2017 IEEE 30th International Symposium on Computer-Based Medical Systems

Novel Method for Storyboarding Biomedical Videos for Medical Informatics

Sema Candemir, Sameer Antani, Zhiyun Xue, George Thoma
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, NIH, Bethesda, MD, USA
(sema.candemir, sameer.antani, xuez, george.thoma)@nih.gov

***Abstract*—We propose a novel method for developing static storyboard for video clips included with biomedical research literature. The technique uses both visual and audio content in the video to select candidate key frames for the storyboard. From the visual channel, the Intra-frames are extracted using FFmpeg tool. IBM Watson speech-to-text service is used to extract words from the audio channel, from which clinically significant concepts (key concepts) are identified using the U.S. National Library of Medicine's Repository for Informed Decision Making (RIDeM) service. These concepts are synchronized with the key frames, from which our algorithm selects relevant frames to highlight in the storyboard. In order to test the system, we first created a reference set through a semi-automatic approach, and measure the system performance with informativeness and fidelity metrics. Results from pilot testing, both subjective visual and quantitative metrics, are promising. It is our goal to conduct a formal user evaluation in the future.**

***Keywords*-Open-i®, biomedical research articles, key frame extraction, key clinical concept, video summarization**

I. Introduction

While there are several techniques proposed for biomedical image retrieval, such as the U.S. National Library of Medicine's (NLM) multimodal (image + text) biomedical literature search engine called Open-i®, there is a paucity of techniques for retrieving biomedical videos with similar ease. Open-i uses text processing, image analysis, and machine learning techniques to retrieve relevant articles from the indexed Open-Access biomedical literature [1], [2], [3]. It indexes 1.2 million biomedical articles and 3.7 million figures which include a wide range of clinical imaging modalities, in addition to graphs, charts, photographs and other illustrations. In addition to the images, there are also more than 1100 video clips on biomedical topics. Some of the video clips are not associated with a research article, therefore, lack an abstract. This makes it difficult for the user to search for the videos. Further, many videos range from 15 to 25 minutes in duration, making it cumbersome to select one that is relevant to the query. In such cases, a short version of the video (video abstract) that can be quickly browsed will reduce the burden on the user.

There are two main approaches to obtain a video abstract [4]: static and dynamic storyboards. In a static video abstraction the video content is represented by a grid of extracted key frames. In a dynamic video abstraction the important content in the video is condensed into a shorter video clip. The effectiveness of both these approaches, however, depends on smart selection of key frames. The goal of this study is to develop a novel technique for developing a static storyboard for the video clips included with biomedical articles.

The proposed approach uses both visual and audio content of the video by extracting selected internally coded frames from the visual content, and key clinical concepts from the automatically transcribed audio content. The algorithm selects relevant video frames from the extracted set that are synchronous with concepts from speech in the audio channel. These frames are developed as a storyboard and present a meaningful summary of the clips.

II. Related Work

Several techniques have been proposed in the literature to address the video abstraction [4], [5], [6]. One of the general approaches is finding the significant visual discontinuities in the video, defining consecutive frames between the breaks as shot, and selecting key frames for each shot [7], [8], [9]. Shot boundaries are detected by measuring the variation of image features (e.g. color histograms, motion vector) of consecutive frames. Depending on the sampling approach, the key frame could be selected as the first or center frame of shots [10], [11], or more sophisticated approach could be used to define the key frame such as finding the most representative component of feature space of shots [8]. Another approach is clustering the similar frames, and then selecting frames closer to cluster centroids as key frames [12], [13], [14].

Major contributions of our study are incorporating the semantic primitives of the video through audio content, extracting key clinical concepts, and using the concepts to find the prior key frames.

III. Method

The proposed approach consist of 3 main stages: (i) extracting Intra-frames, (ii) extracting Concept-frames, (iii) refining Intra- and Concept-frames.

A. Intra-Frame Extraction

Typically biomedical video clips included with the research articles including those found in Open-i are

2372-9198/17 $31.00 © 2017 IEEE
DOI 10.1109/CBMS.2017.57

MPEG [15] encoded. MPEG videos are encoded as a set of frames called *Group Of Pictures (GOP)*. Each GOP starts with an Intra-frame (I-frame) which is a frame that is coded independently of other frames. Therefore, it contains content important to the GOP. Other frames within the GOP reference this I-frame to provide data compression advantages.

For our method, instead of applying a video shot detection algorithm, we consider each GOP as a shot, and the I-frame as its representative frame. Arguably, one could reconstruct another frame from within the GOP and use that as a representative frame also. We used the FFmpeg tool [16] which contains libraries and programs for processing multimedia data and extract the I-frames from the video (Figure 1).

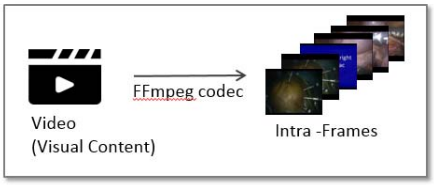

Figure 1. Illustration of Intra-Frame Extraction

B. Concept-Frame Extraction

In addition to the visual content, the audio channel is also an important information carrier in videos, especially in biomedical videos. The spoken word can help categorize the video according to biomedical concepts, provide an improved search capability, and also assist in identifying the key frames in the video. For our technique, we used audio content to find the significant frames in the video. We first use the IBM Watson Speech-to-Text service [17] to transcribe the speech in the video using web-based queries. For this, we resampled the audio from the video file to 16kHz, and submit it to the speech-to-text service. The resulting transcribed text was used to generate the video speech transcript. In addition to providing the transcribed text, the IBM Watson Speech-to-Text service also provides time stamps of each word which are useful for synchronizing the text with the visual key frames. As a part of our research, we also tested Microsoft Bing Speech-to-Text service [18], but found it cumbersome to use due to various limitations imposed on the user i.e., lack of time stamps and the requirement to submit data in small chunks.

After obtaining the transcripts, the next stage is selecting the words that are considered clinically significant. For this, we use NLM's clinical decision support system service - Repository for Informed Decision Making (RIDeM) [19]. This service extracts the key clinical concepts from clinical text resources using NLM's MetaMap [20] and Unified Medical Language System [21], respectively.

In our technique, we submit the speech transcripts to RIDeM service to extract key clinical concepts in the transcript. Using the time stamps of key concepts, we synchronize and tag frames that are associated with the utterance of the concept. We call these tagged frames as *concept frames (C-Frames)* in the video sequence (Figure 2).

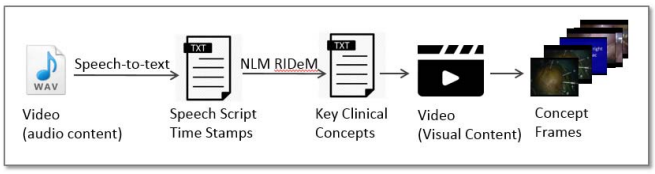

Figure 2. Illustration of Concept-Frame Extraction

C. Elimination of redundant frames

After obtaining I-frames and C-frames, we combine all frames into a set. The set, however, also contains several redundant frames such as complete black/white frames, repetitive frames, fade and other scene transition frames. Some examples of such redundant frames are shown in Figure 3. The goal of this stage is to eliminate these frames using low level visual features.

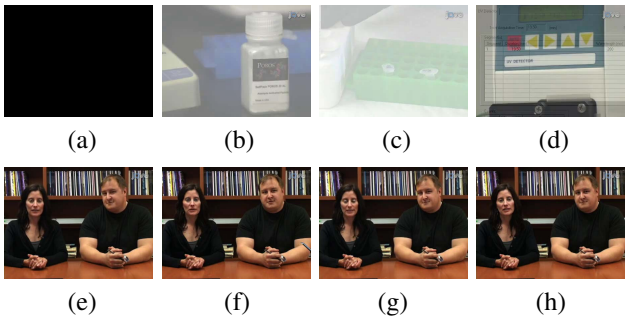

Figure 3. Some example redundant frames: a) complete black frame, b,c) faded frames, d) transitive frame, e,f,g,h) repetitive frames.

In order to eliminate the repetitive frames, we develop a technique inspired by the method presented in [22][23]. We divide the frame into 9 sub-regions, and compute intensity and gradient magnitude histograms of each region. Then we generate a feature vector by concatenating the histograms. Next, each frame histogram vector is compared with other frames in the I- and C-frame set. If histogram difference of frames (e.g. Euclidean distance) is smaller than an empirically determined threshold, then one of the frames is eliminated.

To address the problem of non-informative frames (e.g. pure or mostly white/black frames), we compute the vertical profile of frames, and take the derivatives of the profile. If the frame is non-informative, the derivatives of the profile will tend to be nearly zero. Once all the key frames (I-

and C-Frames) are selected, they are presented sequentially (time-order) to generate the storyboard .

IV. Analysis and Discussion

As a pilot evaluation, we tested our technique on ten randomly selected videos indexed in the Open-i search engine. The selected videos are 25-30 minutes long in duration, and encoded at approximately 30 frames per second (fps); and have approximately 30,000 frames (Table I). Our algorithm extracts the I- and C-frames, and refines the selected frames using intensity and gradient magnitude histograms as described in Section III. The refined set of key frames are used to construct the storyboard.

Evaluating the performance of a video abstraction algorithms is challenging due to lack of a reference key frame set and absence of standard evaluation metrics. Some questions that can be asked of the technique are: (i) How many frames were extracted as key frames? (ii) Do the key frames adequately summarize the video content? (iii) Are there any erroneous/unnecessary frames selected as key frames? (iv) Are any significant frames missing? (v) Are there misalignment problems - i.e., should a different frame from within the GOP be considered the key frame? (vi) Are all relevant topics identified? (vii) Are all identified topics relevant?

In the literature, we find subjective evaluation mechanisms in which a group of evaluators manually select (or detect the position of) the important frames in a candidate video sequence creating a reference set. Then, automatically extracted key frames are compared with the reference set. This simple technique can address the accuracy and subjective aspects without teasing out specific problems in the method, if any. The challenge is, however, to find a sufficiently large set of evaluators, and then normalizing their selected frames to develop the reference standard. Another evaluation metric is to let evaluators view the video and then provide qualitative scores (e.g., good, acceptable, bad) frame by frame based on their observation of how representative the extracted key frame [24] are. In this case also, normalizing the evaluations across a large population of evaluators is challenging, and often not practical. In addition to subjective evaluation, automated evaluation approaches are also proposed. One of them is by measuring *fidelity* which is the similarity comparison of key frame and the shot frame sets [25], [26].

Given the size of our data and the early stage of our research, a manual selection of reference key frames is infeasible, and is deferred as future work. As an alternative and a pilot evaluation strategy, we follow a semi-automatic mechanism to build the reference key frame set. We used the k-means clustering algorithm, and cluster all frames into 250 clusters using RGB color intensity histograms of frames. Then, we check each cluster and manually select the representative frames. The aim of this effort is to create a reference set for each video which contains all non-similar and informative frames. For example, the reference frames that we semi-automatically select for the test video numbered '3169267_jove-50-2096' and titled 'A method for murine islet isolation and subcapsular kidney transplantation' are shown in Figure 4. The number of selected reference frames for test videos are listed in Table II. Note that the reference set is not the optimal key frame set which needs careful selection based on the video content. Our reference set contains only "visually distinct" (informative) and non-repetitive frames.

The technique extracts 10-35 key frames (depending on the length of the video and it's content) for each video. Figure 5 shows example result on test video numbered '3169267_jove-50-2096' and titled 'A method for murine islet isolation and subcapsular kidney transplantation'. (The reference set for this video is shown in Figure 4). We measure the *informativeness* of the extracted key frame set with the following equation:

$$\text{informativeness} = \frac{\text{\# of informative frames}}{\text{\# of extracted key frames}} \tag{1}$$

If the frame is repetitive, transitional (two different scenes are overlapped), and/or does not contain any information, we consider it uninformative. We check the extracted key frames visually and determine if each frame is informative. The informativeness ratio for each video is listed in Table II. The average informativeness of the whole test set is 68.16%. Note that our informativeness ratio does not consider the missing key frames that are not among the extracted key frames, but in the reference set. We will consider a better performance measure for informativeness that identifies the missing key frames in the extracted key frame set.

We use the fidelity measure to compute the similarity between the extracted key frame set and the reference set, which is formulated as below:

$$\text{fidelity} = 1 - \frac{d(F^R, F^K)}{\max_i(\max_j(d_{ij}))} \tag{2}$$

where $d(F^R, F^K)$ is the dissimilarity between reference frame set F^R and key frame set F^K; d_{ij} is the dissimilarity matrix of the reference set. The dissimilarity between the reference set and key frame set is the maximum of the minimal distance between key frames and reference frames, and measured as below:

$$d(F^R, F^K) = \max_i(\min_j(d(F^R, F^K))) \tag{3}$$

where j is the index of the reference frame set; i is the index of key frame set. Higher fidelity is the indication of similarity of the key frame set to the reference set due to smaller $d(F^R, F^K)$. However this measurement is easily influenced by outlier frames. For example, assume all extracted frames are very similar to reference frames,

Video No	Research Title	V. Length	Resolution	F. Rate	#of F.
2762330_jove-32-1398	A lectin HPLC method to enrich selectively glycosylated peptides from complex biological samples	20m 23s	448×336	29.92fps	36623
3169267_jove-50-2096	A method for murine islet isolation and subcapsular kidney transplantation	17m 42s	448×336	29.97fps	31814
3182659_jove-47-2383	Pseudomonas aeruginosa and Saccharomyces cerevisiae biofilm in flow cells	17m 30s	448×336	25.00fps	26256
3197026_jove-52-2068	Fixed volume or fixed pressure: a murine model of hemorrhagic shock	16m 31s	448×336	29.97fps	29725
3197307_jove-49-2538	High-efficiency transduction of liver cancer cells by recombinant adeno-associated virus serotype 3 vectors	16m 31s	448×336	29.97fps	34248
3217647_jove-54-3324	Derivation of enriched oligodendrocyte cultures and oligodendrocyte/neuron myelinating co-cultures from post-natal murine tissues	18m 17s	448×336	29.97fps	32883
3227187_jove-56-3159	Isolation & characterization of Hoechst(low) CD45(negative) mouse lung mesenchymal stem cells	16m 55s	448×336	29.97fps	30446
3399499_jove-60-3774	Preterm EEG: a multimodal neurophysiological protocol	19m 34s	448×252	25.00fps	29360
3490322_jove-68-4093	Expansion of embryonic and adult neural stem cells by in utero electroporation or viral stereotaxic injection	19m 45s	512×384	24.00fps	28456
3577868_jove-70-50124	A research method for detecting transient myocardial ischemia in patients with suspected acute coronary syndrome using continuous ST-segment analysis	18m 11s	512×384	29.97fps	32709

Table I
TEST VIDEOS. V: VIDEO; F: FRAME. THE TEST VIDEOS ARE AVAILABLE THROUGH OPEN-I AND CAN BE REACHED BY SEARCHING FOR THE RESEARCH TITLE.

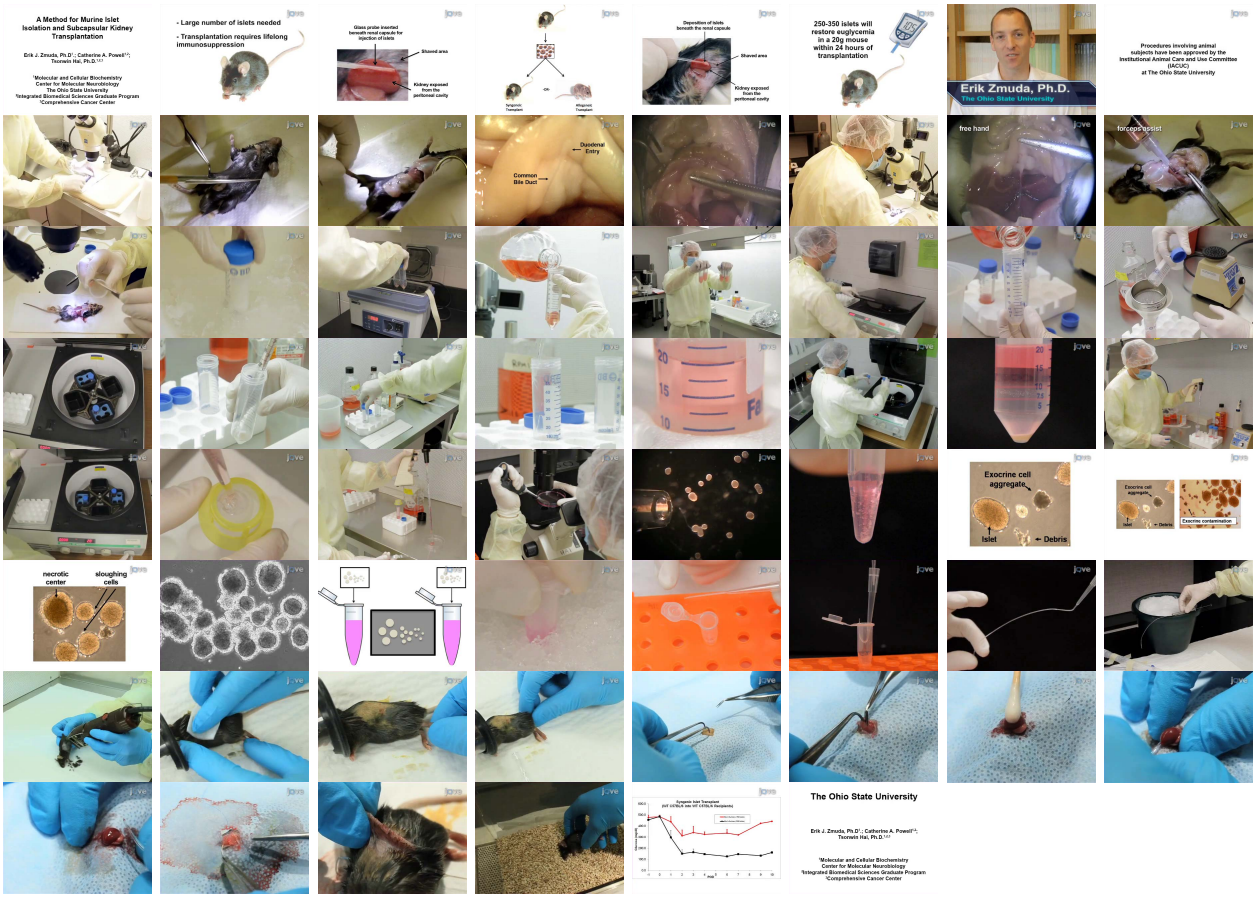

Figure 4. Reference frames for a test video titled 'A method for murine islet isolation and subcapsular kidney transplantation'. The frames are in time sequence. The video is available through Open-i [1] by searching for the research title.

except for one redundant frame. This redundant frame does not show any similarity to any of the frames in the reference set, so will have the maximum $d(F^R, F^K)$. Note also that as the number of key frames increase, the fidelity value also increases, since the probability of finding more similar frames for each frame in the reference set increases.

Video No	#of Ref. F.	Informativeness	Fidelity
2762330_jove-32-1398	109	58.82	0.1925
3169267_jove-50-2096	62	81.25	0.6451
3182659_jove-47-2383	43	66.66	0.7876
3197026_jove-52-2068	67	68.75	0.6828
3197307_jove-49-2538	105	75.00	0.6039
3217647_jove-54-3324	71	65.00	0.2424
3227187_jove-56-3159	77	58.62	0.5236
3399499_jove-60-3774	97	65.38	0.4248
3490322_jove-68-4093	105	66.66	0.3322
3577868_jove-70-50124	78	69.23	0.5172

Table II
NUMBER OF REFERENCE FRAMES, INFORMATIVENESS, AND FIDELITY VALUES OF TEST VIDEOS.

Figure 5. Detected key frames for test video titled 'A method for murine islet isolation and subcapsular kidney transplantation'. The frames are in time sequence.

V. CONCLUSIONS AND FUTURE WORK

We have presented our novel technique for developing a static storyboard for biomedical video clips included with the biomedical research literature. The proposed approach uses both visual and audio content of video to select the key frames. From the visual channel, the Intra-frames are used as candidate key frames. From the audio channel, clinically significant concepts (key concepts) are extracted and used to identify and extract Concept-frames. The major contribution of our technique is the combination of visual primitives in video and semantic concepts from speech in the audio channel to select the key frames and develop a video summary.

The technique is tested on example videos downloaded from the Open-i. Judging by the obtained key frames, the results are promising, but there is room for improvement at all stages of the method. One improvement would be to add another constraint such as video segment duration, or better methods to minimize fade/transition frames to improve the key frame selection.

We have evaluated our results with informativeness and fidelity measures. However, these metrics are not adequate to evaluate the video summarization. Encouraged by the positive results from our pilot evaluation, as a next step, we aim to conduct a thorough evaluation which considers missing frames, transcript errors introduced by the speech-recognition system, and concept errors introduced due to lack of a decision module that discards unrelated concepts from RIDeM output.

ACKNOWLEDGMENT

This research is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications. In addition, we thank Suchet Chachra, Dina Demner-Fushman and Leif Neve for their assistance with RIDeM, MetaMap, and UMLS services.

REFERENCES

[1] National Library of Medicine, Open-i Project, https://openi.nlm.nih.gov, [Accessed: 13-January-2017].

[2] D. Demner-Fushman, S. Antani, S. Chachra, M. Kushnir, S. Gayen, Open-i imaging informatics, natural language processing and multi-modal information retrieval - research and development, in: Technical Report to the LHNCBC Board of Scientific Counselors, 2016.

[3] D. Demner-Fushman, S. Antani, M. Simpson, G. R. Thoma, Design and development of a multimodal biomedical information retrieval system, Journal of Computing Science and Engineering 6 (2) (2012) 168–177.

[4] Y. Li, T. Zhang, D. Tretter, An overview of video abstraction techniques, 2001.

[5] B. T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM Transactions on Multimedia Computing, Communications, and Applications 3 (1) (2007) 1–37.

[6] A. G. Money, H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation 19 (2) (2008) 121–143.

[7] S. S. H. J. Zhang, C. Y. Low, Video parsing and browsing using compressed data, Multimedia Tools Application 1 (1995) 91–113.

[8] H. Bredin, D. Byrne, H. Lee, N. E. O'Connor, G. J. Jones, Dublin city university at the trecvid 2008 bbc rushes summarisation task, in: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, 2008, pp. 45–49.

[9] W. Abd-Almageed, Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing, in: 15th International Conference on Image Processing, IEEE, 2008, pp. 3200–3203.

[10] Y. Rui, T. S. Huang, S. Mehrotra, Exploring video structure beyond the shots, in: International Conference on Multimedia Computing and Systems, IEEE, 1988, pp. 237–240.

[11] M. J. Pickering, D. Heesch, R. O'Callaghan, S. Rüger, D. Bull, Video retrieval using global features in keyframes, in: Proceedings of the 11th Text Retrieval Conference, 2002.

[12] Y. Zhuang, Y. Rui, T. S. Huang, S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, in: International Conference on Image Processing, Vol. 1, IEEE, 1998, pp. 866–870.

[13] V. Chasanis, A. Likas, N. Galatsanos, Video rushes summarization using spectral clustering and sequence alignment, in: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, 2008, pp. 75–79.

[14] E. Asadi, N. M. Charkari, Video summarization using fuzzy c-means clustering, in: 20th Iranian Conference on Electrical Engineering, IEEE, 2012, pp. 690–694.

[15] The Moving Picture Experts Group, http://mpeg.chiariglione.org/, [Accessed: 2-February-2017].

[16] FFmpeg, https://ffmpeg.org/, [Accessed: 13-January-2017].

[17] IBM Watson Speech-to-Text, http://www.ibm.com/watson/developercloud/speech-to-text.html, [Accessed: 13-January-2017].

[18] Bing Voice Recognition, https://www.microsoft.com/cognitive-services/en-us/speech-api, [Accessed: 2-February-2017].

[19] Repository for Informed Decision Making, https://ceb.nlm.nih.gov/ridem/, [Accessed: 13-January-2017].

[20] Semantic Knowledge Representation, https://skr.nlm.nih.gov/, [Accessed: 13-January-2017].

[21] Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/, [Accessed: 13-January-2017].

[22] N. Ejaz, T. B. Tariq, S. W. Baik, Adaptive key frame extraction for video summarization using an aggregation mechanism, Journal of Visual Communication and Image Representation 23 (7) (2012) 1031–1040. doi:10.1016/j.jvcir.2012.06.013.

[23] J. Almeida, N. J. Leite, R. d. S. Torres, Online video summarization on compressed domain, Journal of Visual Communication and Image Representation 24 (6) (2013) 729–738.

[24] T. Liu, H.-J. Zhang, F. Qi, A novel video key-frame-extraction algorithm based on perceived motion energy model, IEEE Transactions on Circuits and Systems for Video Technology 13 (10) (2003) 1006–1013.

[25] H. S. Chang, S. Sull, S. U. Lee, Efficient video indexing scheme for content-based retrieval, IEEE Transactions on Circuits and Systems for Video Technology 9 (8) (1999) 1269–1279.

[26] C. Gianluigi, S. Raimondo, An innovative algorithm for key frame extraction in video summarization, Journal of Real-Time Image Processing 1 (1) (2006) 69–88.