**LITRATURE REVIEW**

1. [Video summarization using textual descriptions for authoring video blogs](https://www.semanticscholar.org/paper/Video-summarization-using-textual-descriptions-for-Otani-Nakashima/fd961668b885fb6e421af897ed72a7f99408bf5a#citing-papers) by *Mayu Otani, Yuta Nakashima, Tomokazu Sato, N. Yokoya*

Authoring video blogs requires a video editing process, which is cumbersome for ordinary users. Video summarization can automate this process by extracting important segments from original videos. Because bloggers typically have certain stories for their blog posts, video summaries of a blog post should take the author’s intentions into account. However, most prior works address video summarization by mining patterns from the original videos without considering the blog author’s intentions. To generate a video summary that reflects the blog author’s intention, we focus on supporting texts in video blog posts and present a text-based method, in which the supporting text serves as a prior to the video summary. Given video and text that describe scenes of interest, our method segments videos and assigns to each video segment its priority in the summary based on its relevance to the input text. Our method then selects a subset of segments with content that is similar to the input text. Accordingly, our method produces different video summaries from the same set of videos, depending on the input text. We evaluated summaries generated from both blog viewers’ and authors’ perspectives in a user study. Experimental results demonstrate the advantages to the proposed text-based method for video blog authoring.

1. [Video Summarization via Semantic Attended Networks](https://cdn.aaai.org/ojs/11297/11297-13-14825-1-2-20201228.pdf) by *Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, Chen Yao*

The goal of video summarization is to distill a raw video into a more compact form without losing much semantic information. However, previous methods mainly consider the diversity and representation interestingness of the obtained summary, and they seldom pay sufficient attention to semantic information of resulting frame set, especially the long temporal range semantics. To explicitly address this issue, we propose a novel technique which is able to extract the most semantically relevant video segments (i.e., valid for a long term temporal duration) and assemble them into an informative summary. To this end, we develop a semantic attended video summarization network (SASUM) which consists of a frame selector and video descriptor to select an appropriate number of video shots by minimizing the distance between the generated description sentence of the summarized video and the human annotated text of the original video. Extensive experiments show that our method achieves a superior performance gain over previous methods on two benchmark datasets.

1. [Key Frame Extraction in the Summary Space](https://ieeexplore.ieee.org/document/7968322) by *Xuelong Li, Bin Zhao, Xiaoqiang Lu*

Key frame extraction is an efficient way to create the video summary which helps users obtain a quick comprehension of the video content. Generally, the key frames should be representative of the video content, meanwhile, diverse to reduce the redundancy. Based on the assumption that the video data are near a subspace of a high-dimensional space, a new approach, named as key frame extraction in the summary space, is proposed for key frame extraction in this paper. The proposed approach aims to find the representative frames of the video and filter out similar frames from the representative frame set. First of all, the video data are mapped to a high-dimensional space, named as summary space. Then, a new representation is learned for each frame by analyzing the intrinsic structure of the summary space. Specifically, the learned representation can reflect the representativeness of the frame, and is utilized to select representative frames. Next, the perceptual hash algorithm is employed to measure the similarity of representative frames. As a result, the key frame set is obtained after filtering out similar frames from the representative frame set. Finally, the video summary is constructed by assigning the key frames in temporal order. Additionally, the ground truth, created by filtering out similar frames from human-created summaries, is utilized to evaluate the quality of the video summary. Compared with several traditional approaches, the experimental results on 80 videos from two datasets indicate the superior performance of our approach.

1. [Cross-language Search: The Case of Google Language Tools](https://digital.library.unt.edu/ark:/67531/metadc96824/m2/1/high_res_d/Chen-first_monday.pdf) by Jiangping Chen, Yu Bao

This paper presents a case study of Google Language Tools, especially its cross-language search service. Cross-language search integrates machine translation (MT) and cross-language information retrieval (CLIR) technologies and allows Web users to search and read pages written in languages different from their search terms. In addition to cross-language search, Google Language Tools provides various language support services to multilingual information access. Our study examines the functions of Google Language Tools and the performance of its cross-language search. The results and analysis show that Google Language Tools are useful for Web users. Its cross-language search service provides quality query translation while the automatic translation of result pages needs further improvement. The paper suggests that cross-language search could be used by different types of Web users. The authors also discuss the strategies and important issues with regard to implementing multilingual information access services for information systems.

1. [ASoVS: Abstractive Summarization of Video Sequences](https://ieeexplore.ieee.org/ielx7/6287639/8600701/08664480.pdf) by *Aniqa Dilawari, Muhammad Usman Ghani Khan*

In today’s age, a massive amount of videos are produced every day, which contains audio, visual, and textual data. This constant increase is due to the ease of recording service in portable devices, such as mobile phones, tablets, or cameras. The major challenge is to understand the visual semantics and convert it into a condensed format, such as caption or summary to save storage space, enables users to index, navigate, and help gain information in less time. We propose an innovative joint end-to-end solution, Abstractive Summarization of Video Sequences, which uses the deep neural network to generate the natural language description and abstractive text summarization of an input video. This provides a text-based video description and abstractive summary, enabling users to discriminate between relevant and irrelevant information according to their needs. Furthermore, our experiments show that the joint model can attain better results than the baseline methods in separate tasks with informative, concise, and readable multi-line video description and summary in a human evaluation.

1. [Video to Text Summary: Joint Video Summarization and Captioning with Recurrent Neural Networks](https://www.semanticscholar.org/paper/Video-to-Text-Summary%3A-Joint-Video-Summarization-Chen-Chen/fb3bce3a6221eb65451584efa898ecbe211bdab6) by *Bor-Chun Chen, Yan-Ying Chen, Francine Chen*

Video summarization and video captioning are considered two separate tasks in existing studies. For longer videos, automatically identifying the important parts of video content and annotating them with captions will enable a richer and more concise condensation of the video. We propose a general neural network configuration that jointly considers two supervisory signals (i.e., an image-based video summary and text-based video captions) in the training phase and generates both a video summary and corresponding captions for a given video in the test phase. Our main idea is that the summary signals can help a video captioning model learn to focus on important frames. On the other hand, caption signals can help a video summarization model to learn better semantic representations. Jointly modeling both the video summarization and the video captioning tasks offers a novel end-to-end solution that generates a captioned video summary enabling users to index and navigate through the highlights in a video. Moreover, our experiments show the joint model can achieve better performance than state-of-the-art approaches in both individual tasks.