Creating Cover Photos (Thumbnail) for Movies and TV Series with Convolutional Neural Network

Mahmut Çakar[1], Kazım Yıldız[2], and Önder Demir[3]
[1] Marmara University Institute of Pure and Applied Sciences, Computer Engineering, Istanbul, Turkey
[2] Marmara University Technology Faculty, Computer Engineering, Istanbul, Turkey
[3] Marmara University Technology Faculty, Computer Engineering, Istanbul, Turkey
Contact author e-mail: mahmutcakar@marun.edu.tr, kazim.yildiz@marmara.edu.tr, odemir@marmara.edu.tr

***Abstract*— The latest version and the use of video streaming platforms is increasing day by day, adding new ones. Competition is also increasing in the development of platforms for expanding film and TV series. The purpose of replicating these platforms is to achieve more, better quality on some platforms, and to keep them on one platform even better. Before that, film and TV series platforms use artificial intelligence algorithms. In this study, the aim is to create more attractive cover photos for the user by finding suitable frames from a movie or TV series and converting these frames into cover/thumbnail images on the platform. It is based on eliminating frames that are useless according to closed eyes, blurred frames or non-face images. Also, deep learning used for labeling images with objects and face's emotion and identity.**

Keywords—Deep Learning, Artificial Intelligence, Video Streaming Platforms, Convolutional Neural Network

I. INTRODUCTION

Technological advances in recent years have enabled internet content to be viewed as video content with mobile phones anytime and anywhere. Thus, there is a huge explosion in watching video content. According to YouTube [1], there are over two billion users on YouTube and one billion hours of video is consumed on YouTube every day. Therefore, more and more videos are produced day by day. For this reason, it is very important to determine the title and thumbnail (a thumbnail is a compressed preview of the original used as a placeholder) of the video for selecting them.

The number of online TV series and movie streaming platforms are also increasing today. At the end of 2018, Netflix, one of the leading platforms, has 139 million paid memberships [2], while Amazon Prime has 100 million [3] and Hulu has 28 million [4]. While these platforms emerge as competitors over time, they aim to make users spend more time on their platforms. For this purpose, it should aim to increase the number of contents on the platform and present the existing content to its users better.

The aim of this study is to produce and label pictures of a film that will fit the cover art. In this way, to create more original cover photos with labels that can attract both attention of the users. Users' interests can be their favorite celebrity, happy moments or wild animals. For this purpose, the algorithm is developed to find faces and discover their identification, emotion and other objects on scene with convolutional neural networks. In addition, the closed eyes are detected on frame to eliminate and several parameters of frame that can give information are calculated. According to these outputs, the frame is labeled to select the most suitable thumbnails for the users.

Thumbnail selection algorithms are mostly based on clustering frames as mentioned in Section II. In this study, there are two main steps: down sampling and convolutional neural network. It is aimed to eliminate blurry, repeating, non-face and closed-eyes images and finding required parameters in Section III.A. It provides eliminating useless images for next steps. In Section III.B, emotion of faces and other objects is found on frame with convolutional neural network. Results are shown in Section IV.

II. RELATED WORK

There are many cover art extraction algorithms, and some of them select the most appropriate one by clustering the frames [5-8]. There are some studies to score the aesthetics of the pictures. These studies aimed to measure the aesthetics in the paintings with parameters such as blur, lightness, colors, scene, composition and object existence [9-11].

Another related work is the AVA system of Netflix, which is also the inspiration for this work. Netflix both has its own system and uses it at the same time. AVA consists of three basic steps for evaluation. The first step, Visual Metadata, finds the visual properties like brightness, color, contrast, and motion blur. Contextual Metadata find the elements in the frame which are face detection, motion estimation, camera shot identification, and object detection. Final step, it is based on some of the core principles in photography, cinematography and visual aesthetic design in Composition Metadata. After these steps, image is ranked according to face identification, calculation of different camera angles for visual diversity and filters for maturity [12].

In another work [13], there are five main steps which are down sampling, filtering, feature extraction, ranking and enhancement. Down sampling has four operations; removing the first and last ten percent of the input video, sampling at one frame per second, eliminating similar frames, applying to sort the shots by length. Filtering step includes calculating sharpness, saturation brightness and contrast, eliminating embedded subtitles, presence of characters and setting threshold. In this study, down-sampling is applied according to presence of face and sampled every frame, compared and eliminated with similarity.

978-1-7281-9136-2/20/$31.00 © 2020 IEEE

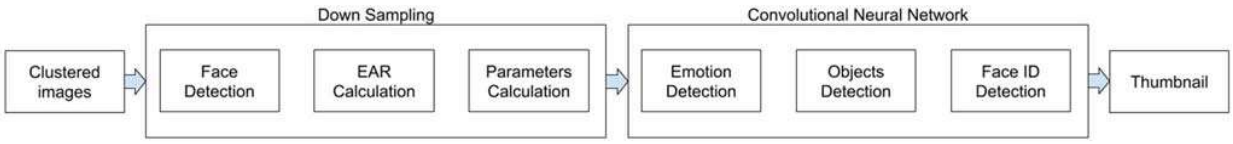

Fig. 1. Flow chart of proposed method

III. Material and Method

After the frames are taken from the video, there are two stages which are selection of appropriate frames and usage of convolutional neural network as shown in Fig. 1. In the first, it was aimed to add less pictures to the model by eliminating useless pictures and some required parameters are also calculated. In the second stage, convolutional neural network was used to find objects and faces with emotion and identity.

A. Down Sampling

First, it is very important to eliminate the unnecessary frames in a movie. There are about two hundred thousand frames in a two-hour film. Inserting all frames into artificial neural networks is much more costly so the frames with faces were only selected using Haar Cascade. Although Histograms of Oriented Gradients (HOG) were used for eye opening detection, Haar Cascade was used first for a more sensitive and faster face detection. Face detection time with Haar Cascade method is between 16 ms and 50 ms but HOG is between 340 ms and 410 ms. Therefore, non-face pictures are eliminated. In addition, the variance of the Laplacian Filter [14] was used to eliminate blurred frames. However, this filter may not be suitable for every movie or frame since the threshold value used may be different for each frame, either a different threshold value for each film should be calculated or used for sieving afterwards. However, eliminating them later can cost time because of applying following steps.

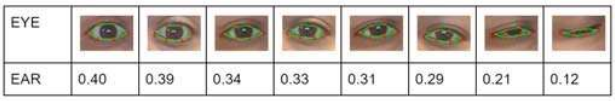

Fig. 2. Eye Aspect Ratio for several eye

In frame with faces, face detection is performed again with HOG and eye opening is determined for the faces found. HOG is used because it allows you to identify 68 different points with 6 points for each eye found. Thus, the Eye Aspect Ratio (EAR) can be calculated with 6 points corresponding to one eye. As it is shown in Fig. 2. [15], 6 points shown with red dots and they are p_1 (most-left), p_2 (top-left), p_3 (top-right), p_4 (most-right), p_5 (bottom-right), and p_6 (bottom-left). Eye Aspect Ratio is calculated (1) and it is assumed closed eyes according to threshold value.

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

There are many more parameters to be calculated that can give us information in the frame. These are brightness, dominant color and blur. Blur parameter was calculated by variance of Laplacian filter before. This parameter is used to find the sharpest in similar frames and to eliminate blur images. For brightness, the RGB (Red, Green, Blue) values in the picture are converted to HSV (Hue, Saturation, Value / Brightness) format and calculated by averaging the value of each frame. In order to calculate the dominant color, the K-means has calculated by grouping the colors by Clustering algorithm and selecting the most dominant value [16]. The purpose of calculating this value is to create hints from the frame and to prevent similar colors from overlapping with the film's logo by calculating the dominant color of the logo to be placed.

B. Convolutional Neural Network

The aim of this section is determining the emotions of the faces in the frame by using convolutional neural networks and to identify and associate the objects in the frame. Convolutional Neural Network, a special type of Neural Network, is made up of neurons with learnable weights and biases and used effectively for image recognition, classification and object detection. Convolutional Neural networks needs a large set of N labeled images $\{x, y\}$ which refer input as x and discrete variable indicating the true class as y. A loss function is used to compare the output of model and the true class y. Filters in the convolutional layers, weight matrices in the fully-connected layers and biases, the parameters of the network, are trained by back propagating the derivative of the loss with respect to the parameters throughout the network, and updating the parameters via stochastic gradient descent [17].

Fer2013 dataset was used to determine facial emotions [18] and the generated model can detect emotions with 91% accuracy. Fer2013 dataset has labeled seven emotion: neutral, sadness, surprise, happiness, fear, anger, and disgust.

Detecting facial emotions is very important about the frame in the film and then grouping. It is also important for the type of frame to be able to detect other objects. For example, the presence of a pet is friendship, a weapon or a sword is an action, a musical instrument is musical, a ball can determine that it is related to sports. They are more important to the user's interest than to provide information about the entire film. The person who is interested in music may be more likely to choose that movie when there is a frame with a guitar.

Emotion detection model has 5 main layers as shown in Fig. 3. In first layer, it is applied 3x3 convolution with 64 filters and rectified linear unit (ReLU) activation function to bound output values on 48x48 input image with three input channels. After that, it is performed batch normalization to stabilize learning, max-pooling to summarize and simplify the network and dropout which is dropping out hidden and visible units to reduce

overfitting [19] and output shape becomes 24*x*24*x*64. These steps applied multiple times with increasing convolution filters until output shape becomes 2*x*2*x*512 and flatten is applied. It is decreased with Dropout and Dense until its output shape is 7 since there are 7 type of emotion.

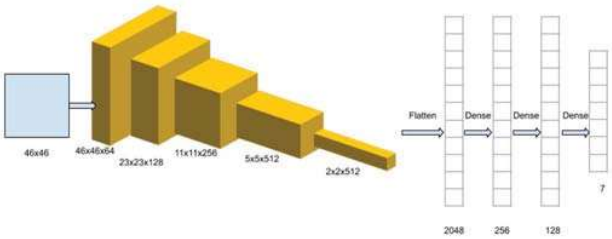

Fig. 3. Emotion Detection Network

YOLOv3 (You Only Look Once) algorithm [20] and Google OpenImages dataset [21] were used to detect the objects. Open Images dataset contains approximately nine million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. It also has a total of 16M bounding boxes for 600 object classes on 1.9M images. YOLO algorithm based on prediction grids and anchors. In this work, 13x13 grids and 5 anchors are predicted and eliminated duplicates with non-maximal suppression approach. Although YOLOv3 algorithm produces both fast and accurate results as it shown in Table 1, it has caused the OpenImages dataset to be lower than expected. Therefore, 600 classes were firstly organized and then the classes were grouped, and the classes were reduced.

TABLE I. COMPARISON MAP (MEAN-AVERAGE PRECISION) AND TIME OF OBJECT DETECTION METHODS ACCORDING TO [22]

Method	mAP-50	time(ms)
SSD321	45.4	61
DSSD321	46.1	85
R-FCN	51.9	85
SSD513	50.4	125
DSSD513	53.3	156
FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

It is also important to determine who belongs to the faces in the square found. Thus, the user is more likely to choose movies with his/her favorite actor or actress films. For this purpose, it is used dlib's pre-trained face detector [23].

Finally, 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.

IV. RESULTS

Emotion detection model in this study is compared with VGG16 and VGG19 [24] networks. Fig. 4 shows train loss and accuracy. Accuracy of network is calculated about 0.91 in training, 0.67 in validation. Accuracy of VGG16 and VGG19 networks are about 0.65 and 0.50.

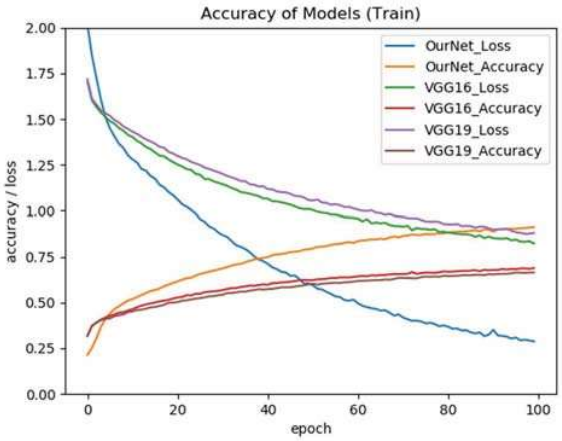

Fig. 4. Train Loss and Accuracy Values of Networks

As shown in Fig. 5 and Fig. 6, the loss graph of 17 classes model more accurate than 19 grouped classes model. 17 classes model's mAP (mean Average Precision) is 34.28 % and other one is 30.56 %. Since 19 grouped classes the model did not reach the expected values, training was discontinued.

Frames in the results obtained can be finally reviewed under human control and avoiding meaningless pictures or incorrectly grouped or tagged pictures. As a result, the 200,000 frames in a movie may have been reduced to 100 frames. In addition, the selection of similar pictures but similar labels can be used to give the user different times.

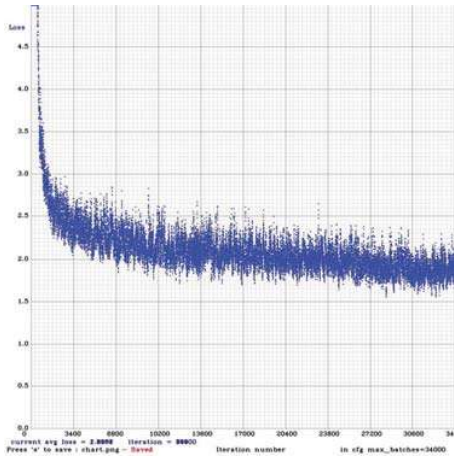

Fig. 5. 17 classes YOLO model Loss Graph

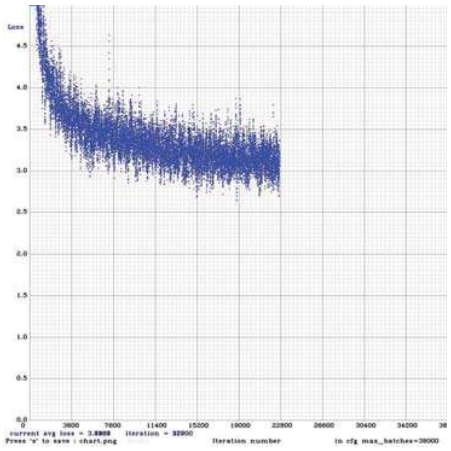

Fig. 6. 19 grouped classes YOLO model Loss Graph

Fig. 7. Sample Image Output Results of this study from Joker Movie [25]

As it is shown in Fig. 8, this study suggests a frame with face and the logo is placed dynamically according to faces. Results of this frame is in Fig. 9. First section is about faces and its location, emotion and identity. Parameters section includes brightness, blur estimation, and dominant color. Final section is yolo model objects and their confidences.

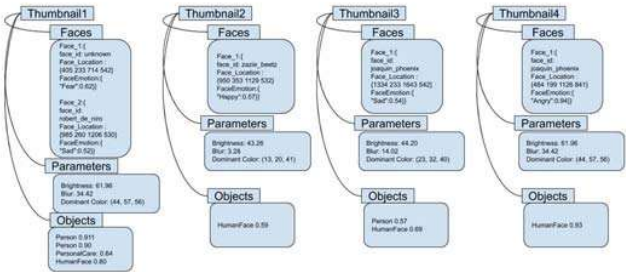

Fig. 8. Sample Parameters Output Results of this study according to Fig. 7

V. Conclusion

It was aimed to creating cover photographs of a movie and a TV series with image processing and convolutional neural networks and to achieve dynamic results for the user on movie streaming platforms. Also, it was aimed to select the features that may interest the user by labeling the created frames.

Also, many features can be added in the future. These may include additional information about faces, determination of action on the scene (walk, swim, fire, etc.), determination of the photographic composition of the frame. In addition, the creation of a new dataset for the development of the models can be added. The fact that the new dataset is more suitable for the study will provide more accurate results.

References

[1] "Youtube for Press.", Youtube. Retrieved February 22, 2020 from www.youtube.com/about/press/

[2] U.S. Securities and Exchange Commission. (2019). "Annual Report Pursuant To Section 13 or 15(D) of The Securities Exchange Act of 1934", Commission File Number: 001-35727, Retrieved from https://s22.q4cdn.com/959853165/files/doc_financials/annual_reports/2018/Form-10K_Q418_Filed.pdf, pp. 1

[3] Bezos, J. P. (2018). Retrieved February 22, 2020 from https://www.sec.gov/Archives/edgar/data/1018724/000119312518121161/d456916dex991.htm

[4] Hulu, "Hulu Tops 25 Million Total Subscribers in 2018". Retrieved February 22, 2020 from https://www.hulu.com/press/hulu-tops-25-million-total-subscribers-in-2018/

[5] Zeng, X., Hu, W., Li, W., Zhang, X., & Xu, B. (2008). Key-frame extraction using dominant-set clustering. Proceedings - IEEE International Conference on Multimedia and Expo, 1285–1288.

[6] Avila, Y. S. E. F., Lopes, A. P. B., Luz, A., Jr., & Araújo, A. A. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters, 32(1), 56–68.

[7] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," Proc. IEEE Int. Conf. Image Processing, vol. 1, pp. 866-870, 1998.

[8] C. Sujatha and U. Mudenagudi, "A study on keyframe extraction methods for video summary," Proc. Int. Conf. Computational Intelligence Communication Networks, 2011,pp. 73-77.

[9] Deng, Y., Loy, C. C., & Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4), 80–106. https://doi.org/10.1109/MSP.2017.2696576.

[10] S. Ma, J. Liu, and C. W. Chen, "A-Lamp: Adaptive layout aware multi-patch deep convolutional neural Network for Photo Aesthetic Assessment," Proc. IEEE Conf. Computer Vision Pattern Recognition, 2017, pp. 722–731.

[11] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," Proc. European Conf. Computer Vision, 2006, pp. 288-301.

[12] "Ava: The Art and Science Of Image Discovery At Netflix" Netflix Technology Blog - https://medium.com/netflix-techblog/ava-the-art-and-science-of-image-discovery-at-netflix-a442f163af6

[13] C.-N. Tsao, J.-K. Lou and H. Chen, "Thumbnail Image Selection For VOD Services", 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019 https://doi.org/10.1109/MIPR.2019.00018

[14] Pech Pacheco, Jose Luis & Cristobal, Gabriel & Chamorro-Martinez, J. & Fernandez-Valdivia, J. (2000). "Diatom autofocusing in brightfield microscopy: A comparative study. Pattern Recognition, Proceedings". 15th International Conference on. 3. 314-317 vol.3. 10.1109/ICPR.2000.903548

[15] T. Soukupova and J. Cech, "Real-time eye blink detection using facial landmarks", in 21st Computer Vision Winter Workshop (CVWW2016), 2016, pp. 1–8.

[16] Likas, A., Vlassis, N., & Verbeek, J. (2003). The global K-means clustering algorithm. *Pattern Recognition, 36(2)*, 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2.

[17] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional neural networks". In ECCV, 2014.

[18] Carrier, P.-L., Courville, A., Goodfellow, I. J., Mirza, M., & Bengio, Y. (2013). *FER-2013 Face Database. Technical report*, 1365. Universit'e de Montr'eal..

[19] Gulli, A., & Pal, S. (2017). *Deep Learning with Keras*. Packt Publishing Ltd.

[20] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You only look once: Unified, real-time object detection*. CVPR.

[21] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale". arXiv:1811.00982, 2018

[22] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[23] Davis, E. (2009). King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10, 1755–1758.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, 2014.

[25] "Joker.", Warner Bros., http://www.jokermovie.net/