**Literature Review: *Creating Summaries from User Videos***

**Original Paper authors:** Michael Gygli, Helmut Grabner, Hayko Riemenscneder, Luc Van Gool.

**Original Paper link:** <http://grabner.family/helmut/papers/Gygli2014CreatingSummariesfrom.pdf>

**Summary of Paper:**

This 2014 paper proposes an automated method for video summarization. Their attempt at video summarization begins by segmenting the video into portions they call “*superframes*”.

At first these *superframes* are equal length clips of the original video that are to be cut and scored individually. However their algorithms first determine where spikes in “*interestingness*” occur inside the clips. They determine i*nterestingness* levels from a number of factors including; Human attention (using temporal gradients), aesthetics/quality (colour, contrast, edges), presence of landmarks (image recognition), faces/persons (outline detection) and objects followed by the camera (motion pattern/segmentation). Each frames *interestingness* is then given a score according to the value and occurrence of each above event recognised.

Using this, they can give each frame of video a numerical value based on how interesting it is deemed to be. The start and end time code positions of each *superframe* (marks where the original video will be cut into *superframes*) are then adjusted to ensure that trends of interestingness (think spikes of the *interestingness* values where plotted against timecode in a graph) are kept within the same cut. Once the cut times are adjusted, and the video is segmented into *superframes*, only the *superframes* with the highest levels of average interestingness are kept while remaining cuts are discarded. The end result is a new cut of the original video, with only the segments of video that were determined to be interesting left in.

The recorded test results (using user videos – think personal holiday recordings, etc) show that they were able to reduce videos with average lengths of 2m40s down to an average of 4.9s. They make note of the accuracy of simple qualifiers for *interestingness* levels such as change in colour. They stated that their method was able to create automatic video summaries often reaching a performance level comparable to humans. Though they acknowledge that methods of video summarization is still in its beginnings.

**Relevance of Paper to our Project:**

The efficiency of our final analysing algorithm is of great importance to the future applications of our project. Even if runtime is not an overly important factor with regards to the scope of our capstone project, it is something we will no doubt have to be mindful of and consider when discussing the future use and implementation of our application. Thus creating effective video summaries is just one way we can help cut down on overall runtime waste for our application.

Using some of the methods in this report we could effectively create video summaries as a pre-screening technique. If we can analyse which portions of recorded video contain no general detected movements of interest. Then we can tell our final, less efficient, algorithm (that detects, analyses and categorizes specific arm/hand movements for each side of the body) to skip the selected timecodes that we aren’t interested in analysing further.

Something like this could significantly cut down on runtimes which when analysing thousands of frames worth of data is necessary if our project is to ever see real world application.

**Literature Review: *Textually Customized Video Summaries***

**Original Paper authors:** Jinsoo Choi, Tae-Hyun Oh, In So Kweon.

**Original Paper link:** <https://arxiv-org.ezp01.library.qut.edu.au/pdf/1702.01528.pdf>

**Summary of Paper:**

This 2017 paper generates video summaries that are customized by user text descriptions (think ‘tags’ effectively). They tackle their project in this way in order to deal with the overarching issue that different users want to extract different information from videos. What one user may classify as important data that they want in their summary, another user may want that particular type of data left out. So this tagging system essentially allows different users to benefit from this method of summarisation. Entering a text description (in natural language) of the actions they are looking for will generate a custom summary containing only the footage of those actions detected that are related semantically to the input text.

They create their method by training a deep architecture to learn semantic embeddings of video frames in a progressive and residual manner. This allowed them to map both video frames and sentences to a common embedding space. Once the relevant frames have been selected they are able to generate the custom summary by combining the clips together and temporally align them through use of the hidden Markov model.

Some of the experiments conducted for creating their summaries involved footage taken from a head mounted camera lasting 3-5 hours each and featured every-day scenarios such as eating, working, driving, shopping, cooking etc. Their method is able to approach human levels of performance, and compete with other current methods of video summarization even exceeding those in some cases.

**Relevance of Paper to our Project:**

As before, the efficiency of our final analysing algorithm is of great importance to the future applications of our project. Even if runtime is not an overly important factor with regards to the scope of our capstone project, it is something we will no doubt have to be mindful of and consider when discussing the future use and implementation of our application. Thus creating effective video summaries is just one way we can help cut down on overall runtime waste for our application.

The approach used in this paper seems quite similar on a surface level to what we are trying to do. It implements deep learning techniques to learn associations between input text and processed images. We are trying to do the same thing by categorizing specific arm and hand movements and then allowing the end user to filter through the video where each of the tagged actions occur. It could be possible to apply their approach to the way in which we wish to detect specific actions.

While it can be used to create video summaries to save time when pushed through our final analysis algorithm, it can also be used after the entire video has been analysed simply as a method of creating short, compilation style video summaries containing only the recorded footage of movements associated with the tags selected by the end user when filtering the data.

Due to the range of ways in which this can be applied to our project I believe it could be beneficial to look more in depth at the way they have managed this. It can be used to save on execution time or simply provide the end user with additional options to view the data they are after. However it is also worth noting that this approach could require more human input when designing.