2020 IEEE International Symposium on Multimedia (ISM) | 978-1-7281-8697-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISM.2020.00033

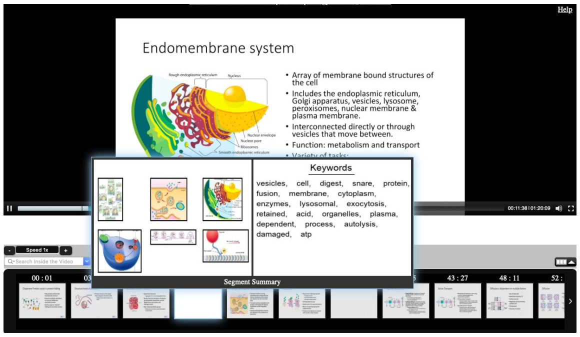
Visual Summarization of Lecture Video Segments for Enhanced Navigation

Mohammad Rajiur Rahman   Shishir Shah   Jaspal Subhlok
Department of Computer Science, University of Houston, Houston, TX 77204
Email: mrahman13@uh.edu, sshah5@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 visual summarization of lecture video segments to improve navigation. A lecture video is divided into segments based on the frame-to-frame similarity of content. The user navigates a lecture video assisted by single frame visual and textual summaries of segments. The paper presents a novel methodology to generate the visual summary of a lecture video segment by estimating the importance of each image in the segment, computing similarities between the images, and employing a graph-based algorithm to identify the most representative images. The summarization framework developed is integrated into a real-world lecture video management portal called Videopoints. An evaluation with ground truth from human experts established that the algorithms presented are significantly superior to random selection as well as clustering based selection, and only modestly inferior to human selection. Over 65% of automatically generated summaries were rated at *Good* or better by the users. Overall, the methodology introduced in this paper was shown to produce good quality visual summaries that are practically useful for lecture video navigation.**

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. Students value lecture videos for allowing them to review at their own pace and typically report a positive impact on grades and overall course satisfaction [1], [2]. This research is conducted in the context of the Videopoints project (www.videopoints.org) at the University of Houston whose central goal is to ease navigation of lecture videos. Conventional video format inherently lacks non-linear navigation support like indexing and content search. Videopoints overcomes these limitations by developing a lecture video framework with innovations in indexing, search, and captioning [1], [3], [4]. Figure 1 illustrates topic based indexing, a unique feature of Videopoints, as well as summarization of lecture video segments. An index panel is situated on the bottom of the player; each index frame represents a segment of the video containing a new subtopic. Users can navigate different topical segments of the video by clicking on these index frames. When a user hovers over an index frame, a summary frame appears as illustrated in the figure. The figure consists of a visual summary, that is the subject of this paper, and a text summary discussed in another contribution [5]. The main goal of the research presented in this paper is to build a visual summary to provide a natural way to connect to a lecture video segment. These visual summaries are employed to index lecture video segments to improve navigation.



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

In general, lecture video summarization techniques focus on finding unique transition frames or minimizing the number of frames of videos across different types of presentations like powerpoint lectures and blackboard handwriting [6], [7]. The system developed in this work goes well beyond identification of transition frames. A closely related work [8] extracts and classifies visual content of a lecture video and presents direct links in the player timeline to help non-linear navigation. Another related project [9] generates an interactive visual summary based on identifying and analyzing text close to the images in a lecture video. The research presented in this paper is unique in developing visual summaries based on an analysis of the similarity and the importance of visual objects. This work leverages existing methods for detecting and matching interest points to establish a measure of visual similarity between images. Several methods for detecting interest points in images such as SIFT, SURF, GLOH, and their variants have been proposed [10]. These have been used in a range of image analysis applications including image matching problems.

II. Visual Content Summarization

The main technical objective of this paper is to identify a subset of the images on the frames of a lecture video segment that best represents the content of the segment. Following steps are taken to reach this objective:

1) Extract all images from the frames in a video segment.
2) Compute the "distance matrix" between all pairs of images based on (dis)similarity between the images.

978-1-7281-8697-9/20/$31.00 ©2020 IEEE
DOI 10.1109/ISM.2020.00033

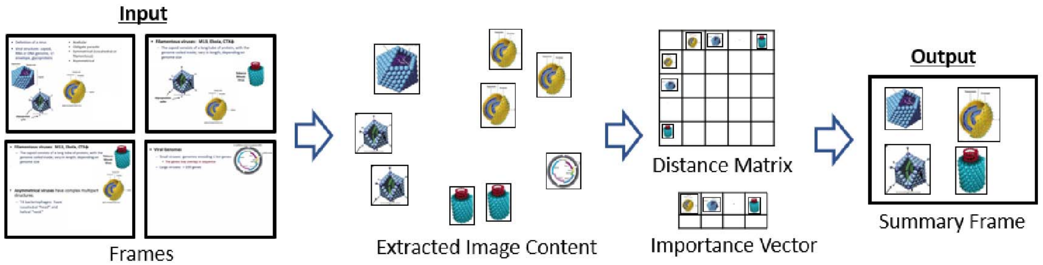


Fig. 2. Steps in generating a visual summary frame

3) Compute the "importance" of individual image for inclusion in the visual summary.
4) Select a subset of representative images based on similarity and importance.

These steps are illustrated in Figure 2 and detailed in this section. The final selected images are placed in a single frame on uniformly sized cells. A more visually appealing arrangement beyond the scope of this work.

A. Extracting Images from a Lecture Video Segment

A lecture video or screencast typically consists of a small number of unique video frames, each displayed from a few seconds to several minutes. These unique *transition frames* are identified by tracking where the scene in the video changes significantly [3], [11]. Next, the text regions in *transition frames* are identified and removed from consideration for image analysis. Then images are identified by scanning the video frame with a sliding window protocol for regions where the pixels change contiguously; images are regions surrounded by a border with no visual content.

B. Image Distance Matrix

An important consideration in deciding whether an image should be included in a visual summary is to quantify how similar (or different) it is to other images. In this work, we calculate the similarity between each pair of extracted image objects and create a distance matrix, where *Distance = 1 - Similarity*.

There are many metrics and algorithms to measure image similarity. Global measures include holistic image properties such as color and texture computed from the image, often represented as histograms. In contrast, local measures rely on identifying parts or points within an image that are unique to an image. Finding similarity of visual objects in a lecture video segment is a unique problem as the images are typically synthetically created and may not represent real-world objects. Images often contain illustrations like diagrams, charts, and graphs. An image may have a specific meaning in a particular domain only. Often one image is a rotated, scaled, or cropped version of another. Based on these considerations and our practical experience, we chose to use SIFT [12] to extract local interest points (keypoints) and the corresponding feature descriptors from an image. To compute a measure of similarity between two images, we measure:

- *KeypointsScore:* The fraction of unique keypoints that match between the pair of images; and
- *TransformScore:* The degree to which one image is a geometric transformation of the other.

The percentage of keypoints matched provides an indication of local similarities between two images. When a large fraction of keypoints match, a finer analysis is conducted based on an affine transformation of matched keypoints. The second image is transformed and aligned with the first image using the computed transformation and a pixel-wise normalized difference is measured to provide an indication of the global similarity between two images. The results are reported as the transformation score. Based on our experience, if at least half the keypoints match, then the transformation score is relevant. The final *Similarity* score is estimated to be simply the *KeypointsScore*, if that is less than 0.5, and the average of the *KeypointsScore* and *TransformScore*, otherwise.

C. Image Importance

Desirability of an image to be included in a visual summary is an independent consideration from similarity to other images. Factors that potentially contribute to this *importance* are the size of the image, information density in the image, and the duration for which the image is visible in the lecture video. The information density is captured by the number of keypoints per unit area. We estimate the importance as follows, where all factors are normalized to 0-1 range:

$$Importance = Size * InfoDensity * Duration \quad (1)$$

D. Selection of Representative Images

Suppose a lecture video segment has n images, $V_1, V_2, V_3, ..., V_n$. The goal is to construct a visual summary consisting of m representative images, $R_1, R_2, R_3, ...R_m$. An nxn $Distance$ matrix is available where $Distance_{ij}$ captures the visual difference between images V_i and V_j. A vector $Importance$ of size n is provided where $Importance_i$ captures the importance of the corresponding image V_i.

For identifying representative images for the summary, we apply two considerations: i) minimize the distance between each image not in the summary to the closest representative image in the summary, and ii) prioritize images that have more importance. Quantitatively our optimality criterion is to identify a set of representative images for which the maximum of $Distance_{ir} * Importance_i$ over all images is minimized, where $Distance_{ir}$ is the distance between image V_i and the image V_r in the summary that is closest to V_i.

An exact solution to this problem has been shown to be NP-hard. We employed a heuristic algorithm outlined as follows. Initially, the visual summary consists of all images in the segment. In the following step, the cost $cost_k$ of removing each image V_k from the summary is computed as follows:

$$cost_k = I_k * D_{k,p} \tag{2}$$

where V_p is the image in the summary that has the least distance (or is most similar to) V_k, I_k is the importance of image V_k, and $D_{k,p}$ is the distance between V_k and V_p. The image with the lowest cost is removed. This step is repeated until the desired number of images are left in the summary. The algorithm is efficient in practice and almost always produces the optimal result for practical cases.

III. Evaluation and Results

Visual summarization framework was implemented in the context of Videopoints, a real world lecture video portal, and evaluated with ground truth provided by the users. Results are presented and compared against K-Medoids clustering algorithm, random selection of summaries, and human expert selection of summaries.

A. Ground truth collection

A web-based survey tool was developed and employed to collect the ground truth.

Dataset: 40 segments from lecture videos in Biology, Geoscience, Computer Science, and Chemistry were selected. The segments were approximately 15 minutes long on average, contained approximately 12 images on average, with a minimum of 5 images.

Survey: Participants were asked to select up to 4 images from all distinct images extracted from a video segment. They also provided reasons for not selecting each of the remaining images; whether it was similar to a selected image or it was not important. Finally, the participants judged the quality of the algorithm generated summary on a 4 points scale with 1 as "Very Good" and 4 as "Poor".

Participants: A total of 30 students and instructors participated in the survey. The participants self-reported subjects that they were familiar with, and were assigned segments accordingly. Each segment was surveyed by 5.75 participants on average, with 6 being the maximum and 3 the minimum.

B. Evaluation Methodology

Evaluation of a visual summarization algorithm is challenging for a number of reasons: i) Multiple images can express the same concept and hence a user may consider two or more images equally representative of a concept, ii) users often differ significantly on their assessment of the best set of images that summarize a video segment, and iii) different sets of images may represent the content equally well. The evaluation methodology is designed to address these factors.

During ground truth collection, different survey participants often pick different sets of images as the visual summary of a segment. Evaluation was performed with the following:

- *Top-K Selected.* Images selected most often by participants constitute the ground truth. In the experiments presented in this paper, the set of 4 most selected images was used. The maximum size of the visual summary generated by our algorithms was also set to 4 images.
- *All Selected.* An image selected by any participant becomes a part of the ground truth.

In the survey, users indicated if an image was not selected because a similar one was already selected for the summary. We use this information to *group similar images*. Results are presented that take this similarity into account. That is, if ground truth contains image X, and the user indicates that they did not select image Y because it is similar to X, the algorithm is scored identically if it selects X or Y as part of the visual summary.

C. Results

We present results for an automatically generated 4 image summary against following formulations of the ground truth:

- Top-4 selected images.
- All selected images.
- Top-4 selected groups of images.
- All selected groups of images.

Figure 3 presents accuracy, precision, recall, and F1-measures for these scenarios. We make a few observations: i) The scores are significantly higher with grouping indicating the role played by user selected similar images, ii) precision is significantly higher when all participant choices are added to the ground truth, but F1-measure is virtually unchanged and iii) precision reaches a high of 0.94 for "All Selected Groups" implying that most algorithm choices found some agreement with at least one survey participant.

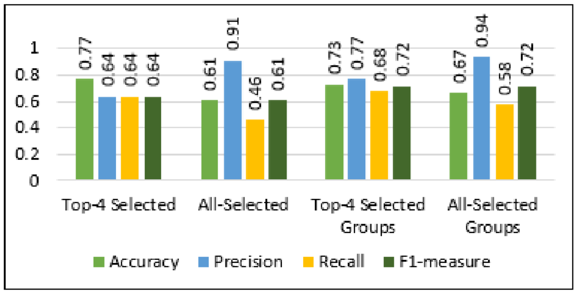


Fig. 3. Performance of graph based visual summarization algorithm

The results from the graph based algorithm were also compared against the following:

- *K-medoid clustering:* K-medoid [13] groups similar images and selects a representative for the summary. This is a natural clustering algorithm for this scenario as it has been used in the context of image retrieval and face recognition [14].

- *Random:* The summary images were selected at random.
- *Human:* Selections from a random survey participant for each segment in the ground truth collection process was the selection for this "Human" algorithm.

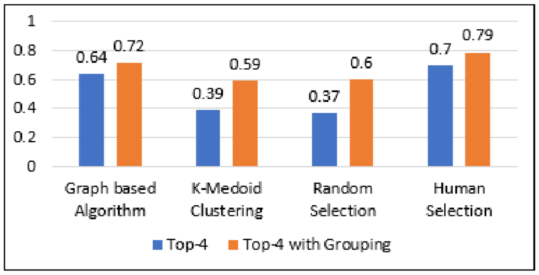


Fig. 4. F-1 measures compared to K-medoid clustering, Random selection and Human selection for the Top-4 selected with grouping ground truth

The results are presented as F1-measures in Figure 4. The graph based algorithm performs significantly better than random selection and K-medoid clustering. The random selection can be considered to be the lower bound for an effective algorithm. The relatively weak performance of the K-medoid algorithm can be attributed to the fact that clustering selection is strictly based on similarity and not on importance of images. Results of human selection are modestly better than the graph based algorithm. The human selection can be considered the upper bound for any algorithm, since an algorithm cannot be expected to do better than a human expert in this scenario. This is also an indirect measure of (dis) agreement between human users; if survey participants were always in agreement, the human results will be perfect.

Summary results from user rating of algorithm generated summaries are plotted in Figure 5. Around 65% of user ratings were "Good" or better, while 85% of the summaries were rated as "Very Good" by at least one user.

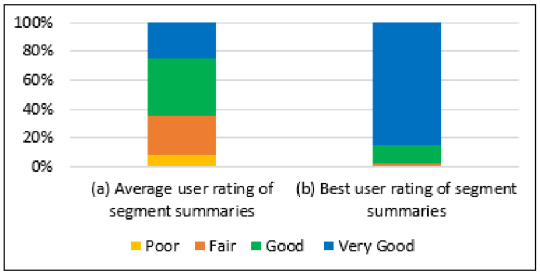


Fig. 5. User perception of quality of algorithm generated summary

IV. Conclusions and Future Work

This paper presents a novel approach to use low-level image features to create a summary of visual content extracted from a lecture video segment. The algorithms developed are implemented in a real-world lecture video management system. The results are encouraging based on quantitative metrics as well as the user perception of the quality of visual summaries.

Ongoing work is identifying and classifying the underlying causes of errors in visual summaries. Future work will focus on improving the quality and relevance of extracted summaries. Research directions under consideration include i) alternate image similarity measures, ii) analysis of high level semantic features, iii) integrated text and image analysis, and iv) enhanced understanding of image importance. We also plan to substantially expand the ground truth with additional surveys.

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] 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.

[2] 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.

[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] R. S. Koka, F. N. Chowdhury, M. R. Rahman, T. Solorio, and J. Subhlok, "Automatic identification of keywords in lecture video segments," in *22nd IEEE International Symposium on Multimedia*. Virtual: IEEE, Dec 2020.

[6] C. Choudary and T. Liu, "Summarization of visual content in instructional videos," *IEEE Transactions on Multimedia*, vol. 9, pp. 1443 – 1455, 12 2007.

[7] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, "Summarization of videotaped presentations: automatic analysis of motion and gesture," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 686–696, Sep 1998.

[8] K. Yadav, A. Gandhi, A. Biswas, K. Shrivastava, S. Srivastava, and O. Deshmukh, "ViZig: Anchor points based non-linear navigation and summarization in educational videos," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, 2016, pp. 407–418.

[9] Y. Wang, Y. Kawai, and K. Sumiya, "iPoster: Interactive poster generation based on topic structure and slide presentation," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 30, no. 1, pp. 112–123, 2015.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[11] T. Tuna, J. Subhlok, and S. Shah, "Indexing and keyword search to ease navigation in lecture videos," in *Applied Imagery Pattern Recognition(AIPR)*, 2011, pp. 1–8.

[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 11 2004.

[13] T. S. Madhulatha, "Comparison between k-means and k-medoids clustering algorithms," in *Advances in Computing and Information Technology*, D. C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, and D. Nagamalai, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 472–481.

[14] S. Parui and A. Mittal, "Similarity-invariant sketch-based image retrieval in large databases," in *European Conference on Computer Vision (ECCV)*. Zurich, Switzerland: Springer International Publishing, 2014, pp. 398–414.