Video Summarization with Sparse Representation

Yuan Song(yson3253@uni.sydney.edu.au)

Supervisors: Dr. Zhiyong Wang and Prof. David Feng

School of Information Technologies

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. Introduction

Motivation

Video is a effective visual medium to deliver information and it has been extensively applied through internet. However, due to its nature, it is time-consuming to obtain information from videos. For example, it takes 30 minutes to watch a 30-minutes news program. Video summarization is to produce a condensed version of a video. It has great potential in may application domains:

- Video retrieval: quickly spot whether the returned videos are of interest.
- Video streaming: only delivery key information to clients such as smartphones.
- In video surveillance domain, video summarization can be used to detect specific event, human face or object appearance. Therefore, it will be much more efficient and easier than manually monitor the video.

Aim

While most of the existing methods heavily rely on heuristic rules such as motion change, we propose to rank each frame in terms of local features which compose video content.

In particular, we utilize sparse coding to derive the sparse representation of local features.

2. Related work

Video summarization can be also considered as key frame extraction which needs to be semantically selected from frame pool. There are many cues can be based on to extract key frames.

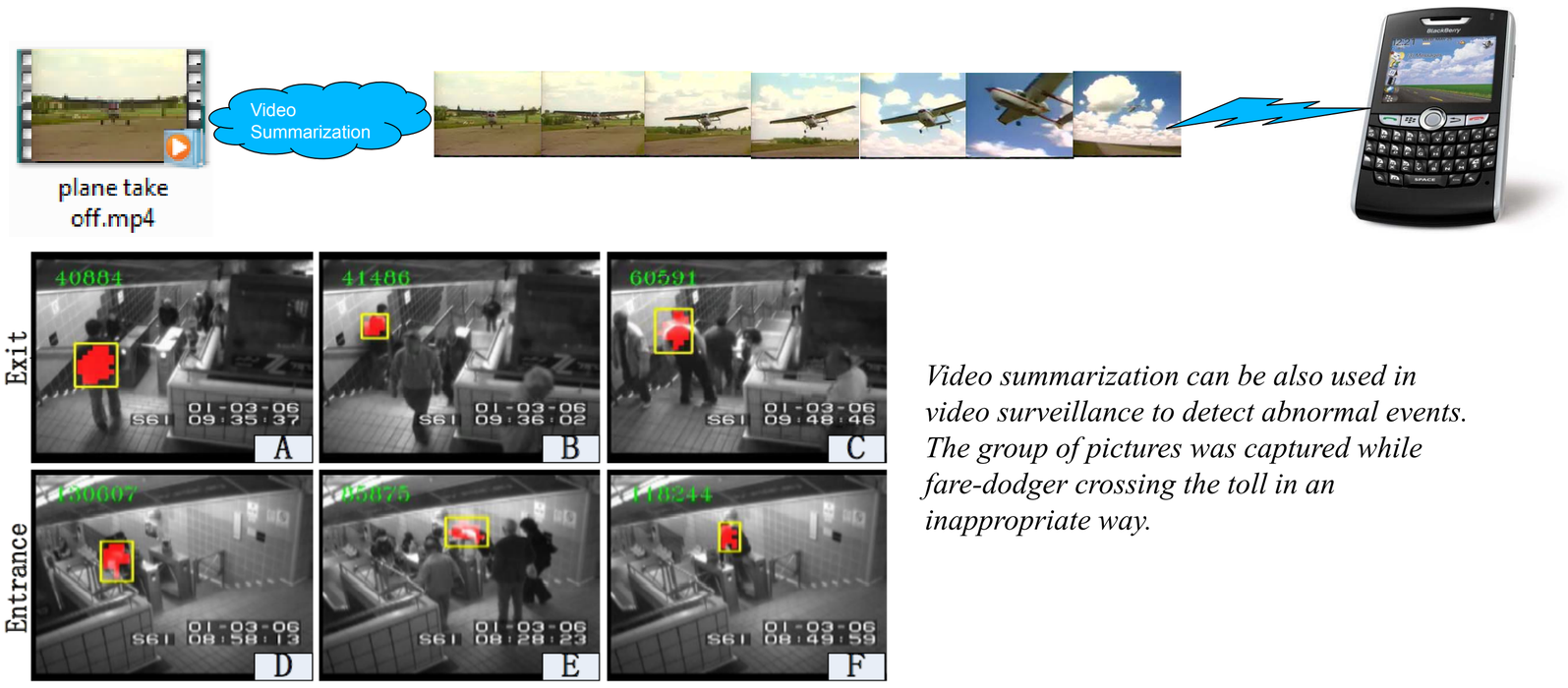
- Textual information which is embedded as subtitles in video or been tagged short titles. It gives immediate explanation on video's content such as game score, participator and so on in sport reporting[1].

- Audio is one other cue to summarize video. It provides extra information about the video that visual information may not capable of. For example, videos which recorded in a speech or sport with audience participated can be summarized by detecting [2] audience's clapping sound even the camera does not showing audiences' happy faces or clapping hands.

- Image feature such as color, salient object contained, motion changes between still images, texture and shape can be also analyzed to obtain key frames. For example [3], [4] proposed extracting key frames by analyzing color histograms and motion changes.

3.Video Summarization Framework

Feature descriptor



Video summarization can be also used in video surveillance to detect abnormal events. The group of pictures was captured while fare-dodger crossing the toll in an inappropriate way.

In order to capture more semantic information contained in each frame, we adopted local feature descriptor and is relatively better than global feature descriptor which roughly describes original video frames in structure.

Sparse coding

Sparse coding was initially used in neurobiology to interpret the phenomenon which only a few neurons will be activated while human eyes receiving visual information.
The process to project original feature space into a new feature space in order to keep most values in new feature space to be zero and remaining non-zeros to be "activated" called sparse coding.

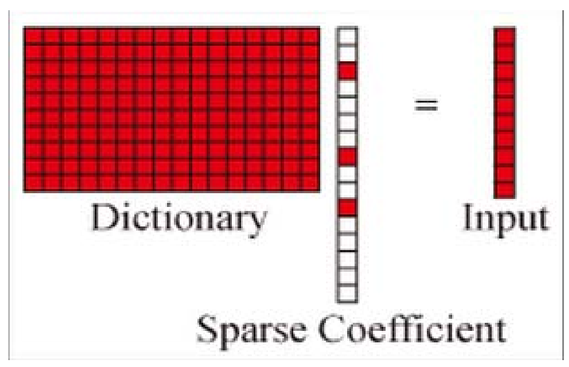
Following this idea, video summarization can be achieved through sparse coding algorithm since they share the same property which need to keep the representation as sparse as possible.

Through group Lasso[5] which is a mathematical model, the dictionary can be iteratively trained and sparse vector for each frame will also be calculated.

Key frame selection

Since each frame has been represented by a sparse vector, the length of i_{th} vector can be calculated by l_2 norm: $||\alpha_i||$. Then a curve of video frame weight can be generated and key frame can be extracted by detecting peaks within the weighted curve.

SIFT: a local feature descriptor

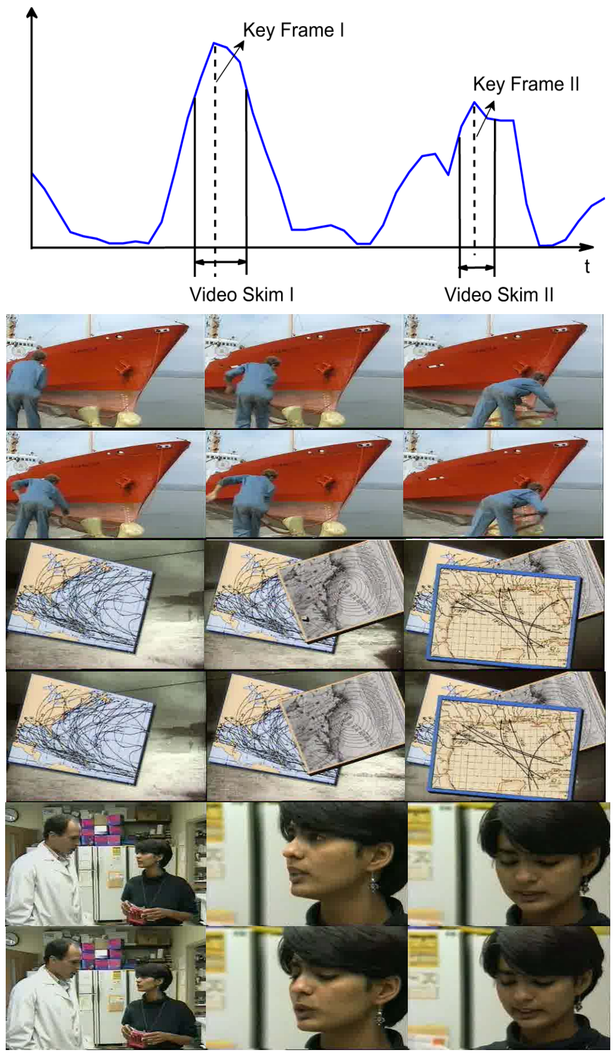


$$\alpha = \arg\min_{\alpha} \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1$$

X is input, D denotes dictionary and α is sparse representation vector. λ is a tuning parameter.

5. Result and conclusion

Result is generated by given number of frames need to be extracted. Hence, we assign the same number as human labeled ground truth for better evaluation. We tested proposed method in 10 data sets and compared with baseline algorithm[7]. Our proposed algorithm obtained 63% overall accuracy and 70% for baseline algorithm. However, it can be improved and optimized from many aspects. For example, we can add color information since SIFT[6] descriptor does not include it, extraction from weighted curve can be more adaptive by segmenting curve into parts.



First row of frames are ground truth;
Second row of frames are extracted key frames by proposed algorithm

References

[1] D. Zhang, S.-F. Chang, Event detection in baseball video using superimposed caption recognition, in: Proceedings of the 10th ACM International Conference on Multimedia, Juan-les-Pins, France, 2–8 November, 2002, pp. 315–318.
[2] M. Xu, C. Maddage, C. Xu, M. Kankanhalli, Q. Tian, Creating audio keywords for event detection in soccer video, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03), vol. 2, Baltimore, USA, 6–9 July, 2003, pp. 281–284.
[3] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.
[4] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: From humans to computers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 289–301, Feb. 2009.
[5] M.Yuan andY. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc.: Series B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.
[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[7] Y. Cong, J. Yuan, J. Luo "Towards scalable summarization of consumer videos via sparse dictionary selection" IEEE Transactions on Multimedia, 99 (2011)