Thumbnail Extraction from Slide-based Instructional Videos

Jae Hyun Kim
Stanford University
Stanford, CA
jakekim@stanford.edu

Sef Kloninger*
Stanford University
Stanford, CA
sef@cs.stanford.edu

Abstract

Thumbnails provide viewers with indices as well as summary of the video, serving a vital role for students taking online courses. Various methods were proposed for video summarization, from the simplest to the most sophisticated. However, distinct characteristics of slide-based instructional videos that come from their information-conveying purpose ask for an extractor that can extract frames with real importance. In this project, we propose a new method of thumbnail extraction using character recognition and clustering algorithm. Lecture videos are divided into slides and no-slide videos. Then each is processed with different algorithms to produce thumbnails. Finally we demonstrate, compare, and evaluate the performance of the method.

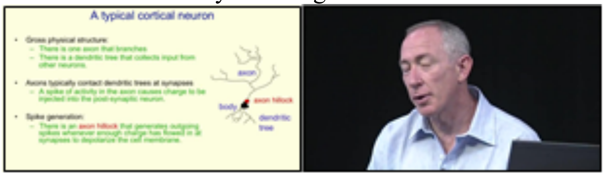
1. Introduction

The Massive Open Online Course (MOOC) movement has produced a great demand for various compute vision techniques to process course videos. Among them is thumbnail extraction which serves a crucial role of providing the viewer with indices of the video as well as a summary. This project aims at selecting thumbnails based on context of the video images, choosing only the most important moments of the lecture.

The project starts from a ready-made thumbnail extractor called 'Kelvinator', which was designed to be a part of Class2go project. (http://class.stanford.edu/) Kelvinator is a simple thumbnail extractor that selects frames with difference to the next frame greater than a given threshold. Kelvinator shows great simplicity, but unfortunately, along with poor performance. Kelvinator often produces blank frames, identical frames, and overlapping frames that do not contain valuable information. Another downside is that Kelvinator does not have a sense of importance. When told to select only a few frames, Kelvinator fails to select the most important ones.

Observations on Kelvinator led us to break the problem into two parts: removing meaningless frames, and selecting important ones. Various methods were proposed on the topic of video summarization, each with its own way of selecting the most 'important' frames. However, conventional high-level methods were not suitable for our purposes. The video file we aim to summarize are mostly based on slides, which are much more highly formatted compared to random videos. Using a method aimed at summarizing ordinary videos would be inaccurate as well as inefficient. For example, frames filled with instructor's face is not so likely to contain valuable information, even if they occupy a very long interval. On the other hand, frames filled with characters (letters) are likely to contain information that users need. The fact that our targets are slide-based lecture videos allows us to implement much simpler methods for summarization.

Our approach is to divide the video into slide-containing frames (which will be called 'slide frames' from now on) and slide-not-containing frames ('no-slide frames'), and then use different methods for each to select thumbnails. (See the figure below.) We assume that most important information is conveyed through slides.



A typical cortical neuron

- Gross physical structure:
  - There is one axon that branches
  - There is a dendritic tree that collects input from other neurons.
- Axons typically contact dendritic trees at synapses
  - A spike of activity in the axon causes charge to be injected into the post-synaptic neuron.
- Spike generation:
  - There is an axon hillock that generates outgoing spikes whenever enough charge has flowed in at synapses to depolarize the cell membrane.

Figure 1. Example of slide frame, no-slide frame

Video data came from Class2go and Coursera. The two open online course platforms provide a vast number of slide-based lecture videos with various lengths. SVM was used to train the classifier that separates slide frames from no-slide frames. After the thumbnail selection, the result was compared to human-selected thumbnails for evaluation.

2. Background & Related Work

Various methods of context-based extraction have been proposed in the field of video mining. A simple method focuses on the difference of consecutive frames and select the ones with difference greater than the threshold, just as Kelvinator does. A more complicated one generates clusters called "shot links" using K-means algorithm and select an index from each of the cluster. After the initial selection of frames, various features of the frames can be used to analyze

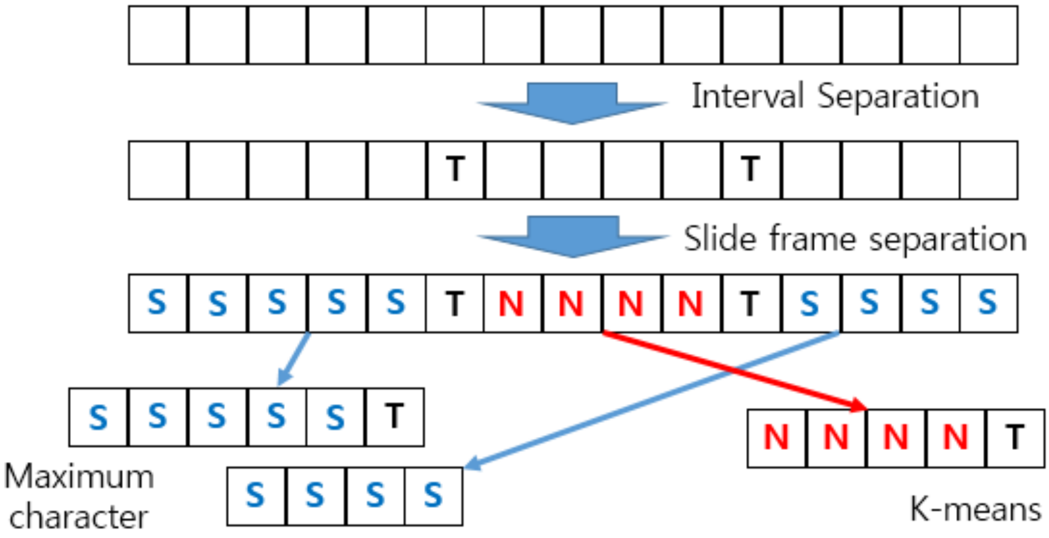


Figure 2. Representation of the process

them. [1] Our approach uses K-means clustering for no-slide frames.

For the analysis of slide frames, character recognition is essential. Tesseract is an open-source optical character recognition (OCR) engine that is widely used, though may fall behind leading commercial engines. [2] We use Tesseract OCR Engine to analyze slide-containing frames and find the most important ones among them.

3. Approach & Implementation Details

In this paper, we first divide the video into slide frames and no-slide frames, using a model trained by SVM. Then, we use Tesseract to select the slide frames with most characters, and K-means clustering to select the most "central" no-slide frames. The result is compared to human-chosen frames.

3.1. Interval Separation

Before slide frame separation, the video frames were separated into intervals. Frames with difference to their next frames higher than transition threshold were labeled transitioning frames. An interval consists of frames between two transitioning frames. By selecting one frame from each interval, we prevented selecting identical or similar frames, as well as transitioning frames. Intervals shorter than 3 seconds were discarded because they may not be static enough to contain valuable information. This process is not required for K-means clustering, because K-means clustering itself has effect of separating frames into K intervals.

3.2. Slide Frame separation

After interval separation, each interval was labeled either slide frame interval or no-slide frame interval. For slide separation grayscale histogram was used instead of original RGB pixel values. Comparing all pixel values are computationally expensive, as well as error-prone. Small shivers of camera or unconscious movement of instructor, as well as noise in videos can seriously damage the comparison results.

Liblinear was used for SVM training of slide-frame separation classifier. Initially, two features, frame's difference to its next frame and normalized sum of grayscale histogram values, were used to train SVM. However, after observing the results, the latter feature was taken out because the difference value by itself produced better separation results.

An interval was labeled a slide interval if there were more slide frames than no-slide frames in that interval. Else, the interval was labeled a no-slide interval.

3.3. Thumbnail Extraction from Slide Frames

Slide frames have following properties: 1. Slide frames have lots of characters. 2. Most of the slide frames have very low difference values to their next frames. 3. Characters for a slide often do not appear all at once, due to animation effect. The most important frame is likely to be the one with the highest number of characters. Because most of the information is conveyed using characters, finding frames with most characters is more effective than finding the transitioning frames or applying K-means clustering.

3.4. Thumbnail Extraction from No-Slide Frames

For selecting no-slide frames, K-means clustering gives high speed and allows us to choose the number of thumbnails. Because the frame closest to the centroid of each cluster is selected, it is guaranteed that K selected frames will be as different to each other as possible, and that none of them will be transitioning frames.

Random frames were selected as initial centroids. This method has possibility of falling into local minima. To minimize such possibility, we ran K-means clustering 10 times, each with different random initial centroids. Then, the one with the smallest distortion function value was selected. [3]

Because most information is contained in slide frames, priority was given to slide frames before no-slide frames. Among the slide frames, priority was given to frames with more characters. That is, if only a few thumbnail is to be chosen, no-slide frames will not be chosen, and only the slide-frames with most characters will be chosen.

4. Experiments

Data used for this project are from Class2go and Coursera. Videos were selected from Class2go's 'Psych 30: Perception' by Prof. Kalanit Grill-Spector, 'An Introduction to Computer Networks' by Prof. Nick McKeown and Prof. Philip Levis, Coursera's 'Probabilistic Graphical Models' by Prof. Daphne Koller, Coursera's 'Neural Networks for Machine Learning' by Prof. Geoffrey Hinton, Coursera's 'Bioelectricity: A Quantitative Approach' by Prof. Roger Coke Barr, and Coursera's 'A History of World Since 1300' by Prof. Jeremy Adelman.

4.1. Slide Separation Classifier Model Training

SVM done to the combination of videos gave accuracy of 91.35%. This accuracy would not be high enough if we were to separate the frames one by one. However, the unit of separation is an interval, not a single frame. Each frame in the interval gets a vote, and if more frames are classified slide frames, the whole interval is classified as a slide interval. (same for the no-slide interval) Thus, even if the interval has the shortest possible length of 3, the probability of classifying it wrong goes down to approximately 2%, which is reasonable. This probability decreases even more as the length of the interval increases.

4.2. Observations

Videos 'A History of World Since 1300 – Atlantic Ocean' and 'An Introduction to Computer Networks – What is Network?' were used for the thumbnail extraction. Transition threshold was empirically set after trials and errors.

1. Interval separation
   From slide intervals, 6 transition frames were obtained and two transition frames were lost. From no-slide intervals, many transition frames were lost and two redundant transition frames were included. Thus, using histogram difference as a means of separating intervals works quite well for slide frames while it shows poor performance for no-slide frames. This supports our usage of K-means for no-slide frames, which does not require pre-separated intervals.
2. Slide separation
   The slide frame separation classifier successfully separated all slides in 'Networks' video and 4 out of 5 slides in 'History' video.
3. Thumbnail extraction
   Thumbnail extraction result is discussed in 5.4 and 5.5.

5. Evaluations

5.1. Evaluation Metric

5 pairs of slide video and no-slide video were analyzed. The metric for evaluation are accuracies of interval frame separation, slide frame separation, and thumbnail extraction. For thumbnail extraction, slide frames and no-slide frames will be evaluated separately. The basis of comparison is human-chosen frames. For slide frame separation and interval separation, sensitivity and specificity were calculated. For thumbnail extraction, we directly compared thumbnail images.

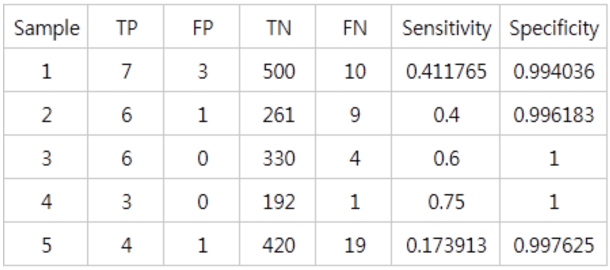
5.2. Interval Separation

Sensitivity and specificity of interval separation method were calculated. True positive is defined as a correctly identified transitioning frame. False positive is a normal frame that is classified as a transitioning frame. False negative is a transitioning frame that is classified as a normal frame. True negative is a normal frame identified as normal frame.

sensitivity = TP/(TP+FN)

specificity = FN / (TN+TP)

Below is the calculated values for the five samples.



Sample	TP	FP	TN	FN	Sensitivity	Specificity
1	7	3	500	10	0.411765	0.994036
2	6	1	261	9	0.4	0.996183
3	6	0	330	4	0.6	1
4	3	0	192	1	0.75	1
5	4	1	420	19	0.173913	0.997625

Table 1. Interval separation performance

We can observe that sensitivity is low; i.e. many transition frames are not caught by our method. This indicates that separating intervals based on a fixed transition threshold is not a good idea. Sensitivity of 0.179 for sample #5 may seem un-usable. However, most of the error comes from no-slide frames, and our method does not use information about intervals when analyzing no-slide frames. (We use K-means instead.) Thus, this does not seriously harm the performance of our method.

5.3. Slide Frame Separation

Sensitivity and specificity of slide frame separation method were calculated similarly.

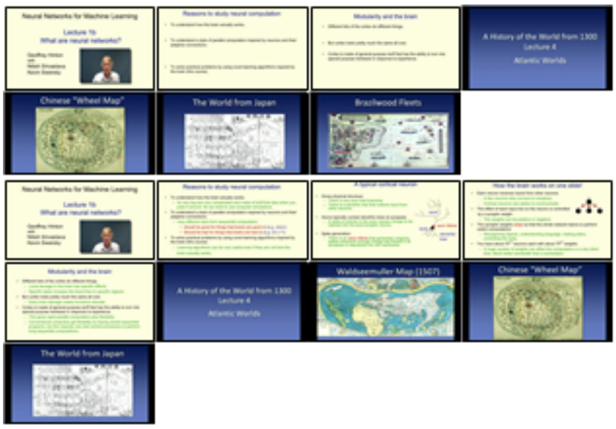
Sample	TP	FP	TN	FN	Sensitivity	Specificity
1	0.542	0.05	0.390	0.017	0.969	0.886
2	0.616	0	0.373	0.011	0.983	1
3	0.403	0.041	0.518	0.038	0.913	0.926
4	0.735	0	0.265	0	1	1
5	0.394	0.016	0.550	0.041	0.907	0.972
Average	0.538	0.021	0.419	0.021	0.954	0.957

Table 2. Slide frame separation performance

We can see that the slide separation method has both high sensitivity and high specificity. Errors are due mainly to two reasons. First, there are slide frames that are too short to be classified as slide frames. Our method requires the slide interval to be at least 3 seconds long to be detected as a slide interval. Second, some slide frames have small video at the corner showing instructor's face. This video usually shows small difference, but sometimes it changes rapidly, increasing the histogram difference, and confusing the classifier.

5.4. Thumbnail Extraction for Slide Frames

Below are thumbnails extracted from slide intervals for the video combination mentioned in 4.2.


Neural Networks for Machine Learning
Modularity and the brain
A History of the World from 1300
Chinese "Wheel Map"
The World from Japan
Brazilwood Fleets
Neural Networks for Machine Learning
Reasons to study neural computation
A typical cortical neuron
How the brain works on one slide!
Modularity and the brain
A History of the World from 1300
Waldseemuller Map (1507)
Chinese "Wheel Map"
The World from Japan

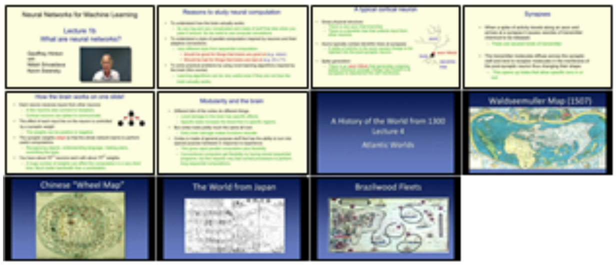

Neural Networks for Machine Learning
Reasons to study neural computation
A typical cortical neuron
Synapses
How the brain works on one slide!
Modularity and the brain
A History of the World from 1300
Waldseemuller Map (1507)
Chinese "Wheel Map"
The World from Japan
Brazilwood Fleets

Figure 3. Thumbnail extraction example, by Kelvinator(top), our method, human-chosen(bottom)

Clearly our method produces better results than the original Kelvinator. Two more frames were selected, and furthermore, each slide frame captured contains more information than ones by Kelvinator. Although our method showed great improvement, it still missed frame #4 and frame #11. These are due to the error in interval separation. Their histogram difference did not reach the transition threshold and were not considered a new interval.

Another example is shown below.

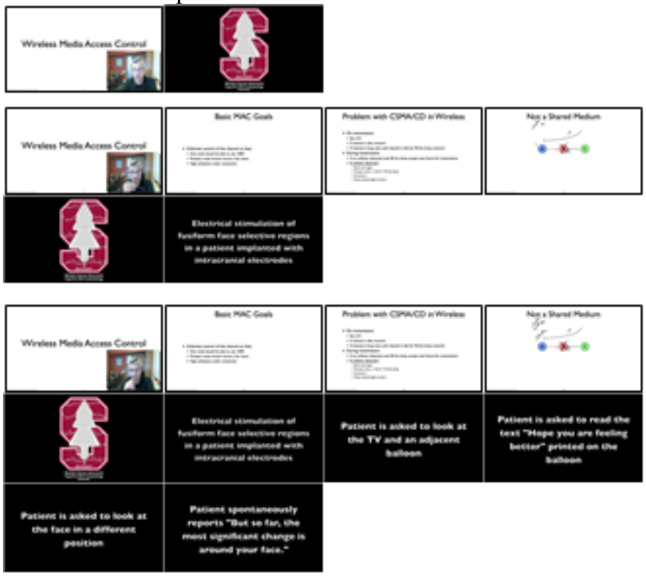
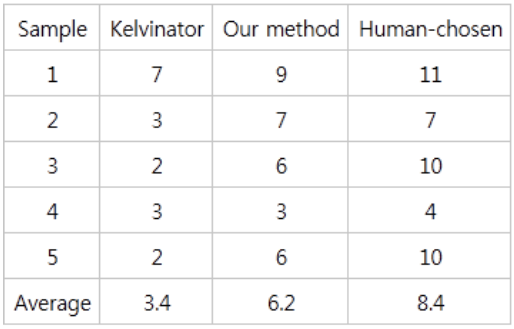

Wireless Media Access Control
Wireless Media Access Control
Basic MAC Goals
Problem with CSMA/CD in Wireless
Not a Shared Medium
Electrical stimulation of fusiform face selective regions in a patient implanted with intracranial electrodes
Wireless Media Access Control
Basic MAC Goals
Problem with CSMA/CD in Wireless
Not a Shared Medium
Electrical stimulation of fusiform face selective regions in a patient implanted with intracranial electrodes
Patient is asked to look at the TV and an adjacent balloon
Patient is asked to read the text "Hope you are feeling better" printed on the balloon
Patient is asked to look at the face in a different position
Patient spontaneously reports "But so far, the most significant change is around your face."

Figure 4. Another thumbnail extraction example

Here Kelvinator shows very poor performance, selecting only two thumbnails with almost no useful information. The performance was greatly enhanced using our method. However, our method also missed four frames. The four missed frames were in intervals shorter than 3 seconds, and were discarded before selection. Usually intervals shorter than 3 seconds are transitioning intervals or some short videos. However, as this case shows, sometimes short intervals do contain important information.



Sample	Kelvinator	Our method	Human-chosen
1	7	9	11
2	3	7	7
3	2	6	10
4	3	3	4
5	2	6	10
Average	3.4	6.2	8.4

Table 1. Number of chosen frames in slide intervals

Above table shows how number of frames chosen in slide intervals increases using our method.

5.5. Thumbnail Extraction for No-Slide Frames

Below are thumbnails extracted from no-slide intervals for the video combination mentioned in 4.2.

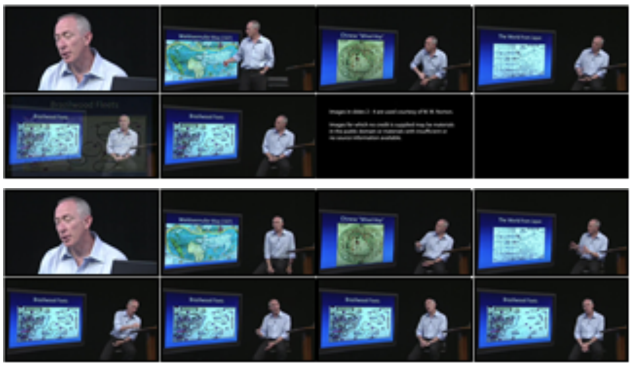


Figure 5. Thumbnail extraction on no-slide intervals, Kelvinator(top), our method(bottom)

Frame #5 of Kelvinator is a transitioning frame, and frame #7, 8 are blank frames. Result produced by our method do not have such frames. Frames #6, 7, 8 are very similar, but this is due to our setting k=8 to compare with Kelvinator. By decreasing k we can easily remove redundant frames.

6. Future Work

Each of the three stages of the process, slide frame separation, interval separation, and thumbnail extraction, has a chance of improvement. Currently only histogram difference value was utilized for training the slide frame classifier. More features could be added to improve performance. Using fixed value of transition threshold showed reasonable performance for conducted experiments, but this method is prone to errors. As mentioned before, a small instructor video placed at the corner of the slide can cause great histogram difference, wrongly breaking the interval into many intervals. This could be s olved by either finding a way to change the transition threshold as frames change, or using a totally different way of determining intervals. We have also seen in the previous sections that sometimes important information is contained in short intervals. Here is a trade-off between sensitivity and specificity. If we decrease minimum required length to be chosen as a thumbnail, it becomes more likely that meaningless frames will be chosen as thumbnails.

For this project we have only assumed that the number of characters contained in the frame is the only factor that determines the importance of the frame. However, this is not true. Some frames do not contain characters at all, but contains important diagrams. Some instructors purposely place the most important concepts by itself, with no description, so that students can concentrate. Many other features including diagrams, graphs, tables, and colored letters could be identified and used as a feature for machine learning. These approaches will eventually allow us to get closer to the definition of importance.

7. Conclusion

In this paper, we have proposed and analyzed a new method of thumbnail extraction using character recognition and K-means clustering algorithm. Blank frames, overlapping frames, and redundant frames were removed. Moreover, greater number of more important frames were captured compared to conventional histogram difference threshold method. Nevertheless, the new method still fails to capture some of the important frames. We expect more features to be added in the future.

References

[1] V.T. Chasanis, A.C. Likas, N.P. Galatsanos, Scene Detection in Videos Using Shot Clustering and Sequence Alignment, IEEE Transactions on Multimedia, Vol. 11, No. 1, 2009.

[2] R. Smith, An Overview of Tesseract OCR Engine, Proc. Ninth Int. Conference on Document Analysis and Recognition, IEEE Computer Society (2007), pp. 629-633

[3] A. Ng, The K-means Clustering Algorithm, CS229 Lecture Notes

8. Appendix

This is a part of joint project for CS229 and CS231A. Image processing techniques are used to analyze videos, and various algorithms covered in class, including clustering and classification algorithms, were used. Sef Kloninger is the mentor and corresponding author, and there is no student partner.