# Analyzing Global Feature for Duplicate Video Retrieval using CNN and PCA

**N. Gayathri, K. Mahesh**

***Abstract- At present, Duplicate video retrieval has attained the interest of researchers because of the vast amount of online videos. This video retrieval has extensive applications like online video monitoring, copyright protection, and automatic video tagging. Some local features constitute primary building blocks in this video retrieval algorithms; with this, most researchers use local information for feature representation. Moreover, this local knowledge-based representation eliminates more prominent information regarding global distribution. However, the discriminative power of local descriptors is diminished by feature quantifiers. The ultimate goal is to use universal features to categorize similar keyframes into the same class that is essential to enhance video retrieval performance. Here, CNN features acquire global geometric distribution of video, from which discrete features are considered for computation. Discretization is performed with principal component analysis. These kinds of features maintain geometry transformation with reduced noise. Next, an integration strategy known as k-NN is used to merge these features with global VR features for enhancing recognition accuracy. Experimentation has been carried out with available datasets to show that the anticipated model outperforms existing approaches in VR applications.***

***Index terms- Video retrieval, video tagging, CNN, global geometric features, discretization, k-NN***

## I. INTRODUCTION

In multimedia applications, digital video content can be easily redistributed and edited. As an outcome, a massive amount of similar and non-similar video clips are generally found in websites for sharing videos and personal video collections [1]. Identical videos are termed as duplicates where video clips that are included for one transformation are termed as near-duplicate video clips. Recognizing NDVC is a target necessity for numerous multimedia applications comprising of monitoring media usage, associating content over the web, meta-data propagation for annotation reasons, intellectual protection property, and redundancy mitigation in video searching outcomes. Detecting NDVC attempts to recognize every match among video clips and query video clips in the referral video database [2]. To diminish computational matching complexity and to mitigate communication overhead, video clips are generally specified by lower dimensionality video-basedSignatures. Those video signatures traditionally comprise of lower-level visual features that are extracted from specific keyframes, for example, demonstrating color or intensity information. This work initiates a novel approach for NDVC predicting the benefits of numerous semantic features (that is, semantic ideas like 'beach,' 'face,' 'sky').

The ultimate target is to utilize global features to classify similar keyframes to similar classes that are significantly essential to improve video retrieval performance [3]. First, content transformation is based on low-level visual feature modification; they pretend to maintain semantic information attained from original video content. Next, based on various prevailing approaches, the user view towards NDVC is effectual detection for measuring semantic similarity. Thirdly, grammatical features hauled out from NDVC detection purpose is reused for annotation purposes. Subsequently, not even a single visual feature type are seems to be more robust towards all probable content transformation, providing space for various experimentation with other kinds of features. In this perspective, it is essential to validate that some of the usages towards semantic features are complementary for visual feature usage providing added discriminative power. The enhancement in the semantic process towards the NDVC detection task is needed for resolving the below-given research factors. 1) acquiring higher semantic coverage: semantic ideas are generally recognized by model-based techniques. Moreover, with the provided model-based semantic idea detection utilizes explicitly a limited amount of classifiers trained with experts, model-based semantic idea reorganization is merely competent to offer limited semantic coverage [4]. Henceforth, to eradicate the need for training those experts is to provide higher semantic coverage, this semantic model to fulfill NDVC based prediction task is based on model-free semantic idea recognition, considering benefits of collective knowledge in image folksonomy (that is, unstructured collection of user constrained tags and images). Finally, folksonomy images are retrieved with the visually similar specification of keyframes and project appropriate tags from folksonomy images to keyframes, 2) Adaptive computation of semantic similarity: While utilizing model-free usage of the semantic idea for prediction, the sum of recognized grammatical concepts, alike of its similarity and nature, it may change based on stronger video shot to video shot. Henceforth, to validate semantic similarity among video shots, Convolutional Neural Network and Principle component analysis are used. The researchers show their interest in global geometric distribution. Therefore geometric transformations are measured to reduce noise [5]. Henceforth, to validate semantic similarity among every video shot, k-NN based integration of global features has to be carried out. Here, grammatical content complexity is measured by facilitating similarity among weighted feature signatures of various dimensionality (therefore, utilization of ‘adaptive measurement’ of research description.

**Revised Manuscript Received on March 02, 2020.**

**N. Gayathri,** Ph.D. Scholar, Department of Computer Applications, Alagappa University, Karaikudi, India. E-mail:gayathri.researchscholar@gmail.com

**Dr. K. Mahesh,** Professor, Department of Computer Applications, Alagappa University, Karaikudi, India.

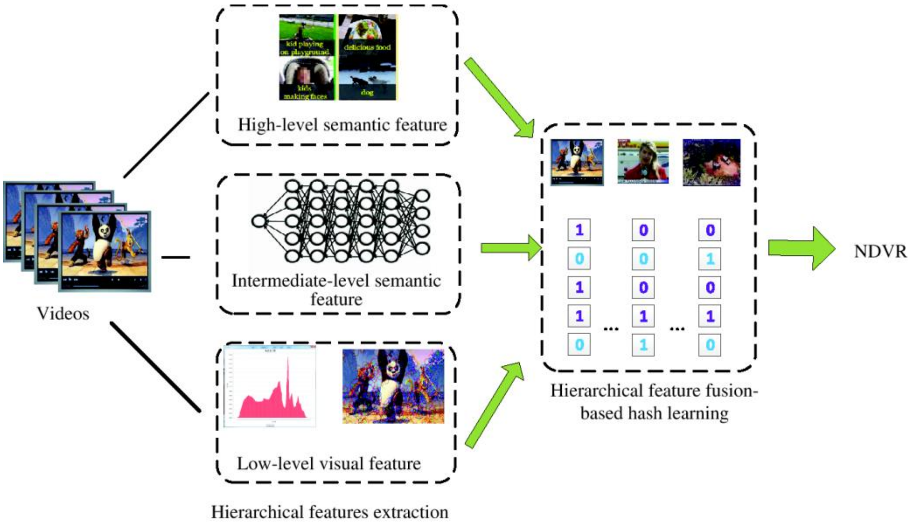
High-level semantic feature
kid playing on playground
delicious food
kids making faces
dog
Intermediate-level semantic feature
Low-level visual feature
Videos
Hierarchical features extraction
Hierarchical feature fusion-based hash learning
NDVR

**Fig 1: Generic NDVR Framework**

3) Acquiring effectual semantic idea for prediction: investigation carried out in this work shows specific attention towards valid impacts on a semantic view-based model prediction on diverse approaches for NDVC prediction as in fig 1, provide with reduced content-based image retrieval efficiency and ambiguity of video tags.

To examine the feasibility of semantic idea based NDVC prediction, experiments are carried out in publicly available videos such as sources from youtube. The experimentation is carried out to provide model-free semantic concept detection and resourcefully used for recognizing NDVC. With the use of model-based semantic concept reorganization, feature evaluated for scale-invariant feature transformation, global feature merging. This work enhances and improves prevailing practice as integrated with experimental outcomes that are more rigorous and extensive. However, additional investigations with the following factors:

1) More appropriate content transformation, 2) use of various content descriptors to provide image that is visually alike of keyframes, 3) use of diverse ground similarity functions; 4) manipulation of collective knowledge towards NDVC prediction; 5) influence of number of shots in query video clip on effectual NDVC detection; and 6) time complexity of matching and generating semantic video signatures.

## II. RELATED WORKS

Various kinds of feature classification to enhance video representation, investigators have also analyzed more appropriately to show how multiple information resources are utilized to improve searching performance. For example, analyze various media kinds of outcomes and queries like text documents, audio, images, and video in NDVR. In this video retrieval, context information related to web videos (for instance, time durations, thumbnail images, and the total amount of views) is merged with video content to enhance retrieval performance. More appropriately, the time duration of videos is utilized to quickly; however, coarsely recognize primary clusters of NDV's. Therefore, seed video is chosen from every group dependent on color histograms of thumbnail images, and its view counts. The last step of NDV detection is diminished to evaluate thumbnail images of candidate videos with chosen seed videos. This technique can acquire around 164 fold speedup, with moderate loss in retrieval accuracy. However, it can merely be utilized to acquire web videos, owing to the use of web context information, which is regrettably not available always in other video films.

Numerous prevailing global features dependent on near-duplicate video retrieval approaches highlight the superior recognition of near-duplicate videos. These approaches are incredibly effectual in dealing with similar or almost similar videos. In prevailing models [6], for instance, investigators utilize HSV to specify keyframes and also accumulate every HSV information in the video to generate a global video signature. This approach acquires superior retrieval accuracy along with quicker retrieval speed with these datasets. In [7], the author depicts discriminative video specifications for all detection over a vast scale video dataset when merely limited hardware resources are accessible. In [8], the author anticipated an appropriate image search with multi-scale contextual evidence; they demonstrate that CNN features are complementary to SIFT because of its semantic awareness and evaluates appropriate numerous other descriptors like HSV, GIST and so on. Moreover, constraints are that global features based approaches generally turn to be very effectual when it seems to process near-duplicate videos with variations and distortions. However, universal features dependent strategies appropriately based on various kinds of chosen elements. In [9], the author anticipated an algorithm for modeling the hierarchical structure to handle the benefits of both local features and global features. In accordance with the color histogram, the author initially hauls out specific videos and then cast off a pairwise comparison approach to fix the significant factors among keyframes. In [10], the author depicted Bag of color approaches for enhancing image search. Even though these techniques improve performance, the pairwise comparison approach utilized to match interest points among keyframes is still more crucial for massive scale video datasets, as the computational cost is enormous [11]-[12]. However, global feature utilization (color histogram) can only filter out a considerable proportion of videos that are not appropriate, for specific near-duplicate videos may have directly measuring techniques or novel features (like spatial, temporal features) to enhance performance [13].

Nevertheless, these investigations are single feature techniques. Specific approaches have been measured to resolving the scalability crisis; they cast off keyframes to query inside large scale video datasets [14]-[15]. Moreover, their essential concept is to recognize diverse frame samplings of reference video dataset to compute probable trade-off among accuracy and scalability.

## III. PROPOSED METHODOLOGY

This section helps in providing suitable CNN features to acquire global geometric distribution of video, from which discrete elements are considered for computation. Discretization is performed with principal component analysis. These kinds of features maintain geometry transformation with reduced noise. Next, an integration strategy known as k-NN is used to merge these features with global VR features for enhancing recognition accuracy.

### A. Video Feature Extraction

In various prevailing investigations, pre-trained CNN approaches are used for extracting visual features for intermediate convolutional layers. These features have to be evaluated through forwarding propagation of video image over the CNN network and utilization of aggregating functions on all convolutional layers.

This work experiments deep CNN architectures: With this architecture, received images have 224 x 224 inputs. For all computation, input images are resized to fulfill all these dimensions. To haul out frame descriptors, the following process has to be carried out. Here, a pre-trained CNN network is used with a total amount of convolutional layer specified as $L^1, L^2, \dots, L^L$. Image-based forward propagation produces a total amount of feature maps, defined as $M^l \in R^{n_d^l\, X \quad and\ X\, c^l} (l = 1, \dots, L)$, where $n_d^l\, X$ and is dimensions of all channel for the convolutional layer (based on input image size) and $c^l$ is the total amount of channels.

To haul out a single descriptor vector for each layer, an aggregation function termed ask-NN is employed in feature mapping. Specifically, use max pooling on each channel of feature map to haul out a single value. The extraction procedure is designed as in Eq. (1):

$$v^l(i) = \max M^l(.,.,i), \qquad i = \{1,2,..c^l\} \qquad (1)$$

Where layer vector is dimensionality vector that is acquired from max-pooling functions of all channel-based feature maps, layer vectors are normalized with $L2$ norm to unit length after extraction.

Here, image descriptors are extracted from activation in intermediate layers, as this work attempts to develop visual representation that maintains local structure in various scales. Fully connected layer activations are not utilized. A positive side impact of the decision is outcoming descriptors as it is compact, decreasing the total processing time and storage necessity. In existing approaches, various uses of initial layers activations are features, as these layers seem to acquire extremely primitive image features. It leads to false matching. For extracting image descriptors, CNN offers a pre-trained model of all CNN networks.

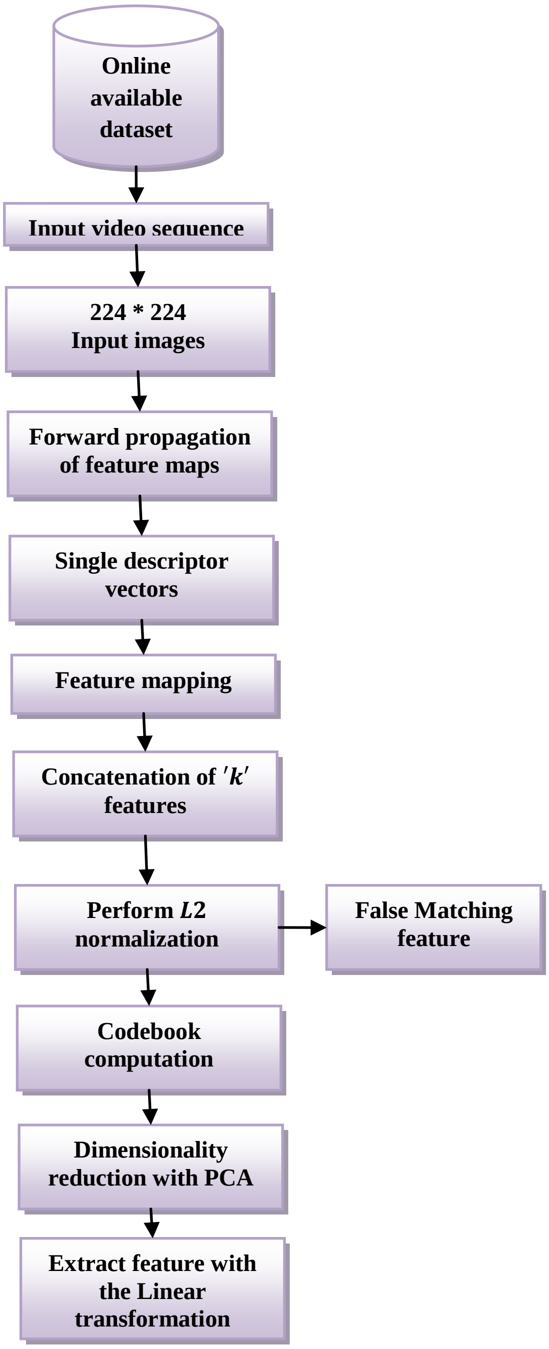
Online available dataset
Input video sequence
224 * 224 Input images
Forward propagation of feature maps
Single descriptor vectors
Feature mapping
Concatenation of 'k' features
Perform L2 normalization
False Matching feature
Codebook computation
Dimensionality reduction with PCA
Extract feature with the Linear transformation

**Fig 2: K-NN based Feature Aggregation**

Video-based image features are generally local, where direct feature matching is not possible as it is computationally costly. Every video image comprises a huge amount of local features that have to match against a huge amount of features in every database. This offers a global specification of images. The most common aggregations comprise some flaws related to the encoding of a single vector.

Therefore, a visual image dictionary with $k-$ visual words is constructed with the k$-NN$ approach. This visual codebook is partitioned into feature space and aggregates statistics of local features to clustered center. Assume a set of dimensionality based local features that belong to a video image.

The difference between the feature related to codebook and visual words are merged and evaluated. Feature space is divided into $'k'$ diverse cells that offer a subset of features. The concatenation of these $'k'$ features is evaluated using Eq. (2):

$$V_i = \sum_j f_j - c_i \qquad (2)$$

Where $V_i$ is vector related to visual data that is specified in $d = k * d$ dimensional vectors, final descriptors are provided with $L_2$ normalization.

By training these codebooks, some subset of extracted features is considered. Moreover, a training codebook with a considerable amount of higher dimensional descriptors is computationally expensive, as well with lower values of $'k$ .' To resolve this crisis, some fractions of all open descriptors are considered. Descriptors of all images are chosen from the training of visual dictionaries. This provides approximately 200K to 300K descriptors with every training dataset. Some specific values of $'N'$ contains more attractive descriptors. Even though it is probable to train feature extraction that is prepared individually for all datasets. Therefore, once if codebook training is completed, image descriptors are generated as per image. Thus, the matching module has to be recognized with similar images. Feature extraction and aggregations can also be done during an offline image process.

**Algorithm 1:**

**Input:** Near video frames<br>
**Input:** Videos from an available online source<br>
**Output:** Extracted video frame

1. Input video to CNN for computation
2. Layers are CNN that are validated for testing and training video images.
3. Perform single descriptors based feature vector extraction as in Eq. (1)
4. Perform layer vector normalization with $the\ L2$ norm.
5. Encounter false matching features
6. Aggregate features with k-NN
7. Visualize the codebook base feature vector using Eq. (2)
8. Trained codebook vectors provide multiple descriptors.
9. Perform discretization with PCA
10. Reduce feature dimensionality
11. Evaluate lower dimensionality space with Mean and covariance as in Eq. (3) &Eq.(4)
12. Perform Linear transformation to provide eigenvector
13. Extract dimensionality reduced features
14. Reduced dimensionality with $10\% - 90\%$
15. Compute matching feature functionality with Eq. (6)
16. Extract Video information from near-duplicate video

### B. Feature Discretization

After aggregating features, discretization based dimensionality reduction is performed with Principle Component Analysis. This facilitates to transform a set of probable correlations among variables to set uncorrelated variables to increase possible variability in the dataset. It is specifically used in image compression and recognition. It facilitates to show data to lower dimensionality space. To perform this, the covariance matrix is evaluated with Eq. (3) & Eq. (4):

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (3)$$

$$Covariance = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \qquad (4)$$

Where $\{x_1, x_2, .. x_n\}$ is a feature set, and $\mu$ is mean of feature characteristics, and covariance specifies the covariance matrix.

Therefore, linear transformation $V^T, W_{PCA}$, is an eigenvector as computed in Eq. (5):

$$W_{PCA} = argmax\, |V^T convariance\, V| \qquad (5)$$

Here, PCA is used to reduce feature dimensionality with a range of 10%-90%.

The near-duplicate image retrieval is hauled out to classify every video. The descriptor features are considered for every intersecting point. Next, global IR features are utilized by position vectors of image. Here, IR normalization is used to classify spatial distribution of all aspects. The k-NN shows better discriminative power, which advantages in the case of local descriptors. Whilst the IR feature exploits global spatial distribution. To display the complete usage of these two features, global IR is used to enhance feature extraction. With the use of PCA, the distance among two features is computed with $'x,'$ and $'y'$ in certain image reflection in Euclidean distance $d\,(x, y)$ is smaller. Therefore, the distance of these IR features with descriptors and k-NN with Euclidean space is extremely smaller. Here, descriptors are provided with $q(x)$ and $b(x)$, where $'q'$ is Quantizer, and $'b'$ is the IR feature. Matching feature functionality is provided by Eq. (6):

$$f_{IR}(x, y) = \begin{cases} \left(tf - idf\left(q(x)\right)\right)^2 & if\ q(x) = q(y), d\left(b(x), b(y)\right) \le h_t \\ 0 & otherwise \end{cases} \qquad (6)$$

Where $'d'$ is Euclidean distance and $h_t$ is Threshold, Quantizer is considered to be high to fulfill $'x'$ and NN match. $h_t$ is extremely lower to filter out more inappropriate points that rely on similar image vectors.

## IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed method execution is carried out in MATLAB 2018a, 64-bit operating system, Intel Core i5 processor, 8GB RAM, respectively.

In this section, the outcome of the proposed model is examined and compared with existing approaches. The efficacy of the proposed method is discussed along with the following performance metrics and comparative charts.

Here, some benchmark datasets are considered for validation, where it comprises referral videos from YouTube, Google, and Yahoo videos. Average precision is used for computing the NDVR dataset. This is provided to show better trade-off among the precision and recall of these datasets. So as to measure the performance of the anticipated model over these videos, evaluation metrics for this video retrieval is analyzed. Normalization detection is cast-off to measure detection costs for all video transformation. This is depicted in Eq. (7):

$$Normalization\ detection = \frac{FN}{N_{target}} + \frac{C_{FA}}{C_{miss} * R_{target}} + \frac{FP}{T_{reflection} \quad * T_{query}} \qquad (7)$$

Where FP and FN are a number of false positives and false negatives, $_{Target}$ is a number of target videos, $C_{miss}$, and $C_{FA}$ is the corresponding cost of false alarms and miss detection. $T_{refdata}$ is total length and $T_{query}$ of complete reference of dataset correspondingly. Least detection cost specifies superior retrieval performance.

**Table I: Average Precision and Threshold Computation**

| Method | Average Precision | Average Threshold |
|---|---|---|
| BOF | 0.89 | 0.60 |
| BOF + HE | 0.9 | 1.25 |
| MFH | 0.95 | 0.63 |
| G-MFH | 0.96 | 0.80 |
| FF | 0.98 | 1.76 |
| TBOW | 0.99 | 1.85 |
| CNN+PCA+k-NN | 1.0 | 0.70 |

Mean processing time is computed for computationally effectual comparison, where mean time to carry out a query for localization and retrieval as in Table I. Execution time for feature extraction and shot detection is providing with this computation.

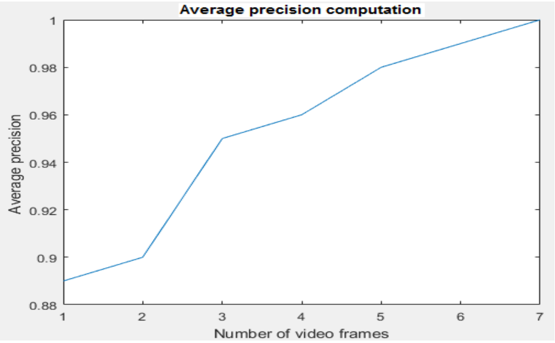

**Fig 3: Average Precision Computation**

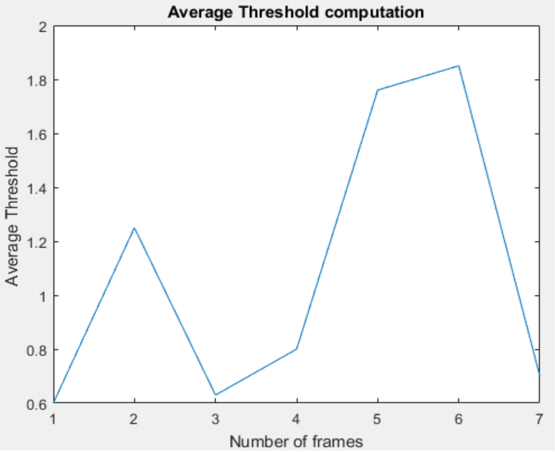

**Fig 4: Average Threshold Computation**

**Table II: Execution Time Computation**

| MAP | Time (s) |
|---|---|
| 0.80 | 0.053 |
| 0.85 | 0.056 |
| 0.84 | 0.0592 |
| 0.86 | 0.0632 |
| 0.87 | 0.0645 |
| 0.88 | 0.0678 |
| 0.86 | 0.073 |
| 0.85 | 0.0735 |

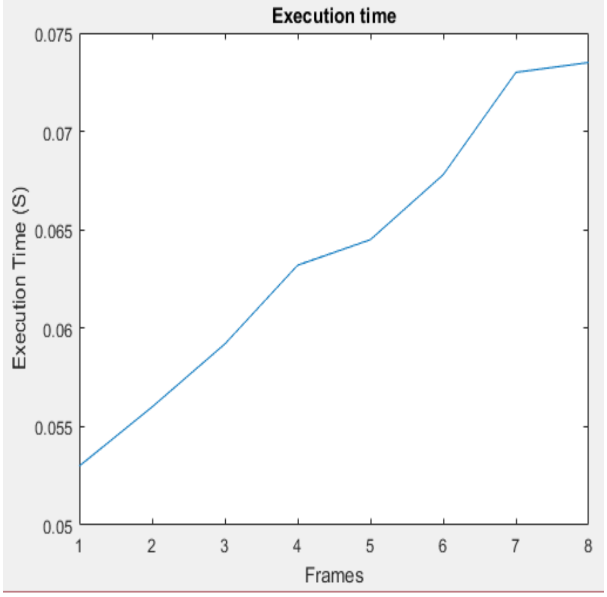

**Fig 5: Execution Time (S) Computation**

Fig 2 and Fig 3 depict average precision and average threshold computation of the anticipated method. This shows better trade-off in contrary to current approaches. Fig 4 illustrates the execution time of the proposed model in terms of seconds. Table I depicts values attained during computation, and Table II shows the execution time of the anticipated model. False-positive and False-negative values are computed with Equation explained above.

## V. CONCLUSION

Here, a novel duplicate video retrieval framework is anticipated for IR features from video content. To acquire a global geometric distribution of information is presented as a holistic representation of the video. Here, PCA and k-NN model is provided along with CNN for reducing dimensionality and measure global descriptors from locally available features. The anticipated model is analyzed in publically accessible datasets for validation; it also proves that this approach outperforms prevailing approaches with near-duplicate VR. However, certain drawbacks are encountered in this approach. Near duplicate, videos are executed in a high speed fast-forwarding, image transformation, and forwarding. This shows a new direction for future research extension to provide better effectiveness in the anticipated system. In the future, temporal correlation is validated with keyframes of similar video to enhance performance.

## ACKNOWLEDGEMENT

This article has been written with financial support of RUSA–Phase 2.0.

## REFERENCES

1. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, “Multiple featurehashing for large real-time scale near-duplicate video retrieval," pp. 423–432, 2011.
2. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, “Effective multiplefeature hashing for large-scale near-duplicate video retrieval,” IEEETrans. on Multimedia, vol. 15, no. 8, pp. 1997–2008, 2013.
3. Liu, L. Huang, C. Deng, J. Lu, and B. Lang, "Multiview complementary hash tables for nearest neighbor search,” pp. 1107–1115, 2015.
4. Shen, F. Shen, Q. Sun, and Y. Yuan, "Multiview latent hashing for efficient multimedia search,” pp. 831–834, 2015.
5. Ranjan, N. Rasiwasia, and C. V. Jawahar, “Multi-label cross-modal retrieval,” pp. 4094–4102, 2015.
6. Chou, H.-T. Chen and S.-Y. Lee, “Pattern-based near-duplicate video retrieval and localization on web-scale videos,” IEEE Trans. On Multimedia, vol. 17, no. 3, pp. 382–395, 2015.
7. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, “A survey on learning to hash,” arXiv preprint ar X iv:1606.00185, 2016.
8. Yang, T. Zhang, and C. Xu, “Cross-domain feature learning in multimedia,” IEEE Trans. on Multimedia, vol. 17, no. 1, pp. 64–78, 2015.
9. Gao, T. Mu, and M. Wang, “Local voting based multiview embedding,” Neurocomputing, vol. 171, pp. 901–909, 2016
10. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. Tao Shen, “Optimal graph learning with partial tags and multiple features for image and video annotation,” pp. 4371–4379, 2015
11. Z. Xu, Y. Yang and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1798-1807.
12. Zheng, S. Wang, J. Wang, and Q. Tian, "AccurateImage Search with Multi-Scale Contextual Evidence," International Journal of Computer Vision, 120(1): pp. 1-13, October 2016.
13. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, "Near-duplicate video retrieval: Current research and future trends," ACM Computing Surveys (CSUR), 45(4): pp. 44, 2013.
14. W. Z., J. S., and H. Q., "Near-duplicate video matching with transformation recognition," in ACM Multimedia, 2009.
15. Zheng, S. Wang, L. Tian, H. Fei, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1741-1750.

## AUTHORS PROFILE

**Gayathri. N** is a Ph.D student in the Department of Computer Applications, Alagappa University, Karaikudi. Tamilnadu, India. Her area of interest includes Video Indexing and Retrieval and Image Processing.

**Mahesh. K** is a Professor in Department of Computer Applications, Alagappa University, Karaikudi, India. He has published many papers in Peer-Reviewed and Reputed Journal and has 28 years of experience in teaching. His research interests are Video Segmentation, Video Processing and Image Processing.