NMF Clustering and Pipeline Improvement in Cross Cultural Analysis

Zhengyi (Skye) Chen and Yifei (Natalie) Chen

Directed Research Report – Spring 2021
Adviser: John R. Kender

COLUMBIA UNIVERSITY

2021

Abstract

The booming of videos and related industries makes it easier for people to express their perspectives thoroughly, utilizing both visual images and language expressions. As a consequence, we use news videos to analyze the cultural differences among countries. We develop a pipeline to extract useful information from videos. Natural language processing tools are used in processing text documents, while the ones for computer visions are used in analyzing frames. To build a relationship between texts and images, which are in different modalities, an embedding model called Deep Canonical Correlation Analysis (DCCA) is utilized to map texts as well as images to a joint latent embedding space and give us a correlation matrix as the result. In the end, matrix decomposition techniques like Non-negative Matrix Factorization (NMF) helps in decomposing the correlation matrix into two sub-matrices for text and visual and providing us with insights on significant features or characteristics of a specific culture.

Continuing on the works by previous research assistants, we test and improve the pipeline until the step of DCCA. Besides, we manually check all input news videos about AlphaGo and use the idea of topic modeling in developing a tool to filter out unrelated videos. What's more, NMF is utilized in image clustering on Chinese and English videos, where two types of NMF and two different stopping criteria are used. NMF based image clustering performs better on English videos compared to Chinese videos.

Table of Contents

Chapter 1: Introduction and Background

With the development of technology and the Internet, people now have much more ways and channels to express themselves. This phenomenon provides researchers with more resources and more creative approaches to one of the most popular topics in human beings' history. That is the cultural difference. It is ubiquitous as it appears in almost all aspects of life, which provides researchers with a diverse and abundant set of approaches. In this project, we primarily focus on one media that can transmit opinions of people, which reveal features or characteristics in culture.

As a media comprised of visual images and language expressions, video enables presenters to thoroughly express themselves in both modalities. However, due to the diversity of videos and the lack of standardization in video sources, analyzing all videos would introduce noises and create difficulties in the discovery of cultural differences. As a consequence, news video is the target of our project as it shows cultural features to some extent while at the same time, maintaining the highest standards, including language usages, idea generations and expressions, and the quality of the presentation.

Currently, we focus on analyzing the cultures of two countries, the United States and China. Due to the cultural differences, those two countries have different news exposure levels on various topics. To minimize the influence of topics on the culture analysis, we target one topic that is frequently discussed by news broadcasts and channels from both US and China, which is AlphaGo. The topic of AlphaGo focuses on the development of artificial intelligence and the competitions between the computer program AlphaGo and human professional Go players. It becomes the first computer program that defeats a Go world champion [1]. As a consequence, it receives lots of media attention from both China and the US.

A pipeline is designed to discover the cultural differences shown in news videos, in which natural language processing tools are used in processing speech documents and computer vision

tools are utilized in extracting keyframes of the videos. After that, the idea of Deep Canonical Correlation Analysis is applied to draw the relationship between the processed text documents and the frames, whose final result is a correlation matrix. Matrix decomposition methods, including Non-Negative Matrix Factorization, are then used to decompose the acquired correlation matrix to draw insights about some key features unique to a culture and compare the results for China and the US.

The general structure of the pipeline has already been built and tested except for the last few steps including the application of DCCA and NMF. During this project, we continue working on testing the existing pipeline and finish the application of DCCA on our data. Besides, after manually checking the existing news videos, we notice the inappropriateness of some videos, including the format of the videos, sources, and topics. Therefore, we clean our input video sources manually and develop a tool to automatically filter out the irrelevant videos.

Other than the pipeline, we are also interested in the image clustering patterns among different cultures and the function of NMF in image clustering on our data. We apply two types of NMF, which are Sequential NMF and ALS NMF, and try two types of stopping criteria in image clustering, which are Dispersion Coefficient and PCA.

Chapter 2: Video Source and Pipeline Improvement

2.1 Video Source for AlphaGo

The original video dataset on the topic of AlphaGo is stored on Google Drive, containing 128 Chinese videos and 200 English videos. As mentioned before in the introduction, we would like to select news videos in two cultures to discover the cultural differences. To achieve a rigorous result, we have a close look at the videos in both cultures and find that some of the videos are not qualified with our standards. In the dataset of Chinese videos, 30 of the 128 videos are either under the irrelevant topic or not in news format. For English videos, 162 of them are either under the irrelevant topic or not in news format. Meanwhile, 33 of them, even though in English, are not generated from the culture of the U.S. Most of them came from South Korean or Chinese international news channels. Hence, using the videos from those non-US culture channels does not help reveal the cultural differences. For those videos with only parts of it meeting our requirements, we manually select the relevant clips.

As a consequence, there are only 7 of the original English videos and 99 of the original Chinese videos left. Even though the small number of English videos would not affect the whole pipeline for this project significantly, we still need to find more news videos about AlphaGo to enlarge our English dataset. Before, the videos are downloaded from YouTube with the searching query of *AlphaGo News*. Also, the length of a video is restricted to be less than 4 minutes. Hence, to obtain more usable videos, we add news channels to the searching query, for example, instead of using *AlphaGo News*, we use *AlphaGO CNN*. We also change the video length limitation to be less than 8 minutes. We then use Youtube API for searching as well as downloading videos and found 15 English videos in total. Since the competitions of AlphaGo happen in the year 2017, which is 4 years ago from now, it is a little outdated, causing some of the news reports to be no longer

available online. However, from our perspective, 15 videos are enough for now since we are still in the stage of finalizing the pipeline and utilizing NMF based clustering method on several videos for exploration. Updated English and Chinese video datasets are uploaded to Google Drive [1].

2.2 Unrelated Video Filtering: Topic Modeling

Since manually checking the downloaded videos is time-consuming, we want to develop a tool to help filter out the irrelevant videos based on the extracted transcripts. Topic modeling is an unsupervised learning and statistical modeling method that discovers topics in a collection of documents. The collection of documents, in our case, is the set of transcripts extracted from news videos. Latent Derelicht Analysis (LDA) is one of the most popular approaches for topic modeling, which treats the documents as bags of words and ignores the order of words. The assumption for LDA is that the documents are generated by selecting a set of topics and a set of words from each topic. Therefore, this probabilistic model uses two probabilities: $\mathbf{P}(topic|document)$ and $\mathbf{P}(word|topic)$. With LDA, we can find the topic word list for each transcript. By checking whether the topic words list contains our target words, we can estimate the relevance of the video.

The procedure for the methodology is as follows:

1. Data preprocessing on transcripts:

    - Tokenization: split the text documents into sentences and then split the sentences into words. Change all the words into lower cases and remove all punctuations.
    - Lemmatisation: words in the past and the future tenses are changed into the present tenses. All words in the third person are changed into the first person.
    - Stem: words are transformed back to their root formats.

2. Set up a target word list, which are the words that are, from our perspective, highly related to the target topic and must appear if the video is relevant. For instance, for the topic of AlphaGo, we select the target list to be: AlphaGo, Google, Go, Computer, Intelligent, Machine,

[1] Google Drive:https://drive.google.com/drive/folders/1HgYGKydsT6IhqBabDHEDNOHl-cPJpRWt?usp=sharing

and Match. These target words can be added or deleted after observing their performances on a training set of documents.

3. Generate topic words for each transcript from LDA and check if the target word appears.

4. If none of the target words appears, then we define it as irrelevant and remove the video from the dataset. If at least one of the target words shows up, then we keep the video.

Since we have already manually checked all English videos and decided which ones of them are irrelevant, we can use the manually checked unrelated list as the ground truth to evaluate the performance of the filtering method. There are 162 irrelevant videos after our manual checking. The result for the filtering algorithm is as following: it filtered out 98 transcripts (documents) and all of them are in the unrelated list. It indicates the good performance of the filtering method.

When downloading videos using the searching queries, it is very likely to have some irrelevant videos. Adding them to the dataset would significantly influence the following analysis. Applying the step of filtering helps double-check the video sources to avoid completely irrelevant videos. If the topic changes in the future, the method can also be applied as long as the target words list is changed correspondingly.

2.3 Pipeline Brief Instruction and Improvement

Due to the unique formats of videos, we design a pipeline to extract useful information from news videos and use it to discover possible differences among cultures. This pipeline starts from the step of downloading and filtering videos to applying NMF in gaining insights on the key features of culture.

Overall, the pipeline contains the following steps:

1. Download and filter videos

2. Extract and process audio transcripts

3. Select video frames and extract key features

4. Find duplicates in keyframes/features of two cultures

5. Transform text documents to vectors

6. Apply DCCA to maximize the correlation between processed transcripts and selected frames in one culture

7. Utilize NMF in decomposing the gained correlation matrix from Step 6 to draw meaningful information

Detailed explanations on the algorithm and reasons lying behind the pipeline could be found in the report [2] by Liu. Instructions on setting up the pipeline are specifically explained by Liu and Zeng in their report [3]. The updated pipeline [1] and codes for video filtering as well as NMF image clustering [2] could be found on the GitHub repositories.

Based on the works done by previous research assistants, Steps 1 - 4 have already been set up and tested. Due to the long-running time of DCCA, step 6 has been set up but has not been tested and modified based on our data.

We continue on their works. After filtering the input videos and downloading new videos mentioned in Sections 2.1 and 2.2, we rerun and make small adjustments on Steps 1 – 4 to ensure their performances on the new data. As for the step of DCCA, we aim at mapping both the vector of extracted frame features and the processed speech documents to a latent space where the correlation between two vectors is maximized. To do so, the pre-processed text documents are transformed into vectors, using the technique called Word2Vec. Then, both vectors are imported into the step of DCCA. In the end, we gain a correlation matrix containing the maximized correlation between text data and frames. Originally, the file of running DCCA is for external use, whose structure is built but has not been tested on our dataset. After adding step 5 and modifying the current code for step 6, we can successfully get the correlation matrix as the result.

[1] Updated Github Repository: https://github.com/YifeiNatalieChen/Cross-Cultural-Analysis2
[2] GitHub Repository: https://github.com/Skyeczy/CrossCulturalAnalysis

Chapter 3: NMF Clustering Methodology

To observe the cultural differences embedded in the video frames, we want to see whether the videos in Chinese and English show different image clustering patterns and whether the different clusters reveal some distinctions on the cultural level. Hence, we are facing an unsupervised learning clustering problem where the targets are frames extracted from videos.

Non-negative matrix factorization (NMF) is a set algorithm in multivariate analysis where a non-negative matrix $\mathbf{V}$ is factorized into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ [4], i.e.:

$$\mathbf{V} = \mathbf{WH}. \tag{3.1}$$

Let us denote n be the dimension of the data vectors, m be the number of examples in the data set, and r be the order of the NMF algorithm, then $\mathbf{W} \in \mathbf{R}^{n\times r}$ and $\mathbf{H} \in \mathbf{R}^{r\times m}$, where r is much less than n and m. Hence, NMF factorizes the original matrix into two matrices with significantly reduced dimensions. And the approximation of the original matrix $\mathbf{V}$ can be achieved by the multiplication of the two factorized matrices $\mathbf{V} \simeq \mathbf{WH}$. Figure 3.1 is an illustration of the NMF algorithm.

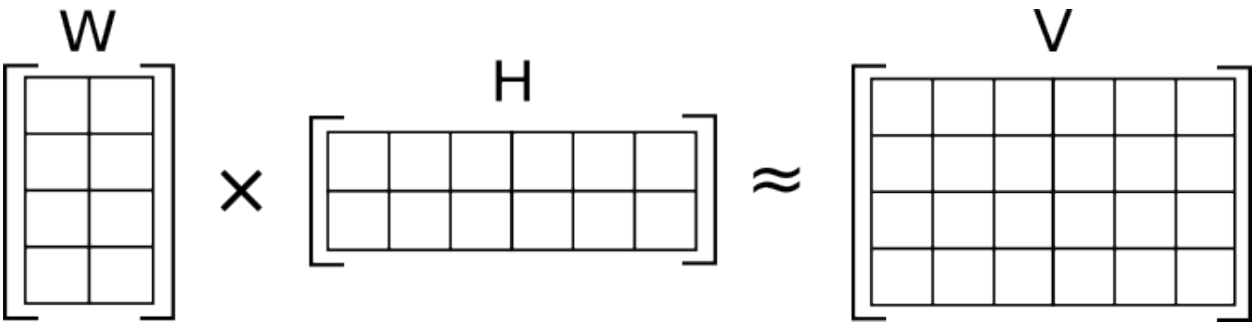

Figure 3.1: Illustration of approximate non-negative matrix factorization

The matrices **W** and **H** are calculated by iterative update to minimize the cost function and the cost function is defined as:

$$||V - WH||_F, \text{ subject to } W \geq 0, H \geq 0. \tag{3.2}$$

Ding et al.(2001) [5] proves that by imposing an orthogonality constraint on **HH** , i.e. $\mathbf{HH}^T = I$, the minimization of the cost function is equivalent to Kernel K-means clustering. Therefore, the NMF algorithm has an inherent clustering property. The clustering membership is determined by observing the matrix **W**: if $\mathbf{W}_{kj} > \mathbf{W}_{ij}$ for all $i \neq k$, then the input data v_j belongs to k^{th} cluster. The cluster centroids can be found in matrix **H**, where k^{th} row gives the centroid of k^{th} cluster. Even though the orthogonality constraint $\mathbf{HH}^T = I$ is not explicitly imposed, the orthogonality holds to a large extent, so is the clustering property [6].

Since we want to find clusters for the frames in each video, the input should represent each frame image. We define two types of input. The first one is to directly input the frame image pixel data and the other one is to input the frame image features extracted from a neural network [3]. For the feature extractor, we use the same one as the pipeline, which is the VGG-19 network. In the extracting process, the frame resolution is resized to be adapted into the VGG-19 network.

3.1 Non-negative Matrix Factorization

NMF is an iterative approach where many different kinds of methods or solvers could help updating the two sub-matrices of the original matrix in the iterative step. In our project, we focus on the study of two types of NMF algorithms, Sequential NMF and Alternating Least Squares(ALS) NMF. In the following sessions, we will introduce and present the results of these two NMF algorithms respectively.

3.1.1 Sequential NMF

NMF algorithm involves a random initialization for the iterative approach, which means that all components are treated as equally important. Besides, Sequential NMF constructs the components sequentially, which means that the $n + 1^{th}$ component is calculated based on all the previous n components [7]. Furthermore, the order of the components represents the importance of the components. This sequential method is first introduced by Ren et al.(2018) to solve an astronomy problem. In this project, the implementation of the NMF algorithm follows the implementation in Ren et al.'s paper, where only the $n + 1^{th}$ component is randomly initialized and the first n components are initialized with their previously constructed values [7]. The development of this method is based on the **NonnegMFPy** package in Python.

3.1.2 ALS NMF

Same as the Sequential NMF, the application of Alternating Least Square (ALS) NMF starts with random initialization of the two sub-matrices, W and H, of the target matrix V. However, the initialization does not have to be all randomized. According to Zhou and other researchers , while doing ALS for movie rating matrix, they initialize the sub-matrices by assigning the average ratings for movies in the first row and random small numbers in the following rows [8].

One big difference between Sequential and ALS NMF is that during each gradient descent update step, one sub-matrix is fixed and the other one is updated by minimizing the objective function. Then the updated matrix is fixed, and the other matrix is updated in the same way. The iterative update stops until the stopping criteria are satisfied [8]. Besides, the loss function of the ALS NMF problem involves the regularization term to avoid overfitting. When the regularization matrices for W and H are not singular, the minimization of each sub-matrix with the other fixed has a unique solution [8]. This characteristic provides supports for the iterative update step.

3.2 Stopping Criteria

For a clustering problem, one of the challenges is to determine the number of clusters we want to set for the data, which is also the input for the NMF algorithm. Having the number of clusters being either too small or too large has influences on the clustering result and creates instability. In our project, we propose two methods of choosing the number of clusters. One of them is based on the Principal Component Analysis (PCA) and the other is to use the dispersion coefficient.

3.2.1 PCA

Principal Component Analysis (PCA) is a technique that reduces the dimensionality of the data while minimizing information loss or preserving as much variability as possible [9]. It is usually used as a method to visualize the data in lower dimensions and pointing out the directions or vectors that contain the most significant amount of information or variability of the dataset as eigenvectors of its covariance matrix. The result of PCA is a collection of principal components, which are unit vectors orthogonal to each other.

With the ability to indicate principal components of the dataset's variability, PCA is used as the stopping criteria of our NMF image clustering. According to Lazar and other researchers, the number of most important principal directions indicated by PCA could be treated as the number of clusters in the NMF based image clustering [10]. Although the paper does not mention the definition of "most important" for the principal directions, we use the eigenvalue of 1 as the threshold. The principal components with eigenvalues larger than 1 are treated as "most important".

The power of PCA as the stopping criteria of NMF based clustering is proved by experiments. Without using the number of clusters determined by PCA, we run NMF clustering on both Chinese and English video frames with clustering numbers ranging from 0 to 15. Then we manually look at the resulting clusters of images. Based on our subjective observations, the clustering does the best with the clustering number equal to the one determined by PCA. Similar images or relevant images with common characteristics are grouped without a lot of noise. The threshold, eigenvalue

of 1, gives us a collection of principal components that could explain 70% - 80% variability of the dataset. It ensures that the selected principal components indeed represent the most amount of information of the original dataset and supports future NMF image clustering.

3.2.2 Dispersion Coefficient

Dispersion coefficient is a metric to evaluate the stability of a clustering algorithm with random initialization. The main idea is to compute NMF with multiple initialization for various cluster numbers and then evaluate the dispersion coefficients. This method focuses on the consistency of the clustering algorithm [11]. The method has already been used by Brunet et al. [12] and Kim and Park [13] to determine the unknown number of groups from gene expression data.

First, we denote K being the number of clusters and n being the number of frames. The process for computing the dispersion coefficient for K clusters is as follows:

1. Construct connectivity matrix

   The connectivity matrix $C_k \in \mathbb{R}^{n \times n}$ for n data points is constructed from each execution of sequential NMF. For each pair of frames i and j, assign $C_k(i, j) = 1$ if the i and j are assigned to the same cluster, and $C_k(i, j) = 0$ otherwise where k is the cluster number given as an input to the clustering algorithm.

   Average connectivity matrix $\hat{C}_k$ is computed by averaging the connectivity matrices over trials. Each element of $\hat{C}_k$ stays between 0 and 1. The value $\hat{C}_k(i, j)$ indicates the possibility of two frames i and j being assigned to the same cluster. If the assignments were consistent throughout the trials, each element of $\hat{C}_k$ should be close to either 0 or 1.

2. Compute dispersion coefficient

   General quality of the consistency is summarized by the dispersion coefficient:

$$\rho_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4(\hat{C}_k(i, j) - \frac{1}{2})^2,$$

where $\rho_k \in [0, 1]$ and $\rho_k = 1$ represents the perfectly consistent clustering.

3. Repeat for different K value and obtain ρ_k for various K

   The number of clusters could be determined by the point K where ρ_k drops.

Even though this methods is not a stopping criteria since it requires the methods to be applied in possible cluster values then determine the proper cluster number based on the dispersion coefficient over all cluster values, it can still help determine the correct cluster number for the data set. In our implementation, we set the number of trials for each possible cluster number to be 15. Since all the components are calculated sequentially, the time cost is relatively high. For example, when cluster number is 11, it takes 9-10 minutes to reach a solution. Having 15 trails means that it takes about two and a half hours to compute the dispersion coefficient for 11 clusters. To improve the time cost for this method may be a possible future research direction.

Chapter 4: NMF Clustering Result

In this result session, we are going to present the number of clusters we determined and show some images within the acquired clusters.

4.1 Sequential NMF

For Sequential NMF, there is one thing that needs to be mentioned, which is our approach to determine which cluster a frame belongs to. Since the dispersion coefficient method involves in many (in our project, 15) trials for each cluster number, and each time the clustering results are not identical even with the most stable cluster numbers, how to determine the belonging clusters becomes an interesting and open problem. According to the papers that use this method, the researchers apply the NMF algorithm once after determining the correct cluster number. In this project, we take the majority vote method to determine the cluster for a frame. For example, if the frame is grouped into Cluster 1 ten times and Cluster 2 five times, we set this frame to Cluster 1 in the final result. Whether there is a better method of assigning frames to clusters will be a potential future research direction.

For Chinese Video 19, we first use the method of PCA to determine the number of clusters. The eigenvalue of the matrix goes below 1 after the tenth eigenvalue, hence we select 11 to be the input for the Sequential NMF algorithm. However, when we try to separate the frames into 11 clusters, only six of them have frames and the other five clusters are empty. Observing this interesting clustering behavior, we try to the use dispersion coefficient method to determine the proper cluster numbers. Graph 4.1 shows the change of the dispersion coefficient under different cluster numbers. Since we select the number of clusters where the dispersion coefficient begins to drop, the number of clusters is 6 for video 19 based on the observation from the figure. This result

is the same as the actual clusters we get for video 19 when we tried to separate it into 11 clusters.

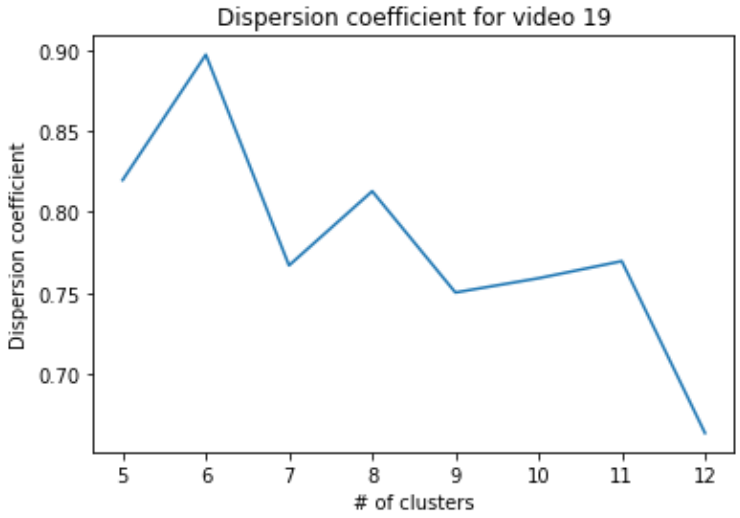

Figure 4.1: Dispersion coefficient of video 19.

We also apply two stopping methods on an English video, which is Video 8 in our data set. For the PCA method, the eigenvalue becomes less than 1 after the ninth eigenvalue, so we pick 9 as the input for the NMF algorithm. After clustering frames into nine clusters, the result gives two empty clusters and one cluster with only two frames. It means that three clusters are unstable. When we apply the dispersion coefficient method, we obtain graph 4.2.

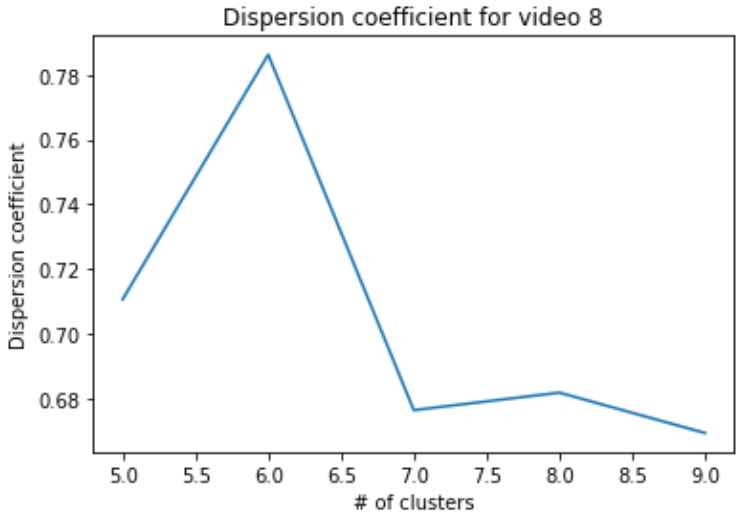

Figure 4.2: Dispersion coefficient of video 8.

From the graph, we observe that the dispersion coefficient begins to drop after the cluster number of 6. Hence we select the cluster number to be 6, which is also the same as the stable cluster numbers given by the PCA method.

Therefore, the PCA method gives an upper bound of the number of clusters in the data. The dispersion coefficient method gives the most stable cluster number that we can obtain. This conclusion also reveals the main idea of the dispersion coefficient, which is emphasizing consistency and stability.

After introducing the cluster numbers determined by the two methods, we present the comparison on the clustering results. In each cluster, we randomly choose 6 frames to show as examples. The left part shows the results for the clustering under PCA methods and the right part shows the ones for the dispersion coefficient method. Since the sequential NMF ranks the importance of component, the Clusters 1, 2 and 3 are the three most significant components in Video 8, which are presented in figure 4.3. The clustering results for Cluster 4, 5 and 6 are presented in 4.4.

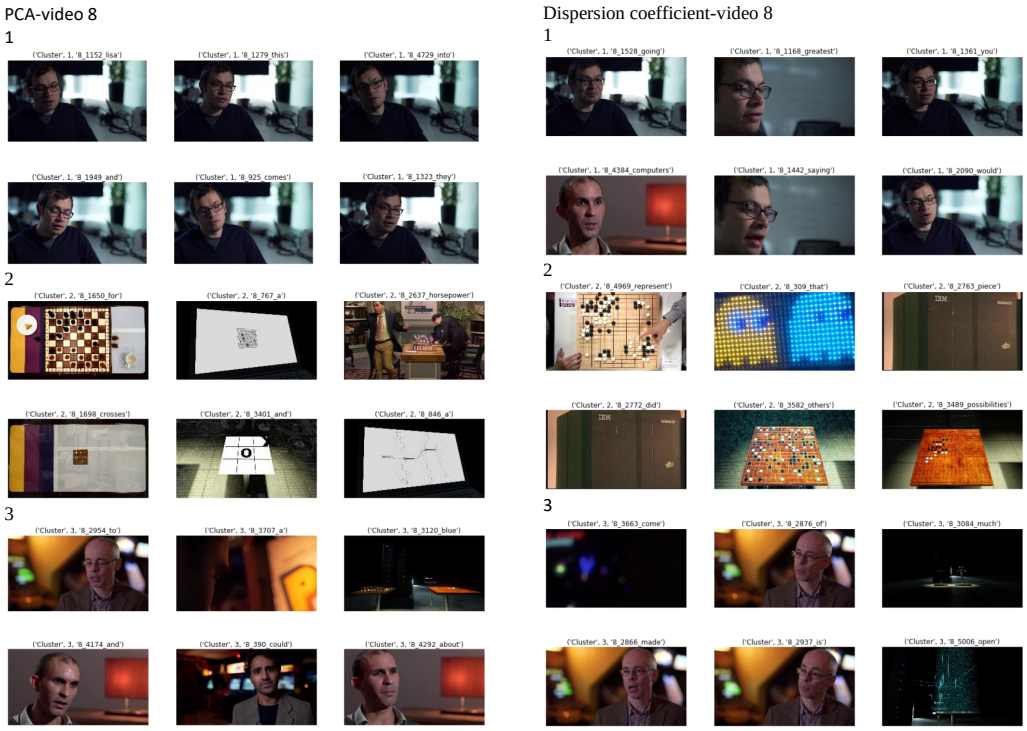

Figure 4.3: Clustering result comparison for Video 8 of Clusters 1, 2 and 3.

We can see that the most significant components for both methods are frames with talking heads, which frequently appear in our dataset. However, the PCA method separates the frames with talking heads into two clusters: Cluster 1 and Cluster 4. The difference is that the frames in Cluster 4 are more zoomed in than the ones in the first cluster. But all of these frames are clustered in the first cluster by the dispersion coefficient methods. The common observation of Cluster 2 is that both of them have a Go board.

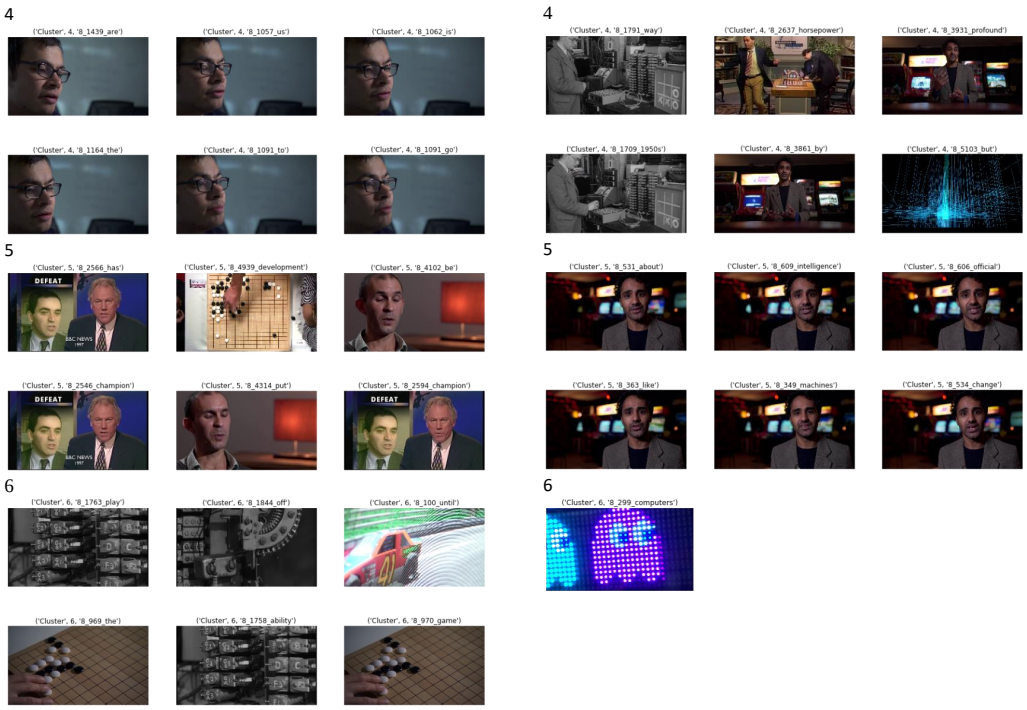

Figure 4.4: Clustering result comparison for video 8 of cluster 4, 5 and 6.

The seventh cluster in PCA methods contains only two frames. Those two frames are shown in figure 4.5. There are also clusters containing the frames relevant to the application of artificial intelligence, such as Cluster 6 for the PCA method and Cluster 4 for the dispersion coefficient method.

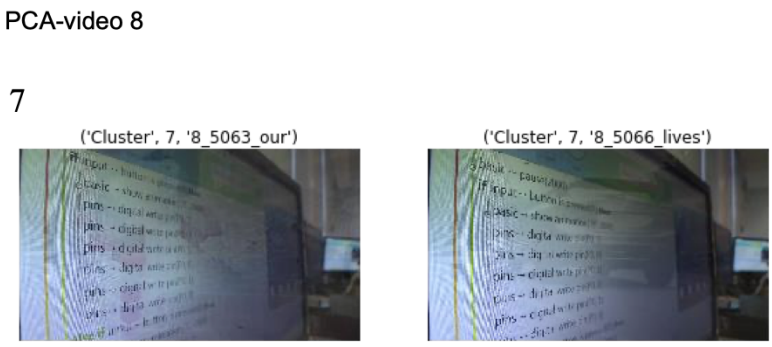

Figure 4.5: Clustering result comparison for video 8 of cluster 7.

In all, from these examples, we can find that the clustering with the dispersion coefficient methods is more stable and efficient. Now, we will introduce the clustering result for video 19, which is under Chinese culture to see the difference in clusters across cultures. The clustering result is produced from the dispersion coefficient method. Figure 4.6 shows six random frames in cluster 1 and 2, Figure 4.7 shows six random frames in cluster 3 and 4, and Figure 4.8 shows six random frames in cluster 5.

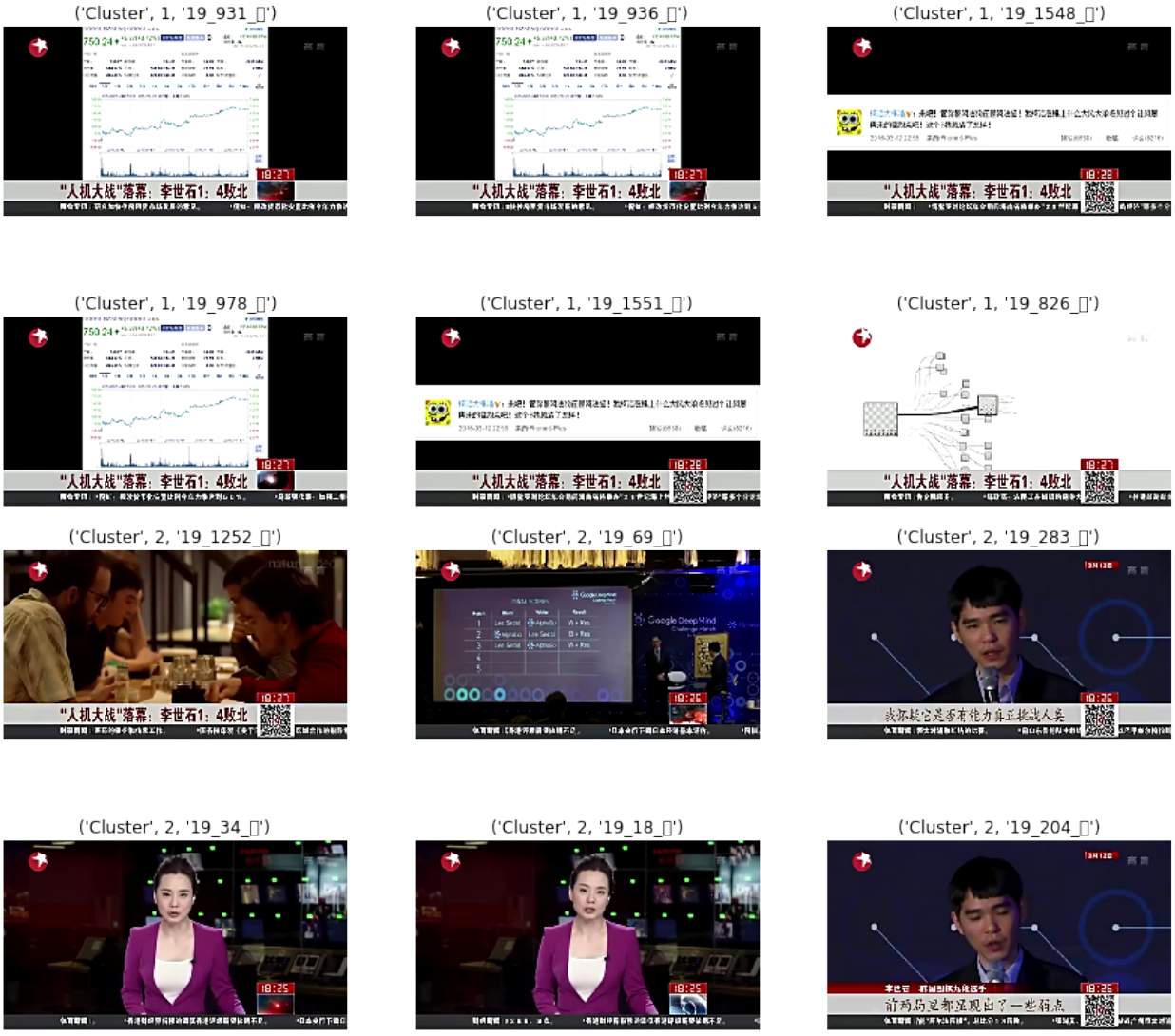

Figure 4.6: Clustering result comparison for video 19 for cluster 1 and cluster 2.

From the figures, we can see that the most significant cluster is about frames of website screenshots. The content behind these website screenshots is that the news is introducing the impact of Alpha-Go on the stock market and also displays some people's online comments about the competition. The second most significant cluster is about frames with hosts, which is very similar to the first cluster in video 8 since they are both news videos. Frames in cluster 3 are mainly about the Go board and the Go board frames in Cluster 5 are from a different angle.

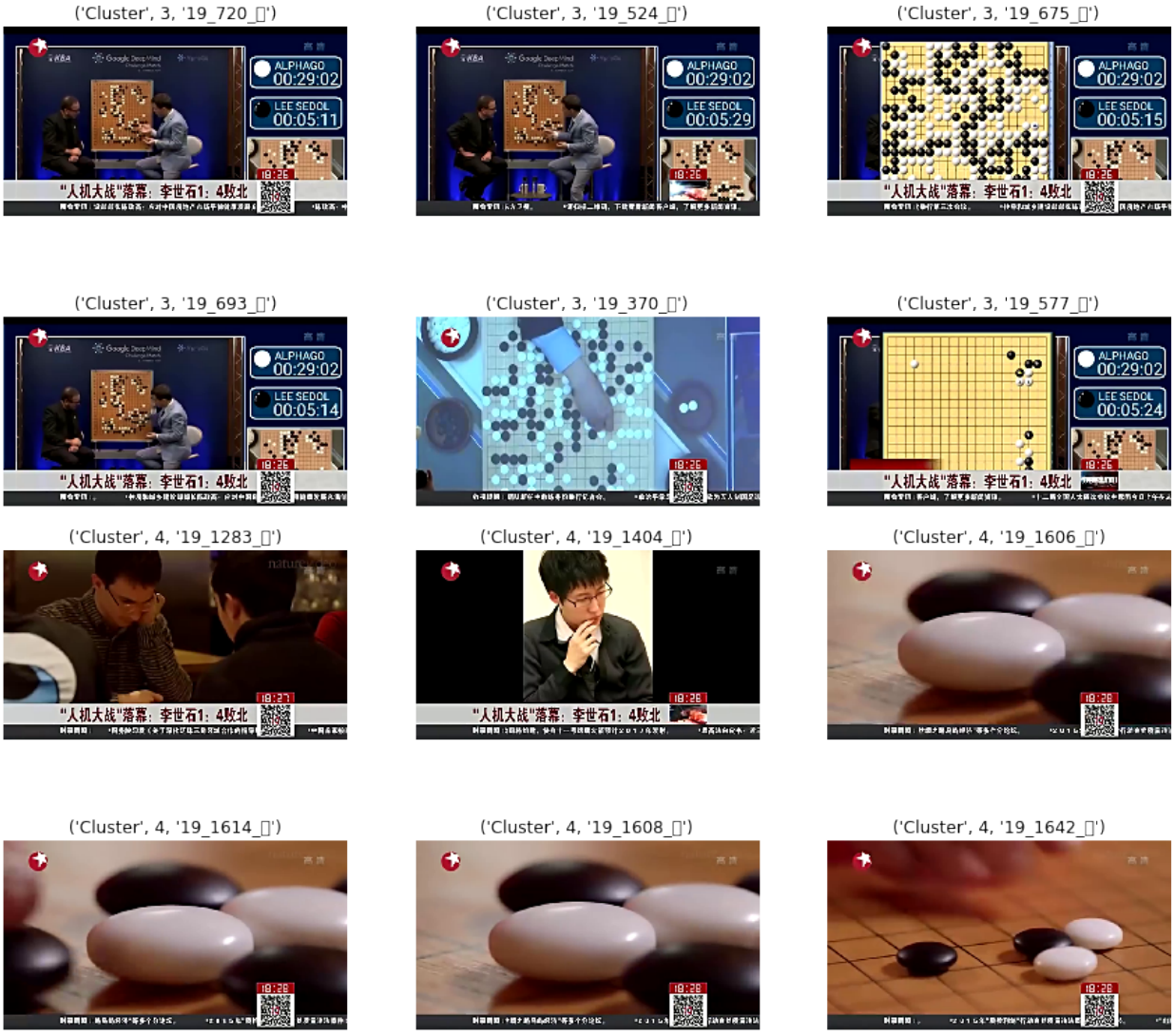

Figure 4.7: Clustering result comparison for video 19 for cluster 3 and cluster 4.

Figure 4.8: Clustering result comparison for video 19 for cluster 5.

From the analysis of the clusters in both cultures, we can see that the U.S. culture news covers the application of AI in other fields and depicts the future of AI. However, for news under Chinese culture, it is more focused on the Go game itself, how this event makes an impact on a social range, and how people react to it. This observation is the same impression people will get when listening to the verbal content in the news. Even though the analysis based on the transcript of the news videos is not presented in this project, we think that the transcripts reveal more significant distinctions between cultures. The clustering based on the transcript will be an extension direction to this project as well.

4.2 ALS NMF

As for the ALS NMF, we use PCA to determine the number of clusters and apply the clustering on pixel images instead of image features used in Sequential NMF. Specifically, two Chinese videos and two English videos, which reveal unique characteristics of the culture from a subjective perspective, are used as input in this step.

Since pixel images are directly used in this step, the execution of both PCA and NMF is long

due to the large size of the pixel images. To reduce the running time, we transform the original images to the color grey and change their sizes to smaller ones. After simple image processing and transformation from pixel images to vectors, PCA is performed to determine the number of clustering. Using the eigenvalue of 1 as the threshold, we could gain the number of "most important" principal components in the collection, which is also the number of clusters we would use in the following NMF based clustering.

First of all, let us look at the PCA results for English Videos shown in Figure 4.9. As we could tell, there are 13 principal components with eigenvalues larger than 1. In total, 79.8% of the variance in the dataset could be explained by those 13 selected principal components.

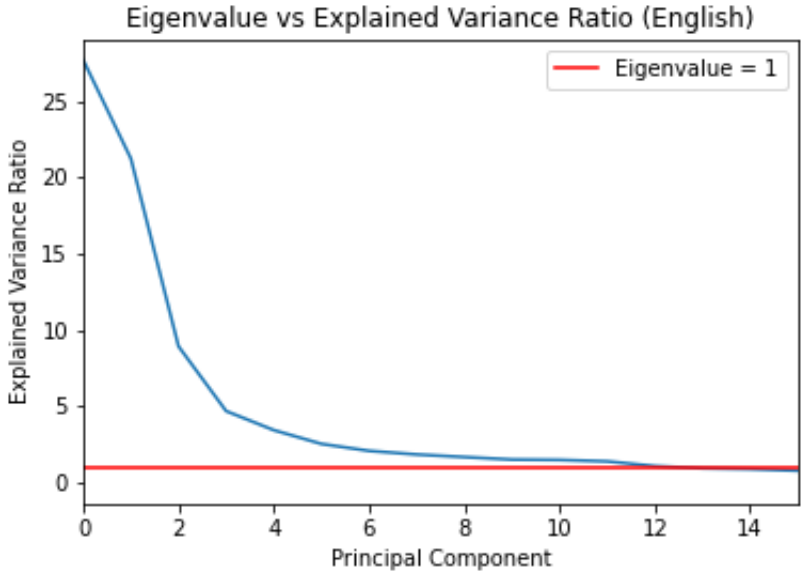

Figure 4.9: PCA Result for English Videos

Next, we perform PCA on Chinese videos. The resulting plot shows that there are 11 principal components that have eigenvalues larger than 1. Those 11 components explain 80.5% of the variance of the Chinese dataset. Both selected collections of principal components of Chinese and English videos cover approximately 80% of the variance in the datasets.

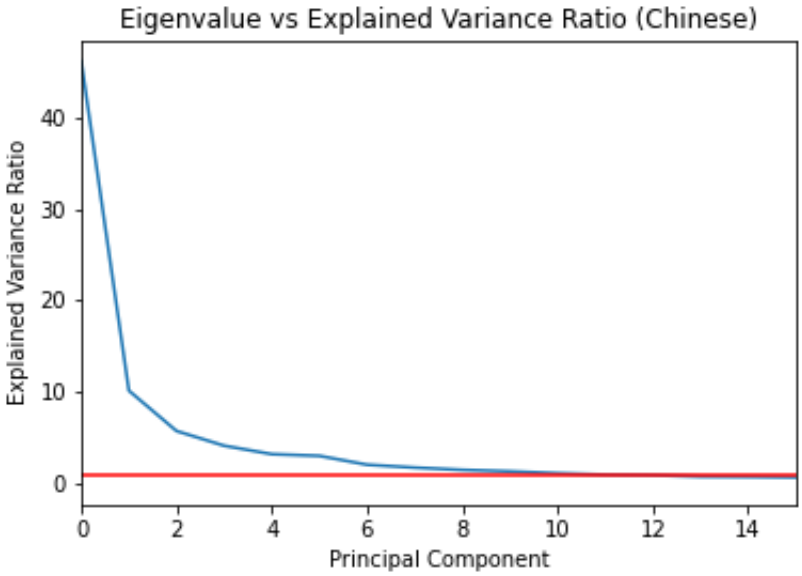

Figure 4.10: PCA Result for Chinese Videos

With the number of clusters in hand, ALS NMF is then used to perform image clustering on Chinese and English videos respectively. We are interested in the differences in image clustering patterns due to cultural differences. Since the order of PCA clusters is decided by the amount of information or variability they contain, the first 6 clusters usually contain more than half of the variability in the dataset. Therefore, similar to the Sequential NMF, we would show some examples of images in the first 6 clusters for only Chinese and only English videos.

First of all, ALS NMF-based clustering does well on English videos. As shown above in Figure 4.11, the clusters group images that share similar visual characteristics, including faces, shapes, lights, etc. Cluster 0, which contains the most amount of variance in the dataset, it captures mostly light shapes in dark backgrounds. For Cluster 1, although it does not include as much variance as Cluster 0, it captures more thorough features of the images. Most images in this group include some shapes of squares and circles like chessboards. Unlike the previous two clusters, Cluster 2 only contains images of an interviewee's face. As the cluster number increases, though there are more noises in each cluster, the images in each cluster start to share common features that are more easily been captured by human beings.

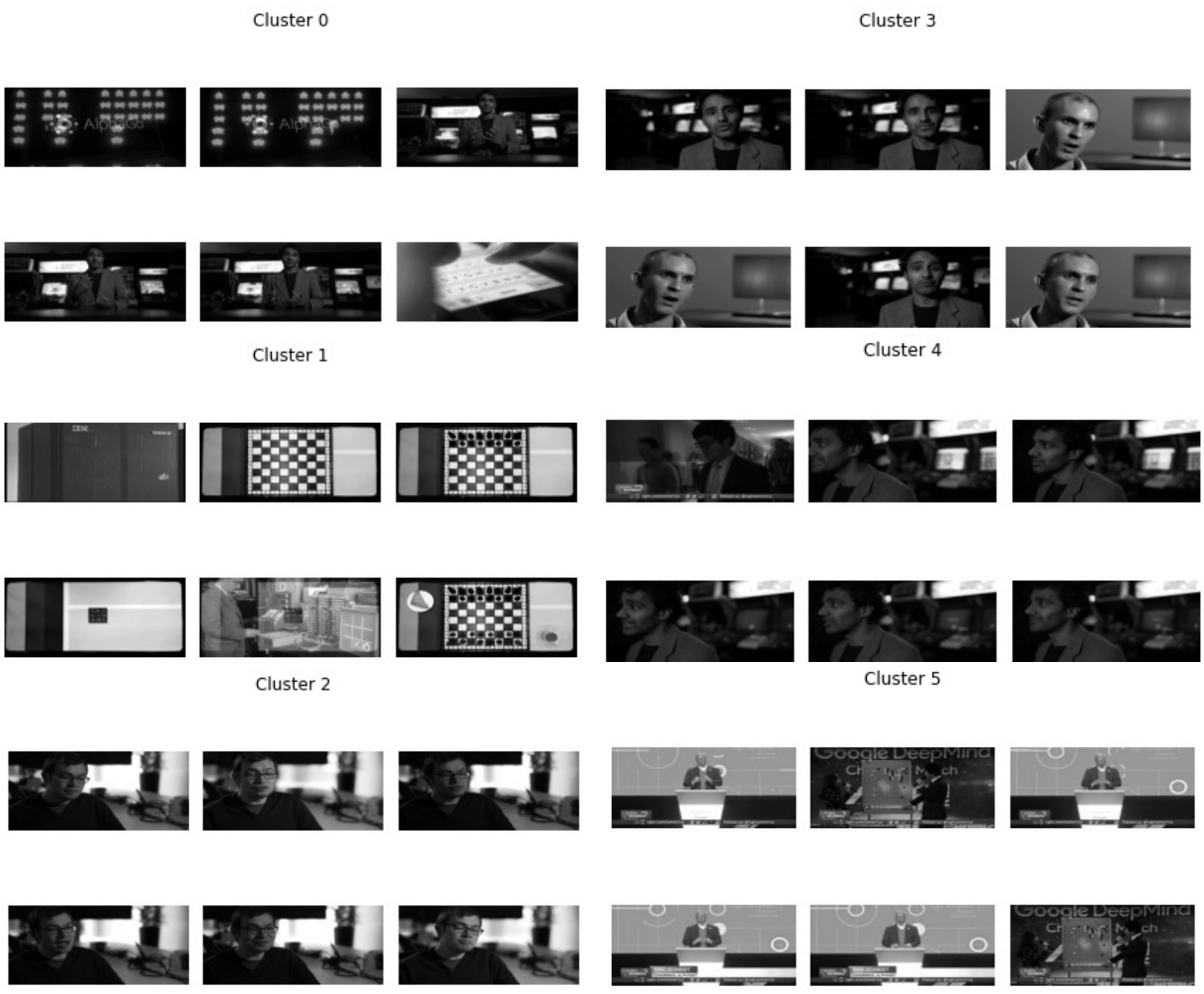

Figure 4.11: ALS NMF Based Image Clustering for English Videos

For instance, Figure 4.12 shows the images in Cluster 11 of English videos. Instead of containing the face of only one person or certain shapes, this cluster contains faces of different people from various angles, even if some of them wear glasses. It is more useful than the cluster containing only the face of a person since it captures the similarities among images that are significant for human beings instead of only focusing on shapes or lights.

As for Chinese videos, the results are not as good as the ones for English videos. Figure 4.13 shows that besides Cluster 0 which only contains frames of a host standing, some selected clusters do not share a lot of common features. For example, Cluster 3 includes frames of a Go player, AlphaGo explanation page, two Go players, and a chessboard in its collection. Those images do

Figure 4.12: Cluster 11 of English videos

not share apparent common features like faces or shapes. This phenomenon does not only happen in Cluster 3. It also applies to other clusters. Based on our observations, we suppose that the reason lying behind it is the format of Chinese news videos.

Most Chinese news videos include logos of broadcasts or captions at the bottom. As we look deep into the images in the same clusters, we could tell that they share the same logos, captions, or locations of them. It indicates the potentially significant influence of logos and captions in the NMF based image clustering. Let us use Cluster 7 for Chinese videos as an example. Figure 4.14 shows some images in the cluster. Those images share two common features. One is containing chessboards somewhere in the images. The other one is having the logo of the broadcast at the top left corner. Some of them even share the same captions at the bottom. With the same captions and broadcast logo, those images may be frames that are in sequence of the same video. Compared to the hypothesis of the algorithm grouping images using chessboards as its factor, the speculation that it groups all images in a sequence of the video as a cluster seems to be more reasonable.

If this speculation is true, it would significantly affect our results in discovering the difference in the image clustering patterns due to cultural differences. To improve this step in the future, the action of minimizing the influence of logos and captions is one of the priorities.

Figure 4.13: ALS NMF Based Image Clustering for Chinese Videos

Figure 4.14: Cluster 7 of Chinese videos

In general, ALS NMF based image clustering does a good job for English videos. It not only captures the visual features of the images but also can group images that share similar contexts or features from a broad view. Besides, PCA shows its potential in helping in the decision of image clustering problem.

Chapter 5: Conclusion and Future Work

In this project, we mainly contribute in two directions. One is improving the video source and pipeline. The other one is utilizing NMF in video frame clustering.

For the pipeline improvement part, we take a careful look at the previous Alpha-Go video dataset, filter out the irrelevant videos, and implement a topic modeling model to help automatically find the irrelevant videos based on the transcripts. As for the NMF clustering part, we apply Sequential NMF and ALS NMF methods on two types of video frame data structures: pixel images and features of frames extracted from neural networks. To determine the correct cluster number, we applied PCA and Dispersion Coefficient methods, the differences of those two methods are analyzed. We perform clustering on both U.S. and Chinese cultural videos. The observations and analyses are presented in the result section.

For future works, there are several potential directions:

1. With the Sequential NMF algorithm, after determining the cluster number, the method of deciding which specific cluster each frame belongs to can be further researched and improved. In this project, we use the majority vote method. However, more stable and advanced methods might be helpful in the cluster selection step. At the same time, we could also look into the methods that help in reducing the running time of the Sequential NMF based clustering.

2. Although we have grouped each frame to the corresponding cluster, it would be interesting and helpful to analyze and visualize the average feature/pixel frame for each cluster after the application of the NMF based clustering.

3. Remove or minimize the influence of logos and captions in Chinese videos while doing NMF based image clustering.

4. Perform clustering on a group of videos based on the transcripts instead of frames. From our perspectives, the transcripts of videos carry much useful information and have the potentials to reveal important cultural differences.

5. Complete the implementation of NMF in decomposing the correlation matrix gained from DCCA and finish the implementation of the pipeline.

Acknowledgements

Zhengyi(Skye) Chen and Yifei(Natalie) Chen acknowledge the guidance and helpful advice on this project from Professor John R. Kender during the Spring 2021 semester.

References

[1] D. Silver et al., "Mastering the game of go without human knowledge," Nature, vol. 550, pp. 354–359, 2017.

[2] Y. Liu, "Tagging and browsing videos according to the preferences of differing affinity groups."

[3] G. Liu and Y. Zeng, "Cross cultural analysis system improvements."

[4] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," vol. 13, T. Leen, T. Dietterich, and V. Tresp, Eds., 2001.

[5] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in Proceedings of the 2005 SIAM DM, pp. 606–610. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611972757.70.

[6] Wikipedia, Non-negative matrix factorization — Wikipedia, the free encyclopedia, http://en.wikipedia.org/w/index.php?title=Non-negative%20matrix%20factorization&oldid=1021588505, [Online; accessed 06-May-2021], 2021.

[7] B. Ren, L. Pueyo, G. B. Zhu, J. Debes, and G. Duchêne, "Non-negative matrix factorization: Robust extraction of extended structures," The Astrophysical Journal, vol. 852, no. 2, p. 104, 2018.

[8] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," Jun. 2008, pp. 337–348, ISBN: 978-3-540-68865-5.

[9] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philos. Trans. R. Soc. A: Mathematical, Physical and Engineering Sciences, vol. 374, 2016.

[10] C. Lazar and A. Doncescu, "Non negative matrix factorization clustering capabilities; application on multivariate image segmentation," 2009 International Conference on CISIS, pp. 924–929, 2009.

[11] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," Georgia Institute of Technology, Tech. Rep., 2008.

[12] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," Proceedings of the national academy of sciences, vol. 101, no. 12, pp. 4164–4169, 2004.

[13] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," Bioinformatics, vol. 23 12, pp. 1495–502, 2007.