

# Thumbnail Selection with Convolutional Neural Network Based on Emotion Detection

*Duygu Algılamaya Dayalı Evrişimli Sinir Ağı ile Küçük Resim Seçimi*

Mahmut ÇAKAR<sup>1</sup> , Kazım YILDIZ<sup>2</sup> , Önder DEMİR<sup>2</sup> 

<sup>1</sup>Marmara University Institute of Pure and Applied Sciences, Computer Engineering, Istanbul, Turkey

<sup>2</sup>Marmara University Technology Faculty, Computer Engineering, Istanbul, Turkey

## Abstract

The use of video broadcasting platforms is increasing day by day. The competition for developing platforms for the broadcasting and sharing of movies and TV series is increasing. The purpose of reproducing these platforms is to increase the quality and to trace them on a single platform. Film and TV series platforms use artificial intelligence algorithms for these shares. The aim of this study is to create more attractive cover photos for users by finding suitable frames from a movie or TV series. First, the frames that were transformed into covers/small pictures on the platform were obtained. Unnecessary frames which consist of closed eyes, blurry frames, or faceless images have been removed. Also, deep learning is used to label images with objects and emotions based on the identity of the face. The thumbnails with the most repeating faces were selected by developing a face recognition model at each step. The experimental results showed that the emotion model was successful.

**Key words:** Thumbnail, Emotion Detection, Video Streaming Platforms, Convolutional Neural Network.

## Öz

Video yayın platformlarının kullanımı her geçen gün artmaktadır. Filmlerin ve dizilerin yayınlanması ve paylaşılması için platformlar geliştirme rekabeti artıyor. Bu platformların çoğaltılmasındaki amaç, kaliteyi artırmak ve tek bir platform üzerinde takip etmektir. Film ve dizi platformları bu paylaşımlar için yapay zeka algoritmaları kullanır. Bu çalışmanın amacı, bir film veya diziden uygun kareler bularak kullanıcılar için daha çekici kapak fotoğrafları oluşturmaktır. Öncelikle platform üzerinde kapak / küçük resim haline getirilen çerçeveler elde edildi. Kapalı gözler, bulanık çerçeveler veya yüzüzsüz görüntülerden oluşan gereksiz çerçeveler kaldırıldı. Ayrıca derin öğrenme, görüntüleri yüzün kimliğine göre nesnelere ve duygularla etiketlemek için kullanılır. En çok tekrar eden yüzlere sahip küçük resimler, her adımda bir yüz tanıma modeli geliştirilerek seçildi. Deneysel sonuçlar duygu modelinin başarılı olduğunu gösterdi.

**Anahtar Kelimeler:** Küçük resim, Duygu Algılama, Video Akış Platformları, Evrişimsel Sinir Ağı.

## I. INTRODUCTION

Developments in technology enable internet content to be accessed as video content on mobile phones anytime and anywhere. For this reason, there is a huge increase in watching video content. YouTube [1] has more than two billion users and a billion hours of video is consumed every day. For this reason, more and more videos are produced every day. Due to the circumstances mentioned, it is very important to choose the title and thumbnail of the video (the thumbnail is a compressed preview of the original version used as a placeholder). The number of online TV series, movies and shows watching platforms is increasing. According to the last quarter report of 2020, Netflix, one of the media services provider, has approximately 204 million paid subscriptions [2], 150 million globally in Amazon Prime [3] and the number of ad-free subscribers in Hulu is 36 million [4] according to 2020 data. Online watching platforms have emerged as competitors over time by aiming for users to spend more time in their application. It is aimed to increase the number of content on the platform and to present the existing content to users better and more attractive.

In the study, it is aimed to produce and label the pictures of the movie in a way that fits the cover art. Thus, it is expected that more original cover photos will be created with labels that may attract users' attention. Users' interests can be famous people, happy or exciting moments. After finding faces on the scene, a convolutional neural network-based algorithm has been developed to discover their identities, emotions, and other objects. In addition, closed eyes on the frame are detected and eliminated, and various parameters of the frame that can provide information are calculated. Based on the information obtained, the frame is tagged to select the most suitable thumbnails for users.

There are many studies on the selection of cover images which proceed by clustering the frames and choosing the most suitable one [5-9]. It is aimed to measure the pictures with parameters such as blur, clarity, colors, scene, composition and the presence of objects in the studies carried out for aesthetically scoring [10-12].

The source of inspiration for another study on the subject is Netflix's AVA system. Netflix also uses its own system and evaluates the AVA system. AVA is a collection of tools and algorithms designed to extract high quality images from videos on Netflix. There are three basic steps in the evaluation phase. Visual metadata such as contrast, color, brightness and motion blur are collected in the first step. Contextual metadata includes elements such as face detection, motion prediction, camera shot identification and object detection in the frame. Finally, photography, cinematography and visual aesthetic design processes are discussed in composition Metadata. Afterwards, sorting by image, face recognition, calculation of different camera angles for visual diversity and filters for maturity are made [13].

An algorithm consisting of five main steps including down-sampling, filtering, feature extraction, sorting and optimization has been developed [14]. In the down-sampling process, there are four steps: removing the first and last ten percent of the promotional video, sampling one frame per second, eliminating similar frames, and sorting the shots by length. In the filtering step, it performs the calculation of saturation, brightness, sharpness and contrast. It also includes removing subtitles embedded in the frame, finding characters, and adjusting the threshold.

Yu and colleagues expanded the video to include the title, description, and audio to define the content. The information obtained was used in selection models. It samples the developed model squares equally over time. Returns the highest aesthetic scores in the subset with a double column convolutional neural network to avoid the computational burden of rendering all frames. Frame properties extracted from VGG16, text properties from ELECTRA and sound properties of TRILL are obtained by the developed model [15].

Huang and friends discussed the task of highlighting the video and thumbnail selection from a different perspective. thumbnails and video clips were selected using the bulleted display, a new feature appearing on

online video streaming sites popular in East Asia. The proposed method was compared with a KKStream which is East Asia's popular streaming service provider. Experimental results show that most participants are satisfied with the thumbnails and video clips chosen in their own way. Therefore, the cover screen can be a valuable resource for understanding the video [16].

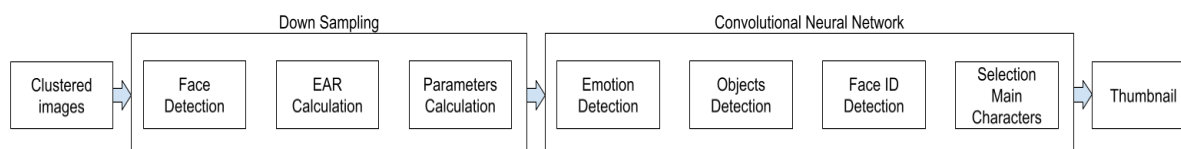
In another study, automatic thumbnail selection was performed for movies and TV shows. Thumbnail selections are automated using the classifier. The performance of different convolutional neural networks (CNNs), namely VGG-19, Inception-v3 and ResNet-50, has been compared. Hybrid approach is designed which performs best in thumbnail selection. In the hybrid model, CNN feature extraction was used in genetic programming classification. ResNet-50 CNN performed better than other CNN models [17].

The face recognition model was developed by preserving the method in our previous study [23]. In addition, by selecting thumbnails with more repetitive faces, the thumbnail pool is narrowed, and it is aimed to select thumbnails that are more eye-catching and out of context.

Section 2 contains literature information about the study. In Section 3, the steps of the developed methodology are explained and the flow chart of the algorithm is given. In section 4, the results obtained are given and interpreted. Conclusion part is given as finally.

## II. MATERIAL AND METHOD

Figure 1 shows the flowchart of the developed model. After taking the frames from the video, there are two stages: selecting the appropriate frames and using the convolutional neural network. Eye aspect ratio is made after detecting the faces in the frame in the down sampling part. The scope of this stage is to eliminate the unnecessary images via face detection to add less pictures to the model. This step is completed with parameter calculations. In order to find faces and objects related to the emotion and character, the convolutional neural network was used as the following stage. In addition, face recognition was performed and the thumbnail of the most repetitive faces was selected. Thus, it was aimed to reduce the number of results as well as identify the characters in the movie.



**Figure 1.** The schematic diagram of proposed method

### 2.1. Down Sampling

In a movie, the elimination process for unnecessary frames is important. When a two-hour film is analyzed, there are approximately two hundred thousand frames. The use of all these frames together with artificial neural networks is quite costly. In order to eliminate this memory problem, only the frames which have face images were selected using the Haar Cascade.

Comparing the Haar Cascade and Directed Gradients (HOG) in the face detection process, the detection time was measured as 16-50 ms in the haar cascade method and 340-410 ms in the other one. For this reason, Haar Cascade was used in the study for faster and more precise face detection. Histograms of Directed

Gradients (HOG) were selected for eye opening detection. Faceless images were eliminated at this stage, and the variance of the Laplacian Filter [18] was used to eliminate blurry frames. The threshold value which may not be suitable for every movie or frame, so it has to be chosen different. For each frame a different threshold value should be calculated.

Figure 2 shows eye aspect ratio samples for several eye images. Face recognition is performed with HOG in the frame of the faces and the eye opening is determined. HOG is used because it allows you to define 68 different points with 6 points for each eye found. Thus, Eye Aspect Ratio (EAR) can be calculated as seen in Figure 2 with 6 points corresponding to one eye [19].









EYE								
EAR	0.40	0.39	0.34	0.33	0.31	0.29	0.21	0.12

Figure 2. Eye Aspect Ratio samples for several eye images

The 6 dots represented by red dots are p1 (far left), p2 (top left), p3 (top right), p4 (far right), p5 (bottom right), and p6 (bottom-left) respectively. Eye Aspect Ratio is calculated as in equation 1 and eyes are assumed to be closed according to the threshold value.

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|} \tag{1}$$

Many more parameters can be calculated in the framework that can inform us. These are brightness, dominant color and mist. The Laplacian filter was used to calculate the blur parameter, which is used to find the clearest in similar frames and eliminate the blurry ones. For the brightness value, RGB (Red, Green, Blue) values are converted to HSV (Hue, Saturation, Value / Brightness) format and calculated by taking the average of the value of each square. The K-means algorithm has been used to calculate the dominant color [20]. In K-means, a single cluster is created by choosing k as 1 and the center point is determined as the dominant color. With this value, it is aimed to prevent overlapping of similar colors with the film logo and to calculate the dominant color of the logo.

### 2.2. Convolutional Neural Network

Convolutional Neural networks require a large set of N-tagged images {x, y} specifying the discrete variable representing the real class as y, as opposed to the input x. It uses a loss function to compare the output of the model with the actual class value (y). It trains the weight matrices in the fully connected layers and parameters of the network by spreading the loss derivative backward according to the parameters in the network using filters in the convolutional layers and updating the parameters by stochastic gradient descent

[21]. Convolutional Neural Networks consist of neurons with learnable weights and biases. It is also used effectively in situations such as image recognition, classification and object detection. Using convolutional neural networks, it is ensured that the emotions of the faces in the frame are detected and the objects are identified and associated.

In the study, it was aimed to detect emotions by using the Fer2013 [22] data set in order to determine the emotions on the face. 7 different emotions, including neutral, sadness, surprise, happiness, fear, anger and disgust, are tagged in the data set. The process of perceiving emotions is an important stage in terms of framing and grouping. It is also important that other objects can be detected in the frame. The presence of the pet can be considered as friendship. Finding a weapon or a sword in the frame can be seen as an action. Having a ball in the frame can determine that it is related to sports. Rather than providing information about the entire movie, this information is more important to the user's interest. For example, if someone interested in extreme sports sees a picture of nature on the cover art of the movie, it may increase the probability of choosing that movie.

Fig.3 shows the emotion model architecture which has improved according to our previous study [23]. The new emotion model is based on another study [24] but it's optimizer from ADAM to SGD (Stochastic gradient descent) replaced. Residual network is used with 224x224x3 shaped input. After flattening RESNET output layer, Dense and Dropout with ReLU activation function is applied to model [25]. Output shape is decreased with Dense with Softmax activation function to 7 since there are 7 type of emotion.

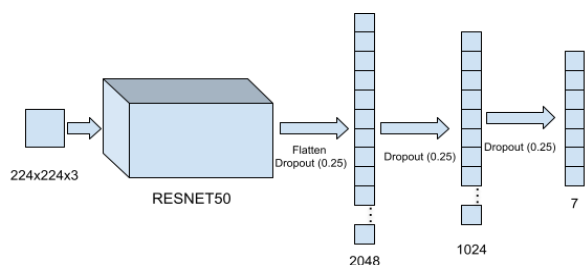


Figure 3. Architecture of Emotion Detection Network

YOLOv3 (You Only Look Once) [26] and Google OpenImages dataset [27] were used for object detection. The Open Images dataset includes about nine million images. It contains complex scenes like object bounding boxes, localized narratives, object segmentations, and more. YOLO algorithm based on prediction grids and anchors. In this study non-maximal suppression approach used to predicted and eliminated duplicates with 13x13 grids and 5 anchors. Although the YOLOv3 algorithm produced fast and accurate results, the OpenImages dataset achieved lower performance than expected. 600 classes in the data set were grouped and reduced.

It is important to identify who owns the faces in the frame to increase the likelihood of the user choosing movies featuring their favorite actor or actresses. Due to this aim, lib's pre-trained face detector [28] is used. In addition to the our previous study [23], a new feature has been added to face recognition. Every face in the frames is added to the face recognition pattern. After the process was finished, the frames with the most repeating faces were selected. As a result, logos are placed on the frames that do not coincide with the face parts. Since this step can also be performed better under human control, it is produced with and without logo.

### III. RESULTS

Emotion detection model's accuracy has improved on training from 0.91 to 0.997 and on validation from 0.67 to 0.692. Figure 4 shows train loss and accuracy of the model.

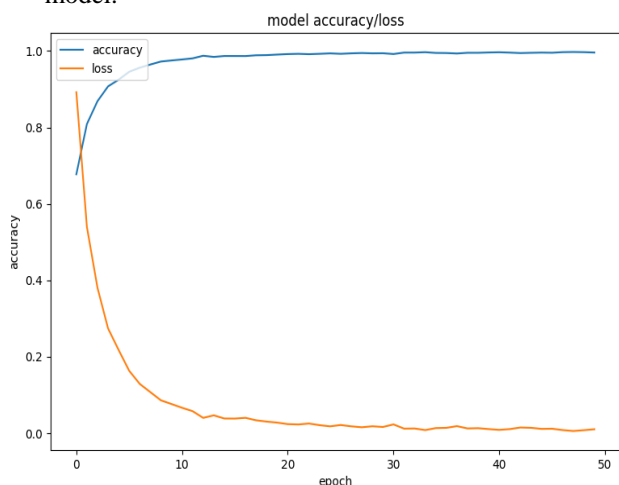


Figure 4. Train Loss and Accuracy Values of Model

Fig.5 and Fig.6 show the loss graph of 17 and 19 classes model loss graph respectively. It can be seen that the 17 classes model more accurate than 19 one. 17 classes model's mAP (mean Average Precision) is 34.28%. Training session was stopped because of 19 grouped classes the model did not reach the expected values. The resulting frames are manually examined so meaningless pictures, incorrectly grouped or tagged are eliminated. Thus, 200,000 frames in the film are reduced to 100 frames In addition, the selection of similar pictures and labels is made at different times and shown to the user.

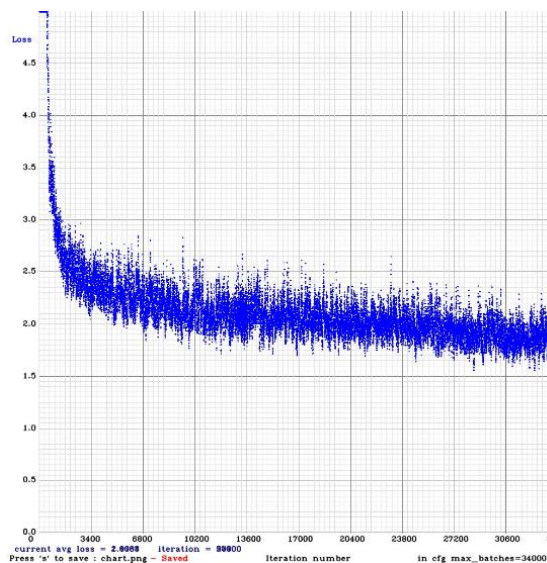


Figure 5. 17 classes YOLO model Loss Graph

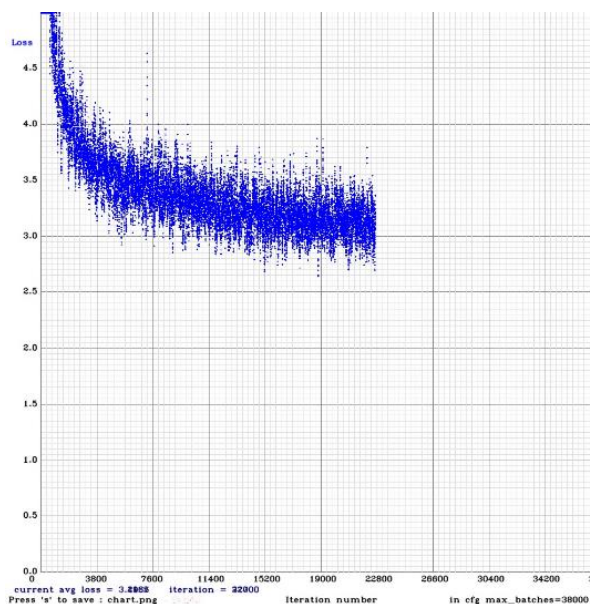


Figure 6. 19 grouped classes YOLO model Loss Graph

Figure. 7 shows sample image output results which suggests a frame with face and the logo of the film is placed dynamically according to faces.





Figure 7. Sample Results from The King of Comedy Movie [29]

## IV. CONCLUSION

The aim of this study is to create cover photos of a movie or TV series using image processing and convolutional neural networks. A successful emotion model was used in this study. In addition, in order to reduce the results, a face recognition model was developed at each step and thumbnails with the most repeating faces were selected. With the developed method, dynamic results were obtained for users on online watch platforms. Selected frames were labeled and features that might attract the attention of the user were determined. In subsequent studies, obtaining detailed information about faces, determining the movement at the scene (running, walking, swimming), thus interpreting the photographic composition in the frame. It is aimed to create a different data set to improve the performance of the models. Thus, more accurate results can be obtained by working with different data sets.

## REFERENCES

- [1] "Youtube for Press.", Youtube. Retrieved February 22, 2020 from [www.youtube.com/about/press/](https://www.youtube.com/about/press/)
- [2] Julia Stoll. (2021). Retrieved February 22, 2021 from <https://www.statista.com/statistics/250934/quar-terly-number-of-netflix-streaming-subscribers-worldwide/#:~:text=Netflix%20had%20203.67%20million%20paid,Netflix's%20total%20global%20subscriber%20base.>
- [3] Digital Commerce (2021). Retrieved February 15, from [https://www.digitalcommerce360.com/article/a-mazon-prime-membership/#:~:text=Amazon.com%20Inc.%20has%20added,Intelligence%20Research%20Partners%20\(CIRP\).](https://www.digitalcommerce360.com/article/a-mazon-prime-membership/#:~:text=Amazon.com%20Inc.%20has%20added,Intelligence%20Research%20Partners%20(CIRP).)
- [4] Hulu (2021), Retrieved January 30, from <https://www.businessofapps.com/data/hulu-statistics/>
- [5] Zeng, X., Li, W., Zhang, X., & Xu, B. (2008, June). Key-frame extraction using dominant-set clustering. In *2008 IEEE international conference on multimedia and expo* (pp. 1285-1288). IEEE.
- [6] De Avila, S. E. F., Lopes, A. P. B., da Luz Jr, A., & de Albuquerque Araújo, A. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1), 56-68.
- [7] Zhuang, Y., Rui, Y., Huang, T. S., & Mehrotra, S. (1998, October). Adaptive key frame extraction using unsupervised clustering. In *Proceedings 1998 international conference on image processing, icip98 (cat. no. 98cb36269)* (Vol. 1, pp. 866-870). IEEE.
- [8] Sujatha, C., & Mudanagudi, U. (2011, October). A study on keyframe extraction methods for video summary. In *2011 International Conference on Computational Intelligence and Communication Networks* (pp. 73-77). IEEE.
- [9] Gharbi, H., Bahroun, S., & Zagrouba, E. (2019). Key frame extraction for video summarization using local description and repeatability graph clustering. *Signal, Image and Video Processing*, 13(3), 507-515.
- [10] Deng, Y., Loy, C. C., & Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4), 80-106.

- [11] Ma, S., Liu, J., & Wen Chen, C. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4535-4544).
- [12] Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006, May). Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision* (pp. 288-301). Springer, Berlin, Heidelberg.
- [13] Riley, M., Machado, L., Roussabrov, B., Branyen, T., Bhawalkar, P., Jin, E., & Kansara, A. (2018). AVA: The Art and Science of Image Discovery at Netflix. *The Netflix Tech Blog*.
- [14] Tsao, C. N., Lou, J. K., & Chen, H. H. (2019, March). Thumbnail image selection for VOD services. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 54-59). IEEE.
- [15] Yu, Z., & Shi, N. (2020). A Multi-modal Deep Learning Model for Video Thumbnail Selection. arXiv preprint arXiv:2101.00073.
- [16] Huang, Y. Y., Kuo, T. Y., & Chen, H. H. (2020, April). Selecting Representative Thumbnail Image and Video Clip from a Video via Bullet Screen. In *Companion Proceedings of the Web Conference 2020* (pp. 48-49).
- [17] Pretorius, K., & Pillay, N. (2020, July). A Comparative Study of Classifiers for Thumbnail Selection. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.
- [18] Pech-Pacheco, J. L., Cristóbal, G., Chamorro-Martinez, J., & Fernández-Valdivia, J. (2000, September). Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (Vol. 3, pp. 314-317). IEEE.
- [19] Cech, J., & Soukupova, T. (2016). Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, 1-8.
- [20] Likas, A., & Vlassis, N. (2003). The global k-means clustering algorithm, *Pattern Recognit.*
- [21] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham.
- [22] Carrier, P.-L., Courville, A., Goodfellow, I. J., Mirza, M., & Bengio, Y. (2013). FER-2013 Face Database. Technical report, 1365. Université de Montréale.
- [23] Çakar, M., Yıldız, K., & Demir, Ö. (2020, October). Creating Cover Photos (Thumbnail) for Movies and TV Series with Convolutional Neural Network. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-5). IEEE.
- [24] J. Jayalekshmi and T. Mathew, "Facial expression recognition and emotion classification system for sentiment analysis," *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*, Thiruvanthapuram, 2017, pp. 1-8, doi: 10.1109/NETACT.2017.8076732.
- [25] Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.
- [26] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
- [27] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... & Ferrari, V. (2020). The open images dataset v4. *International Journal of Computer Vision*, 1-26.
- [28] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10, 1755-1758.
- [29] The King of comedy (1983, February 18). Retrieved March 15, 2021, from <https://www.imdb.com/title/tt0085794/>