IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 25, 2020, accepted October 1, 2020, date of publication October 9, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029834

Event-Oriented 3D Convolutional Features Selection and Hash Codes Generation Using PCA for Video Retrieval

AMIN ULLAH[1], (Student Member, IEEE), KHAN MUHAMMAD[1,2], (Member, IEEE), TANVEER HUSSAIN[1], (Student Member, IEEE), SUNG WOOK BAIK[1], (Senior Member, IEEE), AND VICTOR HUGO C. DE ALBUQUERQUE[3,4,5], (Senior Member, IEEE)

[1]Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, South Korea
[2]Department of Software, Sejong University, Seoul 143-747, South Korea
[3]University of Fortaleza, Fortaleza CE 60811-905, Brazil
[4]LAPISCO, Federal Institute of Education, Science and Technology of Ceará, Fortaleza CE 62930-000, Brazil
[5]ARMTEC Tecnologia em Robótica, Fortaleza CE 60811-341, Brazil

Corresponding author: Sung Wook Baik (sbaik@sejong.ac.kr)

This work was supported by the Institute for Information and Communication Technology Planning and Evaluation (IITP) through the Development of Integrated Cross-Model Data Processing Platform Supporting a Unified Analysis of Various Big Data Models funded by the Ministry of Science and ICT (MSIT), South Korea, under Project 2019-0-00795.

ABSTRACT The extensive video surveillance networks gather an enormous amount of data exponentially on a daily basis and its management is a challenging task, requiring efficient and effective techniques for searching, indexing, and retrieval. The employed mainstream techniques are focusing on general category videos, where the important events in surveillance require fine-grained events retrieval. In this paper, we introduce an event-oriented feature selection mechanism by utilizing the intermediate convolutional layer of a pre-trained 3D-CNN model, that is selected after deep investigation of its weights and response to a particular event. The extracted exclusive features represent an event semantically and effectively eliminate those neurons which do not respond to an event. Furthermore, the event-oriented convolutional features are of very high-dimensions, requiring additional storage, and take more time in feature comparison for retrieval. Therefore, we generate compact binary codes from these features using principle component analysis (PCA) algorithm. This makes our system more efficient to retrieve videos from large scale database. We evaluated our approach on the challenging events of UCF101 and HMDB51 datasets for original features and generated compact codes to achieve reduced execution time and better precision and recall scores.

INDEX TERMS Deep learning, feature selection, video retrieval, video analytics, hash codes, surveillance event analysis.

I. INTRODUCTION

The amount of videos since the birth of Internet is increasing on daily basis, where large number of videos are recorded, uploaded, and downloaded from world wide web. Video retrieval, since last three decades, particularly after the birth of Internet has drawn the attention of researchers due to its wide range of applications and extreme need in multimedia information processing domain. The automatic technique of retrieving user' interest videos is called content-based video retrieval (CBVR), which ensures the relevancy of both, the query and the retrieved items [1]. Videos preserve richer contents than images, containing huge amount of redundant contents at the same time. Similarly, the video processing and analysis require significant computational complexity for its effective usage including browsing and retrieval [2]. Therefore, compared to image retrieval, extracting similar videos from a huge repository is challenging from several aspects. Video retrieval manually for humans is a tedious and time-consuming task. Alongside, humans are much error prone, thus there is a chance of incorrect retrieval results. Automatic techniques for video retrieval are need of the current technological era to carry out smooth usage of available videos for the Internet as well offline videos users. There are many potential applications of video retrieval systems in diverse areas of Data Science including news, advertising,

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

entertainment, education, video archiving, and most importantly the medical domain. Another emerging application of CBVR are recommender systems used in many websites such as YouTube to provide user with their most relevant contents such as movies, soccer highlights, etc.

As a video preserves meaningful contents that can be used to differentiate between two videos and their similarity metric can lead to proper indexing. The contents of video indicate different features such as color, intensity, texture, edges, objects trajectories, motion, shapes of objects, optical flow, saliency, etc. The videos are compared for similar contents by matching any of these feature vectors. The features such as texture, color, and intensity can be extracted through existing algorithms such as scale invariant feature transform SIFT [3], speeded up robust features (SURF) [4], and oriented FAST and rotated BRIEF (ORB), etc. Hence these features are local low-level and have very limited information about the actual contents of the video, therefore, they lead to an under-representation of the input video. Such low-level features usage for video retrieval from big repositories is not reliable. The positive point of low-level features usage is its faster execution which could lead to an efficient video retrieval system. In contrast, there exists mid-level features which include motion, saliency, etc. and achieve better results than low-level for contents representation leading to comparatively an enhanced video retrieval system. The best option for consideration in terms of suitable representation of contents are high-level features which extract every slight detail from frames such as edges, shapes, motion information, flow of objects, events, etc. High-level features lead to satisfactory representation of video and alternatively yielding a supreme video retrieval system. The strategies of representative methods of all the discussed features are given with supported references in the next paragraphs.

Bag-of-features (BoF) are evaluated by Jiang *et al.* [5] using different factors that govern the performance of BoF to choose the optimal one. The main focus of this research is to replace the global features through several variants of BoF for semantic contents representation of an image. The authors presented comparison over different datasets using various kernels and highlighted the best performance of retrieval. Objects' interest point matching in videos using SIFT features for objects-based video retrieval is presented by Huel *et al.* [6]. Similarly, SURF descriptor is used in [7] for video retrieval problem. In order to achieve lower dimensions of SURF descriptors, the authors utilized stochastic dimensionality reduction method to have an efficient CBVR system. The retrieval accuracy is 78%, where the authors also analysed the comparative performance of lower dimensionality. Zhu *et al.* [8] presented a novel technique for large-scale similar videos retrieval using temporal-concentration SIFT features. The authors efficiently encoded SIFT features with temporal information via tracking to generate temporal-concentration SIFT. These are compressed local features which reduce visual redundancy. Frame level processing is used in these methods to retrieve similar videos contents. A detailed discussion about other similar methods can be found in this chapter [9]. The low-level features-based methods have several disadvantages alongside a positive point of reduced computational complexity. These features are limited to represent the overall scenario of a scene or a complete video. The main objective of retrieval techniques is effective and efficient video contents representation, where the low-level features for retrieval are not sufficient or accurate enough for consideration in practical applications. Therefore, most of the current mainstream methods of the CBVR avoid the usage of low-level features, until there is an extreme need of specific system.

A long video caption mechanism advanced to big video data retrieval is presented in [10]. This technique focuses on shortening of a video through video segmentation to decrease the retrieval time via extracting only interesting clip of the video. Finally, this technique generates video caption through long short-term memory (LSTM) that is used for retrieval of similar videos. An image query based video retrieval system is presented in [11] by utilizing an intelligent fusion of CNNs and bag of visual word module to design a single model. The model is capable of video frames information extraction and their effective representation, where visual weighted inverted index and its co-algorithms are used to enhance the retrieval process. The experiments presented for this method are very limited and the authors only decreased the computational complexity without any improvement in the accuracy. Similarly, the authors in [12] proposed a new dataset and optical character recognition based high-level semantic features along with autoencoders for image retrieval. A spatial and language-temporal tensor fusion based network is proposed in [13] for video moments retrieval. Video retrieval is used in vast domains of computer vision. For instance, a deep CNN based framework for surgery videos retrieval is introduced in [14]. Similarly, there are a lot of matching techniques based on various types of algorithms including deep learning and statistical methods to achieve CBVR task effectively [1], [15]–[17].

The employed literature of CBVR exposes various limitations of the existing techniques from different perspectives. The foremost aspect to be covered in CBVR literature is the effective representation of events, actions, and other contents of the input video. It leads to an effective retrieval system because most of the videos contain events that happen in sequence of frames. In the existing literature, frames level processing is utilized to represent the contents of a video such as actions/events. Frames level processing is not recommended while dealing with human actions and events because they occur sequentially. Another key limitation of employed CBVR techniques is computational complexity which is very high due to the utilization of large number of parameters for features extraction. As the video retrieval system searches in Big Data repositories to find out similar features and hence consumes a lot of computational resources. Similarly, the existing techniques utilize complex distance matching algorithms for

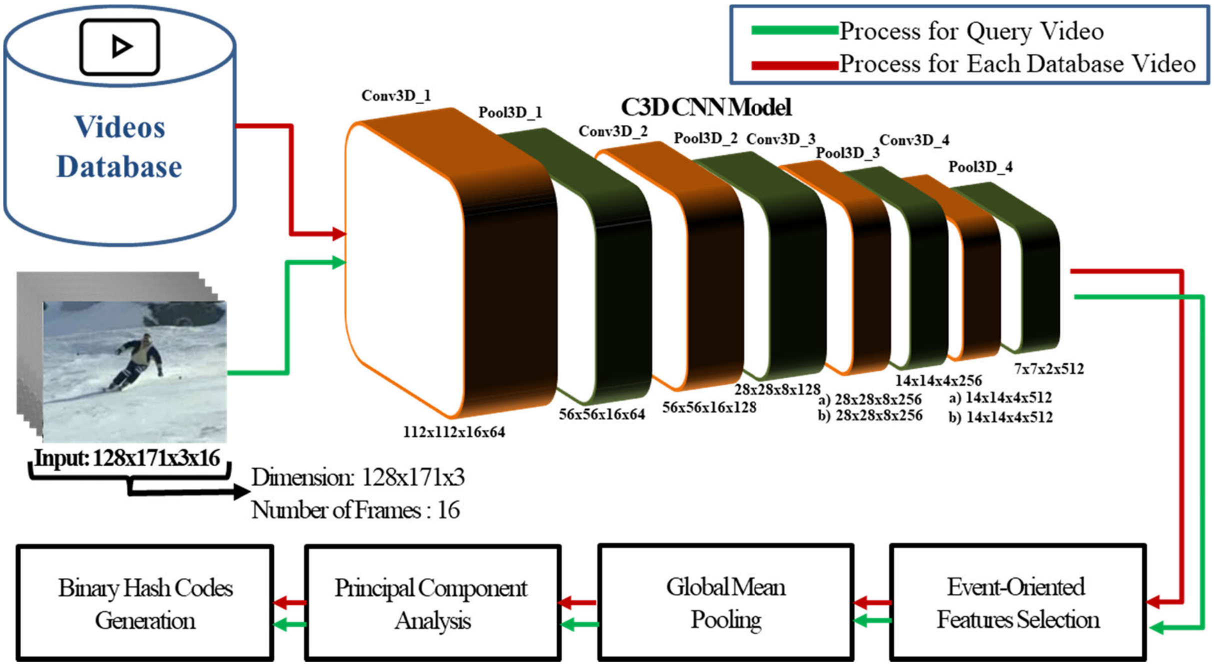


FIGURE 1. The proposed framework for convolutional features selection and compact hash codes generation for event-oriented video retrieval. The process is shown for both database videos and query video. The final step is the comparison of hash codes of query video with all videos of database.

high dimension features similarity computation, which makes the retrieval much slower. In order to overcome the mentioned limitations, we propose a novel technique for video retrieval with the following major contributions to the CBVR literature:

1. The 3D convolutional feature maps of a pre-trained model are investigated to aim at event-oriented activations in order to represent an event effectively and to achieve fine-grained retrieval for surveillance environments.
2. We present a novel feature selection mechanism that chooses only those feature maps where neurons are active for a particular event and eliminate inactive ones. Our event-oriented features are powerful to present event regions in a sequence of frames instead of the background information.
3. The event-oriented convolutional features are very high-dimensional, requiring additional storage and take much extra time in feature comparison for retrieval. Therefore, we propose a unique mechanism to generate compact binary codes from these features using PCA. This makes our system more efficient for large scale video retrieval.

Rest of the paper has three major sections. Section II explains the working of our proposed framework in details. Section III is dedicated to experimental results of our framework with detailed explanation about datasets used for experiments and retrieval results achieved on those datasets. In Section IV, we concluded our paper with discussion about future research works in the CBVR domain.

II. PROPOSED CBVR FRAMEWORK

In this section, we present an event-oriented 3D convolutional feature selection and compact binary hash codes generation framework for fine-grained video retrieval from large scale videos repositories. The proposed framework consists of training process for features selection, convolutional features extraction, and binary hash codes generation using PCA for video indexing and retrieval. Our system can be effectively fine-tuned for any type of surveillance event representation for fine-grained searching and indexing. The details about each step of our framework are provided in the subsequent sections. The graphical representation of our proposed framework and its input leading to output flow is given in **Figure 1.**

A. 2D AND 3D FEATURE MAPS ANALYSIS

The 2D-CNNs have been extensively and successfully applied in many computer vision domains for image data analysis [18], [19]. However, for video analytics the 2D-CNNs are insufficient because they are limited to represent the temporal information of the video data. The 3D-CNNs are introduced to cover both spatial and temporal dimensions for sequential features representation. The 2D-CNNs are also used for video analytics by utilizing some supporting fusion techniques to find temporal information in the output of 2D-CNNs [20]–[22]. In these methods they claimed that it is an efficient way to process high dimensional video data because when they apply 3D filter instead of 2D then the convolutional operation becomes more complex and takes extra processing time. For instance, in our prior work [23], we utilized 2D-CNN and autoencoder for human action recognition. We first extracted the deep features from pretrained VGG-16 CNN model from sequence of frames which are then passed to the optimized deep autoencoder for learning sequential patterns. It is proven from the recent studies [24], that the convolutional and fully connected layer's features of a pretrained CNNs are rich informative and can be utilized for pattern representation tasks of other computer vision applications. The convolutional features maps are famous for the local representations in visual data while the fully connected layer features are the global representations of visual data. Ahmad. *et al.* [25] utilized the convolutional features maps of a 2D-CNN model for the representation of surveillance objects. They claimed that the selected feature maps semantically represent the surveillance objects with minimal background influence. Motivated from the aforementioned works, we investigated the deeper convolutional layer of a 3D-CNN model, known as C3D [26] for the events in video data. This deep learning architecture contains eight 3D convolutional, five 3D pooling, and two fully connected layers. The model is trained on large scale Sport1M human actions video classification dataset which contains 487 action categories and more than one million sample video clips. Its parameters are well trained by achieving 85.2% accuracy score for Sport1M dataset. We investigated its deeper 3D pooling layer named as 'Pool3D_4' because it has large receptive field to capture the tiny patterns of spatial and temporal dimension from the event sequence.

B. EVENT-ORIENTED FEATURES SELECTION

The 3D-CNN model comprises of numerous convolutional feature maps due to the presence of temporal dimensions. It is evident from several studies mentioned in the analysis section that most of the feature maps are not important for the events representation task, which degrades the overall performance. While addressing the effective 3D feature maps selection task related to a particular event, we achieved two major benefits. Firstly, it reduces the features dimension for indexing Big Data repositories. Secondly, it provides the effective discriminative features for event's representation which allows us to precisely compare different events for fine-grained video retrieval [27]. Some of the selected features are given in Figure 2 (a) which shows that the highly activated areas of feature maps represent only the location of event in the sequence of frames while Figure 2 (b) has the discarded feature maps which either have no response for the event or represent the background of the event. The important feature maps where neurons are giving higher attention to the portion of an ongoing event in the sequence of frames are more suitable for the events representation. Contrarily, the feature maps which are active for the background information in the frames sequence has lower potentials for ongoing events representation. The proposed event-oriented feature selection mechanism is mathematically explained in Algorithm 1 alongside the used parameters.

Algorithm 1 Activated Event-Oriented Feature Maps Selection

```
Parameters and abbreviations:
        $TS$ = Training set
        $S$ = Sample event sequence
        $E_F$ = Event-oriented feature maps
        $F$ = Outputs of 3D-CNN
        $F_C$ = Number of channels in the Conv layer
        $F_M$ = Output of Conv layer for features selection
        $NAM$ = Neuron's activation matrix
        $h, w, d$ = height, weight, depth of feature maps
  'i' is representing the iterations of variables and parameters i.e., in $TS_i$, 'i' means one element of training set.
Input:
        $TS = \{S_1, S_2, S_3, \ldots, S_n\}$
Output
        Event-oriented feature maps ($E_F$)
Preparation:
        1. Pretrained C3D CNN
        2. Initialize $NAM0$
Steps:
    1. for each sequence $TSi$ in $TS$
        a. $F \leftarrow$ Feed $TSi$ to C3D CNN
        b. $h \times w \times \text{d} \times F_C \leftarrow$ pool3D_4(F)
        c. $F_M \leftarrow$ Concatenate $(d \times F_C)$
           *note: after concatenation the dimensions will become $(h \times w \times F_N)$
        d. for each $Fmi$ in $F_M$
           $F_{Ni}$ = global average pooling $(Fmi)$
           end for
        e. $F_A \leftarrow$ find ($Fmi$ whose $F_{Ni} > t_1$)
        f. Activate $F_A$ indices with 1 in $NAM$ for $TSi$
       end for
    2. Compute histogram of HIST_$F_{Ai}$ for each $Fmi$
    3. Return all $Fmi$ as $E_F$ whose HIST_$F_{Ai} > t_2$
```

Firstly, our framework inputs a set of training videos TS that represent a particular event. Next, for each video S of a training set, we get F output from a pretrained 3D-CNN model. As each CNN model has multiple Convolutional layers, therefore, after an in-depth analysis, we employed the feature maps of pool3D_4 layer. The channels and the depth of selected layer are concatenated for analysis with $h \times w \times \text{d}$ dimensions and stored in F_M. Following this, for each feature map Fmi in $F_{M,}$ global average pooling is calculated and feature maps with Fmi greater than a defined threshold t are selected in event-oriented feature maps.

To see the activated 3D feature maps of the pretrained C3D CNN model, we exploited a training process which learns from the given event sequences of the UCF101 action recognition dataset [28]. This dataset consists of 101 action events, where each category has more than 100 video samples. These video samples are further divided into short sequences of 16 consecutive frames, forming more than 10

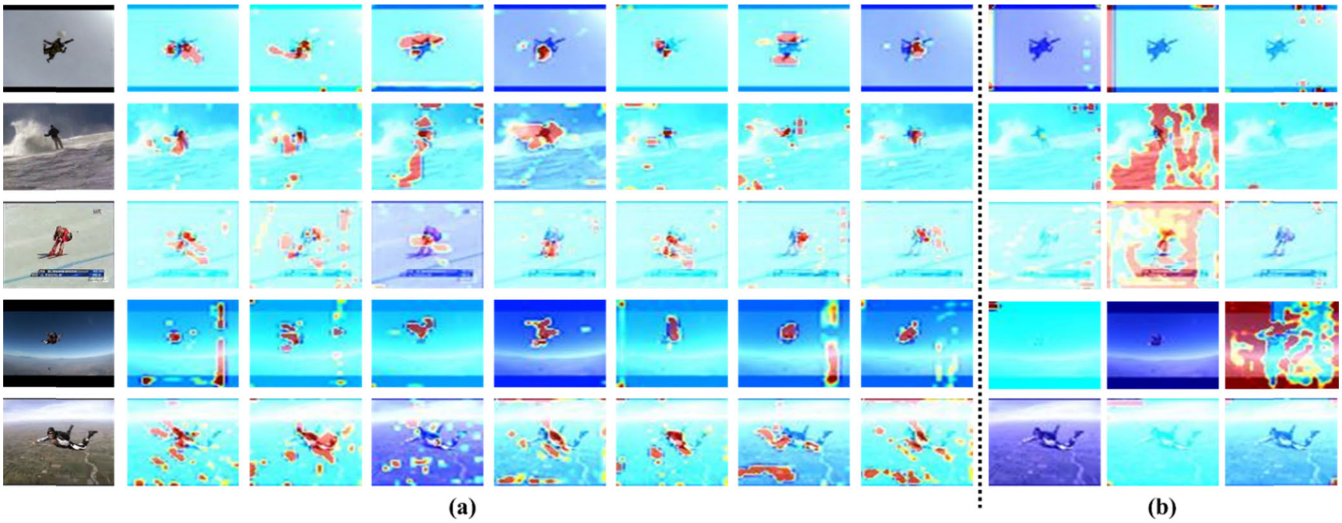


FIGURE 2. (a) The feature maps selected by the proposed mechanism for events representation from a C3D CNN model, where the feature maps are drawn on the actual event to know whether it is responding to the underlying event or not. (b) The feature maps discarded by our algorithm because they are either giving no response or representing the background information of the event.

sequences for each video. It is more suitable for the analysis of 3D event-oriented features because its videos are captured in a real-life environment where the performer of the action is present in the actual locations of happening action. For instance, a soccer player actions are recoded in the soccer ground and surfing is performed on the ocean waves, etc. Thus, in such scenarios it is hard to conclude and analyze the most active feature maps for foreground or background of the ongoing event.

During the feature maps selection, training, and analysis process, we extracted maps of 7 × 7 × 2 × 512 dimensions from pool3D_4 layer of C3D CNN model for $TS = \{S1, S2, S3, \ldots, Sn\}$ training samples. The first two dimensions represent the feature map's width and height, third is temporal dimension information, and the fourth is the number of feature maps in this layer. The N temporal dimensions are concatenated for analysis which give us total of 7 × 7 × 1024 feature maps. Next, for each feature map of training sample i, we calculated the global average pooling. Afterwards, all the feature maps with pooling value greater than a threshold are searched and their indices are checked out to be 1 in our NAM, as shown in Figure 3(b), where the yellow color represents the feature maps that give high activation for the underlaying event in sequence of frames and green ones are the feature maps which do not respond to the event occurrence region in the feature map. Each column in the NAM represents a particular feature map for all the training samples which indicates that significant amount of feature maps provide no neurons activation response for any of the training sample as red box is drawn in Figure 3(b). Thus, features maps without activations can be discarded without adversely affecting the overall performance. Finally, we computed the histogram for the feature maps, NAM, as shown Figure 3(a) and defined a threshold t for the final feature maps selection. Our feature selection mechanism effectively removes the activations which are less responsive for the event in the sequences and the selected ones effectively represent the events for fine-grained video retrieval Some of the selected and discarded feature maps are illustrated in Figure 2(a) and Figure 2 (b).

C. COMPACT BINARY CODES GENERATION

The big multimedia video data require efficient methods for searching and retrieval, where the feature-based matching algorithms such as Euclidean distance has high computational cost. The most prominent solution for addressing this issue is to generate binary hash codes from the original features and measure the distance between the corresponding bits in hamming space, instead of features. This helps in two ways: 1) it reduces the feature storage size; and 2) the binary hash codes matching is very efficient for large scale datasets. However, the performance is comparatively lower when using binary hash codes instead of the original features. Therefore, for precise generation of hash codes, we utilized a multivariate statistical method PCA [29], which is the most prominent feature of factors assessment, that identifies patterns and presents the data in such a way to emphasize similarities and differences inside the given features space. It evaluates correlation among an enormous number of variables and streamlines the complexity of high-dimensional data, while preserving patterns and trends in low-dimensional features space. The lower dimensionality is achieved by transforming the variables to a new set of variables, which are known as the principal components (PCs). The PCA first standardizes the data to a specified range of values so that each variable contributes equally to the analysis. Next, it computes covariance matrix of standardized data, which is a $M \times M$ symmetric matrix (where M is the number of dimensions). For example,

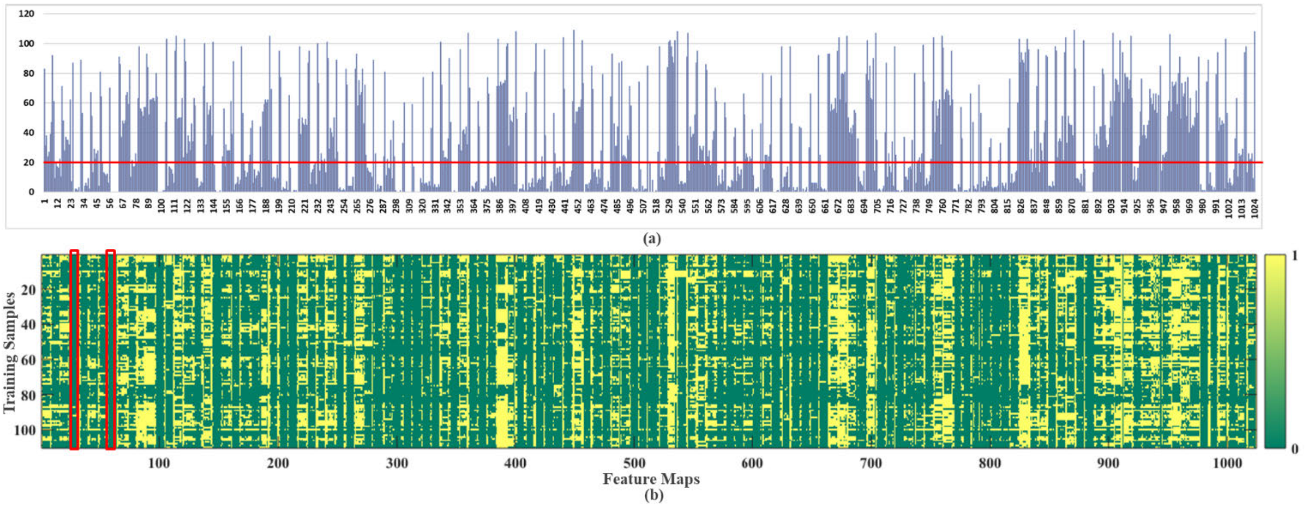


FIGURE 3. (a) The histogram of the neurons' activation responses achieved by an event for all feature maps of all the training samples, the red line indicates the threshold for selecting the feature maps. (b) Neuron's activation matrix (NAM) which shows how many feature maps are active for a particular event, the yellow ones are active and selected for event representation while the green ones are not active and discarded for event representation.

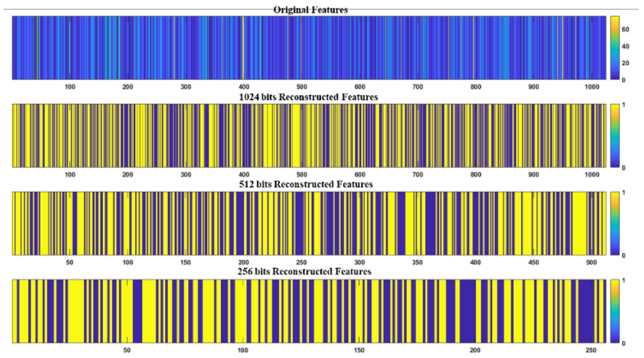


FIGURE 4. Visualization of original convolutional features extracted for a sequence of frames using pretrained C3D CNN model and squeezed to 1024, 512, and 256-bits hash codes.

for a 3-dimensional dataset with 3 variables *x, y*, and *z*, the covariance matrix is a 3 × 3 matrix. Followed by computing the eigenvectors and eigenvalues of the covariance matrix to identify the PCs. The dimensions are reduced by selecting PCs of size of decreasing factor and multiplying it with the original features. More explanation about PCA is out of scope of the paper and could be deeply studied in the referenced paper [29].

The objective of using PCA is its faster and precise conversion rate. In the proposed framework, we extracted 1024 dimensional features for a sequence of 15 video frames which are reduced to 512 and 256 features space and then to binary hash codes, respectively. For retrieval using 1024-bits, we avoided the PCA usage and directly applied a threshold to convert the original convolutional features to hash codes. However, for retrieval using 512-bits and 256-bits, the dimensionality of original features is first reduced using PCA to 512 and 256 features spaces and then we applied a threshold for hash codes generation. The effects of original features and hash codes are visualized in Figure 4. The first row in Figure 4 signifies the original features, where the values are in floating points. From this 1024-dimensional feature vector, the high and low activations of the features are clearly observable. From second row in Figure 4, it is very clear that the transformation from the original feature to 1024-bits using proposed method is very precise. The higher values are converted to 1s and the low values are converted to 0s. Similarly, for 512 and 256-bits almost similar patterns of bits have been achieved. For feature conversion to bits, we utilized different threshold values which we discussed along with results in the experimental section. The detailed analysis of hash codes generation and its performance is discussed in the experimental section.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we explore the comprehensive assessment of our video retrieval algorithm using two benchmark event detection datasets including UCF101 [28] and HMDB51 [30]. We conducted various experiments using original convolutional features, different level of selected features, and binary hash code-based features to evaluate the performance of our framework. Furthermore, we have provided a detailed analysis of the proposed features selection mechanism on the performance of video retrieval in terms of effectiveness and efficiency. The experiments are performed on 8 cores system which contains RTX 2080ti GPU with 11 GB dedicated RAM. Convolutional features are extracted using CAFFE [31] deep learning tool and the features selection and binary hash codes generation mechanisms are implemented in MATLAB 2018.

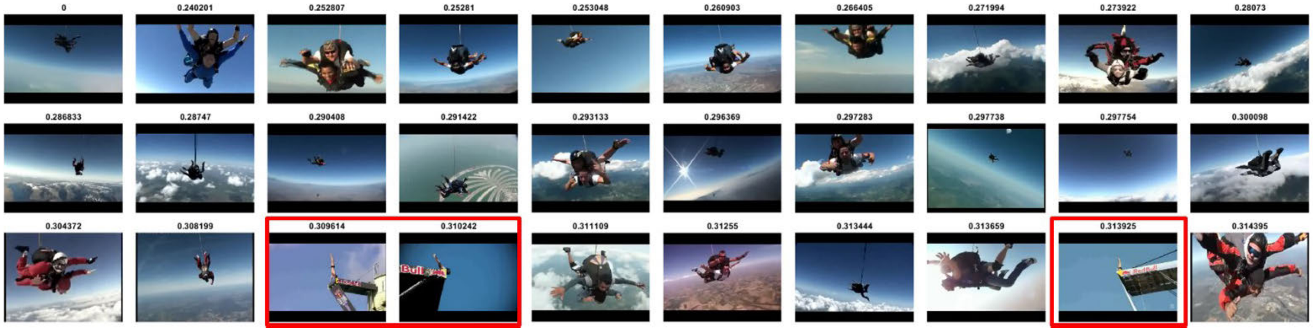


FIGURE 5. Retrieval results for a query video given from sky diving category of UCF101 dataset. The frames having red borders are miss-retrieved results with respect to the input query video.

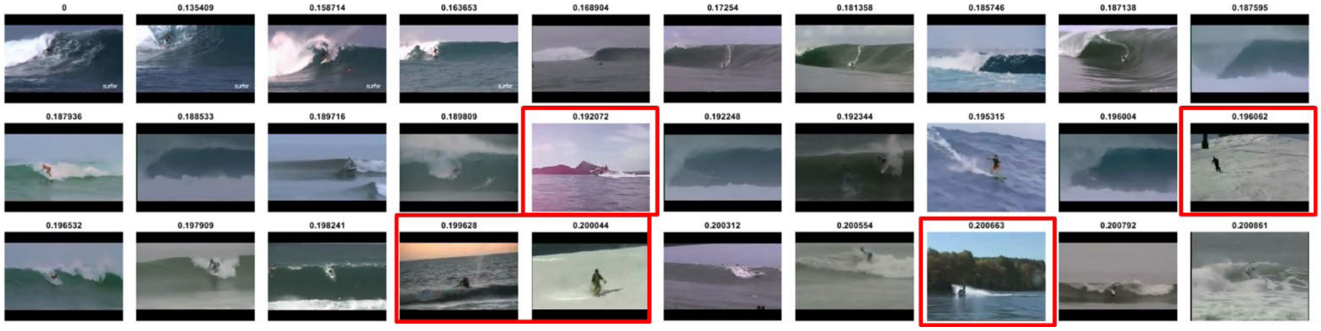


FIGURE 6. Retrieval results for a query video given from surfing category of UCF101 dataset. The red border frames show incorrectly retrieved results.

A. EVALUATION METRICS

The evaluation of information retrieval task is different from the classification problem, where methods are evaluated by crosschecking the prediction and ground truth results. In retrieval problem, when a query is given and its similar videos are retrieved, then the ones which are semantically similar to the query are considered as ''relevant'' and others as ''not relevant''. We utilized precision, recall, cumulative match characteristic (CMC), and mean average precision (MAP) value for the evaluation of our proposed video retrieval framework. Precision is the ratio between the accurate retrieved videos and the total videos user want to retrieve, as calculated in Eq. 1. It is very effective assessment metric which measures the positive predictive values of the retrieval systems. Recall is the ratio between the accurate videos retrieved and total number of relevant videos in dataset for a particular query, as given in Eq. 2. It measures the sensitivity of the retrieval systems, indicating the performance on different levels of recall [38]. Typically, this evaluation is presented by precision and recall graph where precision is calculated for various recall levels as visualized in Figure 7 and Figure 9.

$$precision = \frac{\{relevant\ videos\} \cap \{retrieved\ videos\}}{\{retrieved\ videos\}} \quad (1)$$

$$recall = \frac{\{relevant\ videos\} \cap \{retrieved\ videos\}}{\{relevant\ videos\ in\ dataset\}} \quad (2)$$

The CMC curve is another evaluation metric for the quantitative analysis of CBVR systems. It measures the precision of CBVR at different ranks level. For example, a required query is retrieved by system at which position of the obtained results. In CMC curve, the vertical axis represents the precision percentage and the horizontal axis shows rank of that precision. The details about MAP value calculation can be found from a research in [39].

B. RETRIEVAL RESULTS ACHIEVED ON UCF101 EVENTS DATASET

UCF101 dataset is a collection of 101 different types of realistic human actions videos, gathered from YouTube. It contains 13320 net events videos recorded in different real-life scenarios such as human interaction with objects i.e. soccer and baseball or playing musical instruments, etc. We used this dataset for evaluation because it offers a broad assortment of events with a combination of different types of scenarios such as illumination, object size and pose, camera motion and viewpoint, and it imposes various challenges for video retrieval systems. However, the proposed features selection mechanism is able to deal with the aforementioned challenges by selecting the feature maps with active neurons. For the evaluation, we have randomly selected test queries from different classes of the dataset and calculated their precision values for different recall levels. The performance using CMC curves on UCF101 dataset is shown in Figure 8.

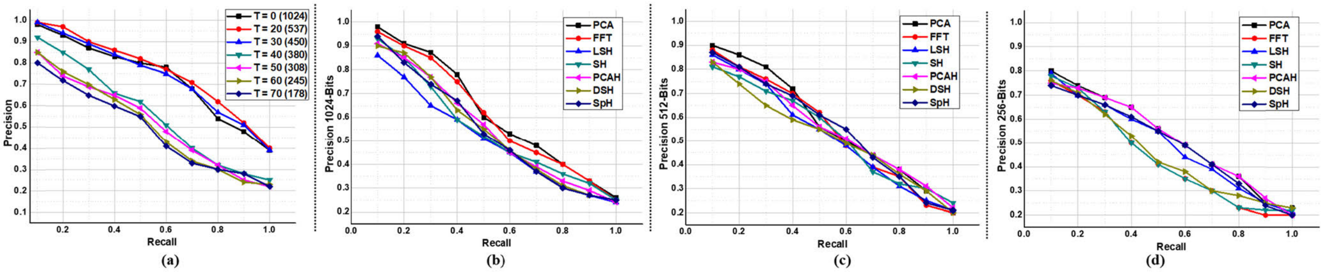


FIGURE 7. Performance of the proposed technique on UCF101 dataset using precision and recall graphs for (a) selected feature maps based retrieval using different thresholds. Comparison of proposed PCA-based hash codes generation with FFT [32], LSH [33], SH [34], PCAH [35], DSH [36], and SpH [37] using (b) 1024-bits hash codes (c) 512-bits hash codes and (d) 256-bits hash codes.

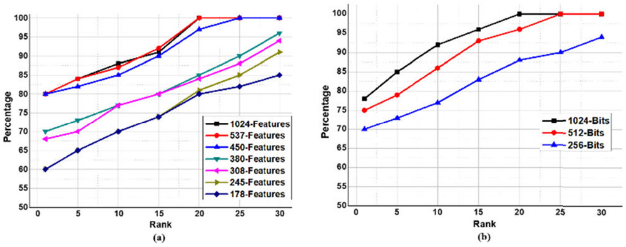


FIGURE 8. The CMC curves for UCF101 dataset (a) results of selected features using different thresholds (b) results achieved using binary proposed hash codes generation method.

TABLE 1. A comparison of the proposed technique with state-of-the-art using MAP values for 1024-bits hash codes-based retrieval.

Methods	MAP (%)	
	UCF101 Dataset	HMDB51 Dataset
FFT [33]	82.14	74.95
LSH [34]	70.62	66.38
SH [35]	74.35	67.54
PCAH [36]	75.42	67.12
DSH [37]	76.91	70.53
SpH [38]	74.65	68.85
Our Method	**83.21**	**75.32**

To begin with, it can be seen from Figure 8 (a) that the selected features using proposed algorithm including 537-, 450, 380-features have similar results compared to using the whole 1024-features. For 1024 dimensional features Rank-1 retrieval start with 80% precision and on Rank 20 it reaches 100% precision. For 308-, 245-, and 178-features, the retrieval performance is lower at initial ranks, because the number of features are very less, however, its precision value reached more than 80% at Rank-30. Furthermore, the performance using proposed hash codes is almost similar to the original features on UCF101 dataset. For instance, the 1024-bits hash codes achieved similar CMC curves as utilizing the original 1024 features. The competence of 512-bits and 256-bits is slightly reduced at lower rank, however, it reached to 100% precision after Rank-15. In addition, it should be kept in mind that time required for bits comparison for finding the similarity is much faster than finding the similarity using double values. Therefore, if we can get faster and real-time results, then the 512-bits are very equidistant to be used for event-oriented video retrieval.

The precision and recall graph for the UCF101 dataset is given in Figure 7 (a). We tested our method by performing different experiments on feature selection mechanism via certain thresholds. The detailed discussion of feature maps selection is given in Section 2.B. The threshold 0 refers to all feature maps selection and other thresholds select the neurons which are active for events in the video. It can be seen from Figure 7 (a) that the features selection at T equals to 20 utilized only 537 dimensions' features that perform much better than the total 1024-dimension features for retrieval task. Our proposed technique removes most of the feature maps that respond to the background of the underlying event or not active for the event. For instance, the dataset contains many categories which are performed in similar background i.e. soccer, baseball, cricket, etc., that are played in green grass. So, if we search for such events using full features then many non-relevant similar background information events are retrieved. The results using T equal to 30 are still similar with the total features results because in this case we discarded only 87 more features maps. The rest of the test thresholds showed poor results for higher recall level however on 0.2 recall, we achieved better precision scores of 0.9, 0.86, 0.85, and 0.8, respectively.

Results for queries from sky diving and surfing class of UCF101 dataset are visualized in Figure 5 and Figure 6. The first image in both the figures is the input query and the labels of the subplot images show their similarity score with the given query, where 0 means 100% sure that it is the same event and the near to 0 value indicates best match. From Figure 5 and Figure 6, it can be seen that the proposed technique has retrieved very similar events, however, in the third row of Figure 5 our method retrieved the jumping event as sky diving. Similarly, in the surfing query it retrieved skiing and skijet event videos. This is because the 3D-CNN models capture spatiotemporal features of the event and if we compare those non-relevant retrieved events to the query, their spatial and temporal patterns are very much similar to each other, therefore such non-relevant events are retrieved as relevant.

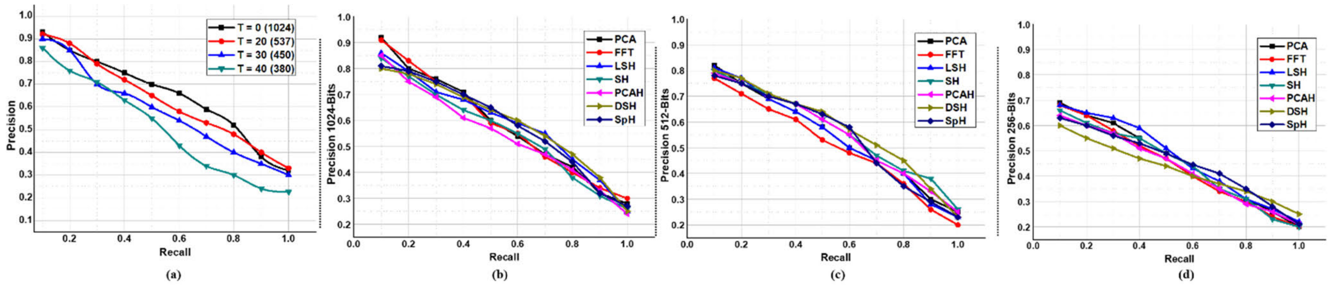


FIGURE 9. Performance graph of the proposed technique on HMDB51 dataset using precision and recall graphs for (a) selected feature maps based retrieval using different thresholds. Comparison of proposed PCA-based hash codes generation with FFT [32], LSH [33], SH [34], PCAH [35], DSH [36], and SpH [37] using (b) 1024-bits hash codes (c) 512-bits hash codes (d) 256-bits hash codes.

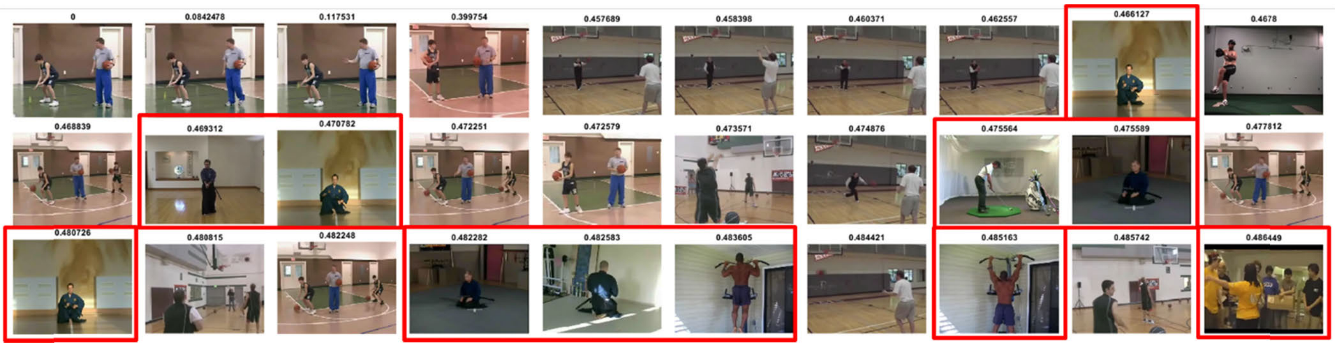


FIGURE 10. Retrieval results for a query video given from basketball dribbling category of HMDB51 dataset. The frames having red borders are miss-retrieved results for the input query of basketball category.

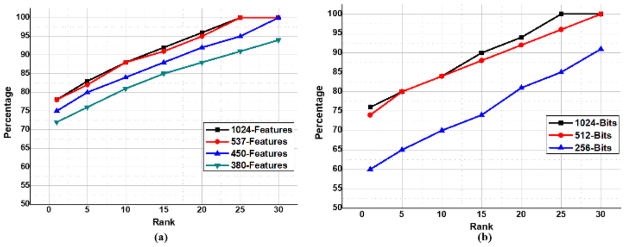


FIGURE 11. The CMC curves for HMDB51 dataset; (a) results of selected features using different thresholds, (b) results achieved using binary proposed hash codes generation method.

The comparison with state-of-the-art using MAP performance for UCF101 dataset is given in column 2 of Table 1. The MAP values are calculated using randomly selected 50 different queries. The proposed method has achieved highest MAP value of 83.21%, while FFT [32], LSH [33], SH [34], PCAH [35], DSH [36], and SpH [37] achieved 82.14%, 70.62%, 74.35%, 75.42%, 76.91%, 74.65% MAP values, respectively.

C. RETRIEVAL RESULTS ACHIEVED ON HMDB51 DATASET

HMDB51 dataset contains 6474 manually annotated video clips made of 51 distinct human actions categories along with some clips included from various sources such as YouTube, movies, and public databases. This dataset is very challenging because its inter class data is very diverse in nature, where each clip is captured in unique scenario, therefore, the results on this dataset are not that promising as UCF101. The precision and recall graph for the HMDB51 dataset is given in Figure 9 (a). We performed the same kind of experiments on this dataset; however, the selected and total features have almost the same performance on lower recall values but on higher recall the selected features have achieved better results. Some of the visual results for challenging queries from ''basketball dribbling'' and ''selfie smiling'' are shown in Figure 10 and Figure 12. It can be seen in Figure 10 that for basketball dribbling the incorrect events are mostly indoor and performing sports activities. For instance, pull-up events are retrieved, where moment patterns are very similar to the basketball jumps. Similarly, for the selfie smiling events in Figure 12 most of the non-relevant facial events are retrieved. For example, in first row of Figure 12, the 3rd retrieved video is of kissing event however, the performer is also smiling at the same time. Our proposed technique has such kind of non-relevant video retrieval, but the overall performance is better, and we achieved higher precision scores on all recall levels. The performance using CMC curves on HMDB51 dataset is shown in Figure 11. It is a very challenging dataset, yet our method has achieved more than 75% precision values on very low rank for selected features and reached to 100% on Rank-25 results. Using proposed binary hash codes, the performance of 1024 and 512-bits is similar to the performance using original features. However, only 256-bits results are very low because most of

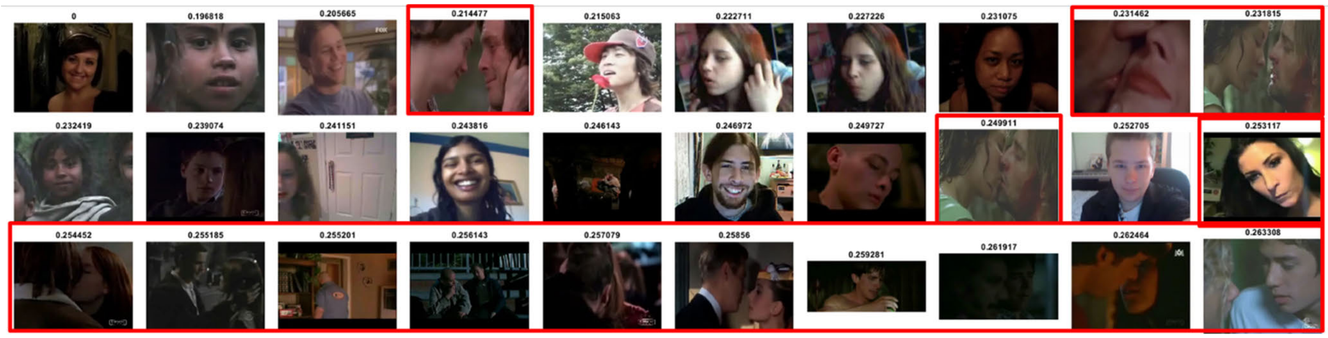


FIGURE 12. Retrieval results for a query video given from selfie smiling category of HMDB51 dataset. Most of the frames are miss-retrieved, as given in red borders.

the information is lost while reducing the features and also due to low bit conversion. Therefore, the retrieval results of 256-bits are comparatively lower. We evaluated our method using MAP value for 50 randomly selected queries from HMDB51 dataset and the comparison with state-of-the-art is given in column 3 of Table 1. As this dataset has verity of challenging videos, therefore, all the methods have achieved MAP values under 80%. The proposed method has achieved highest MAP value of 75.32%, while FFT [32], LSH [33], SH [34], PCAH [35], DSH [36], and SpH [37] achieved 74.95%, 66.38%, 67.54%, 6712%, 70.53%, 68.85% MAP values, respectively.

D. PERFORMANCE USING BINARY HASH CODES

The proposed technique is experimentally evaluated using three levels of binary hash codes including 1024-bits, 512-bits, and 256-bits. The precision and recall scores achieved using various lengths of hash codes for UCF101 and HMDB51 datasets are given in Figure 7-(b, c, and d) and Figure 9-(b, c, and d) for 1024, 512, and 256-bits, respectively. The results of the hash codes-based methods are not as better as the original features however the execution time performance is increased exponentially. For instance, for 0.2 recall level our proposed method achieved 0.99 precision score but for 1024-bits, 512-bits, and 256-bits it achieved 0.94, 0.9, and 0.81 precision scores, respectively. Furthermore, for higher recall level 1 the original features achieved 0.4 precision score, while the 1024-bits, 512-bits, and 256-bits, it achieved 0.3, 0.23, and 0.2 precision scores, correspondingly. Similarly, the storage size for original features of UCF101 dataset (13320 videos) is 123 MB; however, when features are converted to hash codes it is only 1690 KBs. We compared our proposed technique with the recent bi-directional fast Fourier transform (BD-FFT) [32], LSH [33], SH [34], PCAH [35], DSH [36], and SpH [37] based hash coding schemes in Figure 7-(b, c, and d) and Figure 9-(b, c, and d). It can be seen from precision and recall graphs of both datasets that for higher recall level, our proposed technique achieved better results as compared to the state-of-the-art techniques and for lower recall level, our results are higher or overlapped at some points for 1024, 512, and 256-bits code. Similarly, the time required to covert original features of one sample to binary hash codes is faster than BD-FFT, LSH, SH and DSH. For instance, BD-FFT takes 0.12 seconds for 512-bits code generation where our proposed technique takes only 0.0082 seconds for its transformation. The experimental results show encouraging performance of the proposed technique in terms of accuracy as well runtime for event-oriented video retrieval.

IV. CONCLUSION AND FUTURE WORK

The concept of video retrieval, particularly CBVR is widely used in different real-life scenarios with applications to medical, surveillance, entertainment, and many other domains. In this paper, we present an event-oriented 3D-CNN features based CBVR system that is extremely efficient and effective for the retrieval of similar contents from huge video data repositories. We opted middle layer features of a 3D-CNN model after a deep investigation of its effectiveness for representation of sequential frames. The sequential features help in capturing the overall context of events which has a prominent role in contents presentation of a video, as it has chunks of events happening at different time intervals. We exploited the convolutional features selection mechanism which is able to select only those features maps from the CNN layer which are active for the ongoing event in the sequence of frames. In order to squeeze the size of extracted high dimensional features for efficient retrieval and faster storage, we introduced the concept of hashing in our problem. We represented these high-dimensional features in compact binary codes via PCA, which ensures efficient searching and lower storage capacity. We performed experiments over several action events datasets and achieved better accuracy. The experimental results confirm the faster retrieval of our framework and the fine quality of exactness in finding video from huge repository. In future, we aim to use medical and healthcare data [40] for efficient endoscopy video retrieval [41] and its sub-domains, functional in IoT environments [27].

REFERENCES

[1] K. Zhang, H. Sun, W. Shi, Y. Feng, Z. Jiang, and J. Zhao, "A video representation method based on multi-view structure preserving embedding for action retrieval," *IEEE Access*, vol. 7, pp. 50400–50411, 2019.

[2] Z. Dong, J. Wei, X. Chen, and P. Zheng, "Face detection in security monitoring based on artificial intelligence video retrieval technology," *IEEE Access*, vol. 8, pp. 63421–63433, 2020.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, vol. 99, no. 2, pp. 1150–1157.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2006, pp. 404–417.

[5] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. 6th ACM Int. Conf. Image Video Retr. (CIVR)*, 2007, pp. 494–501.

[6] X. Hu, Y. Tang, and Z. Zhang, "Video object matching based on SIFT algorithm," in *Proc. Int. Conf. Neural Netw. Signal Process.*, Jun. 2008, pp. 412–415.

[7] S. Asha and M. Sreeraj, "Content based video retrieval using SURF descriptor," in *Proc. 3rd Int. Conf. Adv. Comput. Commun.*, Aug. 2013, pp. 212–215.

[8] Y. Zhu, X. Huang, Q. Huang, and Q. Tian, "Large-scale video copy retrieval with temporal-concentration SIFT," *Neurocomputing*, vol. 187, pp. 83–91, Apr. 2016.

[9] R. C. Veltkamp, H. Burkhardt, and H.-P. Kriegel, *State-of-the-Art in Content-Based Image and Video Retrieval*. Springer, 2013.

[10] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Gener. Comput. Syst.*, vol. 93, pp. 583–595, Apr. 2019.

[11] C. Zhang, Y. Lin, L. Zhu, A. Liu, Z. Zhang, and F. Huang, "CNN-VWII: An efficient approach for large-scale video retrieval by image queries," *Pattern Recognit. Lett.*, vol. 123, pp. 82–88, May 2019.

[12] S. U. Rehman, S. Tu, Y. Huang, and O. U. Rehman, "A benchmark dataset and learning high-level semantic embeddings of multimedia for cross-media retrieval," *IEEE Access*, vol. 6, pp. 67176–67188, 2018.

[13] B. Jiang, X. Huang, C. Yang, and J. Yuan, "SLTFNet: A spatial and language-temporal tensor fusion network for video moment retrieval," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102104.

[14] D. R. Chittajallu, A. Basharat, P. Tunison, S. Horvath, K. O. Wells, S. G. Leeds, J. W. Fleshman, G. Sankaranarayanan, and A. Enquobahrie, "Content-based retrieval of video segments from minimally invasive surgery videos using deep convolutional video descriptors and iterative query refinement," *Med. Imag., Image-Guided Procedures, Robotic Interventions, Model.*, vol. 10951, Mar. 2019, Art. no. 109512Q.

[15] G. N. Kumar and V. Reddy, "Key frame extraction using rough set theory for video retrieval," in *Soft Computing and Signal Processing*. Springer, 2019, pp. 751–757.

[16] R. S. Ram, S. A. Prakash, M. Balaanand, and C. Sivaparthipan, "Colour and orientation of pixel based video retrieval using IHBM similarity measure," *Multimedia Tools Appl.*, pp. 1–16, Jun. 2019.

[17] A. K. Mallick and S. Mukhopadhyay, "Video retrieval based on motion vector key frame extraction and spatial pyramid matching," in *Proc. 6th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Mar. 2019, pp. 687–692.

[18] S. ur Rehman, S. Tu, Y. Huang, and G. Liu, "CSFL: A novel unsupervised convolution neural network approach for visual pattern classification," *AI Commun.*, vol. 30, pp. 311–324, Jan. 2017.

[19] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino, and V. H. C. de Albuquerque, "Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4455–4463, May 2020.

[20] A. Ullah, K. Muhammad, T. Hussain, and S. W. Baik, "Conflux LSTMs network: A novel approach for multi-view action recognition," *Neurocomputing*, 2020.

[21] I. Ul Haq, A. Ullah, K. Muhammad, M. Y. Lee, and S. W. Baik, "Personalized movie summarization using deep CNN-assisted facial expression recognition," *Complexity*, vol. 2019, pp. 1–10, May 2019.

[22] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human behavior understanding in big multimedia data using CNN based facial expression recognition," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1611–1621, Aug. 2020.

[23] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Gener. Comput. Syst.*, vol. 96, pp. 386–397, Jul. 2019.

[24] K. Muhammad, T. Hussain, J. Del Ser, V. Palade, and V. H. C. de Albuquerque, "DeepReS: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5938–5947, Sep. 2020.

[25] J. Ahmad, K. Muhammad, S. Bakshi, and S. W. Baik, "Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets," *Future Gener. Comput. Syst.*, vol. 81, pp. 314–330, Apr. 2018.

[26] 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.

[27] K. Muhammad, J. J. P. C. Rodrigues, S. Kozlov, F. Piccialli, and V. H. C. D. Albuquerque, "Energy-efficient monitoring of fire scenes for intelligent networks," *IEEE Netw.*, vol. 34, no. 3, pp. 108–115, May 2020.

[28] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: http://arxiv.org/abs/1212.0402

[29] J. Shlens, "A tutorial on principal component analysis," 2014, *arXiv:1404.1100*. [Online]. Available: http://arxiv.org/abs/1404.1100

[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.

[32] J. Ahmad, K. Muhammad, J. Lloret, and S. W. Baik, "Efficient conversion of deep features to compact binary codes using Fourier decomposition for multimedia big data," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3205–3215, Jul. 2018.

[33] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB*, 1999, vol. 99, no. 6, pp. 518–529.

[34] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1753–1760.

[35] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image auto-annotation by search," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1483–1490.

[36] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1362–1371, Aug. 2014.

[37] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, "Spherical hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2957–2964.

[38] T. Hussain, K. Muhammad, S. Khan, A. Ullah, M. Y. Lee, and S. W. Baik, "Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers," *J. Artif. Intell. Syst.*, vol. 1, no. 15, p. 2019, 2019.

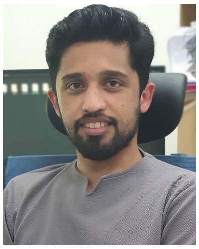
[39] M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, and S. W. Baik, "Integrating salient colors with rotational invariant texture features for image representation in retrieval systems," *Multimedia Tools Appl.*, vol. 77, no. 4, pp. 4769–4789, Feb. 2018.

[40] K. Muhammad, S. Khan, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 30, 2020, doi: 10.1109/TNNLS.2020.2995800.

[41] K. Muhammad, S. Khan, N. Kumar, J. Del Ser, and S. Mirjalili, "Vision-based personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges," *Future Gener. Comput. Syst.*, vol. 113, pp. 266–280, Dec. 2020.

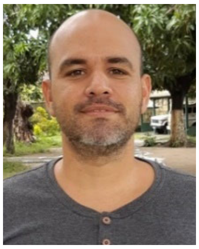
AMIN ULLAH (Student Member, IEEE) received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the M.S. degree leading to the Ph.D. degree with the Intelligent Media Laboratory, Department of Digital Contents, Sejong University, South Korea. He has published several papers in reputed peer-reviewed international journals and conferences, including the IEEE Transactions on Industrial Electronics, the IEEE Transactions on Industrial Informatics, the IEEE Internet of Things Journal, IEEE Access, *Future Generation Computer Systems* (Elsevier), *Neurocomputing* (Elsevier), *Multimedia Tools and Applications* (Springer), *Mobile Networks and Applications* (Springer), and *Sensors*. His major research interests include human-actions and activity recognition, sequence learning, image and video analytics, content-based indexing and retrieval, the IoT and smart cities, and deep learning for multimedia understanding.

KHAN MUHAMMAD (Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2018. He is currently working as an Assistant Professor with the Department of Software and a Lead Researcher with the Intelligent Media Laboratory, Sejong University. His research interests include intelligent video surveillance (fire/smoke scene analysis, transportation systems, and disaster management), medical image analysis, (brain MRI, diagnostic hysteroscopy, and wireless capsule endoscopy), information security (steganography, encryption, watermarking, and image hashing), video summarization, multimedia, computer vision, the IoT, and smart cities. He has filed/published over seven patents and 120 articles in peer-reviewed journals and conferences in these areas. He is serving as a Reviewer for over 90 well-reputed journals and conferences, from IEEE, ACM, Springer, Elsevier, Wily, SAGE, and Hindawi publishers. He acted as a TPC Member and a Session Chair at more than ten conferences in related areas. He is also an Editorial Board Member of the *Journal of Artificial Intelligence and Systems* and a Review Editor of the Section ''Mathematics of Computation and Data Science'' in the journal *Frontiers in Applied Mathematics and Statistics*.



TANVEER HUSSAIN (Student Member, IEEE) received the bachelor's degree (Hons.) in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the joint master's and Ph.D. degrees with Sejong University, Seoul, South Korea. He is also a Research Assistant with the Intelligent Media Laboratory (IM Laboratory). He has filed/published several patents and articles in peer-reviewed journals and conferences in reputed venues, including the IEEE Transactions on Industrial Informatics, the IEEE Internet of Things Journal, *Network Magazine*, *Pattern Recognition* (Elsevier), *Neurocomputing*, *Pattern Recognition Letters*, the *International Journal of Energy Research* (Wiley), the *International Journal of Distributed Sensors Networks*, and *Multimedia Tools and Applications* (Springer). His major research domains are features extraction (learned and low-level features), video analytics, image processing, pattern recognition, medical image analysis, multimedia data retrieval, deep learning for multimedia data understanding, single/multi-view video summarization, the IoT, the IIoT, and resource-constrained programming. He received the Gold Medal from the Islamia College Peshawar. He provides a Professional Review services in various reputed journals, such as the IEEE Transactions on Cybernetics, the IEEE Transactions on Industrial Informatics, and the IEEE Journal of Biomedical and Health Informatics. He is serving as an Associate Editor for the *Journal of Biological Sciences*. For further activities and implementations, visit: https://github.com/tanveer-hussain

SUNG WOOK BAIK (Senior Member, IEEE) received the Ph.D. degree in information technology engineering from George Mason University, Fairfax, VA, USA, in 1999. He was with the Intelligent Systems Group, Datamat Systems Research Inc., as a Senior Scientist, from 1997 to 2002. Since 2002, he has been a Faculty Member with the Department of Digital Contents, Sejong University, Seoul, South Korea. He is currently a Full Professor with the Department of Digital Contents, a Chief with the Sejong Industry, Academy Cooperation Foundation, and the Head of the Intelligent Media Laboratory (IM Laboratory), Sejong University. He has published over 100 articles in peer-reviewed international journals on top venues of these domains, including the IEEE Internet of Things Journal, the IEEE Transactions on Systems, Man, and Cybernetics: Systems, *IEEE Communications Magazine*, the IEEE Transactions on Industrial Electronics, the IEEE Transactions on Industrial Informatics, IEEE Access, *Neurocomputing* (Elsevier), *FGCS*, *PRL*, *MTAP* (Springer), *JOMS*, *RTIP*, *Sensors* (MDPI), and so on. He is involved in several projects, including AI-Convergence Technologies and Multi-View Video Data Analysis for Smart Cities, Effective Energy Management Methods, Experts' Education for Industrial Unmanned Aerial Vehicles, Big Data Models Analysis, and so on. He was supported by the Korea Institute for Advancement of Technology and the Korea Research Foundation. He holds several Korean and internationally accepted patent in disaster management, image retrieval, and speaker reliability measurement. His research interests include image processing, include image indexing, retrieval for various applications and video analytics, video summarization, action, and activity recognition, anomaly detection, CCTV data analysis, image processing, pattern recognition, video analytics, big data analysis, multimedia data processing, energy load forecasting, the IoT, the IIoT, and smart cities. He serves as a Professional Reviewer for several well-reputed journals and conferences, including the IEEE Transactions on Industrial Informatics, the IEEE Transactions on Cybernetics, IEEE Access, and *Sensors* (MDPI).



VICTOR HUGO C. DE ALBUQUERQUE (Senior Member, IEEE) received the degree in mechatronics engineering from the Federal Center of Technological Education of Ceará (CEFETCE), in 2006, the M.Sc. degree in teleinformatics engineering from the Federal University of Ceará (UFC), in 2007, and the Ph.D. degree in mechanical engineering from the Federal University of Paraíba (UFPB), in 2010. He is a currently a Professor and a Senior Researcher with the University of Fortaleza, LAPISCO/IFCE, and ARMTEC Tecnologia em Robótica, Brazil. His research interests include the IoT, machine/deep learning, pattern recognition, and robotic.