Music Video Retrieval

Georgios Batsis

June 22, 2022

1 Introduction

Nowadays, the development of technological tools related to the online storage and distribution of data have led to a rapid increase in the availability of multimedia data. For example, on YouTube video streaming platform, 500 hours of videos are published every minute, while the videos already published have a total of 15 billion daily views. In terms of music, 60% of US citizens between 35-55 years old visit the platform at least once a week to watch Music Videos [1]. In addition, over 60,000 songs are posted daily on Spotify and over 4 billion playlists have been created so far [2]. It would not be an exaggeration if we claim that such digital platforms are the most widespread way in which the music industry makes its products, suchs as music tracks of music videos, available to listeners.

As the available multimedia data grows exponentially, there is a need to improve the daily users' interaction with these digital applications. The starting point for facilitating users' search within large data collections is effective Music Video content analysis which is a challenge for the engineers. Therefore, it is required to create systems that aim to find and immediately make available data similar to the content that users search for or choose to watch most frequently. These systems are called Information Retrieval Systems and their most important application is Recommender Systems which filter huge amounts of digital information based on user preferences. In this regard, these stystems create a set of recommended data personalized for each individual user [3, 4].

Music Video Retrieval (MVR) process is by its nature a complex problem because it requires the combination of two modalities: music and video. Music Information Retrieval (MIR) systems are used to analyze and search music data based on audio information. The main idea of these systems is to convert the user's selection into a specific semantic representation and then search a database to extract the association information. MIR systems have many applications, which among others are: Copyright issues, development of Recommender Systems and matching music data based on melody, genre or emotion [5]. A similar methodology is followed by engineers which develop Video Information Retrieval (VIR) systems. Both audio and video are characterized by temporal variability, with the difference that the latter are composed of sequences of images and it is necessary during

their analysis to capture information about movement and scenes alternation [6]. Music Videos consist of song and the visual content with which an artist chooses to combine it. In other words, a hybrid approach and analysis of multimodal data is required in order to develop an MVR system.

In this project, an MVR system is implemented which frames an application in which the user enters a url from a Youtube Music Video and five Music Videos are suggested to him based on similarity. Retrieval options are given either by visual or acoustic representation or by fused modalities. The first step, however, is the creation of a database from which the retrieval procedure is performed. This database contains not only the digital video but also features related to both the audio and video domains. The extraction of audio features is performed at the Segment level and are related to the time and frequency domain. For instance, some of these features are Zero Crossing Rate, Spectral Entropy, MFCCs and more. Despite the fact that in various scientific publications color, shape and texture features are extracted from images or videos [7], in this word this process is performed via Pretrained Deep Convolutional Neural Networks (CNN). In particular, for each video we find the keyframes which are essentially images fed to the CNN and extract the feature representation from the last layer. Finally, we aggregate the features by modality and fuse them into a single representation. The same steps are followed to generate the audio-visual representation of the Music Videos for which the user wishes to obtain recommendations, and retrieval is achieved based on similarity with the corresponding from the database using distance metrics.

The retrieval process and thus the development of the Recomender System could be based on both examining the set of past user searches and examining textual information associated with Music Videos. The first approach does not lead to the desired results because it is difficult to handle the new data posted on streaming platforms every day. The second approach concerns the modality of the textual information which is often added by the users themselves, for example tags and description, and in many cases this type of information is either incorrect or even not available. Although in these days the combination of the aforementioned methods leads to the development of efficient MVR systems, in this project the main objective is to examine how content-based information retrieval is performed. In this regard, we attempted to develop an MVR system that recommends Music Videos to users based on the similarity of the audio and visual features of their options in comparison with the Music Videos in the database.

This project is structured in the following sections: Section 2 presents a brief review of related work and publications. Moving on to Section 3, we analyze the methodology, focusing on multimodal feature extraction and similarity-based retrieval. The results of the proposed method and the performance of the different distance/similarity metrics are reported in Section 4. Finally, the conclusions and summary of the process are reported in Section 5.

2 Related work

The rapid growth of multimedia data in recent years has led engineers to develop information retrieval systems in both the music and video domains and obviously to implement a multitude of multimodal approaches. In [8] the authors present a hybrid approach of a Movie Recommender System which combines Handcrafted and Deep feature extraction, three temporal aggregation methods, Multimodal Fusion via Canonical Corellation Analysis and Feature Weighting in order to compute the proposed similarity metric. A similar methodology is followed by the Authors of the paper [9] when performing experiments on the new Dataset they published, which consists of user ratings data in addition to audio and image.

Regarding the field of applications related to this work, a VIR system is developed in [10] in order to both classify the content of music tracks through audio features and to identify artists through visual information, in this case face recognition. In other words, they demonstrate that in several cases the genre to which a Music Video belongs may be identified exclusively through the image features and optical flow. Furthermore, in the work of Nemati et. al. [11] emphasis is placed on the association of each modality with the corresponding music genre during the operation of the Multimodal Music VIR they present. The scientific publication whose content is similar to the workflow of this work is [12]. In particular, a similar methodology is followed in terms of feature extraction but the retrieval of Music Videos using only Euclidean is performed in two stages: Retrieval based on similarity of the audio content and subsequently retrieval via visual feature similarity.

Heterogeneous data fusion is a challenge for engineers when developing Deep Learning systems using video data [13, 14]. In [15] the Music Video classification and retrieval process is approached using Deep Neural Networks (DNN) with an emphasis on emotion recognition patterns, a field which shows a lot of interest especially in the field of audio and music. Finally, the Authors of the paper [16] present a novel Multimodal Data Fusion Method combining Supervised Learning, DNN and Canonical Correlation Analysis in order to create a unified vector representation of audio and video and in the end to develop a novel and efficient VIR system.

3 Proposed method

In Figure 1 we present the flowchart of the overall MVR system development process. Initially, the Music Video collection is created in order to develop a database from which the retrieval is performed. We chose to collect videos from two different music genres, in this case Pop and Hip Hop. The collection consists of about 200 videos, 100 for each genre, and at the end we kept 10 videos, 5 per genre, for validation purposes and to select the appropriate distance metric. Last but not least, we iterate over each video in the database in order to extract audio and visual features which are also stored in the form of a special data structure mentioned below. When the user enters the Url of a Music Video

and chooses the desired modality, the audio, visual or multimodal features are extracted and similar videos are retrieved from the database depending on the distance/similarity metric.

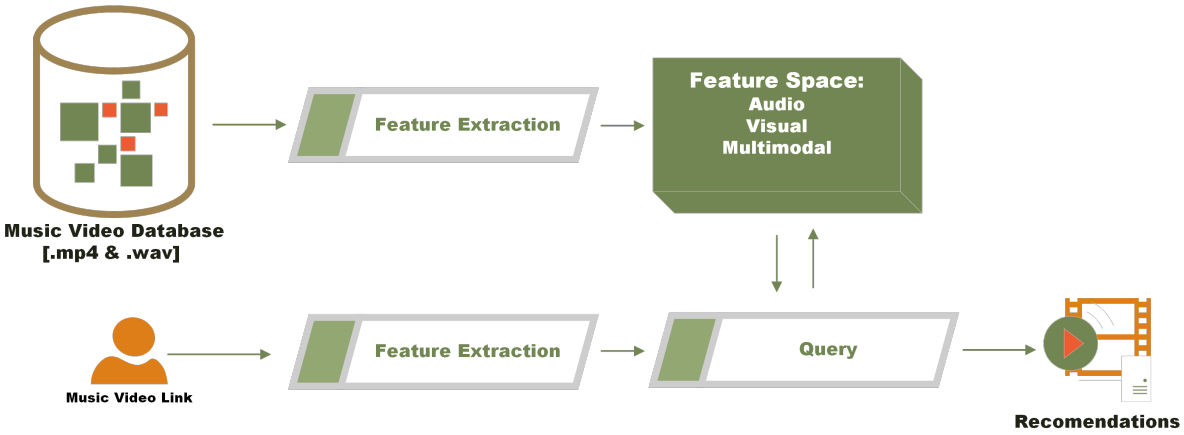

Figure 1: Overview of Retrieval Process.

3.1 Audio Features

The starting point for extracting features from audio signals is to divide the signal into small windows - frames and extract features in each of them. The process is extended by defining Segments, each of which contains a sequence of short-term feature vectors and statistics such as mean and std are calculated. In this way, on the one hand, the temporal variability of the short-term windows is captured, and on the other hand, the audio signal is transformed into a new representation of a set of features at the segment level which includes useful information for the implementation of Pattern Recognition applications [17].

The aforementioned steps were implemented with the help of the PyAudioAnalysis library [18]. The library provides the necessary tools to extract the following features: zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, Mel Frequency Cepstral Coefficients (13 features), chroma vector (12 features) and chroma deviation. For each short-term window the above 34 features and correspondingly 34 deltas are calculated, which results in a vector of 68 features. At the segment level, the statistics for the frames included within each are calculated, the feature vector is doubled after the mean and the variance are calculated. Therefore, for each segment a vector of 136 attributes is obtained. As for the size of the segments and short-term windows, it was set to 1 second and 50 milisecond respectively, while the step size is the same as the window size for both types of signal separation. PyAudioAnalysis features are described in the following table [18]:

Index	Name	Description
1	Zero Crossing Rate	*The rate of sign-changes of the signal during the duration of a particular frame.*
2	Energy	*The sum of squares of the signal values, normalized by the respective frame length.*
3	Entropy of Energy	*The entropy of sub-frames' normalized energies.*
4	Spectral Centroid	*The center of gravity of the spectrum.*
5	Spectral Spread	*The second central moment of the spectrum.*
6	Spectral Entropy	*Entropy of the normalized spectral energies for a set of sub-frames.*
7	Spectral Flux	*The squared difference between the normalized magnitudes of the spectra of the two successive frames.*
8	Spectral Rolloff	*The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.*
9–21	MFCCs	*Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.*
22–33	Chroma Vector	*A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).*
34	Chroma Deviation	*The standard deviation of the 12 chroma coefficients.*

3.2 Visual Features

Transfer Learning of popular CNN models which are pretrained on large-scale datasets, such as ImageNet, has been shown to be an effective approach to a Machine Learning problem when there are not enough data available for training from scratch to achieve high performance. In addition, Transfer Learning could be a useful tool if an alternative yet efficient feature extraction method is needed. A characteristic property of pretrained models on datasets of large number of classes is the high discriminative ability which implies the possibility of extracting informative features. Therefore, they become a powerful tool in managing smaller data sets even if they belong to a different domain [19].

One of the state-of-the-art DNN topologies are Residual Networks (ResNet), a family of CNNs which are architecturally organized in Convolutional Blocks and are available in various forms depending on their depth. The basic structural unit of ResNet are Residual units based on the idea of skipping blocks via shortcut connections, i.e. a connection between the input and output of a block. The interior of a block consists of a sequence of Convolutional Layers, Activation Funcitons and Batch Normalization and obviously the Residual connections. The latter appear in two different variants: directly connecting the block's input to the output or connecting them via a single Convolutional Layer. This particular trick allows the construction of very deep CNNs which allow the efficient solution of complex problems while optimizing the learning parameters to avoid the risk of vanishing gradients [20].

In this work, we use ResNet-152 for feature extraction and for the interpretation of visual information of videos. By recalling the model, we remove the last Fully Connected Layer, which is for the classification of input image into one of the ImageNet classes, and we keep and freeze the remaining 151 Layers. Thus, a tool for visual feature extraction from images is available. In other words, given an input image to this network, a new representation vector of 2048 dimensions is generated.

Videos as a data format consist of a series of sequential images-frames. The problem that arises in this case is that for reasons of computational and time complexity we cannot treat each frame as a single image and make predictions using ResNet, especially with videos of high number of Frames Per Second (Fps). We handle this difficulty by finding the keyframes of a video, i.e. the representative frames from each scene, using the ffmpeg tool. In this way we have for each video a set of images much smaller than the total number of frames and in the end each Keyframe is entered into ResNet to receive the extracted features.

3.3 Similarity-based Retrieval

The application of the aforementioned methods of feature extraction from audio and video generated a new vector representation of the Music Video. As far as audio features are concerned, pyAudioAnalysis extracts a vector of dimensions 136 x number of segments.

In addition, for visual features a vector of dimensions 2048 x the number of keyframes is available. However, both the number of segments and keyframes depend on the duration and scene structure of the Music Video respectively and obviously a vector of the same number of dimensions is not generated for all Music Videos within the database. That is, it is required for each modality to proceed to some temporal aggregation technique. In this work, we chose to the audio to obtain the median value of each feature, as suggested in [21], while we derive the mean and std of the ResNet predictions for each keyframe. Finally, we perform multimodal fusion and a single vector of $136 + 2 * 2048 = 4234$ dimensions is constructed.

This technique is applied both to Music Videos that exist in the database and to those for which the user wishes to find similar ones. Similarity is found through distance and similarity metrics. The appropriate metric is selected through a tuning process using a small number of videos selected from database. In particular, we tested our system using the following metrics: euclidean, cityblock, cosine, correlation, hamming, jaccard, jensen-shannon, chebyshev, canberra, braycurtis and sokalsneath. After selection, we create a cdist object using scipy library which is usefull for the calculation of the given distance between each pair of the two collections of inputs. For each modality we create a cdist object to give users the option for recommendations through Music Video via Audio, Video or Multimodal retrieval. At this point the Music Video database has been converted into a large matrix and using cdist, retrieval process is simplified and query submition is performed directly and efficiently.

4 Results

As mentioned in the previous section, retrieval experiments were performed using a small set of Music Videos for two reasons: To verify the methodology that was implemented and to select the appropriate distance/similarity metric. In particular, we tested the MVR system for all the metrics mentioned above for all 3 modalities. The performance of the system is calculated based on the correct recommendations for the validation Music Videos. A correct recommendation is considered to be one that belongs to the same music genre as the video query. The mathematical formula by which the metric performance of the system is calculated is:

$$metric = \frac{\textbf{number of correct recommendations}}{\textbf{number of query videos}*\textbf{number of retrieved videos (=5)}}$$

As for the audio modality, the retrieval results for each distance metric are presented in Figure 2. We note that the euclidean and chebyshev metrics are these that achieve the highest scores during the retrieval process, in this case 36/50 and 35/50. Especially the second metric seems to capture the similarity of features more effectively compared to euclidean when is used in audio and image data Information Retrieval tasks [22].

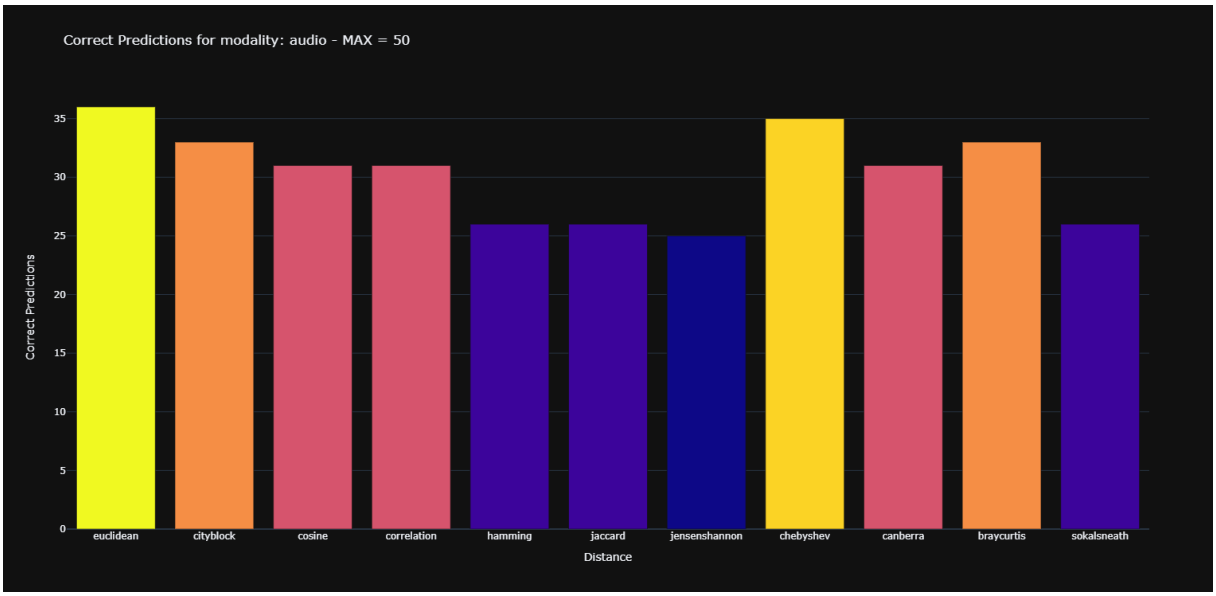

Figure 2: Performace of Audio-based retrieval per distance metric.

The nature of Deep Visual Features, allows to a variety of metrics to achieve efficient results in information retrieval. Only for three of the total of distance/similarity metrics the number of correct recommendations is low, as observed in Figure 3. The remaining metrics achieve 33/50, while cityblock is at 31/50. Although these metrics seem to have good results both if handcrafted visual or Deep visual features are extracted [23], the results of the visual modality are lower than the audio one.

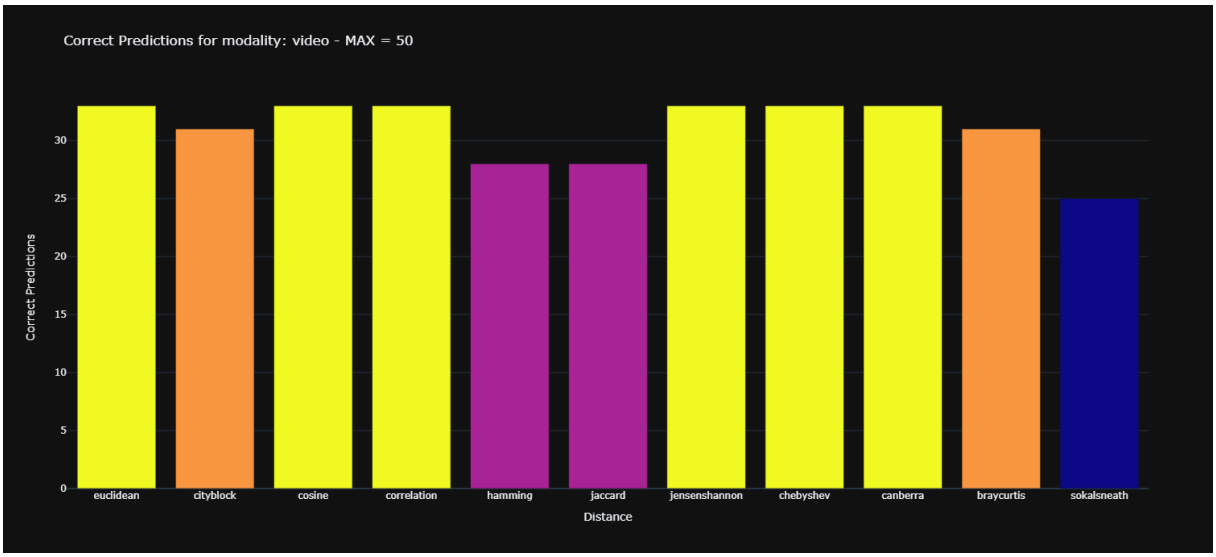

Figure 3: Performace of Video-based retrieval per distance metric.

By integrating the information from both modalities and creating a fused representation space of the Music Videos, we proceeded to the implementation of the MVR system through audio and visual content. As observed in Figure 4, metrics such as euclidean are no longer appropriate for similarity based retrieval. Chebyshev and cosine remain stable in terms of performance, but cityblock, camberra and similarity based on the Bray-Curtis algorithm achieve the highest percentage of correct recommendations. The latter in particular is the metric we chose when the system is to perform multimodal retrieval because according to the literature it is an efficient similarity metric when the complexity of the data increases due to the fact that they are generated from different sources [24]. Finally, after selecting the most efficient metrics by modality we created a Dashboard application in which the user has the possibility to type the Url from the Music Video he/she wishes and to receive recommendations of the selected modality.

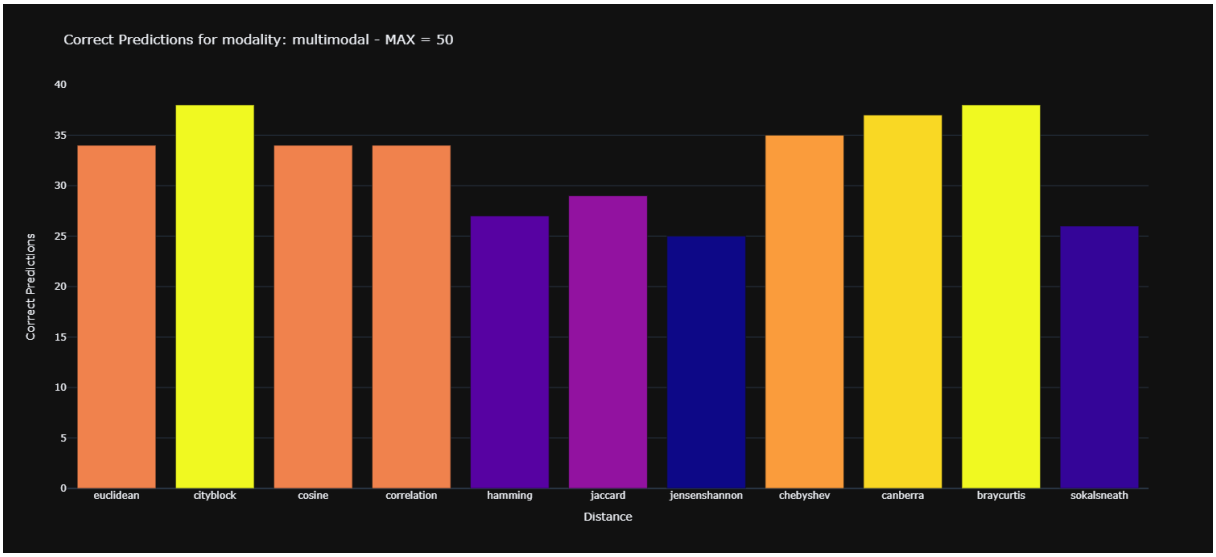

Figure 4: Performace of Multimodal retrieval per distance metric.

5 Conclusion

In this work, we presented a multimodal approach for an MVR system. Within the system, feature extraction from both audio and video domain was performed in order to perform audiovisual content-based retrieval when the user wants to receive recommendations for the selected Music Video. Initially, we collected 200 Music Videos of two music genres, 100 for each genre, creating the database from which information retrieval is performed. A precondition for the implementation of the system was the extraction of features for both modalities. Regarding audio, we extracted Handcrafted features at segment level, while in video data we extracted Deep Visual features from the selected keyframes. The same

procedure is followed when a new Music Video is given to the system in order to return recommendations from the database. This process is implemented per modality and multimodal via distance/similarity metrics. Although this methodology is characterized by high time complexity and computational resources necessity, it is still an effective method for meaningful content analysis of multimodal data because it focuses on the patterns that each modality represents individually.

References

[1] *160 amazing YouTube statistics and facts: By the numbers*, 2022, available at https://expandedramblings.com/index.php/youtube-statistics/.

[2] *30+ Spotify statistics and Facts: By the numbers*, 2022, available at https://appinventiv.com/blog/spotify-statistics-facts/.

[3] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, "A unified framework for video summarization, browsing & retrieval: with applications to consumer and surveillance video," 2005.

[4] R. Burke, A. Felfernig, and M. H. Göker, "Recommender systems: An overview," *Ai Magazine*, vol. 32, no. 3, pp. 13–18, 2011.

[5] R. Typke, F. Wiering, R. C. Veltkamp, J. D. Reiss, G. A. Wiggins *et al.*, "A survey of music information retrieval systems," in *Proc. 6th international conference on music information retrieval*. Queen Mary, University of London, 2005, pp. 153–160.

[6] D. Brezeale and D. Cook, "Automatic video classification: A survey of the literature," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 38, pp. 416 – 430, 06 2008.

[7] Y. A. Aslandogan and C. T. Yu, "Techniques and systems for image and video retrieval," *IEEE transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 56–63, 1999.

[8] Y. Deldjoo, M. Ferrari Dacrema, M.-G. Constantin, H. Eghbal-zadeh, S. Cereda, M. Schedl, B. Ionescu, and P. Cremonesi, "Movie genome: Alleviating new item cold start in movie recommendation," 01 2019.

[9] Y. Deldjoo, M. G. Constantin, M. Schedl, B. Ionescu, and P. Cremonesi, "Mmtf-14k: A multifaceted movie trailer feature dataset for recommendation and retrieval," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018.

[10] A. Schindler and A. Rauber, "A music video information retrieval approach to artist identification," 10 2013.

[11] S. Nemati and A. R. Naghsh-Nilchi, "An evidential data fusion method for affective music video retrieval," *Intell. Data Anal.*, vol. 21, no. 2, p. 427–441, jan 2017. [Online]. Available: https://doi.org/10.3233/IDA-160029

[12] S. Hou and S. Zhou, "Audio-visual-based query by example video retrieval," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–8, 2013.

[13] C. Jin, T. Zhang, S. Liu, Y. Tie, X. Lv, J. Li, W. Yan, M. Yan, Q. Xu, Y. Guan, and Z. Yang, "Cross-modal deep learning applications: Audio-visual retrieval," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 301–313.

[14] D. Yasin, A. Sohail, and I. Siddiqi, "Semantic video retrieval using deep learning techniques," in *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2020, pp. 338–343.

[15] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, pp. 1–19, 01 2021.

[16] D. Zeng, Y. Yu, and K. Oyama, "Audio-visual embedding for cross-modal music video retrieval through supervised deep cca," in *2018 IEEE International Symposium on Multimedia (ISM)*, 2018, pp. 143–150.

[17] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing*, vol. 20, 04 1998.

[18] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[21] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees, "Automatic music tag classification based on blocklevel features," in *In Proc. of the 7th Sound and Music Computing Conference*, 2010.

[22] C. Beecks, "Distance-based similarity models for content-based multimedia retrieval," Ph.D. dissertation, Aachen, Techn. Hochsch., Diss., 2013, 2013.

[23] Y. Mistry, D. Ingole, and D. Ingole, "Content based image retrieval using hybrid features and various distance metric," *Journal of Electrical Systems and Information Technology*, vol. 5, 01 2017.

[24] D. N. Thakur, D. Mehrotra, A. Bansal, and M. Bala, *Analysis and Implementation of the Bray–Curtis Distance-Based Similarity Measure for Retrieving Information from the Medical Repository: Proceedings of ICICC 2018, Volume 2*, 01 2019, pp. 117–125.