Video Summarization using Deep Semantic Features

M. Otani[1], Y. Nakashima[1], E. Rahtu[2], J. Heikkilä[2], N. Yokoya[1]

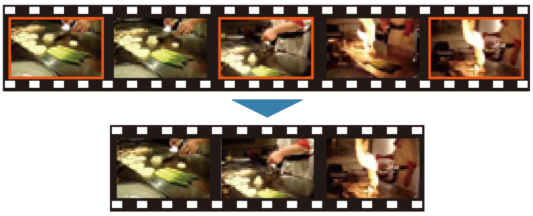
[1] {otani.mayu.ob9, n-yuta, yokoya}@is.naist.jp Graduate School of Information Science, Nara Institute of Science and Technology

[2] {erahtu, jth}@ee.oulu.fi Center for Machine Vision and Signal Analysis, University of Oulu

Introduction

Video Summarization

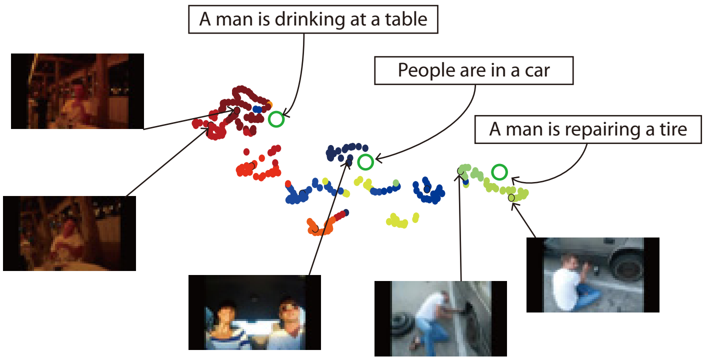
Enable quich review of long videos by automatically extracting short video segments



Motivation

Video summary:
Consists of **semantically representative and diverse** video segments

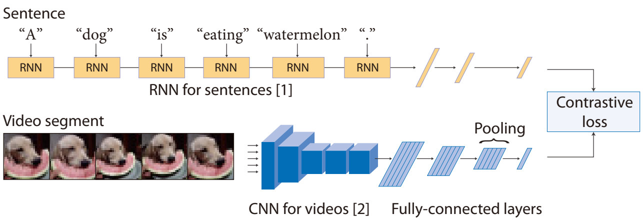
- Map video segments to a sentence-level semantic space
- Sample cluster centers of video segments in a semantic space



Approach

Learning Deep Features

Learn deep features of videos from pairs of sentences and videos

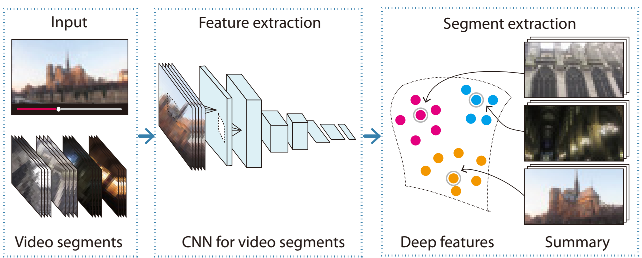


Contrastive Loss

$$loss(X_n, Y_n) = t_n d(X_n, Y_n) + (1 - t_n)\max(\alpha - d(x_n, Y_n), 0)$$

- $d(X_n, Y_n)$: Euclidean distance between video and sentence embeddings
- t_n : Label. $t_n = 1$ if the video and the sentence is relevant,otherwise $t_n = 0$

Generating Video Summaries



Segment selection as k-medoids problem:
Evaluate the representativeness of sampled segments

$$F(\mathcal{S}) = \sum_{X \in \mathcal{X}} \min_{S \in \mathcal{S}} \|X - S\|_2^2$$

Experiment

Create video summaries of videos and compare to manually created summaries

Dataset

SumMe [3] (25 videos)

- Unedited or slightly edited videos
- Provide 15 manually created video summaries for each video

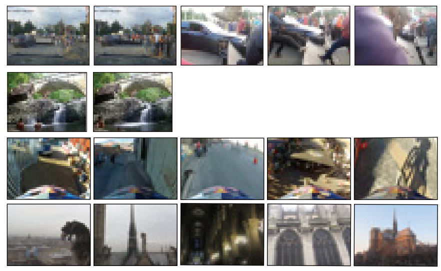
Evaluation Metric

Average of F_1 scores of a summary to each manually created summary

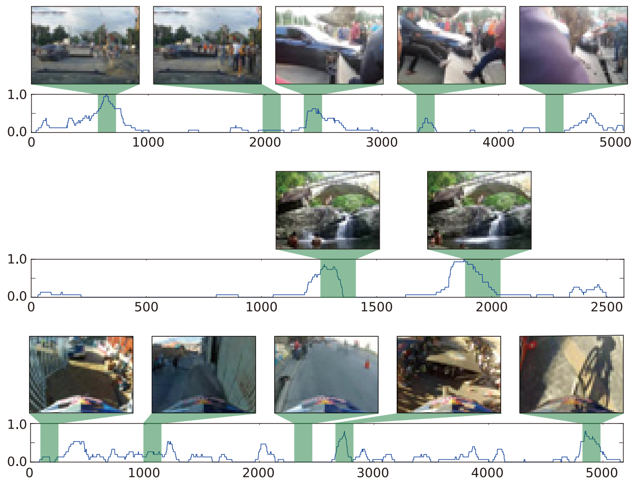
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Generated Summaries

Key frames of video summaries



Sampled Video Segments and Ground Truth Scores



Method	F1 Score	Relative to Human Avg.	Relative to Human Max.
Uniform	0.124	0.398	0.303
VGG	0.127	0.408	0.310
Attention-based [4]	0.167	0.537	0.408
Ours	0.182	0.588	0.447
Supervised [3]	0.234	0.752	0.571
Human	0.311	1.000	0.760

Conclusion

- Our deep features trained to capture sentence-level semantics benefits an unsupervised video summarization technique

Future work:

- Incorporating a video segmentation method
- Expanding the objective function with other criteria such as interestingness

[1] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-Thought Vectors," NIPS, pp. 3276–3284, 2015.
[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recoginition," ICLR, 2015, pp. 1–14.
[3] M. Gygli, H. Grabner, H. Riemenschneider, and L. van Gool, "Creating summaries from user videos," ECCV, pp. 505–520, 2014.
[4] N. Ejaz, I. Mehmood, and S. Wook Baik, "Efficient visual attention based framework for extracting key frames from videos," Signal Process. Image Commun., vol. 28, no. 1, pp. 34–44, 2013.