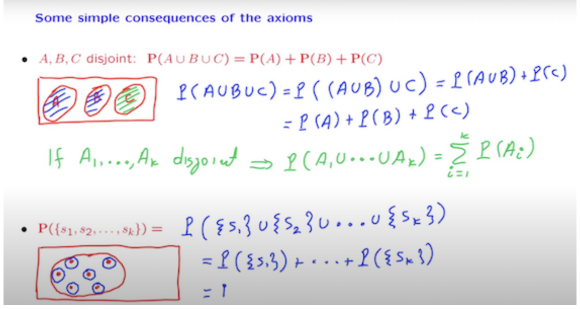
Note preparation from Video lectures for hybrid style educational videos

Written as part of submission for AI 1 seminar
Yogesh Singh[1]
[1]Faculty of Mathematics and Physics, Charles University
Malostranska Namesti 25

Abstract

With the availability of high speed internet, knowledge gathering from videos has become a widespread practice. Before, people used to primarily watch videos for leisure, but lately it has become a way for people to research any topic or just to study and understand a difficult topic. This is most reflected in the popularity of MOOCS and companies like Courses and Udacity creating learning modules primarily through video. Although these videos and course are structured and prepared in ways that it is easy to browse and look for relevant topics, often it is time consuming if you are looking for a specific information or explanation. Moreover, In MOOCS, which are organised and often provide annotated transcripts, the majority of topics are not covered. Students have to rely on classroom style videos like plain recordings of blackboard style teaching or through slides. I want to propose creating an AI system which can provide searchable transcripts with attributes such as equations and pictures extracted with the relevant explanation attached to it. For achiveing this I will use visual recognition and NLP techniques to extract meaningful notes which can then be used to organise the lectures into several subtopics so that they can be navigated effieciently.



Some simple consequences of the axioms

- A, B, C disjoint: $\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$

$$P(A \cup B \cup C) = P((A \cup B) \cup C) = P(A \cup B) + P(C)$$
$$= P(A) + P(B) + P(C)$$

If $A_1, \ldots, A_k$ disjoint $\Rightarrow P(A_1 \cup \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$

- $\mathrm{P}(\{s_1, s_2, \ldots, s_k\}) = P(\{s_1\} \cup \{s_2\} \cup \cdots \cup \{s_k\})$

$$= P(\{s_1\}) + \cdots + P(\{s_k\})$$
$$= 1$$

Figure 1: Example of hybrid style video. Note that the slides contain both the slide type content and handwritten parts which the instructor use to explain.

Introduction

There has been work done for generating such documents from various types of videos which can broadly classified into two types. Blackboard style and slides type. The former is old school way of teaching and consists of a lecturer writing down the information on blackboard and explaining them as he does it. On the other hand in the slide type videos we have a set of slides which are explained by the instructor as he navigates through them. It should be noted that not all type of teaching can be done only with slides. Most of these educational videos are long and lack annotations to quickly navigate the video. Some work has been done in this regrad to create more information from processing these videos. (Biswas 2015, Zhao 2017). But having an option of annotated video greatly help students to selectively choose among the various subsections of these videos. E-learning platform such as coursera and also the vidoes on Youtube rely heavily on the annotations done manually by the content creator themselves if such an option is available. Work has been done to create visual notes in these two style of videos previously (Shin 2015; Xu 2019). They presented a method to generate straightforward lecture notes for the blackboard-style lecture videos which establishes the correspondence between the visuals and the subtitles by time-align technology (Shin 2015).

Moreover, with online teaching becoming more common, more notably in past couple of years, teachers have used various other tools available to them to do the teaching. This tool set consist of using tablets to write with hand as in the case of online teaching and using online short quizzes to assess the understanding of students at various intervals. In this form a teacher uses a hybrid system of teaching which consists of slides and tools to either directly write on top of the slides or use a blank digital board to do so. The audio also is an important part of the content being presented. Adding the transcripts and its time information will allow more features and keywords to be extracted reliably. Chen et al and Zhao et al analysed the audio information and extracted key terms to make the which then were embedded on original slides (Chen, Huang 2010, Zhao; Lin 2017). These works use slide level matching of audio and video cues to create useful systems.

I want to create an AI system which can allow us to create notes from this hybrid system of teaching. The main goal is to determine various parts which belong to either the slide type content and the board type and using methods previously used create a system which can process these videos. Also the previous work has focused more on creating PDF style notes only. Here I would like to make a more usable form by generating HTML pages with embedded links to smaller subsections from the video itself.

Design

Designing any system like this requires 3 broad steps

- **Information extraction:** First step is to extract all the audio and visual entities present in the slide or the written boards. These can be text, equations, graphs, figures etc.
- **Create information maps:** We need to arrange all the extracted entities in groups and also align the audio snippets with each of the visual objects. This involves creating connected objects placed in time.
- **Layout:** After the visual and audio elements are recognised and map we have to arrange them in readable and navigable web pages. They should also involve links to the video snippets where they were extracted from so that they can be referred to if needed.

In the following sections I shall present ways for doing the above.

Visual and Audio Information extraction

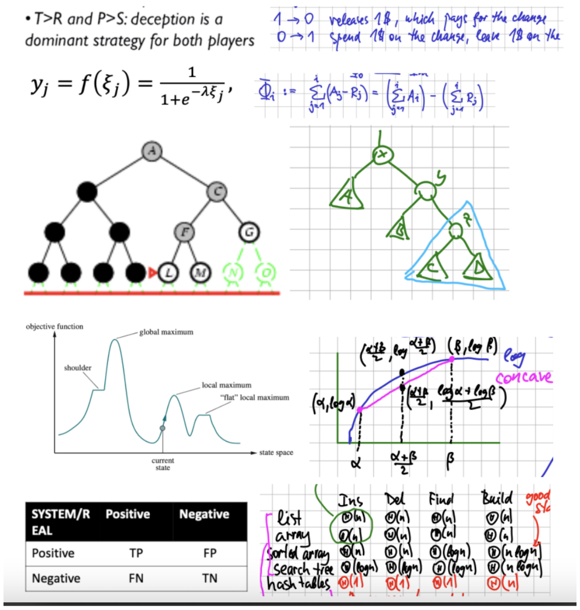
It is the first task for such a system. The information which needs to be extracted could be present in various ways. Some of them are as follows

- Text
- Equations
- Graphs
- Tables
- Diagrams
- Transcripts. (With time information)
- The affect of the audio. (Pitch, Tone information etc.)

The first step is to differentiate between the handwritten part and the computer generated data present on the slides. These shall be processed in separate pipelines. Later they can be placed according to their temporal signature.

Each of the entity is to be detected as an 8-connected segment. Since the segments are usually in horizontal direction, they should also be combined in this direction. Smoothing algorithm like Run-length Smoothing Algorithm (Wong and Casey 2010)will help us combine elements in horizontal direction if they are separated. Some heuristics can be used to combine uncombined segments which might have been separated out. For example comparing the heights and centroid of the bounding boxes found in (Xu 2019).

In case of the written text. We need to ignore strokes like highlighting of the already existing boxes. Once all the boxes are determined they can then be identified to be one of the many categories which we had determined previously.



• T>R and P>S: deception is a dominant strategy for both players

1 → 0 releases 1$, which pays for the change
0 → 1 spend 1$ on the change, leave 1$ on the

$$y_j = f(\xi_j) = \frac{1}{1+e^{-\lambda\xi_j}},$$

$$\Phi_i := \sum_{j=1}^{i}(A_j - R_j) = \left(\sum_{j=1}^{i} A_i\right) - \left(\sum_{j=1}^{i} R_j\right)$$

SYSTEM/REAL	Positive	Negative
Positive	TP	FP
Negative	FN	TN

	Ins	Del	Find	Build
list	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$
array	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$
sorted array	$\Theta(n)$	$\Theta(n)$	$\Theta(\log n)$	$\Theta(n \log n)$
search tree	$\Theta(\log n)$	$\Theta(\log n)$	$\Theta(\log n)$	$\Theta(n \log n)$
hash tables	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	$\Theta(n)$

Figure 2: Examples of computer generated and handwritten types of various entities namely from top to bottom (and Left: digital Right: Handwritten) Text, Equations, Diagrams, graphs, Tables

$$\mathcal{H}_{new} > 2 * \mathcal{H}_{cur} \quad or \quad |\mathcal{O}_{new} - \mathcal{O}_{cur}| > \mathcal{H}_{cur}$$

Figure 3: Example of heuristic to seperate different entities. H_{curr}= Height of current bounding box considered for expansion, H_{new}= Height of bounding box being encountered, O_{curr}= Centroid of current bounding box considered for expansion, O_{new}= Centroid of bounding box encountered

Supervised Image recognition can be used to have various extracted subentries classified.

Various other features of the entities can be used to classify them in various types some of them are

- **Aspect ratio:** It is different for text, Formulae, images and tables.
- **Special Symbols:** Equations have a higher count of special symbols than others.
- **Alignment of characters:**is different in graphs, texts and formulas.
- **Special Characters:**Presence of special structure like Cartesian axes and cells in cases of graphs and tables.

Images can be grouped as a misc type category containing all of entities which were not classified.

For the next type of content where the instructor is purely writing over a blank or almost empty space present in the slides we have to use a more sophisticated pipeline to extract the entities. In this case a **stroke** is defined as a continuous curve made by the instructor without lifting away the pen/stylus/chalk(on boards). Different strokes can be combined to form a compact entity recognition. Start and end

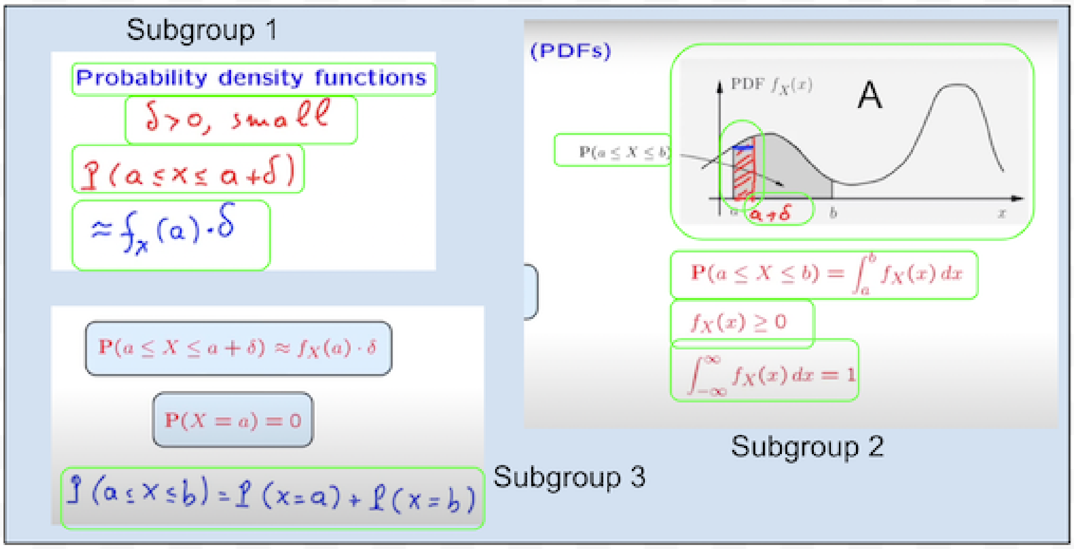


Figure 4: Example of how subgroups will be formed within each of the slide. Each subgroup has various connected entities. Note that some Entity (Like A) can have smaller entity within them

of each stroke is measured and is successively added to a set. Line segmentation algorithm used in (Knuth and Plass 1981). The set is closed if it reaches a stable state and the region for it is defined. (Xu 2019) showed equivalence of this problem to a line-breaking problem– arranging words in a paragaph into a line. We can consider it as grouping a set of elements (strokes or words) into sets(lines or entities). In addition the text written is usually accompanied by verbal communication of the same. This will be important for later parts of our task. In graphic tablets, Instructor uses set of strokes to create the information which they want to convey. One more advantage is that the information is created left to right and from top to bottom. Keeping this in mind we can extract entities at various states and then combine them to a set of sets. Where each set consist of a stable entity which does not change at various stages of construction. We can use heuristics defined in previous work. A horizontal gap expected value is about 100 pixels (Shin and Berthouzoz 2015).

In cases of graphs or diagrams drawn, a large change corresponds to the starting of a new entity and less changes occur towards the end. We can identify this by setting up threshold difference between the foreground and the background.

Many times the lecturer uses pointers, cursors or digital pens on the slides to highlight the areas which they are currently being talked about. These changes can be captured by comparing each slide at different time steps. Each entity will thus be a set of temporal forms. Tracking changes with their time stamps can be used in the later parts when the flow map is created in conjunction with the audio information. These sets of the same segments can be used to create gifs for representations in the final part of output formation.

Each of the type of visual entities should be recognised and extracted from the stills of videos. We should be able to recognise changes within each of the entities locally and have a sequential changes saved.

Next task is to identify audio and generate transcript by using many of the state-of-the art deep learning architectures available such as DeepSpeech (An implementation of CTC loss based architecture). We can use algorithm such as (Rubin 2013) built on the HTK speech recognition. This algorithm takes an audio as an input and outputs the transcript with timestamps for each word from start to end.

These audio can be aligned with the entities extracted. The transcripts can be easily grouped with each slide or a board present in the video without much processing. In general, the time from the appearance of a visual entity(text, formula or graph) to the following one can be recorded, and speech text within this duration can be considered as the explanation for the previous visual entity.

Creating Maps of information

In this we try to recognise information out of the text written or the diagrams presented. Our aim is to create maps between objects like written text, equations, graphs etc with the audio information. This consists of various tasks. First and the most straightforward is the having various silos of sub topics. This can be done by organising various subtopics and entities within as a tree of information. Each entity would be aligned and connected with various snippets of audio.

Videos are generally divided into subtopics by the design of the material. This is most often conveyed explicitly orally or using headings on the boards, which are many times uppercase or underlined or have a larger font than others. I propose segmenting each slide or board into various subgroups. Each subgroup would have multiple entities. Usually the instructor talks about these entities back and forth. We have to keep in consideration that audio always compliments visuals. The audio is present usually in one of the following two ways. First while the instructor is reading any material present in the slides. it is also true when the instructor is drawing or writing something over the board or the slides. These entities can be directly aligned with the information currently being discussed about. Secondly while the instructor is not talking about anything explicitly present in the slide. This audio should be matched with the last known subgroup. Concept of bracketing. Each entity group will have a time bracket attached to it, in which the instructor talks about the group of entities. This bracket will be extended in the tail direction till the recognition of new possible subgroup is done. This way each audio snippet will also have a corresponding entity. This information is stored by means of connected graph. A scoring function can be defined for grouping consecutive entity so that the groups with highest score can be used. The scoring heuristic function should use features from the meta data of entity.

Each entity and subgroup has many keywords and meta data (features like aspect ratio, font size, color etc). These can be also used to make more connections among the entities and groups. Methods such as semantic similarity can be used to match words from transcript with various entity in the sub-group. Word vectors are well suited for this task. Pre-trained vectors such as fastText—Word vectors trained on Wikipedia articles— can be used to find similarity between written text extraction. Certain entities which are not texts such as tables and figures can be matched when the mentioning of words such as "graph, diagram, figure etc" is present. This categorisation has been already done before. Moreover when the Instructor is drawing such graphs and tables it is easier to make connections as the evolution of entity is already attached to the audio at that time. Figure

We have now constructed a map of connections between various entities within a slide. As we know this is a difficult problem to teach to learning algorithm. With some heuristics we can say that we have around two major categories of subgroups. 1) text and equations 2) graphs, diagrams, tables. Broadly these group complement each other and there is less references among the group itself. So if there is a relation found between any group. They should be displayed together with the corresponding transcript. Moreover, the text itself might contain references to other entities with keywords as described above.

Layout Design

The aim here is to produce structured webpages with embedded components and links to smaller video segments containing sub-topics which are presented in the video. Each topic can contain upto multiple slides and handwritten digital boards. More often than not a different topic are present in different slides/boards. Hence we can design the page for each slide or board.

The design will be done at a slide/board level. These can be called blocks. Each entity present in the slide or block is also assigned an importance level. This can also be derived from various metrics like the amount of time spent on each sub-topic, no of connections etc. This importance metrics is usually high for graphs and tables. As the Instructor mostly talks about these a lot. After which equations can be arranged. Text is the least important, as the text present will have a high chance of being said verbatim or described in the audio extracted. The text associated highly with the graph and figure entity shall be part of those entities. The association is done by matching transcripts corresponding to each entity, adjacency of each entity in space, Presence of same fonts and their sizes etc.

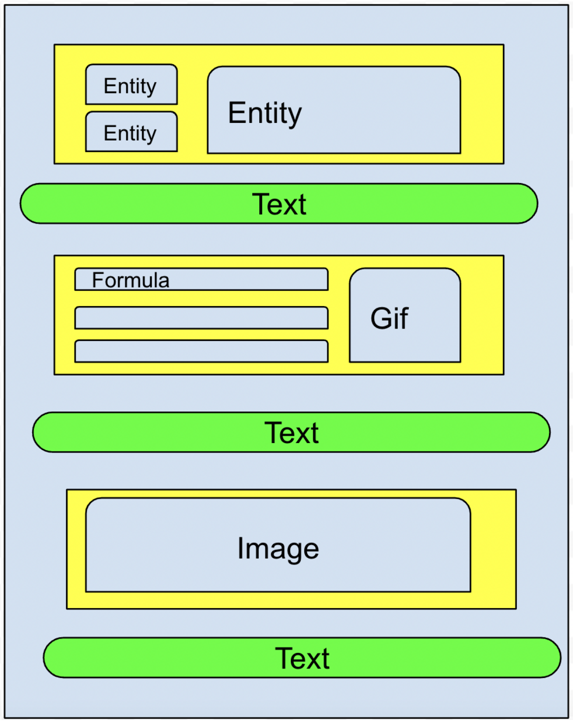


Figure 5: Final layout for generation of blocks. Note that Each block will be a slide or a board. Within this there will be smaller subgroups. The text below will consist of the transcript corresponding to the subgroup

Within this, a recursive algorithm can be used to generate connected representation. Recursion will help with the entities which have smaller entities within them like in case of graphs. Each entity group can then be shown and the transcript corresponding to the same shall be present below it. It has been illustrated in the graph entity in Figure on previous page. Links of each video snippet corresponding to the segment can be present below each subgroup.

At the end matching should also be done again to make sure each slide or board is represented in one block. In case of board style videos, a changes which were determined before should be presented as gif format image to show changes. For example a graph made by instructor shall come a gif after certain threshold of change was done. As the generation would mostly have errors present due to many factors. Hence, at the end of the block the link to the video snippet and the snap shots of stabilised boards be present. This can help student review and go to the video directly if its not clear.

Summary

I presented here a way to make online HTML notes using educational videos. This needs multiple steps. First the extraction of information to make entities representing connected elements. Then organising them in a map with connections derived from audio and visual data extracted. Lastly, it was discussed how to present all the information into an online documents with links. This would be a preliminary process and would need further addition and deletion of ideas suited for creating such notes from video.

Acknoledgements

To illustrate the process, I have taken screenshots from multiple courses and their corresponding videos here at Charles University. They are Data Structures 1, Neural Networks, Artificial Intelligence 1. I have also taken a few screenshots from the MIT Courseware ”Introduction to Probability” course present on Youtube.

REFERENCES

Arijit Biswas, Ankit Gandhi, and Om Deshmukh, 2015, “Mmtoc: A multimodal method for table of content creation in educa- tional videos,” in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 621–630.

Baoquan Zhao, Shujin Lin, Xin Qi, Zhiquan Zhang, Xiaonan Luo, and Ruomei Wang, 2017, “Automatic generation of visual- textual web video thumbnail,” in *SIGGRAPH Asia 2017 Posters*. ACM, p. 41.

Shujin Lin, Baoquan Zhao, Xiaonan Luo, Songhua Xu, and Ruomei Wang, 2017, “A novel system for visual navigation of edu- cational videos using multimodal cues,” *in Proceedings of the 2017 ACM on Multimedia Conference*. ACM, , pp. 1680– 1688.

Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl, 1982, “Document analysis system,” *IBM journal of research and development*, vol. 26, no. 6, pp. 647–656.

Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Fre do Durand, “Visual transcripts: lecture notes from blackboard-style lecture videos,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 240, 2015.

Xu Chengpei, Wang Ruomei, Lin, Shujin, Luo Xiaonan, Zhao Baoquan, Shao Lijie, Hu Mengqiu, 2019, Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos, *IEEE International Conference on Multimedia and Expo (ICME)*, 898-903.

D.E. Knuth, M. F. Plass 1981. Breaking paragraphs into lines. *Software: Practice and Experience 11*, 11, 1119–1184.

Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., AND Agrawala, M., 2013. Content-based tools for editing audio stories. *In Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM, 113–122.