Mode of Teaching Based Segmentation and Annotation of Video Lectures

Yogesh Singh Rawat
School of Computing,
National University of Singapore
yogesh@comp.nus.edu.sg

Chidansh Bhatt
Idiap Research Institute
Martigny, Switzerland
cbhatt@idiap.ch

Mohan S Kankanhalli
School of Computing,
National University of Singapore
mohan@comp.nus.edu.sg

***Abstract*—Online education is becoming more and more prevalent these days. Many universities provide pre-recorded classroom lectures for distance learning and remote users can access these lectures over Internet. With the available indexing techniques, users can search and retrieve videos related to their topic of interest in these stored databases. However, sometimes the 'mode of teaching' impacts the viewer's perception for the retrieved video lecture or snippet. In this work we make use of visual concepts in the video lecture to identify the mode of teaching and generate annotations for the video. The developed approach uses low-level features like color and edges to classify video frames into high level semantic concepts. The system performs frame-by-frame classification and mode of teaching can be inferred for each segment as well as the complete video. Experimental results show high accuracy of proposed method and demonstrate its potential for relevant applications.**

I. Introduction

As online learning using video lectures is becoming popular, the database of these videos is growing rapidly. In order to support remote learning, Massive Open Online Course (MOOC) service providers like, Coursera, Khan Academy, Udacity, etc, provide a large database of recorded video lectures. In this huge database, it is often hard for the user to browse through the video and search specific content in parts of the video without any indexing. The existing indexing techniques mainly make use of the audio features, which can provide them only the subject content of the lecture. Sometimes the 'mode of teaching' may impact the perception of the lecture as boring (only slides without view of the professor explaining) or interesting (slides with view of professor engaging students in explanation with expressions and emotions) [1]. Furthermore, the user may sometimes have specific preference for the mode of teaching used in the video lectures, e.g., presentation slide based teaching over blackboard based teaching (or vice versa). Therefore, high level visual semantic information representing the mode of teaching will be useful in user-centric search and recommendation.

In this work we address the problem of video lecture segmentation and annotation by detecting 'mode of teaching' using data mining approach. We used the dataset from 'MediaMixer/VideoLectures.NET Temporal Segmentation and Annotation' ACM Multimedia Grand Challenge 2013 [2]. Based on the dataset we define visual concepts like, *'Professor Talking', 'Professor Writing On Blackboard'* and '*Professor Explaining Slide*', which are used for annotations indicating the teaching mode of the video. Each frame of the video is tagged with defined concepts based on the classification system. We used Support Vector Machine (SVM) [3] for frame classification. After annotation of video frames we use these concepts to define segmentation boundaries for the complete video. A boundary is defined as the point where we observe change in concepts for consecutive frames. These boundaries are meaningful from the user's perspective as well, as changing from one concept to another may indicate change in teaching mode or even possibly a change in topic or sub-topic [4].

The motivation behind the proposed method is to utilize higher level semantic information (mode of teaching) from the video frames for segmentation. Video lectures can have multiple teaching modes like slides as well as blackboard. For videos with mixed mode of teaching, change in the mode is a potential point for segmentation. Moreover, the defined concepts will augment the annotations with useful information regarding the mode of teaching in video lecture apart from just the subject terms and enable users to search videos and within video segments for their preferable mode of teaching. It will make user-centric search and recommendation more efficient and meaningful. As the main contribution of this work, we automatically detect visual semantic concepts corresponding to mode of teaching for generating segmentation and annotation of video lectures. Also, we presented probabilistic approach to improve the accuracy of proposed method.

II. Related Work

The existing methods for video lecture segmentation, annotation and classification mainly focus on either audio features or low-level visual features. Some of the works make use of high level semantic information like, drawing, erasing, scrolling and explaining [5], writing, erasing and speaking [6] and gesture of instructor [7]. But these methods require manual pose marking on video frames and are targeted for specific type of videos like, either slide mode or blackboard mode. In another work [8] Dorai et al. utilized color moments to classify video frames as narrative or text based (slide, web or whiteboard) using Decision Trees. But their method is limited to videos where the narrator and text does not appear simultaneously.

In a more recent study, Schreer and Masneri [9] used semantic visual information for video classification. Their method make use of face detection and color features to

978-1-4799-3990-9/14/$31.00

classify a frame into visual concept like, '*Professor*', '*Slide*' or '*Blackboard*'. However, it is not always possible to classify a frame into one of these concepts as more than one concept may be present in the same frame. Also, mere the presence of these concepts provide little information as there may be a blackboard in background which is not used at all. These basic concepts are important but, events like 'Professor writing on blackboard', 'Professor explaining slide', etc. provide more useful information where more than one concept may be present. In this work we present a more comprehensive approach which utilizes high level semantic concepts related to instructional videos for segmentation and annotation. Visual concept detection is a well researched area for general videos [10], but here we target for concepts and events like 'Professor writing on blackboard' and 'Professor explaining slide', which are specific to video lectures.

III. Proposed Methodology

The proposed methodology consists of concept detection, video segmentation and annotation. Figure 1 presents the workflow of the proposed methodology.

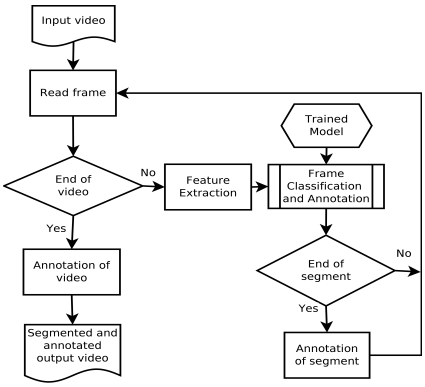


Fig. 1: Workflow of the proposed framework

A. Semantic Concept Detection

In this section we will discuss the methods used for concept detection.

1) Concepts: We define two level of concepts, three basic concepts and three complex concepts. Basic concepts are *'Professor', 'Blackboard'* and *'Slide'*. The basic concepts are mutually inclusive, which means they can occur together in any video frame. The complex concepts are *'Professor Talking to Audience', 'Professor Writing on Blackboard' and 'Professor Explaining Slide'*. The complex concepts are mutually exclusive, which means they can not occur in any video frame at the same time.

2) Feature Extraction: For each frame, we extract four set of feature vectors. We use 96 dimensional RGB histogram (32 bins for each color), 91 dimensional Edge histogram using Sobel Filter (2 degree slot for each bin where 1 bin is for no direction), 64 dimensional Speeded Up Robust Features (SURF) [11] and 72 dimensional Histogram of Oriented Gradients (HOG) [12]. RGB features are used because they are very useful for distinguishing concepts like slide and blackboard. Frames having slides will be mainly bright colors and on the other hand blackboards are generally black colored. HOG and Edge histograms are used to extract features which corresponds to the edges in frames which gives us features related to the shape of objects in the frame. SURF features are based on 2D Haar wavelet response of the image and it corresponds to the interesting points in any image.

3) SVM Based Classification: Since each video consists of large number of frames and each frame needs high dimensional feature vector for representation, we propose a scalable method which can handle high dimensional feature vectors and which is also suitable for large video dataset. We use Support Vector Machine (SVM) as a classification method, because it is well known to perform efficiently for high dimensional feature vectors [13].

The dataset [2] consist of 20 video lectures which are 1-2 hours long. Each video in this dataset has on an average 50K frames. The videos are from diverse areas and use different teaching modes. We used half of the videos (10) for creating test and training dataset. The frame rate of the videos in the dataset varies from 15-30 frame per second. We used manually annotated dataset of frames to create training and test dataset by equally diving the annotated frames into two sets. In total we have 11265 annotated frames and it was equally divided into training and test dataset.

We used Radial Basis Function (RBF) as a kernel for the SVM classifier. We have used five-fold cross validation using grid search for finding the optimal values of the corresponding parameters [3]. We used binary classification for each of the basic concepts. The complex concepts are mutually exclusive and we employ a multi-class classification method for complex concept detection. We employed a four class classification , three classes for complex concepts and one class for unidentified concept.

4) Probabilistic-SVM Approach: We used a two level probabilistic classification approach where we relate the basic concepts with the complex concepts. If a frame is classified as *'Professor Explaining Slide'*, then it must also be classified as *'Professor'* and *'Slide'*. To improve the accuracy of proposed method we use probabilistic classification [14] and exploit the relation between basic and complex concepts. The idea is to use the probabilistic score of basic and complex concepts for deriving the confidence score of the complex concept.

Using the conditional probability we have, Prob(B and A) = Prob(B|A)Prob(A). We used this conditional probability to derive a confidence score for complex concepts. For example, let A = (*'Professor'*) and B = (*'Professor Talking to Audience'*), then we use the above equation to evaluate the confidence score of the complex concept B. Prob(A) comes from the basic classifier for *'Professor'* and Prob(B|A) comes from the multi-class classification of *'Professor Talking to Audience'*. Using this confidence score we improved the accuracy of our proposed method for complex concept detection.

B. Frame Level Annotation

Based on classification of the frames, we use the detected concepts for annotation of frames. The annotations used are summarized in the table I. For frames where no annotation is found (case unknown-X in table I), we used the annotation from previous frame in the video. The occurrence of

TABLE I: Annotations

Basic Concept			Complex Concept	Annotation
Professor	Blackboard	Slide		
0/1	0/1	0/1	PWBB	PWBB
0/1	0/1	0/1	PXS	PXS
0/1	0/1	0/1	PTA	PTA
0	0	0	X	X
0/1	0/1	1	X	S
0/1	1	0/1	X	BB
1	0/1	0/1	X	P

PWBB - 'Professor Writing on Blackboard', PXS - 'Professor Explaining Slide', PTA - 'Professor Talking to Audience', S - 'Slide', P - 'Professor', BB - 'Blackboard', X- 'Unknown, 1/0 - Corresponds to Presence/Absence of corresponding concept.

annotations on each frame is used to compute the overall annotation for each segment as well as complete video lecture by summing up individual occurrence. The segment level and video level annotations represent the fraction of frames tagged with corresponding frame level annotations.

C. Segmentation Using Mode of Teaching

We use the detected concept for each frame and declare the change of concept as a point for segmentation. Algorithm 1 presents the details of segmentation. We proposed segmentation scheme with a sliding window of one second ahead of the current frame to avoid over segmentation. If there is a mismatch of annotation between two consecutive frames then we evaluate the annotation for frames from next one second video (we expect segments to be at least one second long). For this, we employ a voting scheme and used frame-level annotation from each frame. We used an exponential decay function for giving more weight to frames which are closer to current frame. The current point is marked as segmentation point if there is a mismatch in the annotation of previous frame and evaluated next one second window. This will also avoid segmentation due to misclassified frames.

Algorithm 1 Segmentation Point Detection

```
CFrA ← nil % current frame annotation
CSegA ← nil % current segment annotation
NxtWinA ← nil % next 1 second window annotation
while (not end of video) do
    featureExtraction()
    frameClassification()
    frameAnnotation()
    if (CFrA ≠ CSegA or CFrA = 'Unkown) then
        continue
    else
        NxtWinA = GetNextWindowAnnotation()
        if (NxtWinA = CSegA) then
            continue
        else
            GenerateAnnotation() % Segmentation point
        end if
    end if
end while
```

IV. Experiments

The dataset contains videos which are either classroom lectures with black-board/slides or presentation with slides for some conference or events. The videos are from different subject areas and mostly comprised of combination of mode of teaching along with some videos which used only single mode of teaching.

TABLE II: Results for Basic Concepts (%)
(CV - cross validation)

Concept	Frames	Accuracy		Precision	Recall	F-Score
		No CV	CV			
Professor	1403	96.79	97.22	97.76	98.73	98.24
Blackboard	1343	99.18	99.55	98.61	99.30	98.95
Slide	1646	99.31	99.57	99.50	99.91	99.71

A. Frame Classification

We trained the classification model after combining the selected features and use it for classification of the dataset. For each frame we need to extract the 323 dimensional feature vector. After performing grid search for RBF kernel, the derived parametric values are (C=8.0 and γ=0.0078125) for concept 'Blackboard' and (C=32.0 and γ=0.0078125) for other concepts. Table II and III show accuracy results after merging all the features into one feature vector. To improve the accuracy of classification for complex concepts, we use probabilistic approach to evaluate the confidence score of the complex concepts. Table III shows the results after using the probabilistic approach.

TABLE III: Probabilistic approach for Complex Concepts (1268 Frames)

Concept	Recall		F1-Score	
	Normal	Probabilistic	Normal	Probabilistic
'*PTA*'	82.48	85.31	87.95	89.35
'*PWBB*'	99.44	99.44	95.07	96.34
'*PXS*'	93.60	94.70	92.57	93.72

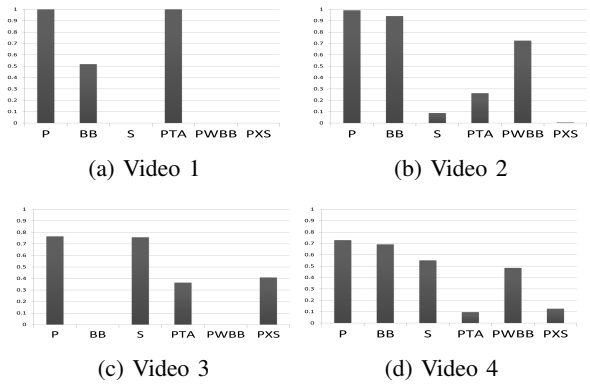
Combined Accuracy	Normal	Probabilistic
	91.72	92.59

B. Results and Discussion

As we have seen in the classification results that the concepts defined in this work are detected with high accuracy in the video lectures. Since these concepts provide us visual clues from the lecture video they can be effectively used for segmentation and annotation. We used the proposed framework for segmentation and annotation of all the 20 lecture videos from the dataset [2].

As we can see from table III, the concept *'Professor Talking to Audience'* has low recall (82.48%) as compared to other complex concepts (94-99%). After analyzing the results we found that frames with the concept *'Professor Talking to Audience'* were classified as *'Professor Writing on Blackboard'* and *'Professor Explaining Slide'*. This was for frames where either blackboard or slide was present in the background. This also explains the low F1-score of *'Professor Writing on Blackboard'* and *'Professor Explaining Slide'* as compared to corresponding recall measure. Also, it is important to note that our approach is complementary to earlier proposed methods which make use of audio recording for segmentation and annotation. The visual and audio features can always be combined together using a multi-modal approach for more useful results as is shown in this work [4].

Figure 2 presents annotation results for videos which were not used during training. The first video is with the professor



(a) Video 1 (b) Video 2

(c) Video 3 (d) Video 4

Fig. 2: Annotations for four different videos (values represent fraction of frames classified as corresponding concept in the video)

sitting on a chair during the whole lecture with a blackboard in the background. In the second video, the professor used a blackboard as a mode of teaching with few slides at times and the third one is with professor using the slides in the complete lecture. In the fourth video, all combination of teaching modes were used by the instructor. As we can see in the chart, the annotations provide a fair idea about the mode of teaching for any video lecture.

C. Comparison

In [9], the authors proposed to detect similar visual cues in video lectures. However, they defined basic concepts like presence of professor, slide or blackboard and do not study the relation between these basic cues which is important from the perspective of lecture videos. Also, they used only face detection and color feature for classification, which was one of the factor for misclassification of video frames (accuracy rate of only 71% for one of the class and average accuracy of 82%). In our study, we follow a more comprehensive approach and describe these basic cues as concepts in frame and relate them to higher level concepts which corresponds to events in a frame. In another work [6], Imran et al. proposed action (writing, erasing, speaking and idle) classification for video lectures. Table IV presents a comparison of overall classification accuracy with other proposed methods.

TABLE IV: Results Comparison

Concept	Schreer et al. [9]	Proposed Method	
	Recall	Recall	F1-score
Professor	86.3	98.64	98.19
Blackboard	99.6	99.91	99.71
Slide	84.5	99.30	98.78
Method	Imran et al. [6]	Schreer et al. [9]	Proposed Method
Accuracy (%)	87-89	71-82	92-99

V. Conclusion and Future Work

In this work we proposed use of visual semantic concepts for segmentation and annotation of video lectures. These concepts are meaningful for user and provide *'action'/'mode of teaching'* based segmentation. Annotation of individual segments enable user to understand mode of teaching in the video without even browsing. These annotations are useful in video search, recommendation, summarization and retrieval as it honors preferred mode of teaching along with lecture content. For future work we plan to use these visual concepts along with other modalities like audio and text. These visual concepts will also be useful for personalized video lecture recommendation system [4].

Acknowledgment

Research work by Yogesh and Mohan was carried out at the SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO. The second author is grateful for partial support by the Swiss National Science Foundation (SNSF) under the AROLES project n. 51NF40-144627.

References

[1] L. Yuan and S. Powell, "Moocs and open education: Implications for higher education," *Cetis White Paper*, 2013.

[2] "Mediamixer/videolectures.net temporal segmentation and annotation grand challenge," *ACM Multimedia 2013*.

[3] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[4] C. A. Bhatt, A. Popescu-Belis, M. Habibi, S. Ingram, S. Masneri, F. McInnes, N. Pappas, and O. Schreer, "Multi-factor segmentation for topic visualization and recommendation: the must-vis system," in *ACM Multimedia*, 2013, pp. 365–368.

[5] T.-J. K. P. Monserrat, S. Zhao, K. McGee, and A. V. Pandey, "Notevideo: Facilitating navigation of blackboard-style lecture videos," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 1139–1148.

[6] A. Imran, A. Moreno, and F. Cheikh, "Exploiting visual cues in non-scripted lecture videos for multi-modal action recognition," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, 2012, pp. 8–14.

[7] J. Zhang, K. Guo, C. Herwana, and J. Kender, "Annotation and taxonomy of gestures in lecture videos," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 1–8.

[8] C. Dorai, V. Oria, and V. Neelavalli, "Structuralizing educational videos based on presentation content," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 2, Sept 2003, pp. II–1029–32 vol.3.

[9] O. Schreer and S. Masneri, "SVM-based video segmentation and annotation of lectures and conferences," in *9th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2014)*, Lisbon, Portugal, 2014.

[10] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, 2011.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008, similarity Matching in Computer Vision and Multimedia.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1.

[13] G. Wang, "A survey on training algorithms for support vector machine classifiers," in *Fourth International Conference on Networked Computing and Advanced Information Management*, vol. 1, 2008, pp. 123–128.

[14] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Dec. 2004.