Acoustic Features to Predict Topic Change in Instructional Video

Vineet Kumar , Sushant Gupta

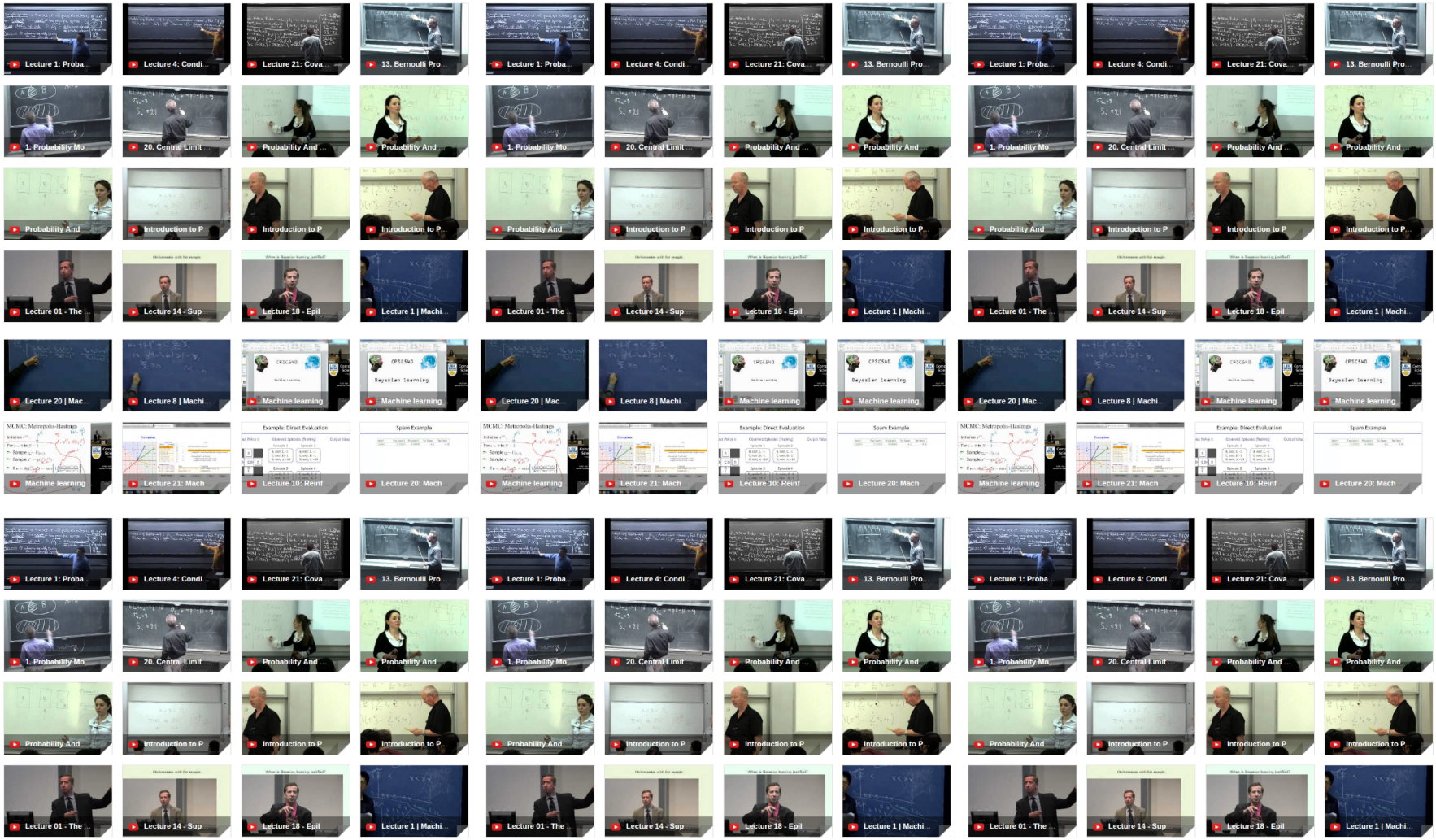
Computer Science & Engineering Department
Jadavpur University
{vntkumar8,sushantgupta1996}@gmail.com

4 March 2017

Motivation

India is one of the largest producer of engineers in the world. The growth of Massive Open Online Courses (MOOCs) are considered as one of the biggest revolution in education in last several decades. One significant impact of the MOOC phenomenon is that they have accelerated the widespread availability of quality Open Educational Resources (OER).

MOOCs have given an open challenge to all current methods of education system. They have high potential of acceptability among all kind of learners. We believe that Instructional Videos are going to be next generation textbooks.



120 Hours of Educational Content from Youtube

Feature Extraction

We wanted to use acoustic features of the videos, hence after an extensive study of sound and audio processing literature and found that MFCC & InterSpeech 2009 Feature Set are general best choice for such acoustic processing tasks.

Feature Extraction

We wanted to use acoustic features of the videos, hence after an extensive study of sound and audio processing literature and found that MFCC & InterSpeech 2009 Feature Set are general best choice for such acoustic processing tasks.

Disambiguity Resolution

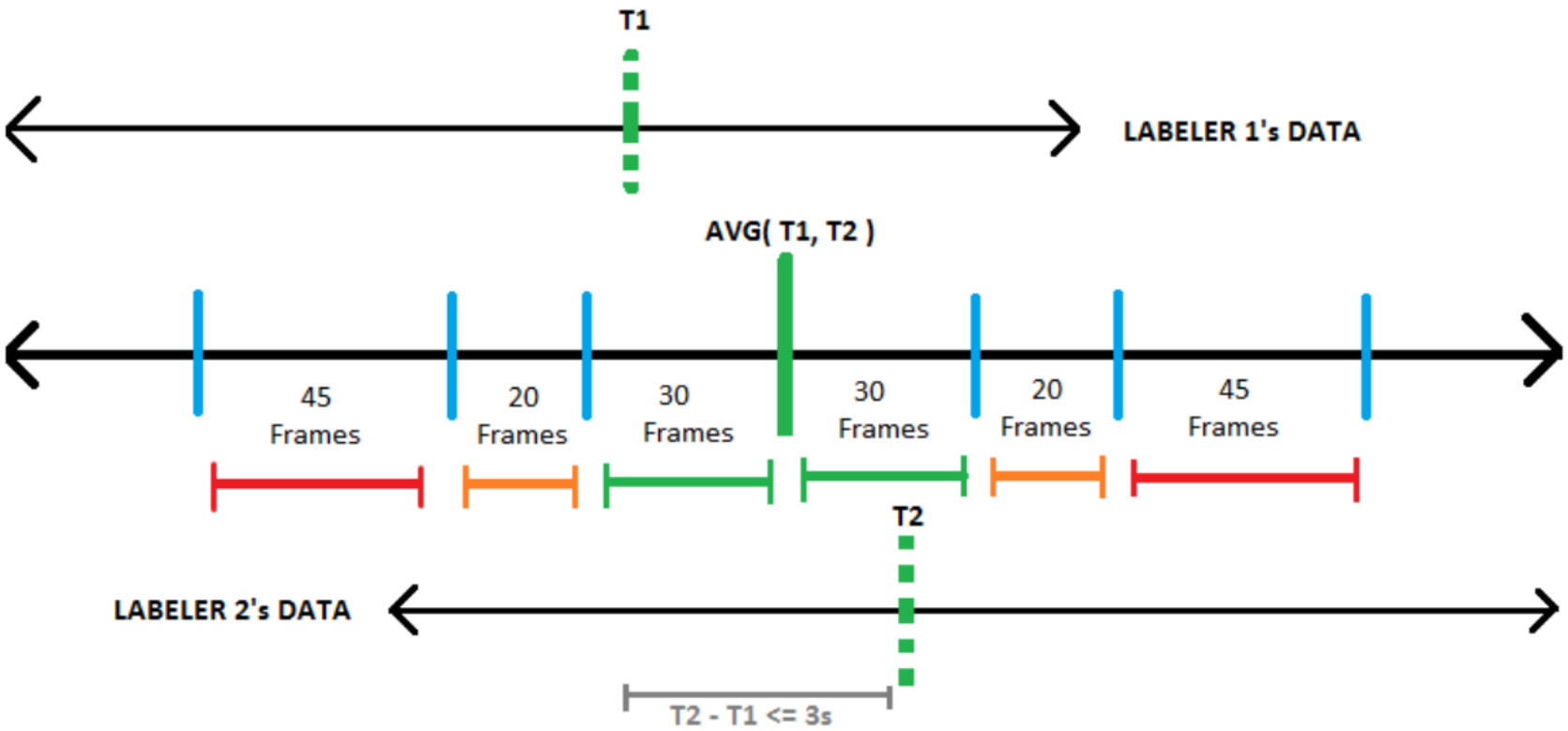


Figure: Labelling Scheme for Variation

Data Set

- Taking out the features for every 50ms of a video, our data set was pretty large with around

Data Set

- Taking out the features for every 50ms of a video, our data set was pretty large with around

  9 Million instances

- In a video, we noticed around 15-20 topic changes in an hour. If we assume topic change to be 1 second transition then in an hour,i.e, 3600 seconds we had only 15-20 seconds where topic has actually changed. This lead to a class imbalance problem. We had used a nice method to tackle this class imbalance problem by down-sampling.

DataSet Labeling

This was the most time consuming task of our project. We had to watch each and every video and manually note down the exact time of topic change as accurate in terms of second. This labelling was done by both of us independently. Labelling Scheme is :-

- 45 instances above are -ve (0)
- 20 instances above are skipped
- 30 instances above are +ve (1)
- **Exact time of topic change** (1)
- 30 instances below are +ve (1)
- 20 instances below are skipped
- 45 instances below are -ve (0)

Dimension Reduction

- Can we reduce the dimension of feature space ?

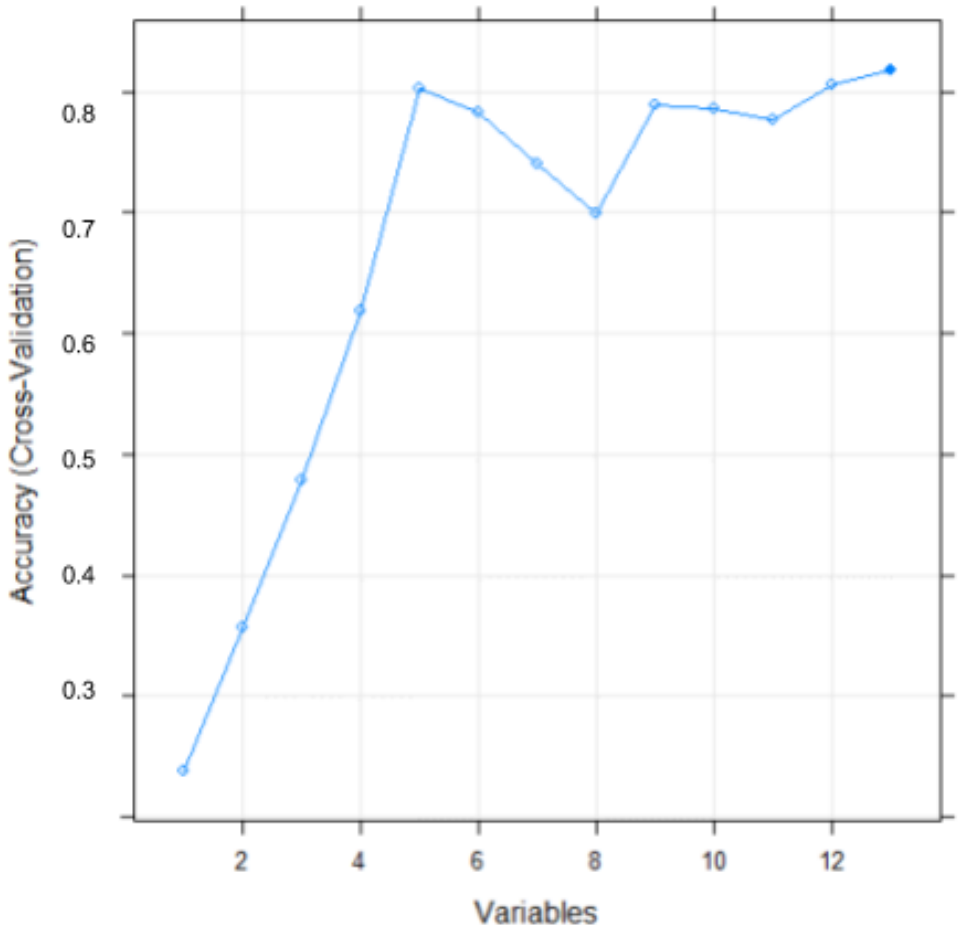
Dimension Reduction

- Can we reduce the dimension of feature space ?
- Classical problem of Dimensionality reduction
- We used Learning Vector Quantization to reduce the feature space. We ranked the features according to their importance.

Dimension Reduction

- Can we reduce the dimension of feature space ?
- Classical problem of Dimensionality reduction
- We used Learning Vector Quantization to reduce the feature space. We ranked the features according to their importance.
- An LVQ system is represented by prototypes $W = (w(i), ..., w(n))$ which are defined in the feature space of observed data. In training algorithms one determines, for each data point, the prototype which is closest to the input according to a given distance measure. The position of this so-called winner prototype is then adapted, i.e. the winner is moved closer if it correctly classifies the data point or moved away if it classifies the data point incorrectly.

Dimension Reduction - I



Model Training

- Now we have manageable number of instances

- Train the model using SVM & Random Forest
- SVM with RBF
- 78% of accuracy is achieved

Future Directions - I

- This work is an aid to help already available state of art table of content generation for video. Visual cues and acoustic features both are combined to accurately pinpoint the topic change transition and build a table of content kind of thing for educational video.

Future Directions - I

- This work is an aid to help already available state of art table of content generation for video. Visual cues and acoustic features both are combined to accurately pinpoint the topic change transition and build a table of content kind of thing for educational video.
- **Improving Accuracy** In order to improve our accuracy we are planning to train our model on Support Vector Machines for Multiple instance Learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances. In the simple case of multiple-instance binary classification, a bag may be labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. From a collection of labeled bags, the learner tries to either (i) induce a concept that will label individual instances correctly or (ii) learn how to label bags without inducing the concept.

Future Directions - II

- Song detection
- war scene detection rather any particular topic of interest

References I

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann.

Support vector machines for multiple-instance learning.

In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 577–584, Cambridge, MA, USA, 2002. MIT Press.

Arijit Biswas, Ankit Gandhi, and Om Deshmukh.

Mmtoc: A multimodal method for table of content creation in educational videos.

In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 621–630, New York, NY, USA, 2015. ACM.

References II

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller.

Recent developments in opensmile, the munich open-source multimedia feature extractor.

In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM.

Teuvo Kohonen.

The handbook of brain theory and neural networks.

chapter Learning Vector Quantization, pages 537–540. MIT Press, Cambridge, MA, USA, 1998.

We are very much thankful to Xerox Research Center India

We are very much thankful to Xerox Research Center India

Thank You !!!