IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 May 2024, accepted 2 June 2024, date of publication 7 June 2024, date of current version 3 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3411101

RESEARCH ARTICLE

Deep Metric Learning for Near-Duplicate Video Retrieval Leveraging Efficient Semantic Feature Extraction

ANIQA DILAWARI[1], **SAJID IQBAL**[2], **FARIAL SYED**[3], **AND QAZI MUDASSAR ILYAS**[2]

[1]Department of Computer Science and Information Technology, University of Home Economics, Lahore 54000, Pakistan
[2]Department of Information Systems, College of Computer Science and Information Technology, King Faisal University, Saudi Arabia
[3]University of Regina, Regina, SK S4S 0A2, Canada

Corresponding author: Sajid Iqbal (siqbal@kfu.edu.sa)

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under Grant A125.

ABSTRACT Video sharing platforms like YouTube, TikTok and Instagram have gained popularity in the online space. Daily several videos are uploaded, which calls for an efficient video retrieval system that could identify near-duplicate videos that offers several advantages in content management, copyright protection, and multimedia retrieval. This will facilitate efficient content management by removal of redundant videos from large repositories to streamline storage resources and improve accessibility of multimedia collections. Additionally, this can help copyright protection and intellectual property allowing right holders to identify unauthorized copies of their original work. Moreover, in applications such as multimedia retrieval and recommendation systems, removal of near-duplicate videos can enhance user experience by providing relevant search results. AI provides a promising solution to this problem. We have proposed an effective system built on deep metric learning that solves the near duplicate video retrieval. This proposed model uses the pre-trained VGG-16 network that contains convolutional and fully connected layers to find video features. These video representations are fed to the deep metric learning framework in the form of triplets which are trained to calculate the accurate distance between similar or near-duplicate videos. For the training of the framework, VCDB dataset was used whereas for the evaluation of the model CC_WEB_VIDEO and TRECVID BBC Rushes 2007 datasets were used. Experiments have shown that mean average precision of 0.985% for the CC_WEB_VIDEO dataset is achieved thus outperforming the state-of-the-art models.

INDEX TERMS Deep metric learning, distance calculations, near copy video retrieval, similarity search.

I. INTRODUCTION

With the advancement in technology, many new gadgets such as mobile phones, video recorders and DSLR cameras have been introduced. Due to these devices, multimedia data has been growing ever since. New videos are generated from thousands to millions daily. Among these videos there are many that are near duplicates or copies with slight modifications or different formats. Therefore, the traditional ways of retrieving video data with the help of text queries are not enough to fulfill the requirements of finding the related videos. Furthermore, the method of finding the relevant videos with the help of image/video queries is also not very scalable. Copy detection and content-based video retrieval using video queries is considered the most realistic answer in these modern days. In the process of copy detection, the basic idea is to find the similarity between the original and copied video.

Near-duplicate video retrieval is a process of finding videos in a large database that are related or identical to an input video in question. This technology is beneficial in many applications, such as plagiarism detection [1], copyright infringement detection [2], [3], [4], video summarization [5], efficient content management, multimedia retrieval and recommendation systems [6]. Traditional methods [7], [8] for near-duplicate video retrieval rely on finding low-level

The associate editor coordinating the review of this manuscript and approving it for publication was Yilun Shang.

features like color histograms, texture descriptors, or keyframes. However, these techniques are often resource intensive and do not capture video semantic information such as people, objects, images, scenery etc.

Artificial Intelligence (AI) techniques have revolutionized various domains, including computer vision and information retrieval. Deep Learning (DL), a subset of AI, has shown great potential in learning complex patterns and representations from data. Convolutional Neural Networks (CNNs) are frequently used in visual tasks and have achieved remarkable results in computer vision, video classification and object recognition. To solve the problems of video retrieval and copy detection, CNN features have been used [9], [10]. In near-duplicate video retrieval using AI, CNNs are employed to extract features from videos. These features consider both low-level visual information and high-level semantic representations. Training these networks requires a huge amount of labeled video data. However, annotating videos manually is a laborious and expensive process. To overcome this limitation, transfer learning can be applied, where pre-trained CNNs on large-scale image datasets, like ImageNet, are treated as a starting point. These pre-trained CNNs learn general visual representations that can be fine-tuned on the target near-duplicate retrieval task.

In this paper, we have proposed a content-based video retrieval system that detects the near duplicate videos. This system uses the concepts of deep learning in feature extraction and video retrieval. In the first step, feature descriptors of layer level are extracted by applying the max pooling to the activations of every convolution layer for the extraction of video features using VGG-16. This scheme is named Maximum Activation of Convolutions [11], [12]. In the second step, framework of Deep Metric Learning (DML) is used for the near duplicates. The basis of DML is the triplet wise scheme and studies have shown that this technique is very effective [13], [14], [31].

Remaining sections of the paper are as follows. The related work in the field of copy detection and video retrieval is discussed in Section II. Methodology and the complete architecture of the proposed system is detailed in Section III. Discussion on the datasets has been done in Section IV. Discussion of the experiments, obtained results, findings are presented and analyzed in Section V. Moving on to Section VI, the conclusion drawn from the experiments and future directions for further research are discussed.

II. RELATED WORK

Near Duplicate Video Retrieval (NDVR) has emerged as a critical area of research due to the exponential growth of video content on the internet. To facilitate the browsing and searching of large images and videos collection, content-based video retrieval systems have been established. Traditional methods for NDVR often depended on handcrafted features and similarity metrics. While these methods provide reasonable results, they were limited in handling complex visual patterns. Early work, such as the bag-of-visual-words model [15] and local feature-based methods [16], laid the basis for retrieval of content-based videos. Low level features are extracted from the frames of videos and analysis is done using multiple texture features, color histograms and other methods. A real-scenario copy detection system [17] for videos detects the near duplicates or partial copies of the complex query from the database of real videos. Since the algorithms of copy detection are not accurate and time efficient, [18] studied two algorithms implemented using Hadoop framework for copy detection. Cluster based similarity search was also studied that is a comparatively fast searching algorithm. These algorithms tend to cope with real-time video detection and retrieval requirements. A novel technique for the implementation of video retrieval system by detecting similar temporal patterns [19] in various videos. Two effective and efficient sequence matching and indexing techniques are unified thus increasing the accuracy and reducing the computational cost.

A method based on Information theory to analyze content-based videos is proposed in [20]. This system is categorized into three parts: detection of shot boundary, hierarchical video summarization and retrieval/indexing of the target video. Evaluation of the system performance and the computation of the results using TRECVID 2006 dataset. An unsupervised content-based retrieval system [21] uses the combined representations of spatio-temporal features. Those videos are retrieved having the same trajectories and moving objects. Quantitative and Qualitative analysis of the two benchmark datasets of UCF50 and MCVS have been done and evaluated.

The advent of deep learning brought about a paradigm shift in NDVR. CNNs have become pivotal for extracting hierarchical features from videos. Two-Stream Convolutional Networks [22] introduced a spatial stream for fine-grained information and a temporal stream for motion information, enhancing the discriminative power of features. Similar networks like Siamese Convolutional Neural Network (SCNN) [23] aimed to handle video information for extracting features from video frames. These networks learn to reduce the distance between alike videos and increase the distance between unrelated ones. CNNs for feature extraction were carried out on a single frame, which is a significant drawback in processing of videos. A video includes the spatial and temporal features which when ignored affect the precision of the retrieval process. This can be resolved by using 3D convolutions [24] which can extract spatial as well as temporal features leading to powerful video embeddings. Attention mechanisms have been introduced to focus on significant parts of videos. The integration of attention mechanisms [25] in video retrieval tasks allow models to dynamically weigh the significance of different parts of a video, hence, enhancing the discriminative power. Generative models, particularly Variational Autoencoders (VAEs) and Video Generative Adversarial Networks [26], have been explored to learn compact and semantically rich video representations. These models generate embeddings that capture the essence of a video, aiding in near-duplicate retrieval.

From the literature survey it was found that many methods have been used to extract the features of videos but there exists a semantic gap in the video retrieval system. It means to convert the query of video or image from the human to low level features. By extracting the intermediate features from the various convolutional layers, the work in retrieval systems for near duplicate videos has been given a new direction. Metric learning plays a vital role in NDVR by defining a suitable similarity metric. This can be demonstrated with triplet and contrastive loss in learning discriminative video embeddings.

III. METHODOLOGY

The methodology for the retrieval of near duplicate videos is defined in two steps. In the first step, the compressed global video representations are developed by extracting the characteristics from the convolutional layers of deep architectures of CNN. In the second step, to compute an embedding function, a Deep Neural Network (DNN) is trained. The DNN helps to figure out the likeliness between two videos by evaluating the distance between them. The proposed architecture for the retrieval of near duplicates is inspired from [31]. Batches of triplets generated from the videos of the VCDB dataset [27] are used to train the model.

A. FEATURE EXTRACTION

Research studies [28], [29] have shown that CNN pre-trained architectures can be used to pull out the features from intermediate convolution layers. The image is propagated forward over the entire framework of CNN. On every convolution layer, an aggregation function such as max pooling is applied to extract the features from each layer. Maximum Activation of Convolutions (MAC) is the name given to this method [12].

The proposed architecture's uniform sampling is implemented to select a single frame per second for each video. Global video descriptors are generated by using the pre-trained VGG-16 [30] model, which consists of total 5 convolutional layers symbolized by conv1, conv2, conv3…conv5. Input of the VGG-16 model is an image having 224 × 224 dimensions. Resizing and zero padding is a part of pre-processing to make the dimensions of input frames equal to 224 × 224. After the forward propagation of the frame over the entire network, a total of 5 feature maps are generated that can be symbolized as $F^n \in R^{a_d^n * a_d^n * e^n}$ $(n = 1, 2, 3, 4, 5)$ where $a_d^n * a_d^n$ denotes the dimension of each convolutional layer's channel and e^n denotes the total number of channels. An aggregation function, specifically max pooling, is employed on each channel of the feature map F^n to extract a singular value.

In this way, a single descriptor is extracted from each layer. Equation 1 shows how extraction process can be formulated.

$$V^n(x) = \max F^n(., ., x) \text{ where } i = 1, 2, 3, \ldots e^n \quad (1)$$

Here $V^n(x)$ is a layer vector having dimensions e^n derived by applying max pooling to each feature map channel F^n. Once the extraction is done, all the layer vectors are concatenated to make a single descriptor. At the end, zero mean and $l2$-normalization is used to normalize the global frame descriptors. This process of feature extraction is not end-to-end as theVGG-16 model pretrained weights are not updated. Figure 1 shows the procedure of how normalized global frame descriptors are extracted.

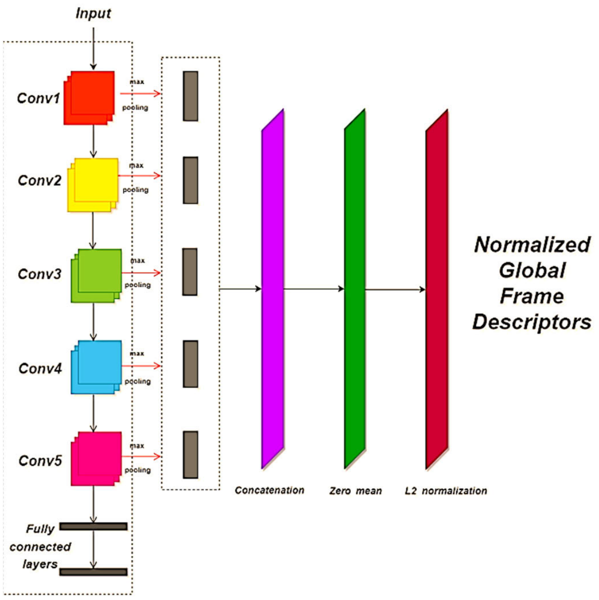

FIGURE 1. Extraction of normalized global frame descriptors.

The normalized global frame descriptors were extracted using VGG-16. Each video frame underwent a series of steps within the architecture. The frames were preprocessed which involved resizing the frame size to 224 × 224 pixels. Then, each frame is fed to the VGG-16 network where there are multiple convolutional and pooling layers followed by fully connected layers. The frame when passed through the network undergoes series of transformations with each layer extracting abstract and discriminative features. The convolutional layers take out local patterns and spatial information whereas the pooling layers combine and down sample the feature maps to reduce dimensions. Lastly, the fully connected layers combine high-level features into a compact representation that captures the global content of the frame. The L2 normalization ensures that the feature vectors produced from the network have a consistent scale across multiple samples for accurate and efficient analysis of videos. L2 normalization is also known as Euclidean norm where the length of the feature vector is same regardless of the original magnitude of features.

B. DEEP METRIC LEARNING

In this approach, we deal with the learning of similarity between two videos with the help of information available from the relations of triplet wise videos. When an input video and a repository of various videos dataset are given, the main objective is to find the likeliness between the query video and each video of the dataset. Then these videos can be sorted in descending order depending on the similarity content. This will help to retrieve the near duplicates from the top ranks. To achieve this, similarity between the two videos A and B, a square Euclidean distance is defined as the space of video embeddings. Equation 2 depicts this formula to compute the Euclidean distance.

$$D_{Euclidean}\left(F\left(A\right), F\left(B\right)\right) = \sqrt{\sum\nolimits_{i} \left(F(A)^{i} - F(B)^{i}\right)^{2}} \tag{2}$$

where F defines the embedding function for mapping each video in the Euclidean space and $D_{Euclidean}$ defines the square Euclidean distance in that space. Moreover, to detect the pair of near duplicate videos, an indicator function $I_{pair}()$ is defined in equation 3.

$$I_{pair}(A, B) = \begin{cases} 1, & if\ A\ and B\ are\ near\ duplicates \\ 0, & otherwise \end{cases} \tag{3}$$

The goal of the training is to assign a smaller distance value to near-duplicates and larger distance values to dissimilar videos. The embedding function F maps the representations of videos to common Euclidean space R^{dim} where dim is the feature embedding's dimension, when the feature vector of a video V, near duplicate video $V+$ and dissimilar video V- are given. The space between the video given as a query and the near

$$\begin{aligned} &D_{Euclidean}(F(V), F(V+)) \\ &\quad < D_{Euclidean}(F(V), F(V-)), \forall \text{V}, \text{V} + \text{and} V - such\ that \\ &\qquad I_{pair}(V, V+) = 1\ and\ I_{pair}(V, V-) = 0 \end{aligned} \tag{4}$$

duplicate is always lesser than the space between the video and the dissimilar video as shown in equation 4.

C. TRIPLET LOSS

During training of the DML architecture [31], triplets $T = \left\{\left(V^{i}, V+^{i}, V-^{i}\right), i = 1, 2, 3 \ldots N\right\}$ containing N instances are created where V is the query video feature vector, $V+$ is the near duplicate video feature vector and $V-$ is the dissimilar video feature vector. The relative similarity between three videos is expressed by a Triplet i.e. V^{i} is more like $V+^{i}$ than $V-^{i}$. For a triplet a loss function is defined named as "Triplet loss" explained in equation 5.

$$\begin{aligned} &Loss\left(V^{i}, V+^{i}, V-^{i}\right) \\ &\quad = \max\{0, D_{Euclidean}(F(V), F(V+)) \\ &\qquad - D_{Euclidean}(F(V), F(V-)) + \Upsilon\} \end{aligned} \tag{5}$$

where Υ defines a parameter for margin to ensure an adequately big variation between similar and dissimilar query distance. The triplets are not penalized if the calculated distance of the videos lies within Υ. The other way, the loss calculated is convex estimation of the loss which calculates the contravention in required distance among video pairs defined by triplets. The loss function is improved by using batch gradient descent as mentioned in equation 6.

$$min_{\theta} \sum\nolimits_{i=1}^{n} Loss_{\theta}\left(V^{i}, V+^{i}, V-^{i}\right) + \Omega\|\theta\|_{2}^{2} \tag{6}$$

where Ω is the parameter for regularization that helps the model from overfitting and n defines the size of mini-batch triplet. During training, the distance between the input video and similar video is decreased by minimizing the loss and the distance between the query video and dissimilar video is increased thus leading to the satisfactory ranking order. Eventually the model will discover an effective video representation thus improving the near duplicate video retrieval solution with the help of triplet generation policy.

D. DML ARCHITECTURE

The network based on triplets has been proposed for the training of the DML model and shown in Figure 2. Triplets T are provided as an input to the network. Three deep neural networks DNNs having similar parameters and architecture are fed separately with the query video feature vector V, similar video V+ and a dissimilar video V-. The embeddings for videos are computed within the DNNs. Three fully connected layers and one normalized layer makes the architecture of all three DNNs. The dimension of the final output vector and the size of every layer depends on the feature vectors of input. Accumulated loss is calculated by sending the computed embeddings from the batch of triplets to triplet loss layer.

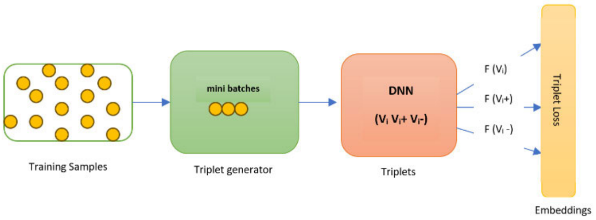

FIGURE 2. DML architecture.

IV. DATASETS

The experiments on the proposed model were executed on two datasets: CC_WEB_VIDEO [32] and TRECVID BBC Rushes 2007 [33]. VCDB dataset [27] was used for the training of the model.

A. CC_WEB_VIDEO

The CC_WEB_VIDEO dataset is mostly used in NDVR that contains 24 queries designed to retrieve data from top favorite and most viewed websites of Yahoo! Video, Google Video and YouTube. The dataset was collected by the combined efforts of group called VIREO from the City University of Hong Kong and Carnegie Mellon University (informedia group). The dataset was collected in the year 2006. The dataset comprises of a sum of 12,790 videos with a duration under 10 minutes. The keyframes of 398,015 are there in the dataset.

B. VCDB DATASET

VCDB dataset was used to train the DML architecture. With over 100,000 videos, this is a vast video copy database (VCDB). 9,000 copied segment pairs are manually annotated. This data set serves as a benchmark in the recent research on copy detection, showcasing the latest advancements in achieving state-of-the-art results. Figure 3 shows the 28 videos and their near duplicates taken from the VCDB dataset.

FIGURE 3. Near duplicate videos from VCDB dataset.

C. TRECVID BBC RUSHES 2007

TRECVID have provided the data of about 100 hours of 5 BBC drama programs for the task of video summarization. This data consists of raw video footages of a detective program, an emergency services program, a historical drama of early 1900's of London, an ancient Greece series, a police drama, and various scenes from different programs. This dataset contains both the indoor and outdoor scenes of everyday situations. Total data is divided into two sections I). Development Videos (50 videos) 2). Testing Videos (42 videos). Development Videos were used for the testing of our trained DML model. These 50 videos were divided into chunks of 1-minute videos forming a total of 961 videos. Out of these 961 videos, 6 videos were taken randomly as query videos and the rest of the database was searched to find the near duplicates.

V. EXPERIMENTAL RESULTS

Experiments have been conducted using the Tensorflow framework using pre-trained model of VGG-16 on imageNet dataset for the extraction of features. Adam optimizer is used with 10-5 learning rate. The size of mini batches is kept to 1000 triplets. Time for the generation of triplets is t = 0.8 and during that time it generates 2000 pairs of near duplicate videos and total of 5M resulting triplets. Parameter for margin Υ is set to 1 and regularization Ω is set to 10-5. All the training and experiments were performed on the system having specifications of Intel (R) core (TM) i7-7500U CPU @ 2.70 GHz x 4 Processor, Nvidia Geforce GTX 950M GPU, 15.1GiB RAM, 64-bit Ubuntu 16.04 and LTS OS.

A. EVALUATION METRICS

The effectiveness of the near duplicate video retrieval system is estimated using the interpolated precision-recall (PR) curve, that measures the accuracy of the system. The precision of a search algorithm can be described as the ratio of related videos retrieved to the total videos retrieved. In contrast, recall represents the ratio of related videos retrieved to the total relevant videos available. Mean average precision (mAP) is another system of measurement that is used to evaluate the performance of a video retrieval system. It calculates the average precision for a group of relevant videos that are related to a given input video. The value of N represents the total relevant videos and R^x is the rank of x^{th} relevant retrieved video.

$$Average\ Precision = \frac{1}{N}\sum_{x=0}^{N}\frac{x}{R^x}$$

B. COMPARISON BETWEEN NDVR STATE-OF-THE-ART

The comparison of our proposed approach with the three state-of-the-art NDVR approaches from the literature includes Color Correlogram (CC) [34], Pattern-based approach (PA) [35], Layer-wise Convolutional Network Networks (L-CNN) [36] and Deep Metric Learning using AlexNet (DML-A) [31].

CC approach proposes a new image characteristic called the color correlogram for image indexing and comparison. It is shown to be more effective than traditional color histogram methods for content-based image retrieval. This correlogram is also suggested as a generic indexing tool for various image retrieval and video browsing applications.

A spatiotemporal pattern-based approach is employed by PA to retrieve and locate near-duplicate videos on a large scale effectively and efficiently. The pattern indexing tree is built using encodings of keyframes, facilitating the retrieval of potential video matches. The localization of near-duplicate segments is achieved through the utilization of the M-pattern-based dynamic programming (mPDP) algorithm.

L-CNN introduced an approach for near-duplicate video retrieval, utilizing intermediate CNN layers and employing a layer-based aggregation scheme to construct video representations. The evaluation involved assessing the effectiveness of this feature aggregation scheme using three different CNN architectures: VGGNet, AlexNet, and GoogLeNet.

DML-A approach proposed a video-level near-duplicate video system using deep metric learning with CNN features from AlexNet and a triplet generation scheme. Triplet-based network architecture for deep metric learning was devised along with hybrid-level and frame-level matching for near-duplicate video retrieval.

Our proposed utilized early fusion employed for generating video descriptors, wherein all the frame descriptors' extracted features were combined into a single vector, followed by the application of the learned embedding function. Two experiments were conducted: the first utilized max pooling on all convolutional layers, concatenating the resulting vectors, and the second extended max pooling to both convolutional layers and first fully connected layer. Table 1 displays the mAP scores of various state-of-the-art NDVR approaches evaluated on CC_WEB_VIDEO and TRECVID BBC rushes 2007 datasets. Our proposed approach demonstrated better performance as compared to previous approaches.

TABLE 1. mAP comparison between NDVR approaches.

Method	mAP
Color Correlogram (CC)	0.944
Pattern-based approach (PA)	0.958
Layer-wise Convolutional Network Networks (L-CNN)	0.974
Deep Metric Learning using AlexNet (DML-A)	0.981
Proposed Model (ours)	0.985

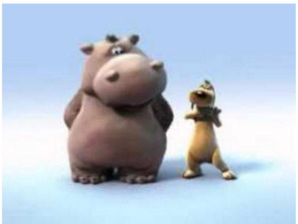

FIGURE 4. CC_WEB_VIDEO dataset query video.

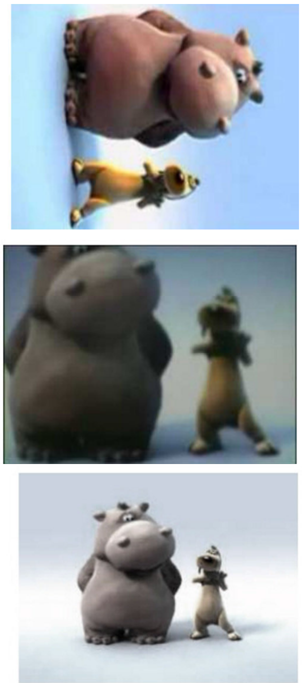

FIGURE 5. Near duplicate videos against the query video.

In Figure 4, a query video passed to our proposed system to check near duplicate videos is taken from CC_Web_Video dataset.

The resultant near duplicate videos from the same dataset are shown in Figure 5, extracted by our proposed model.

Our model leveraged efficient deep learning methods to capture complex patterns and temporal dynamics in a video. Convolutional Neural Networks (CNNs), that is VGG-16 was used to extract high-level features from video frames. These frames are processed through multiple convolutional and pooling layers, which capture spatial hierarchies and significant visual patterns. Max pooling was applied to these features to create a cohesive video representation. Additionally, temporal pooling methods was used to integrate frame-level features across the video sequence, preserving the temporal context. Once the features are extracted, the deep metric learning model, was employed to learn an embedding space where near-duplicate videos are mapped closely together. This approach combined frame-level feature extraction with robust similarity learning, enabling accurate identification of near-duplicate videos while handling variations in video content and quality efficiently.

VI. CONCLUSION & FUTURE WORK

This paper introduced video representations for a near-duplicate video retrieval system, leveraging global features extracted from all convolutional layers and fully connected layers using VGG-16. Additionally, the framework incorporates deep metric learning, generating compact video representations in the form of video triplets. The video triplets are trained to map videos into a high-dimensional embedding space where similar videos are close together and dissimilar videos are far apart. This approach enables accurate and efficient similarity measurement between videos. This network minimized the distance between embeddings of near-duplicate videos. The proposed architecture, implemented using the VGG-16 model, underwent testing on the CC_WEB_VIDEO and TRECVID BBC Rushes 2007 datasets. The evaluation results showcased the highly competitive performance of the proposed model. Comparative analysis with state-of-the-art methods, based on metrics such as mean average precision (mAP) highlighted the effectiveness of the presented approach.

Discussing the future directions, improvements can be made to the model by training this architecture as an end-to-end system i.e. training the VGG-16 architecture instead of using the pre-trained weights of the model. There exist challenges in NDVR, including scalability to large datasets, interpretability of learned representations, and handling diverse video modalities. Furthermore, the integration of multi-modal information, including audio, text, and metadata, can further enhance the robustness and effectiveness of near-duplicate video detection systems. Future research could explore explainable AI techniques, address dataset biases, and enhance the robustness of NDVR models in dynamic and uncontrolled environments.

REFERENCES

[1] E. Thirani, J. Jain, and V. Narawade, "Enhancing performance evaluation for video plagiarism detection using local feature through SVM and KNN algorithm," *Int. J. Image, Graph. Signal Process.*, vol. 13, no. 5, pp. 41–50, Oct. 2021.

[2] D. Y. Zhang, Q. Li, H. Tong, J. Badilla, Y. Zhang, and D. Wang, "Crowdsourcing-based copyright infringement detection in live video streams," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 367–374.

[3] D. Y. Zhang, L. Song, Q. Li, Y. Zhang, and D. Wang, "StreamGuard: A Bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 901–910.

[4] S. Agrawal and A. Sureka, "Copyright infringement detection of music videos on YouTube by mining video and uploader meta-data," in *Proc. 2nd Int. Conf. Big Data Anal.*, vol. 2, Mysore, India. Springer, Dec. 2013, pp. 48–67.

[5] Y. Li and B. Merialdo, "Multi-video summarization based on video-MMR," in *Proc. 11th Int. Workshop Image Anal. for Multimedia Interact. Services (WIAMIS)*, Apr. 2010, pp. 1–4.

[6] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Recommender systems leveraging multimedia content," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–38, Sep. 2021.

[7] K. K. Thyagharajan and G. Kalaiarasi, "A review on near-duplicate detection of images using computer vision techniques," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 897–916, May 2021.

[8] J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, "Near-duplicate video retrieval: Current research and future trends," *ACM Comput. Surv. (CSUR)*, vol. 45, no. 4, pp. 1–23, 2013.

[9] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.

[10] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1798–1807.

[11] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–20.

[12] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, 2016.

[13] W. Zheng, B. Zhang, J. Lu, and J. Zhou, "Deep relational metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12045–12054.

[14] B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi, *Elements of Dimensionality Reduction and Manifold Learning*. Cham, Switzerland: Springer, 2023.

[15] L. Wang, D. Song, and E. Elyan, "Improving bag-of-visual-words model with spatial–temporal correlation for video retrieval," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2012, pp. 1303–1312.

[16] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *Proc. 19th ACM Int. Conf. Multimedia*, Nov. 2011, pp. 423–432.

[17] Y. Zhang and X. Zhang, "Effective real-scenario video copy detection," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3951–3956.

[18] P. Prajapati, M. Mishra, S. Patil, and M. Pawar., "Real-time video copy detection in big data," Tech. Rep., 2017.

[19] J.-H. Su, Y.-T. Huang, H.-H. Yeh, and V. S. Tseng, "Effective content-based video retrieval using pattern-indexing and matching techniques," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5068–5085, Jul. 2010.

[20] H. Yarmohammadi, M. Rahmati, and S. Khadivi, "Content based video retrieval using information theory," in *Proc. 8th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Sep. 2013, pp. 214–218.

[21] C. Chattopadhyay and S. Das, "STAR: A content based video retrieval system for moving camera video shots," in *Proc. 4th Nat. Conf. Comput. Vis., Pattern Recognit., Image Process. Graph. (NCVPRIPG)*, Dec. 2013, pp. 1–4.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[23] Y.-G. Jiang and J. Wang, "Partial copy detection in videos: A benchmark and an evaluation of popular methods," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 32–42, Mar. 2016.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[26] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: A review," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–25, 2022.

[27] Y.-G. Jiang, Y. Jiang, and J. Wang, "VCDB: A large-scale database for partial copy detection in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 357–371.

[28] J. Y. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 53–61.

[29] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian, "Good practice in CNN feature transfer," 2016, *arXiv:1604.00133*.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[31] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "Near-duplicate video retrieval with deep metric learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 347–356.

[32] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 218–227.

[33] P. Over, A. F. Smeaton, and P. Kelly, "The TRECVID 2007 BBC rushes summarization evaluation pilot," in *Proc. Int. Workshop TRECVID Video Summarization*, Sep. 2007, pp. 1–15.

[34] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, "Spatial color indexing and applications," *Int. J. Comput. Vis.*, vol. 35, pp. 245–268, Sep. 1999.

[35] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 382–395, Mar. 2015.

[36] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "Near-duplicate video retrieval by aggregating intermediate CNN layers," in *Proc. 23rd Int. Conf. MultiMedia Modeling*, 2017, pp. 251–263.

ANIQA DILAWARI is currently with the Department of Computer Science and Information Technology, University of Home Economics, Lahore, Pakistan. She has also worked on multiple artificial intelligence research projects. Her area of research interests include image processing, natural language processing, pattern recognition, and deep learning in image/video analysis.

SAJID IQBAL received the Ph.D. degree from the Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan. He is currently an Assistant Professor with the Department of Information Systems, College of Computer. He has published more than 30 papers in local and international journals and conferences. His research interests include medical image analysis, natural language processing, and computer vision.

FARIAL SYED has gained decades of valuable academic experience from lecturing in graduate and post graduate courses in Pakistan, from 2006 to 2016, and continued it afterward overseas. Currently, she is a Research Scholar in computer science with the University of Regina, SK, Canada. She is a dedicated computer science professional with a passion for teaching and research. She strives to create engaging and innovative learning experiences. Her research interests include three-way decisions, data science, machine learning, and artificial intelligence.

QAZI MUDASSAR ILYAS received the M.Sc. degree in computer science from the University of Agriculture Faisalabad, Pakistan, in 2000, and the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, China, in 2005. He has served several educational institutions in Pakistan, including the Ghulam Ishaq Institute, Topi; the University of the Punjab, Lahore; and the COMSATS Institute of Information Technology, Abbottabad. Currently, he is an Associate Professor with King Faisal University, Saudi Arabia. His research interests include machine learning, knowledge management, and information retrieval.