Literature Notes

Adam Lehavi
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089
alehavi@usc.edu

***Abstract*—This outlines Adam Lehavi's notes related to his research with Yoonsoo Nam.**

I. Working Model Idea

Take a video, use [1] to turn the video into key frames, use image summarization as surveyed in [2] to generate image summaries, and use an NLP model to turn a series of image summaries to a description.

II. Core Definitions and Acronyms

-

III. Literature Review

A. Survey of Techniques for Labeling Video, Audio, and Text Data

Past approaches to video data:

1) Automatically annotate news videos through unsupervised learning using mining of similar videos. For a given video with a speech-recognized transcript, it first searches and ranks most similar videos and then mines those.
2) Automatically annotate of people passing surveillance cameras, annotating clothing color, height, and focus of attention.
3) Use TREC video retrieval evaluation.
4) Address the problem of automatic temporal annotation of realistic human actions in video.
5) Use an optimal graph (OGL) from multicues (partial tags and multiple features) to get results superior to other state-of-the-art methods.
6) Propose Interactive Self-Annotation framework, based on recurrent self-supervised learning.
7) State a recursive and semi-automatic annotation approach which proposes initial annotations for all frames in a video based on segmenting only a few manual objects.
8) Use Multiple Annotation Maturation (MAM)
9) Propose a novel Correlative Multi-Labe (CML) framework which simultaneously classifies concepts and models correlations in a single step on the TRECVID dataset

[3]

Note a lot of the papers are from the ACM international conference on Multimedia.

B. Video Summarization Based on ML

Process to summarize a video in terms of some smaller amount of frames/images [1]

C. Survey for Automatic Description Generation from Images

Datasets

- Pascal1K
- Flickr8K
- Flickr30K
- MS COCO

Recommendation is MS COCO, having 164K images each with 5 texts, collected judgements, and partial objects.

To read:

1) Karpathy and Fei-Fei (2015) MultRetrieval on Flickr8K/30K, COCO measured w/ BLEU, Meteor, CIDEr, mRank, R@k
2) Jia et al. (2015) Generation on Flickr8K/30K, COCO measured w/ BLEU, Meteor
3) Yagcioglu et al. (2015) VisRetrieval on Flikr8K/30K, COCO measured w/ Human, BLEU, Meteor, CIDEr

[2]

References

[1] W. Ren and Y. Zhu, "A video summarization approach based on machine learning," in *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2008, pp. 450–453.

[2] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[3] S. Zhang, O. Jafari, and P. Nagarkar, "A survey on machine learning techniques for auto labeling of video, audio, and text data," *CoRR*, vol. abs/2109.03784, 2021. [Online]. Available: https://arxiv.org/abs/2109.03784