abrupt change and gradual change frames. Cirne et al. [12] described a video frame by means of color co-occurrence matrices, then normalized sum of squared differences was used to detect shot boundary. He et al. [14] exploited spoken text, pitch and pause information in audio signal to show that these changes under various speaking conditions. Repp et al. [15] presented a standard linear text segmentation algorithm (LTSA) to segment the transcript into coherent sections for slide transitions detection. Lin et al. [16] combined speech text transcript, audio and video information to design an automated segmentation approach. Speech Recognition [17] can be also applied to segment the audio signal for lecture video summary. Qazi et al. [18] introduced a hybrid technique for speech segregation and calssification using the deep belief network (DBN) model. These approaches are not suitable for mostly lecture video without meta-data. In contrast, our method automatically detects slide transitions without additional data.

High level features, e.g., textual content in lecture videos, have important information and can be extracted by OCR (Optical Character Recognition) techniques [19], [20], [21]. Yang et al. [19] introduced a multi-hypotheses framework to recognize texts in slides, which includes text location, segment, OCR, spell checking and result output. Che et al. [20] reconstructed each texts content structure extracted from OCR results to segment the lecture video. Baidya et al. [21] presented a semantic segmentation method for lecture videos. They identified slide title and detects texts by OCR technology to generate keyframes. Besides, some approaches are proposed to reduce the semantic gap across images[22-24]. Mehmood et al. [22] presented an approach to reduce the semantic gap beteween low-level image features and high-level semantic concepts by collecting dense LIOP features and spatial histograms over four adapted triangular areas of an image. Sarwar et al. [23] introduced a novel BoW model, which perform visual words integration of LIOP and LBPV features to reduce semantic gap across images. However, the text reconstructed and similarity comparision after textural content extraction is complex. Compared with this, our system is simple and effective.

Our work is inspired by general video summarization approaches, such as [25], [26]. Yao et al. [25] splited the input videos into a set of video segments and a deep convolutional neural network with two components was designed to learn the spatial and temporal video representation for each video segment. Finally, highlight for each video segment was obtained by fusing two components output. Qiu et al. [26] proposed a Pseudo-3D Residual Net (P3D ResNet) based on various bottleneck building blocks to extract spatio-temporal video features in deep networks. In our system, a novel ConvNet architecture is introduced to detect slide transition for lecture videos rather than general videos.

Many previous approaches detect visual features to generate the summary by measuring the similarity between adjacent frames. However, if there are camera motions in the lecture video, appearance difference is not always effective. Besides, it is limited to a single PTZ camera and camera switches are not allowed. Moreover, parameter estimation for the similarity threshold is trivial. Other slide transition detection methods usually rely on additoonal meta-data, such as audio signal, transcripts and text dtection results. In contrast, our system performs spatio-temporal residual network to achieve automatic slide change detection without parameter estimation and any additional metadata and it is also applicable to multiple cameras recorded lecture videos. Experimental results show that our method successfully summarizes lecture video by key frames and achieves the best performance on six types of lecture videos.



(a) (b) (c) (d)

Fig. 1. C3D-ResNet pipeline (a) Input video; (b) Deep convolution neural networks; (c) slide transition detection result; (d) video summary.

3. Network Architecture

In this section, we present the C3D Residual block by combining 3D ConvNet and ResNet and analyze the spatio-temporal residual network in detail.

3.1 C3D Residual block

Spatial and temporal information in video frames is crucial for slide transition detection. 3D ConvNet is widely used as an efficient approach for extracting spatiotemporal features in videos. Compared to 2D ConvNet, 3D ConvNet apply 3D convolution and 3D pooling operations to model temporal information. We separate an input video into a set of video segments and each segment contains multiple frames which are represented as a frame volume. After implementing 3D ConvNet, the number of channels increases, but the size of each frame volume decreases. Eventually we obtain a feature map by merging adjacent frames in a video segment, as is shown in **Fig. 2**. The feature maps represent the spatial and temporal information across adjacent frames in a single video segment.

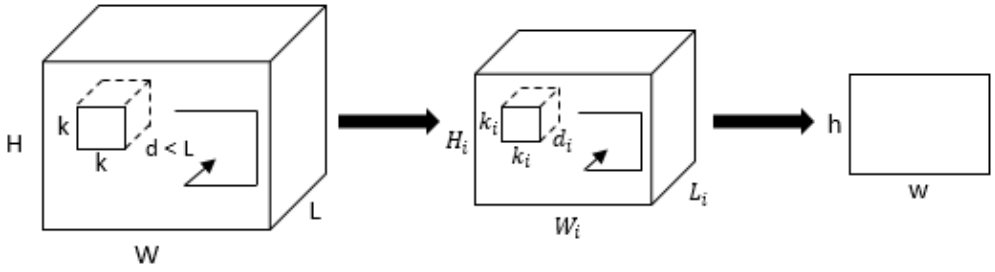
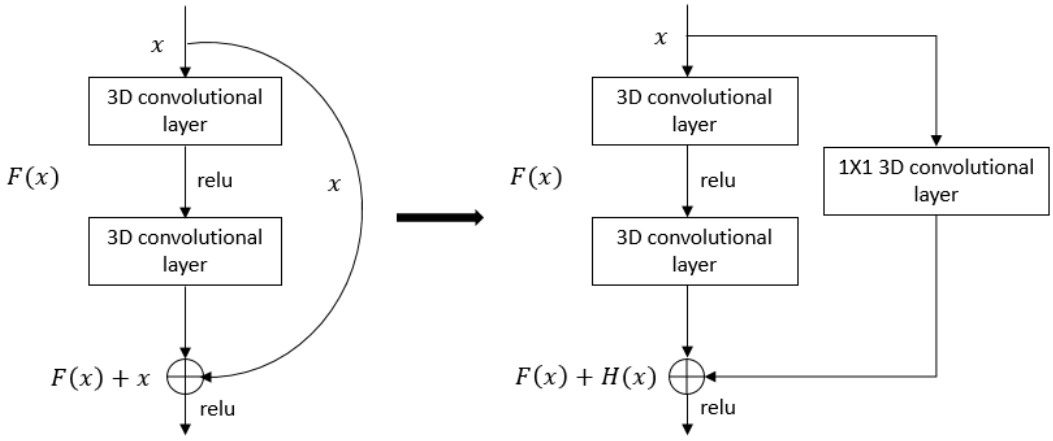


Fig. 2. 3D ConvNet overview

As shown in **Fig. 2**, the size of input frame volume is $H \times W \times L$ where H and W are the height and weight of the frame, and L is the length in number of frames. The kernel size of the 3D convolution is $k \times k \times d$ where k is the kernel spatial size and d is the kernel temporal depth. 3D convolutions simultaneously model the spatial information like 2D filters and construct temporal connections across frames. After implementing 3D ConvNet, the output volume reduces in size and it preserves the temporal information of the input signals. Finally, the feature map captures both spatial and temporal dimensions in the input video segment.

With stacking more network layers, the deep network leads to higher training/test error and the performance of the network degrades rapidly as reported in [27], [28]. There are complex reasons for such a situation, such as optimization, vanishing and exploding gradient problem, which has been largely solved by normalized initialization [29] and batch normalization [30]. With the increasing network depth, the network is not easy to optimize, and the deeper networks lead to saturated or degraded accuracy. The added number of nonlinear layers increases the difficulty of network optimization and it is not easy to find an optimization solution from the shallow network to a deeper network, which causes the degradation problem. Besides, training the deep C3D ConvNets on video dataset cost more memories and is time-consuming. Fortunately, residual learning reduces the deep network complexity by identity mapping [3]. The identity mapping is performed by a shortcut connection and element-wise addition, as is shown in **Fig. 3** If the identity mapping is optimal, the residual

learning across over several layers reduces the depth of network layers and precondition the optimal function for the solver.



(a) (b)

Fig. 3. C3D Residual block

The basic C3D residual learning block includes a 3D convolutional layer and shortcut connection [3] as shown in **Fig. 3** (a). It computes the residual function, e.g., $F(x) = Z(x) - x$, in which x denotes the input and $Z(x)$ denotes the underlying mapping. The output $Z(x)$ is the element-wise addition by $F(x) + x$ and has the same dimension as input x. The residual learning is a preprocessing for network to learn the identity mapping if it is optimal for the solver. Different from the basic C3D residual learning block, if the contained 3D convolutional layer narrows the frames volume size, as shown in **Fig. 2**, the input x and residual mapping $F(x)$ are not in the same dimension. To match the dimension, we add a 1×1 3D convolutional layer to the shortcut connection and get the weighted mapping $H(x)$, as shown in **Fig. 3** (b). The underlying function becomes:

$$Z(x) = F(x) + H(x) \tag{1}$$

The C3D Residual block in **Fig. 3**(b) has two layers and thus, $F(x) = W_2\sigma(W_1x + b_1) + b_2$, in which σ denotes the RELU activation function. The form of $F(x)$ is based on the number of layers and it is flexible to choose. Experiments show that one layer is more effective than others in this paper, which is different from [3]. The weighted mapping function is $H(x) = W_s x$, in which W_s is the weighted value matrix and used for matching dimensions. The element-wise addition operation $F(x) + H(x)$ is performed on features maps, channel by channel. For simplify, the above formulas are calculated on fully-connected layers and they are applied equally to the convolutional layers.

3.2 Spatio-Temporal Residual Network

The Spatio-Temporal Residual Network architecture is based on C3D Residual Network, as shown in **Fig. 4**. The details are described below.

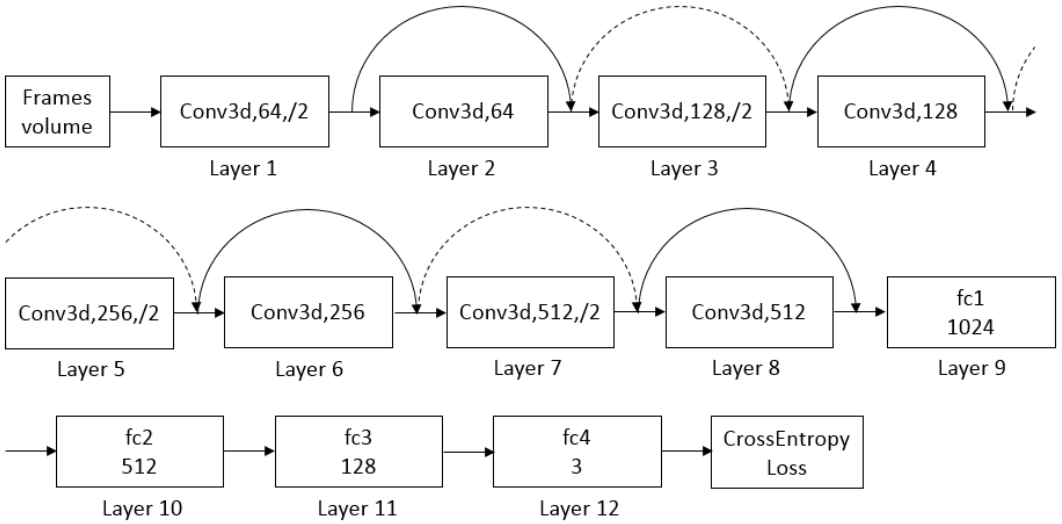


Fig. 4. Spatio-Temporal Residual Network

The network architecture has eight convolutional layers and four Fully-Connected layers, as shown i**n Fig. 4** The network is mainly inspired by the method of VGG nets [28] and it follows two design rules: (i) if the output feature map has the same size with the input, the number of filters remain unchanged; (ii) if the output feature map size is rescaled to the half size of the input, the number of filters is doubled to keep the time complexity. The kernel size in convolutional layer is mostly 3×3 and we apply the BatchNorm layer after each convolutional layer. The BatchNorm layer accelerates the network optimization by normalizing the means and variance of layer output and reducing the dependence of parameters gradients. Furthermore, batch normalization is also a network regularization method. The Relu function is adopted at each activation layer. Especially, max-pooling is performed after activation function at per layer over a 3×3 pixel window with stride 1 to preserve the prominent features. The fully-connected layers contain descending number of nodes over four layers and the last one has three nodes, which means the output includes three classes.

The loss function in the final layer is a CrossEntropy Loss:

$$loss(x, class) = -\log(\frac{e^{x[class]}}{\sum_j e^{x[j]}}) = -x[class] + \log(\sum_j e^{x[j]}) \tag{2}$$

We treat the slide transition detection as a volume classification issue, which is discussed in detail in Sec. 4 and the CrossEntropy Loss is an effective loss function for multi-classification problems. x denotes the output of the network and $class$ denotes the ground truth of classification.

3.3 Implementation

Each video frame is resized to 112×112 for reducing the memory cost and the mini-batch size is set to 128. The number of data loading workers is 1. Adam, a first-order gradient algorithm, is applied for the network optimization with $\beta_1=0.9$ and $\beta_2=0.999$. The network is regularized by weight decay and dropout regularization. The L2 penalty multiplier for

weight decay is set to 5×10^{-4} and the dropout ratio is set to 0.5 for the first three fully-connected layers. The initial learning rate is 0.001 and divided by 10 when the training epoch reaches a multiple of 10. The number of total epochs to run for the model is set to 100 and the default value of validation frequency is 10, which means that after every ten training epochs, the validation phase is implemented.

4. Experiments

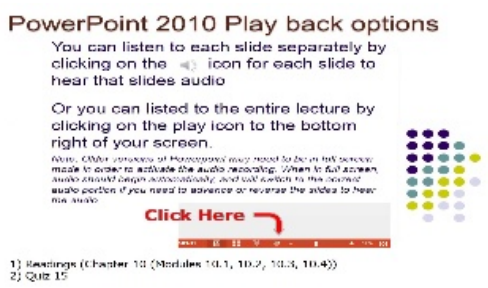
We collect six types of lecture videos from Stanford University Courses, Yale University courses and YouTube to verify the superiority of our model over other approaches. The training dataset contains 67 lecture videos and the testing dataset contains 26 lecture videos. The video length is roughly from 10 minutes to 70 minutes.

Type-A lecture videos contain only the computer slide screen. Type-B lecture videos are recorded by multiple cameras and the speaker and the slide screen are presented simultaneously. Type-C and Type-D lecture videos capture the speaker and projected on-stage screen by a still camera. But the speaker movement interruption in Type-C is produced. Type-E lecture video contains complex camera motion, such as pan, tilt and zoom. Besides, the speaker movement interruption in Type-E lecture video will possibly affect the detection accuracy. Type-F lecture video is recorded with multiple cameras and it allows sudden camera switch, such as the switch from the speaker to slide. Each lecture video is temporally down-sampled to 6 frames per-second and spatially down-sampled to the resolution of 640 by 360. We group several frames into a frame volume and the experimental result shows that two frames in a frame volume is better for slide transition detection, which contains less slide change types. We divide the frame volumes into three classes, as shown in **Fig. 6**: (i) slide transition; (ii) camera switch between slide and the speaker; (iii) frames remain unchanged or contain little changes, such as people movement, camera motion. Therefore, the slide transition detection in lecture videos turns into three-class classification problem for this spatio-temporal residual network model.

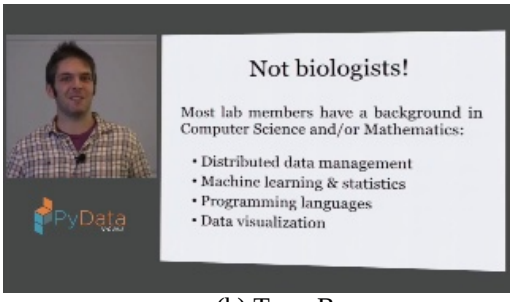
The quantitative evaluation is performed to demonstrate the superiority of our spatio-temporal residual network model. After manually labeling the slide transition groundtruth of each lecture video, *F*-score is applied as the evaluation metric:

$$\begin{aligned} F_1 &= \frac{2\times Precision\times Recall}{Precision+Recall} \\ Precision &= \frac{S_c}{S_t} \\ Recall &= \frac{S_c}{S_a} \end{aligned} \tag{3}$$

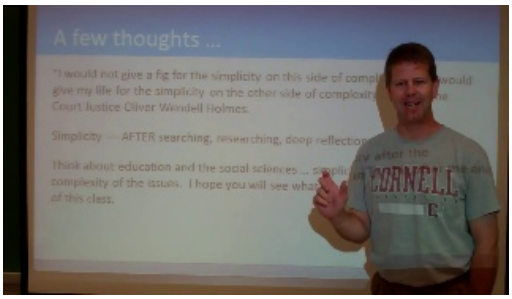
where S_a is the total number of actual slide transitions, S_c is the number of slide transition correctly detected, and S_t is the total number of detected slide transitions.



(a) Type-A



(b) Type-B



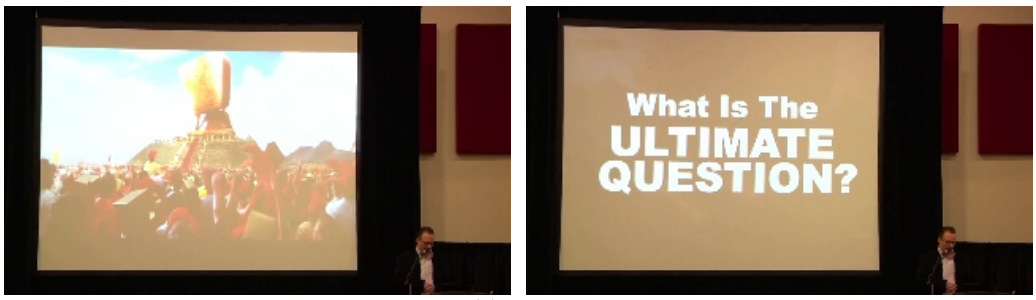
(c) Type-C

(d) Type-D

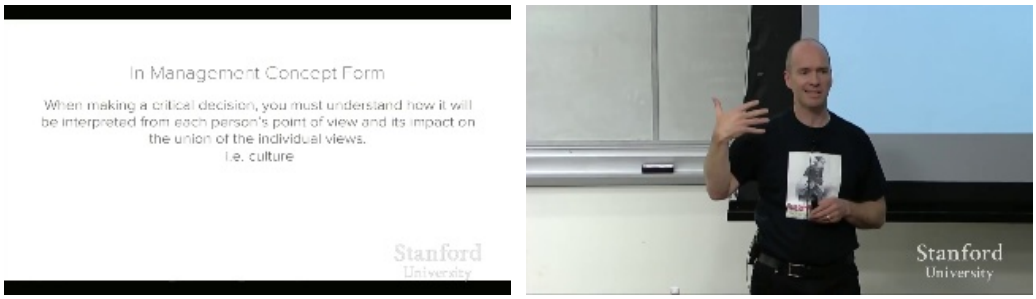
(e) Type- E

(f) Type-F

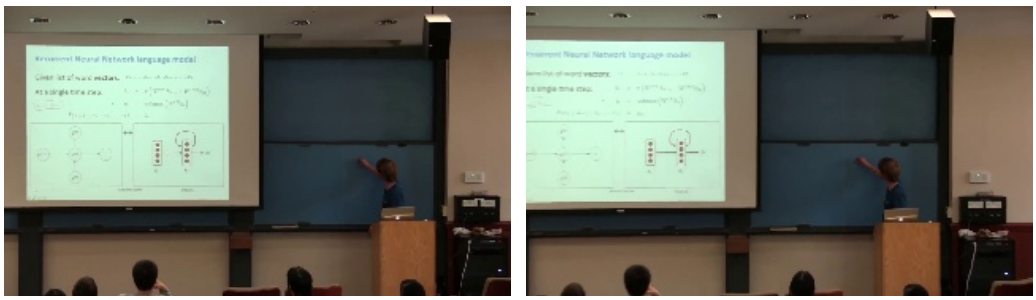
Fig. 5. Six types of lecture videos in our experiments. (a) Type-A presents the computer slide screen only. (b) Type-B presents the speaker and computer screen in two regions simultaneously. (c) Type-C presents the on-stage screen with speaker blocking the projected screen by a single camera. (d) Type-D presents the on-stage screen without speaker blocking the projected screen by a single camera. (e) Type-E presents complex camera motion, such as pan, tilt and zoom. (f) Type-F presents the sudden camera switch between speaker and computer slide screen.



(a) type1



(b) type2



(c) type3

Fig. 6. Three types of frame volume. Type1 presents slide transition. Type2 presents the camera switch between the speaker and slide screen. Type3 show that the camera remain still or contains little changes.

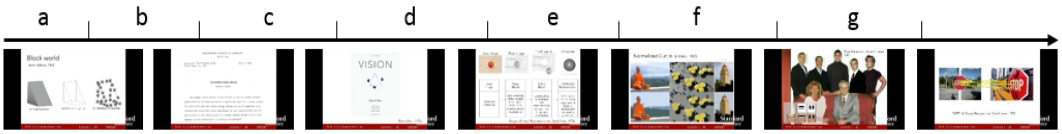


Fig. 7. Slide transition detection result for Type-F by our model. The slide transition time is marked on the timeline and the detected slide change frames are shown below the timeline.

A typical slide transition detection result for Type-F is shown in **Fig. 7**, where the marked ticks on the timeline indicate the automatic detected slide progressions. Result verifies that our model effectively detects the content changed slide frames of lecture videos.

We compare our system with other three approaches on test dataset, namely, Singular Value Decomposition (SVD) [31] on video summarization approach, Frame Transition Parameters (FTP) [11] on shot boundary detection method and analyzing the feature trajectories (SPD) [7] on slide progression detection method. **Table 1** shows the average Precision, Recall, and F-score of different approaches on all types of lecture video, and **Fig. 9** presents the detailed performance on each type of lecture video.

Fig. 8. Slide transition detection result for Type-F by the SPD approach. The camera switch and people movement are mistakenly detected as slide transitions.

Table 1. Average performance of different methods on six types of lecture video.

	Precision	**Recall**	***F*-score**
SVD [31]	0.700	0.745	0.722
FTP [11]	0.748	0.264	0.390
SPD [7]	0.773	0.848	0.810
STR Net (Ours)	**0.868**	**0.965**	**0.914**

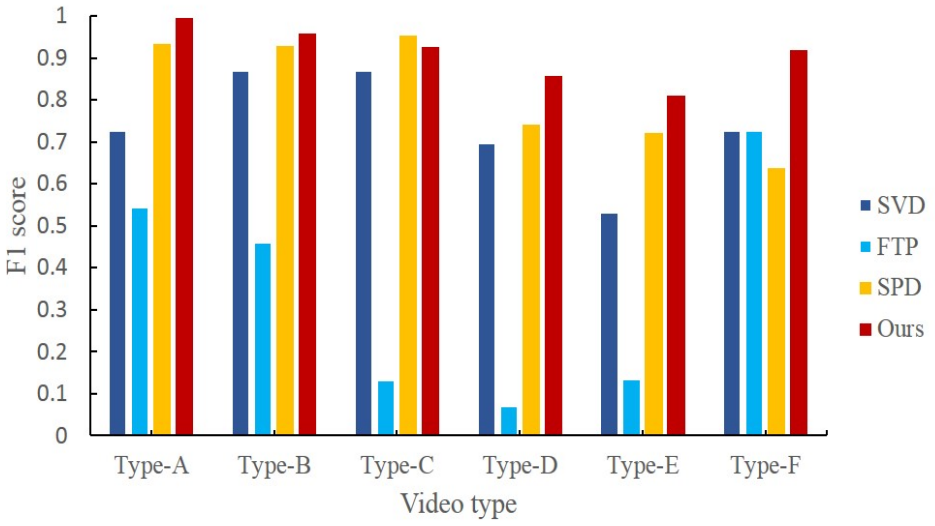


Fig. 9. Detailed performance of different methods on six types of lecture video

As shown in **Table 1** and **Fig. 9**, Our system significantly outperforms these approaches in detecting slide transitions. Compared with the feature trajectory-based SPD approach [7] and singular Value Decomposition (SVD) [31], our system improves the *F*-score by 12.8% and 26.6% on average. In particular, our approach improves the *F*-score up to 30.5% on Type-F against the SPD, where sudden camera switch and frequent people movement are presented. Due to the camera switch and motion, lots of false positives are detected by SPD when dealing with Type-F lecture video, which is also evidenced by **Fig. 8**. The general SVD approach achieves a better precision but fails to detect most of the slide changes as the result of the complex lecture video recording methods. In addition, shot boundary detection method using FTP achieves the worst performance for detecting slide transition. The amount of training data helps the spatio-temporal residual network learn the transition characteristics across video frames. We classify the noises into several types of frame volume and the model avoids the

various noise interruption by classified learning. In contrast, above three methods and some algorithms(e.g., [32]) detect slide transitions with artificial designed features. They focus on the characters of slide changes and ignore to learn the different types of temporal noise characters. Apart from good performance, the system is also superior at scalability. The proposed C3D Residual block in the system is suitable for spatio-temporal feature learning and it is simple, efficient and generic. In addition to the slide transition detection, the C3D Residual block can be also applied to other visual fields, e.g., video retrieval and shot change detection. Other methods are good at extracting spatial features but they are unable to summarize the temporal evolutional among adjacent video frames, which is important for various tasks.

The proposed slide transition detection is performed on a single 1080Ti GPU. In our experiment, the spatio-temporal residual network model is trained on 24K video frames and it takes about 43 minutes in total with 100 epoches, i.e., 0.003s per frame volume. We test the trained model on 53k video frames and the proposed system is processing at 60 fvps (frame volume per sec), e.g., 120 fps. The time complexity of the other three methods respectively is 4.6fps(FTP), 195fps(SVD) and 6.3fps(SPD). Note that we cannot find GPU implements of comparison methods and it is not trivial to implement a parallel version of these algorithms on GPU.

5. Conclusion

In this paper, we present the spatio-temporal residual network approach to detect slide transition in lecture videos. By combining the 3D ConvNet and ResNet, a novel network architecture is constructed. After spliting the input lecture video into video segments named frame volume, we divide these frame volumes into three classes and sent them into the classification model. By classifying each frame volume, the slide transition detection is accomplished. Experimental result shows that our system successfully extracts key frames by slide transition detection over various types of lectures videos and achieves the best performance.

References

[1] Du Tran, Lubomir Bourdev, et al, “Learning Spatiotemporal Features with 3D Convoluntional Networks,” in *Proc. of IEEE International Conference on Computer Vision*, pp. 4489-4497, Feburary 17, 2015. Article (CrossRef Link).

[2] Rehman, A., Abbas, N., Saba, T., Rahman, S.I.U., Mehmood, Z., Kolivand, H. “Classification of acute lymphoblastic leukemia using deep learning,” *Microscopy Research and Technique*, vol. 81, no. 11, pp. 1310-1317,2018.

[3] He K, Zhang X, Ren S, et al, “Deep Residual Learning for Image Recognition,” in *Proc. of Computer Vision and Pattern Recognition*, pp. 770-778, June 26-July 1, 2016. Article (CrossRef Link).

[4] Ma, Di, Agam, Gady, “Lecture video segmentation and indexing,” in *Proc. of The International Society for Optical Engineering*, 8297(1), pp .48, January 25- 26, 2012. Article (CrossRef Link).

[5] Hyun Ji Jeong. Tak-Eun Kim. Myoung Ho Kim, “An accurate lecture video segmentation method by using SIFT and adaptive threshold,” in *Proc. of Conference: Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia*, pp. 285-288, December 03-05, 2012. Article (CrossRef Link).

[6] Jeong H J, Kim T E, Kim H G, et al, "Automatic detection of slide transitions in lecture videos," *Multimedia Tools & Applications*, vol. 74, no. 18, pp. 7537-7554, September 28, 2015. Article (CrossRef Link).
[7] Li K, Wang J, Wang H, et al, "Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp.1233-1246, June 1, 2015. Article (CrossRef Link).
[8] Yousuf, M., Mehmood, Z., Habib, H.A., Mahmood, T., Saba, T., Rehman, A., Rashid, M. "A Novel Technique Based on Visual Words Fusion Analysis of Sparse Features for Effective Content-Based Image Retrieval," *Mathematical Problems in Engineering*, vol. 2018, 2018. Article (CrossRef Link).
[9] Sharif, U., Mehmood, Z., et al, "Scene analysis and search using local features and support vector machine for effective content-based image retrieval," *Artificial Intelligence Review*, vol. 52, no, 2, pp. 901-925, 2019. Article (CrossRef Link).
[10] Subudhi B N, Veerakumar T, Yadav D, et al, "Video Skimming for Lecture Video Sequences Using Histogram Based Low Level Features" in *Proc. of International Advance Computing Conference*, pp. 684-689, January 05-07, 2017. Article (CrossRef Link).
[11] Mohanta P P, Saha S K, Chanda B, "A Model-Based Shot Boundary Detection Technique Using Frame Transition Parameter," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp.223-233, February, 2012. Article (CrossRef Link).
[12] Cirne M V M, Pedrini H, "VISCOM: A robust video summarization approach using color co-occurrence matrices," *Multimedia Tools & Applications*, vol. 77, no. 1, pp. 857-875, 2018. Article (CrossRef Link).
[13] Balagopalan A, Balasubramanian L L, Balasubramanian V, et al, "Automatic keyphrase extraction and segmentation of video lectures," in *Proc. of IEEE International Conference on Technology Enhanced Education*, pp. 1-10, January 3, 2012. Article (CrossRef Link).
[14] He L, Sanocki E, Gupta A, et al,"Auto-summarization of audio-video presentations," in *Proc. of Acm Multimedia*, pp. 489-498, October 30 ,1999. Article (CrossRef Link).
[15] Repp, Stephan, Meinel, Christoph, "Segmentation of lecture videos based on spontaneous speech recognition," in *Proc. of 10th IEEE International Symposium on Multimedia*, pp. 692-697, December 15-17, 2008. Article (CrossRef Link).
[16] Lin M, Diller C B R, Forsgren N, et al, "Segmenting Lecture Videos by Topic: From Manual to Automated Methods," in *Proc. of 11th Americas Conference on Information System*, pp. 1891-1898, August 11-15, 2005. Article (CrossRef Link).
[17] Kanwal Yousaf, Zahid Mehmood, Tanzila Saba, et al., "A Novel Technique for Speech Recognition and Visualization Based Mobile Application to Support Two-Way Communication between Deaf-Mute and Normal Peoples," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1-12, 2018. Article (CrossRef Link).
[18] Qazi KA, Nawaz T, Mehmood Z, et al, "A hybrid technique for speech segregation and classification using a sophisticated deep neural network," *PLoS ONE*, vol. 13, no. 3: e0194151, 2018. Article (CrossRef Link).
[19] Yang H, Siebert M, Lühne P, et al, "Lecture Video Indexing and Analysis Using Video OCR Technology," in *Proc. of Seventh International Conference on Signal-Image Technology and Internet-Based Systems*, pp. 54-61, November 28, 2011. Article (CrossRef Link).
[20] Che X., Yang H., Meinel C, "Lecture video segmentation by automatically analyzing the synchronized slides," in *Proc. of the 2013 ACM Multimedia Conference*, pp. 345-348, October 21 -25, 2013. Article (CrossRef Link).
[21] Baidya, ESHA, Goel, Sanjay, "LectureKhoj Automatic Tagging and Semantic segmentation of online lecture videos," in *Proc. of 7th International Conference on Contemporary Computing*, pp. 37-43, August 07-09, 2014. Article (CrossRef Link).
[22] Mehmood Z, Gul N, et al, "Scene search based on the adapted triangular regions and soft clustering to improve the effectiveness of the visual-bag-of-words model," *Eurasip Journal on Image & Video Processing*, vol.48, 2018. Article (CrossRef Link).

[23] Sarwar, A., Mehmood, Z., et al, "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine," *Journal of Information Science*, vol. 45, no. 1, pp. 117–135, 2018. Article (CrossRef Link).
[24] Mehmood, Z., Rashid, M., et al, "Effect of complementary visual words versus complementary features on clustering for effective content-based image search," *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 5, pp. 5421-5434, 2018.
[25] Yao T, Mei T, Rui Y, "Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization," in *Proc. of Computer Vision and Pattern Recognition*, pp. 982-990, June 26 –July 01, 2016. Article (CrossRef Link).
[26] Qiu Z, Yao T, Mei T, "Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks," in *Proc. of IEEE International Conference on Computer Vision*, pp. 5534-5542, October 22-29, 2017. Article (CrossRef Link).
[27] He K, Sun J, "Convolutional neural networks at constrained time cost," in *Proc. of Computer Vision and Pattern Recognition*, pp.5353-5360, June 7-12, 2015. Article (CrossRef Link).
[28] Simonyan K, Zisserman A, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014. Article (CrossRef Link).
[29] Glorot X, Bengio Y, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, pp. 249-256, May 13-15, 2010. Article (CrossRef Link).
[30] Ioffe S, Szegedy C. "Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift" in *Proc. of 32nd International Conference on Machine Learning*, pp.448-456, July 6-11, 2015. Article (CrossRef Link).
[31] Gong Y, Liu X, "Video summarization using singular value decomposition," *Multimedia Systems*, vol. 9, no. 2, pp. 157-168, 2003. Article (CrossRef Link).
[32] Z. j. Liu, K. Li, L. Q. Shen and P. An, "Sparse time-varying graphs for slide transition detection in lecture videos," in *Proc. of International Conference on Image and Graphics (ICIG)*, pp. 567-576, Sept 13, 2017. Article (CrossRef Link).

Zhijin Liu received the B.S. degree from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2016, where he is currently pursuing the M.S. degree in Key Laboratory of Advanced Displays and System Application. His research interests include video summarization, computer vision and graphics.

Kai Li is an Assistant Professor at Shanghai University, Shanghai, China. He received the B.E. degree (with honor) in electrical engineering from Shanghai University, Shanghai, China, in 2010, and the Ph.D. degree in automation from Tsinghua University, Beijing, China, in 2015. He worked as a Student Intern at Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, from September 2012 to May 2013. His research interests include image and video processing, computer vision and graphics. He is a member of the IEEE.

Liquan Shen received the B. S. degree in Automation Control from Henan Polytechnic University, Henan, China, in 2001, and the M.E. and Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2005 and 2008, respectively. Since 2008, he has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where he is currently an Associate Professor. His major research interests include H.264, Scalable video coding, Multi-view video coding, High Efficiency Video Coding (HEVC), perceptual coding, and multimedia communication.

Ran Ma received the B.S. degree from Yangzhou University, Yangzhou, China, in 1997, and the M.S. and Ph.D. degrees from Shanghai University, Shanghai, in 2000 and 2008, respectively. She is currently an Associate Professor in School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her research interests include error concealment, stereoscopic image and video processing, coding and application.

Ping An received the B.A. and M.S. degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2002. In 1993. She joined Shanghai University. From 2011 to 2012, she was a Visiting Professor with the Communication Systems Group, Technical University of Berlin, Berlin, Germany. She is currently a Professor with the Video Processing Group, School of Communication and Information Engineering, Shanghai University. Her research interests include image and video processing, with a focus on 3D video processing.