Summarisation of casually captured videos

Dissertation Proposal

Submitted in partial fulfilment of the requirements for the award of the degree

Master of Sciences in Mathematics

*(With a specialization in **Computer Science**)*

By

Varshaneya V

Registration No: 15013

Supervised by:

Dr. S Balasubramanian

Department of Mathematics and Computer Science

Sri Sathya Sai Institute of Higher Learning,

Prasanthi Nilayam 515134

Aim:

To study the various techniques of video summarisation of casually captured videos and to develop method(s) to improve the existing technique(s). The emphasis is on automatic video summarisation without any parameter tuning or user intervention.

Objectives:

1. Implementing key-frame based summarisation using k-means and Delaunay clustering.
2. Implementing skim based summarisation using optimization of sub-modular functions.
3. To study scenarios where the aforesaid algorithms work and where they fail.
4. To develop techniques for improvising these algorithms on their area(s) of failure.

Introduction:

Video summarisation is the technique of providing condensed and succinct representations of the content of a video stream through a combination of still images, video segments, graphical representations and textual descriptors. An ideal video summary should capture the remarkable events that constitute the video. Video summarisation techniques produce summaries by analysing the underlying content of a source video stream, condensing this content into abbreviated descriptive forms that represent essence of the original content embedded in the video.

The summarisation techniques can be broadly categorised into three, namely shot based, key-frame based [1], [2] and skims based [3]. In the shot based summarisation individual shots are identified and initial and the final frames from each shot is shown as the summary. The key-frame based technique samples the video and creates clusters using either parameterised or non-

parameterised clustering algorithms. The median of each of the clusters, called the key-frame of that cluster, is extracted and displayed as the summary of the video. The last technique is supervised and the summary is generated using maximisation of sub-modular mixture of objectives. The objectives are interestingness, uniformity and representativeness.

Motivation for study:

Due to the widespread of economic tools for video capturing there has been a growth in the amount of video data that is generated. Also many popular sites like YouTube, Facebook, Twitter, Metacafe, Daily Motion etc. have risen up and are encouraging users to upload and share their videos with others. In this context the amount of video data in the internet is increasing exponentially. In order to enable user to access relevant videos easily without spending a lot of time and bandwidth, video summarisation provide a condensed version of the video capturing the important and the pertinent aspects in the video.

Video summarisation can also be integrated with other applications like interactive browsing and searching systems. Video summarisation also form an integral part of video cataloguing where a huge repository can be categorised and classified based on the summaries generated.

Video summarisation is very useful to those users on low bandwidth network connections. The multimedia data communicated must be information-oriented rather than the actual content itself. This will give the users an opportunity to select the video of their choice for complete viewing later.

A Very Brief Look at Literature:

One popular way to summarise videos is to summarise based on key-frames. The general procedure is to extract frames from the video, cluster them using a suitable clustering algorithm

to get well defined clusters and finally the key-frames are the medians of each cluster which form the summary of the video. The choice of clustering algorithm plays a vital role in generation of summaries.

A generic conceptual framework for video summarisation is presented by the survey paper [4]. This framework distinguishes between video summaries (outputs of video summarisation techniques) and video summarisation techniques (the methods used to process content from a source video stream to achieve a summarisation of that stream). The video summarisation techniques are considered to be within three categories: internal that analyse information sourced directly from the video stream, external that analyse information not sourced directly from the video stream and hybrid that analyse a combination of internal and external information. Video summarisation techniques produce summaries by analysing the underlying content of a source video stream, condensing this content into abbreviated descriptive forms that represent surrogates of the original content embedded within the video.

Video summaries are considered as a function of the type of content they are derived from object, event, perception or feature based) and the functionality offered to the user for their consumption (interactive or static, personalised or generic). There are several audio-visual clues incorporated by video summaries in presenting the user with condensed and succinct representation of the content of a video stream.

The paper on “Video summarisation using clustering” [2] is about summarising videos by using K-means clustering algorithm. K-means clustering is of logarithmic order in time. The input video is sampled and the histograms are generated. Principal component analysis is done to the histograms to reduce the dimensionality. The frames are then clustered using K-means clustering algorithm with the user specifying the number of clusters to be formed. Median of each cluster is the representation of that cluster and is a key-frame. The main drawback with this is that an input parameter, which is the number of clusters required, has to be specified. Finding the right

parameter for every video is a cumbersome task. Hence parameter tuning is very difficult in this case.

As an improvement over this, the paper titled “Key frame based video summarisation using Delaunay clustering” [1] makes use of Delaunay triangulation to create Delaunay diagram. The initial procedure is to sample frames from the video, find histogram and reduce the dimensions using principal component analysis. The Delaunay clusters are created on the reduced feature vector. From this diagram, clusters are formed by identifying and removing the separating edges. The separating edges are those that are longer than the mean of edges incident at that vertex plus the global standard deviation of all the edge lengths. The major advantage is that this method is a parameter-less method of clustering and is suitable for batch processing. This algorithm runs in. This paper also prescribes metrics for evaluation based on the representativeness of the summary generated.

Automatically creating skims is challenging, as even a strongly shortened version should still convey the story of the initial video. Hence an innovative method is introduced in the paper by Michael Gygli et.al [3] which is to jointly optimise multiple objectives like interestingness, representativeness and uniformity globally and to use a supervised approach rather than the previous techniques that relied on unsupervised learning like clustering. This makes this method more applicable to causally captured videos. The dataset used here is called SumMe, which consists of short user videos each annotated with more than 15 user summaries. The method consists of two stages. First is that of a supervised learning stage (training) and inference (testing). Given pairs of videos and their user created summaries as training examples, a combined objective is learnt. The next stage is when a new video is input, the method creates summaries that are interesting, representative and uniform. The summaries are evaluated automatically using a software package called ROUGE (Recall-Oriented Understudy of Gisting Evaluation). Sub-modular functions are useful in representing interestingness, representativeness and uniformity, as sub-modular functions have diminishing returns property.

Tools, methods and techniques used for study:

- **Programming environment:** MATLAB
- **Video datasets:**
    1. Open video project. https://open-video.org/
    2. SumMe dataset. https://data.vision.ee.ethz.ch/gyglim/SumMe/SumMe.zip
- **Version Control:** GitHub. https://github.com/varshaneya
- **Papers Referred:**
    3. Padmavathi Mundur, Yong Rao, Yelena Yesha. *Key-frame based Video Summarization using Delaunay Clustering.* International Journal on Digital Libraries April 2006, Volume 6, Issue 2. Pages 219 – 232.
    4. Tommy Chheng. *Video Summarization Using Clustering.* University of California, Irvine Winter project 2008.Pages 1 – 7.
    5. Michael Gygli, Helmut Grabner, Luc Van Gool. *Video Summarization by Learning Submodular Mixtures of Objectives.* Proceedings of CVPR 2015. Pages 3090, 3092 and 3094.
    6. Arthur G. Money, Harry Agius. *Video summarisation: A conceptual framework and survey of the state of the art.* Journal of visual communication and image representation 2008. Page 121.