Automatic Identification of Keywords in Lecture Video Segments

Raga Shalini Koka   Farah Naz Chowdhury   Mohammad Rajiur Rahman   Thamar Solorio   Jaspal Subhlok
Department of Computer Science, University of Houston, Houston, TX 77204
Email: ragasalini@gmail.com, fchowdhury4@uh.edu, mrahman13@uh.edu, tsolorio@uh.edu, jaspal@uh.edu

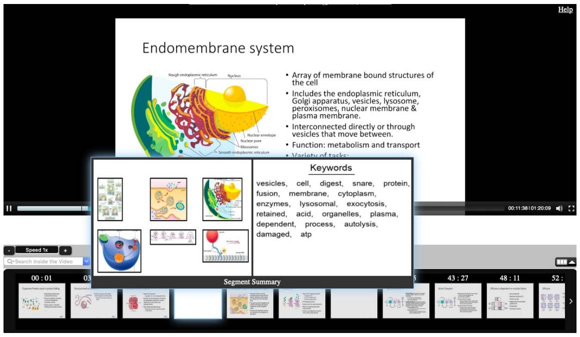
***Abstract*—Lecture video is an increasingly important learning resource. However, the challenge of quickly finding the content of interest in a long lecture video is a critical limitation of this format. This paper introduces automatic discovery of keywords (or tags) for lecture video segments to improve navigation. A lecture video is divided into topical segments based on the frame-to-frame similarity of content. A user navigates the lecture video assisted by visual summaries and keywords for the segments. Keywords provide an overview of the content discussed in the segment to improve navigation. The input to the keyword identification algorithm is the text from the video frames extracted by OCR. Automatically discovering keywords is challenging as the suitability of an N-gram to be a keyword depends on a variety of factors including frequency in a segment and relative frequency in reference to the full video, font size, time on screen, and the existence in domain and language dictionaries. This paper explores how these factors are quantified and combined to identify good keywords. The key scientific contribution of this paper is the design, implementation, and evaluation of a keyword selection algorithm for lecture video segments. Evaluation is performed by comparing the keywords generated by the algorithm with the tags chosen by experts on 121 segments of 11 videos from STEM courses.**

2020 IEEE International Symposium on Multimedia (ISM) | 978-1-7281-8697-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISM.2020.00035

I. Introduction

Lecture videos are widely employed as the core medium for online learning, and as a supplementary tool for traditional face-to-face learning. When classroom lectures are captured on video and made available to students, they make use of them, enjoy using them, and perceive them to be a valuable learning tool. Students value videos for allowing them to review at their own pace and typically report a positive impact on grades and overall course satisfaction [1], [2].

Proposed research has its roots in the Videopoints project (www.videopoints.org) whose central goal is to convert a lecture video into an interactive learning companion by overcoming the difficulty of quickly navigating to the content of interest in a long lecture video. Videopoints developed and evaluated a lecture video framework with innovations in indexing, search, and captioning [2]–[4]. Figure 1 shows a screenshot of the Videopoints player highlighting topic based indexing and summarization. An index panel at the bottom shows the first frames of automatically generated video segments representing subtopics in the lecture. When a user hovers over the index frame of a segment, a summary frame appears as illustrated, consisting of a list of keywords, that is the topic of this paper, and a visual summary, that is discussed in related work [5]. Videopoints has been deployed at the University of Houston for coursework across Biology, Chemistry, Computer Science, Geology, Mathematics, and Physics.



Endomembrane system

- Array of membrane bound structures of the cell
- Includes the endoplasmic reticulum, Golgi apparatus, vesicles, lysosome, peroxisomes, nuclear membrane & plasma membrane.
- Interconnected directly or through vesicles that move between.
- Function: metabolism and transport
- Variety of tasks:

Keywords

vesicles, cell, digest, snare, protein, fusion, membrane, cytoplasm, enzymes, lysosomal, exocytosis, retained, acid, organelles, plasma, dependent, process, autolysis, damaged, atp

Fig. 1. Videopoints player showing topical indexing and summaries

The goal of the research presented in this paper is to build a list of the most relevant keywords or tags to represent the content of a segment (or section) of a lecture video. The keywords improve navigation by giving users an idea of the content in different parts of a lecture video.

The methodology for keyword identification starts with extraction of text from the frames of a video segment with OCR (Optical Character Recognition) tools. A preprocessing phase eliminates extraneous background text, makes spelling and language corrections, removes invalid N-grams, removes stop words, and reduces words to their root word with stemming. Subsequently, the importance of each N-gram is determined by their frequency in the document, font size, the time on screen, and if they exist in the relevant field and English dictionaries. The top ranked words are selected as keywords, with the exact number based on user provided parameters.

Most recent work on identifying keywords in lecture videos has focused on tags for indexing the entire video, and relied on automatically generated audio transcripts to extract keywords [6]–[9]. For example, [8] employed an unsupervised clustering approach to keyword summarization of Massive Open Online Courses (MOOCs) that outperformed state-of-the-art methods such as PositionRank [10], and SingleRank [11]. Our experience in a related project indicates that OCR text is likely to be a better guide to keywords than a speech transcript [3]. Traditional approaches to topic modeling such as latent Dirichlet allocation (LDA) [12] have shown promising results, but are not suitable for very short documents [13], which is

978-1-7281-8697-9/20/$31.00 ©2020 IEEE
DOI 10.1109/ISM.2020.00035

the case for the screen text content of lecture videos.

II. Keyword identification

This section outlines the process of keyword selection based on the text on the frames of a lecture video segment. A keyword may contain 1 to 3 words; it can be a 1-gram, 2-gram, or a 3-gram. The output is a list of keyword N-grams with their *importance* on a 0-1 scale. Additional details are available in [14].

A. Text preprocessing

The input to the keyword generation process consists of all words identified by OCR on the lecture video frames in the relevant video segment, along with their font size, location on the screen, and the duration on the screen. Following steps improve and organize the input text:

- *Removal of background text:* Most lecture videos have some fixed screen content that is not relevant to the topic of discussion. Examples include a course title, name of an organization, or the taskbar of software in use. Clearly, such terms are not likely to be keywords. An algorithm was developed to eliminate such "background" text. The basic premise is that background text occurs repeatedly at the same location on the screen for most or all of the lecture videos. The algorithm identifies and eliminates such text. The implementation allows for small variations in the location and content of the potential background text to account for OCR errors and other minor changes in the presentation material.
- *Error Correction:* The text discovered with OCR often contains errors. The causes range from poor quality of the input image, small or exotic fonts, and unusual color separation between text and background. An automated context based error correction was applied using *Google Spell Check* API; an approach suggested in [15]. This helps correct these two types of errors:
  1) *Non-word errors* occur when the OCR outputs sequences of characters that do not exist in the language dictionary. Such non-words are replaced by another dictionary word a short edit distance from it, if one exists that fits the context.
  2) *Real-word errors* occur when OCR outputs an incorrect but valid word. This work employs *"Did you mean?"* feature from the Spell Check API to replace a sequence of words with an alternate phrase when, most likely, that was the intended phrase.
- *Generation of valid N-grams:* The next step is to discover and validate N-grams from the raw text. Only unigrams, bigrams, and trigrams are considered in this work as the likelihood of a larger n-gram being a keyword is very low. Discovering n-grams from OCR extracted slide text poses some unique challenges as i) OCR scans an image left to right and top to bottom, and does not guarantee contextual relation between consecutive words and ii) content on slides usually does not include proper sentences as it lacks punctuations and sentence boundaries.

  Any N-gram that can be constructed from consecutive words is validated with the *Phrase Finder* API which matches a sequence of words against a list of valid N-grams based on the content of 'Google Books Service' that covers digitized versions of over 5 million books. N-grams that contain repeated words or start and end with stop words are removed, as they are unlikely to be keywords.
- *Stop word removal*: Stop words are frequently occurring and trivial words which help frame sentences but do not represent meaningful content. Articles, Prepositions, Conjunctions, and Pronouns are generally considered stop words. Examples of stop words include: *a, an, the, it, and, as, what, how*. The list of stop words is obtained from an external source [16]. Stop words are not removed if they are part of a valid N-gram, such as "cost of living".
- *Stemming:* A stem or root is the part of word retained after removing its suffixes and prefixes. The stemming process groups all versions of the inflected word to a canonical form. For instance, nouns 'Computer', 'Computers', and 'Computing' are all associated with the stem 'Compute'. This work studied various available stemming algorithms and applied the Snowball stemmer [17]. The stemming process proceeds as follows: i) All occurrences of words are replaced by their stems, and ii) occurrence of each N-gram in the original text is replaced by the most commonly occurring N-gram with a matching stem.

B. N-gram Importance

The text extraction and preprocessing identify a sanitized list of N-grams in a lecture video segment, along with their frequency, font size and the duration for which they were displayed. This section describes the factors that determine the importance of an N-gram for consideration to be a keyword.

- *Term frequency:* The number of times an N-gram occurs in a segment. When the same word occurs in a unigram, as well as a longer N-gram, the occurrence in the longer N-gram is not counted towards the term frequency score in the unigram to prevent double counting.
- *Inverse document frequency:* In addition to the raw frequency of an N-gram, it is also important to consider the *relative frequency* of the N-gram in a lecture video. For illustration, in a course on computer networks, the word "network" is likely to occur often throughout. However, it is unlikely to be an interesting keyword for any segment. Hence, this work considers "Inverse Document Frequency" score defined as follows:

  $$IDFScore = log(TotalSegments/NgramSegments)$$

  where $TotalSegments$ is the total number of segments in a lecture video and $NgramSegments$ is the number of segments where the relevant N-gram occurs.
- *Font size:* Text in larger fonts, that typically includes titles, is more likely to be a good keyword. Font size is part of the OCR output. Font size value is squared and then normalized from 0 to 1. The squaring is motivated

by the fact that fonts typically vary in a small range in a document. The font weight is then averaged across all occurrences of an N-gram.

- *Time on screen:* The amount of time for which an N-gram is displayed is another consideration in determining importance. The hypothesis is that the longer an N-gram is displayed, the more likely it is to be a good keyword. A normalized *time score* is calculated for N-grams.
- *Domain importance:* Important keywords often have a unique meaning or relevance in an academic domain. For example, the word "solution" may be worthy of additional consideration in a chemistry lecture. The domain score of a unigram is simply 1 or 0 based on whether it occurs in the domain dictionary or not. For N-grams with k words in the domain, the corresponding score is 2^k. Oxford Reference Dictionaries were employed to determine if a word belongs to a specific domain.
- *Rare word analysis:* Most words in a lecture video segment are part of the English dictionary and/or part of a domain dictionary. However, some words may not be part of any dictionary, possibly because of an OCR error or use of proper nouns. Such words are less likely to be good keywords. We compute a *Rare word score* based on whether they occur in any dictionary or not. The score for N-grams is computed in a manner similar to the computation of domain importance.

C. Keyword importance computation

The final importance score computation for an N-gram proceeds as follows:

$$Importance = FreqWt * FreqScore +$$
$$FontWt * FontScore + TimeWt * TimeScore$$

representing the sum of the frequency, font, and time scores with parameterized weights. The computations of $Fontscore$ and $TimeScore$ are described in the discussion above. $FreqScore$ is the weighted sum of i) term frequency score, ii) IDF score, iii) domain score, and iv) rare word score described in this section. Currently weighting parameters are selected based on experience with lecture videos with more details in [14]. Finally the importance scores are normalized in a range from 0 to 1.

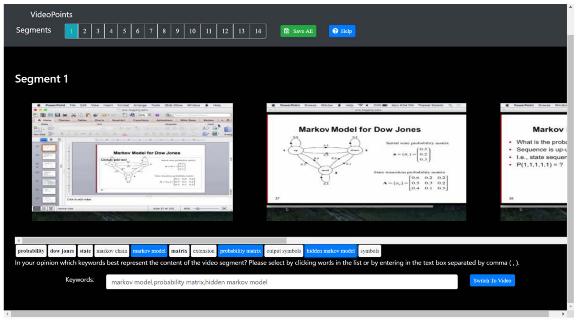
III. Evaluation and Results

The automatic keyword selection portal has been implemented in the Videopoints lecture video portal that is in active use. For evaluation, algorithm keyword selections were compared against the ground truth provided by users.

A. Ground truth collection

Ground truth for evaluation is the keywords selected by a human expert. Ground truth collection was done by surveying instructors and students familiar with the video content.

Survey instrument: A web-based survey tool was developed to allow a human expert to conveniently select keywords in lecture video segments. The main interface for the survey tool is illustrated in Figure 2. The bar at the top represents the segments present in the video lecture; segments 1-14 in this example. The user can click on the segment numbers to navigate to the corresponding video segment. The slides/frames of the current video segment are shown on the main display in the middle. All valid N-grams in the segment are potential keywords that are listed below the video frames as a row of buttons. The user can (un)click the buttons to select keywords. Additionally, a user can add any keywords not in the list in the editable textbox below the labeled keyword buttons. The user can play the video segment at any time by clicking on the button labeled "Switch to Video" at the bottom right.



In your opinion which keywords best represent the content of the video segment? Please select by clicking words in the list or by entering in the text box separated by comma (,).

Keywords: markov model,probability matrix,hidden markov model

Fig. 2. Interface for a user to provide ground truth keywords

Survey data collection: Ground truth was collected for 121 video segments. The subject areas, in decreasing order of the number of selected segments, are Biology, Computer Science and Chemistry. 16 students and instructors in the relevant subject areas participated in the survey. Keywords were provided by one survey participant for 22 segments, 2 participants for 80 segments, and 3 or more participants for the remaining 19 segments.

B. Evaluation Methodology

The parameters in the system were set based on experiments and our experience. For the results presented, the algorithm assigned all valid N-grams an importance score in the range of 0-1 as discussed in section II. The top 20 selections were selected as keywords, with the condition that an N-gram must have at least a score of .4 in the 0-1 importance scale to be considered. Since ground truth keywords are not well defined when humans disagree, and since there can be a partial match with the ground truth for N-grams, the evaluation considered the following scenarios:

- *All or Majority users*: The ground truth is composed of keywords selected by all users who evaluated that segment, or a majority of the users.
- *Strict or Partial scores*: With the strict criteria, 2 N-grams are not a match if even a single component word is different. With the partial criteria, a partial match is included in the analysis. A partial match is the ratio of the number of matching words in predicted and ground truth N-grams, to the total number of unique words in the two N-grams.

C. Results

The performance of the algorithm is plotted in Figure 3. For majority users, with credit for partial scores, Precision, Recall, and F1-measure are 0.52, 0.79, and 0.62, respectively. For a problem of this nature where ground truth is ambiguous, we consider the results to be encouraging and the resultant system to be practically useful.

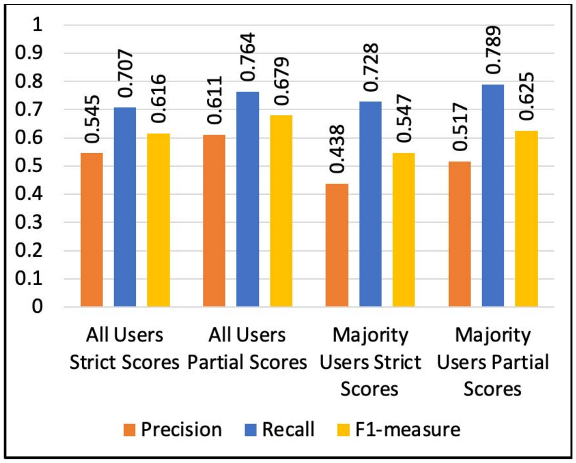


Fig. 3. Performance of keyword selection algorithm

Figure 4 validates that the algorithm selected keywords are far superior to randomly selected keywords for the majority users with credit for partial scores case. Random keywords were selected after the text preprocessing phase discussed in section II; otherwise the random results would be much worse.

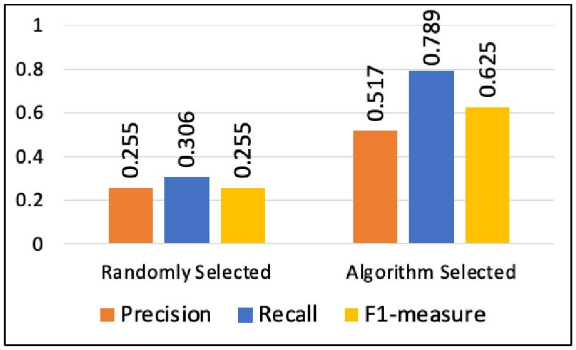


Fig. 4. Comparison of randomly selected and algorithm selected keywords

IV. Conclusions and Future Work

This paper introduced a methodology to identify keywords in a lecture video segment by extracting text from video frames with OCR and applying a diverse set of techniques including autocorrection, removing stop words, stemming, and analysis of term frequency, font size, duration, and presence in field and language dictionaries. Results are presented by comparing the ground truth collected from active users to algorithm generated keywords. The keyword module is integrated with Videopoints, an advanced lecture video portal that is actively used by instructors and students.

Ongoing research is performing ablation studies to develop a detailed understanding of the factors that impact keyword selection. Future work will address combining speech with OCR text to improve the quality of keywords. Finally, a usability study is needed to determine the real world impact of accurate keywords on lecture video navigation.

Acknowledgement

Research was conducted in collaboration with Videopoints LLC. with support, in part, from the National Science Foundation under award NSF-SBIR-1820045.

References

[1] S. D. Sorden and J. L. Ramírez-Romero, "Collaborative learning, social presence and student satisfaction in a blended learning environment," *IEEE 12th International Conference on Advanced Learning Technologies*, pp. 129–133, 2012.

[2] L. Barker, C. L. Hovey, J. Subhlok, and T. Tuna, "Student perceptions of indexed, searchable videos of faculty lectures," in *Proceedings of the 44th Annual Frontiers in Education Conference(FIE)*, Madrid, Spain, Oct 2014.

[3] T. Tuna, M. Joshi, V. Varghese, R. Deshpande, J. Subhlok, and R. Verma, "Topic based segmentation of classroom videos," in *Proceedings of the 45th Annual Frontiers in Education Conference(FIE)*, El Paso, Texas, Oct 2015, pp. 1–9.

[4] T. Tuna, J. Subhlok, L. Barker, S. Shah, O. Johnson, and C. Hovey, "Indexed captioned searchable videos: A learning companion for STEM coursework," *Journal of Science Education and Technology*, vol. 26, no. 1, pp. 82–99, 2017.

[5] M. R. Rahman, S. Shah, and J. Subhlok, "Visual summarization of lecture video segments for enhanced navigation," in *22nd IEEE International Symposium on Multimedia*. Virtual: IEEE, Dec 2020.

[6] A. S. Imran, L. Rahadianti, F. A. Cheikh, and S. Y. Yayilgan, "Semantic tags for lecture videos," in *IEEE Sixth International Conference on Semantic Computing*, 2012, pp. 117–120.

[7] A. Balagopalan, L. L. Balasubramanian, V. Balasubramanian, N. Chandrasekharan, and A. Damodar, "Automatic keyphrase extraction and segmentation of video lectures," in *IEEE International conference on technology enhanced education (ICTEE)*. IEEE, 2012, pp. 1–10.

[8] A. Albahr, D. Che, and M. Albahar, "Semkeyphrase: An unsupervised approach to keyphrase extraction from mooc video lectures," in *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2019, pp. 303–307.

[9] H. Shukla and M. Kakkar, "Keyword extraction from educational video transcripts using NLP techniques," in *6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 2016, pp. 105–108.

[10] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1105–1115.

[11] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge." in *Proceedings of the 23rd National Conference on Artificial Intelligence*, vol. 2, 2008, pp. 855–860.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[13] V. K. Rangarajan Sridhar, "Unsupervised topic modeling for short texts using distributed representations of words," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado, June 2015, pp. 192–200.

[14] R. S. Koka, "Automatic keyword detection for text summarization," Master's thesis, University of Houston, Houston, TX, May 2019.

[15] X. Tong and D. A. Evans, "A statistical approach to automatic OCR error correction in context," in *Fourth Workshop on Very Large Corpora*, 1996.

[16] S. PowerSuite, "https://www.link-assistant.com/seo-stop-words.html," 2018.

[17] M. F. Porter, "Snowball: A language for stemming algorithms," 2001.