# Video Summarization Using Artificial Intelligence

Amina Naidjate, YSP Student, *Al-Noor Academy*
Alejandro Gardinier, YSP Student, *Arlington High School*
Zijia Lu, Ph.D. Student, *Northeastern University*
Yuhan Shen, Ph.D. Student, *Northeastern University*
Dr. Ehsan Elhamifar, Electrical and Computer Engineering, *Northeastern University*

Northeastern University
Center for STEM Education

Northeastern
Khoury College of Computer and Information Sciences

Young Scholars Program

## Abstract

Humans have the remarkable ability to summarize large amounts of visual and textual data, like describing the content of a video or news article, with a few concise sentences. This skill saves both time and energy, so the question must be asked: can machines, which have an increasing presence in our lives, learn to do the same? The purpose of this project is to explore automatic methods based on artificial intelligence to summarize large collections of data in visual media. It involves experimenting with combinations of content detection and recognition codes through software such as SciKit and methods such as K-means clustering. The automatic algorithm produced can determine which frames in a video or movie are the most important and creates a trailer, giving the viewer a summary of the content similar to human summarization.

## Background

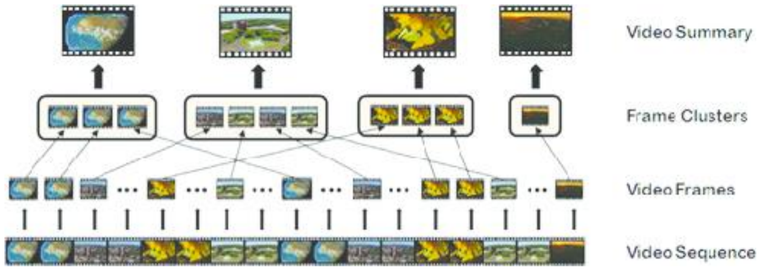


**Figure 1.** A visual outline of the video summarization process. The video is broken into frames which are grouped into scenes based on pixel similarity, and one frame per cluster is chosen to represent the scene.

The main question of this project is: how can computers be programmed to automatically select the frames most representative of a video through deep learning, and what combinations of features like PCA and K-means will provide the highest accuracy and efficiency? Machines cannot intuitively know what is important in a data set though; they need to be provided the logic and rules to determine what should be saved and what is negligible. There are several ways to go about providing this logic, but in this experiment, the K-means clustering algorithm was utilized. As described in Figure 1, the input into the code was a video, and in this case, a COVID-19 video update. The video was then broken down into 151 frames at the rate of 1 frame per second and the K-means method was then called. It automatically read each frame's pixels, including its red, green, and blue values, converted them into numerical values, and then grouped together images with similar numbers. These images were determined to be part of one scene, with the number of scenes chosen by the user. The image in each scene closest to the average of the all the scene's image's values was then chosen as the representative frame. Grouped together, the representative frames provided an insightful preview of the content of the video.

## References

Caetano, C. (2016, January). *Steps for video summarization* [Image]. Research Gate. https://www.researchgate.net/figure/Steps-for-video-summarization_fig1_282423076

Jordan, M. (n.d.). *Figure 1: K-means algorithm* [Image]. Stanford CS221. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html#

CNBC Television. (2020, June 29). *Coronavirus cases in the United States just surpassed 2.5 million infections* [Video]. YouTube. https://www.youtube.com/watch?v=HyttnavZBNU

## Experimental Methods

**K-Means:** K-means is a centroid based algorithm that can classify data points into clusters based on similarity. The user chooses how many clusters, or Ks, they want and unless the locations are specified, the computer will plot the points randomly on the number line, graph, or other data set. Next, the computer will determine which cluster each point is closest to and calculates the average distance of all the points. This average point will then become the cluster's new center, and the process is repeated until the variance between the points and the cluster center is minimal. Different numbers of clusters were experimented with to determine which amount produced the optimal balance of accuracy and efficiency.

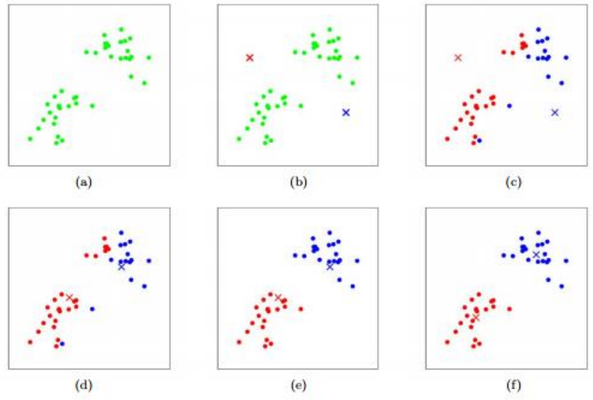


**Figure 2.** K-means algorithm. Data values are shown as dots, and cluster centroids are crosses. (c-f) shows the running of two iterations of the algorithm. In each iteration, we assign each data point to the closest centroid and then we move the centroids to the average of the points assigned to it.

**Principal Component Analysis (PCA)**: PCA is a linear dimensionality reduction technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional subspace. It is mostly used as a dimensional reduction tool to improve the efficiency of code and limit redundant dimensions.

**ResNet-50:** ResNet-50, short for Residual Network-50,, is a deep neural network in a subclass of convolutional neural networks used for image recognition. The network, which consists of 50 layers, solves complex operations that increases the recognition and classification accuracy of images.

**RGB Feature:** uses red, green, and blue to make every color and have integer values up to 255. Each pixel in each frame in the video was assigned a R, G, and B value that was then compared to other frames' RGB values.

## Results

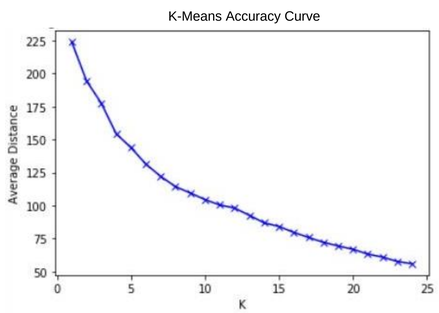


**Figure 3.** K-Means accuracy curve. While it intuitively seems like the highest cluster number, or K, will yield the best result, that is not the case. As K increases, the average distance decreases (variance), but once the graphs hits the "elbow point", or the point where the slope drastically flattens, the average distance lost for each cluster added minimizes. It becomes inefficient for the computer to compute new clusters for minimal gain, so the optimal K is the one found on the elbow point. In this case, it is 14, as there are 14 scenes in the video being summarized.

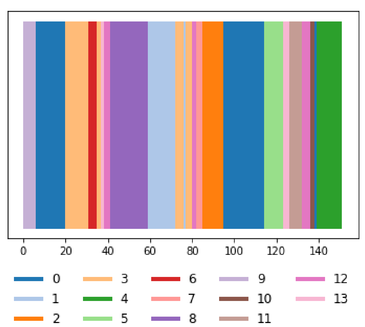


**Figure 4.** Timeline plot for cluster frequency. On the x-axis are the frame numbers (151 total) and each cluster (14 total) is represented by a bar of color (Cluster 0 is dark blue, Cluster 1 is light blue, etc.). Clusters are not always comprised of successive frames in the video; frames 5 to 20 and 95 to 115 are grouped into Cluster 0 though they are far apart. Many times, this is due to inaccuracy on the computer's part, but in this case, it is correct since US Secretary of Health and Human Services Alex Azar is seen talking in frames belonging to both regions.

## Results Continued

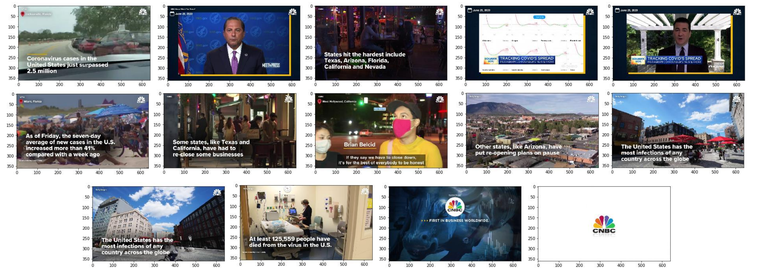


**Figure 5.** Representative Frames. Above are the 14 frames chosen by the computer as the most representative images of the video in a ResNet-50 test run. While the process was not perfect (images 10 and 11 are from the same cluster and there is a scene left unrepresented), it does give an accurate and helpful depiction of what is mentioned in the video.

| | ResNet-50 | PCA-32 | PCA-128 | RGB feature |
|---|---|---|---|---|
| Mean Purity | 0.8039 | 0.8088 | 0.8080 | 0.8094 |
| Variance | 0.00175 | 0.000168 | 0.000109 | 0.000135 |
| Computation Time | 1.17 sec | 16.3 sec | 31.9 sec | 84.8 sec |

**Table 1.** Accuracy and efficiency of different methods. The tested features were the ResNet-50 network, the PCA when limited to 32 dimensions, the PCA when limited to 128 dimensions, and the RGB feature. Though all four methods have such similar accuracies to the point where discrepancies are statistically insignificant, the RGB feature has the highest. However, that comes at the cost of being time consuming. As shown on the table, the RGB feature takes a significantly longer time than the other features to load, largely because it goes through all 691,200 of its dimensions compared to just the 32 of PCA-32, 128 of PCA-128, or 2048 of ResNet-50. On the other hand, the ResNet-50 feature, with a very similar accuracy, was the fastest to determine the representative frames, and nearly 84 times quicker than the RGB feature.

## Conclusions

- Computers can be taught to summarize large amounts of data, specifically videos, through machine learning algorithms like K-means
- The optimal number of clusters, or Ks, for summarization is the value found on the "elbow point" of the accuracy curve. In this case, it is 14.
- Though the RGB feature produced the most accurate results, its accuracy was so similar that of ResNet-50, PCA-32, and PCA-128 that the differences were statistically insignificant.
- The higher the number of dimensions considered is, the longer it takes to compute the representative frames. The feature with the highest performance in relation to its computing time is ResNet-50.
- While this process is not infallible and much can be improved on, it does serve the purpose of giving a glimpse of the video to come. It is a practical method that follows human intuition and performs relatively quickly.

## Acknowledgements

**Department of Electrical and Computer Engineering**
Dr. Ehsan Elhamifar, *Associate Professor, Electrical and Computer Engineering*
Zijia Lu, *Ph.D. Student for Computer Science*
Yuhan Shen, *Ph.D. Student for Computer Science*
**Center for STEM Education**
Claire Duggan, *Director of Programs and Operations*
Salima Amiji and Natasha Zaarour, *YSP Coordinators*

**Young Scholar's Program at Northeastern University**
Claire Duggan, Program Director

This work was supported by the National Science Foundation
Grant #IIS-1657197