Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

An Automated Approach to Retrieve Lecture Videos using Context Based Semantic Features and Deep Learning

N. Poornima and B. Saleena
School of Computing Science and Engineering, VIT Chennai, Tamil Nadu, India

Key words: Video retrieval, keyframes, features, clustering, deep learning

Corresponding Author:
N. Poornima
School of Computing Science and Engineering, VIT Chennai, Tamil Nadu, India

Page No.: 3514-3525
Volume: 15, Issue 20, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: One of the emerging technologies in applications like video recording and video compression holding significant importance over the years is video digitalization. Video retrieval is a popular research topic and various techniques are available in literature for the effective retrieval of videos. This research work presents a deep learning strategy based video retrieval scheme. Initially, the video archive is subjected for the key frame extraction, for extracting useful keyframes from the video. Then, the features have been extracted from the Keyframe and formulated as the feature database. The features are subjected for clustering using the Fuzzy C Means (FCM) algorithm. Then, clustered features have been provided to the deep learner for finding the optimal centroid for the incoming user query. For the experimentation, the research has considered videos from different category and both the text query and the video query have been used for the retrieval. Results from simulations demonstrate the efficiency of the proposed deep learning strategy in video retrieval and its achievement of improved values of 0.98 and 0.9743, respectively for recall, precision and F-measure.

INTRODUCTION

Digitization of the lecture contents by recording it in the form of videos has helped universities and colleges in the improvement of teaching skills[1]. Students prefer learning materials that are in the video format as they are easily available in the online platforms. Thus, the lecture videos have improved the robustness of the study material[2]. Video is the combination of the text, image and sound and hence, the use of the lecture videos as the study material enables the study experience live for the students[3]. Also, universities can upload the study material in their portals to make it readily available for the students. Using lecture videos as study material has been in vogue in recent years in most of the universities. Some colleges record the presentation of the lecturer and upload it in the internet platform[4]. Direct recording of the presentations may increase the multimedia content in the internet and it is extremely difficult for the students to find the actual content from large source of information[5]. The ever increasing demand for lecture videos has given rise to video retrieval system that analyzes the large database and retrieves similar video content related to the user query from the library. As the video archives in the internet are very large, retrieval of similar video contents is a complex task[6, 7].

Most of the video retrieval schemes use the search function for retrieving the video contents. As the size of the video database is large, retrieving the similar videos without the search function is nearly impossible. Also, the user finds it difficult to ascertain the correctness of the retrieved video contents without opening the video file[8]. Sometimes, the video may cover only a very little portion related to the query rather than providing any additional information, making the video retrieval process unsuccessful. Hence, it is necessary to build a video retrieval system for the user with appropriate file contents[9, 10]. Video search engines such as YouTube, Bing and Vimeo, reply with the video files based on the title, genre, person, etc. Most of the metadata information are created manually by the user and may mislead the contents sometimes. Hence, the video retrieval system should automatically generate the metadata based on the contents, to improve the quality of the retrieval process[5]. Literature has provided two different schemes for retrieving the contents of the video. They are described as manual and automatic schemes. The manual approach is considered to be more accurate than the automatic scheme but requires a longer time and cost for the retrieval process[11]. The automatic scheme uses the low level video analysis approach for analysing the contents of the video[12].

The video file is a combination of several text files, audios and images and hence, there is the need for video retrieval to extract the related feature contents for video retrieval[13]. The video retrieval scheme extracts the features from the video through several techniques such as Text-based, Audio-based, Metadata-based and Content-based techniques[14]. Use of metadata type feature extraction helps retrieval of information relating to the type, the title, date, etc., from the video. Meanwhile, the text based technique extracts the text available in the video through the Optical Character Recognition (OCR) based scheme. Audio based scheme uses different speech recognition based techniques in the feature extraction for retrieving the audio contents of the video file. The content based feature extraction technique is said to be a combination of all the above mentioned techniques[15]. Content Based Video Retrieval (CBVR) is considered to be most successful technique for retrieving the videos from large video archives. The CBVR technique retrieves the lecture video contents with a smaller number of keywords. The term 'content' in the CBVR may refer to color, texture, text, or audio. Also, the CBVR technique responds to the image query provided by the user. Despite the CBVR technique having proved its utility for retrieval of the digital video contents, increase in the volume of media contents in the internet has made the retrieval process a complete task. Using more digital libraries or repositories may improve the video retrieval process[16]. By Han *et al*.[17], On-the-fly Video Retrieval has been presented for retrieving the video contents.

This study proposes a video retrieval strategy using the deep learning based scheme. Initially, the key frames from the input video frames are generated and then, the feature database is constructed by extracting the keywords, semantic words, contextual features, together with the image texture which is extracted using Local Directional Pattern (LDP). The features extracted are clustered using Fuzzy C Means (FCM) for the indexing and are used for training the deep learner with respect to the relevant clusters. Finally, the features are given as input to the trained deep learning for the output query to find the relevant cluster or relevant videos.

The major contribution of this research work is the design and development of the deep learning based video retrieval strategy for the retrieval of the lecture videos from a large database. This works specifically extracts the contextual features along with other features for retrieval purpose.

Literature review: An automatic video indexing approach for retrieving video contents was developed by Yang and Meinel[18]. They have adopted techniques such as OCR and automatic speech recognition, for retrieving the features from the database. The feature extraction extracts useful keyframes for the retrieval process. Even though the technique has an improved performance, the presence of noise reduces the overall performance. Li *et al*.[19] have presented an automated video retrieval system for capturing and detecting similar videos in the video archive. The retrieval was done by analysing the text contents and the keywords in the video. The system yielded reduced performance while using multi-videos. Baidya and Goel[20] have proposed an automated video retrieval scheme using the OCR technique. Using the OCR, the important information was extracted from the video and further, the technique collected the embedded information in the video slides. The scheme had reduced performance with the high recall rate. Nguyen *et al*.[21] have presented a video retrieval scheme with document analysis. The scheme has adopted the text detection and graphic localization methods for extracting the keywords from the video. The scheme performed well while using the multimodal and cross-modal videos. The system exhibited errors in retrieval.

Araujo and Girod[22] have proposed an asymmetric comparison technique for video retrieval. The scheme explored the database by incorporating Fisher vectors. The technique works in a flexible retrieval environment. Rahmani and Zargari[23] have proposed a feature vector for the video retrieval process which involves analysis of the motion structure of the video sequences. Besides

improved results, the scheme faces complexity issues during the analysis of large video contents. Lin *et al*.[24] presented the deep learned global descriptors for the video retrieval. The deep learned global descriptors depend on the invariance theory. The authors have further proposed the Nested Invariance Pooling (NIP) scheme for analyzing the pooled descriptors in the video. The scheme has complementary effects on the handcrafted descriptors. Rouhi and Thom[1] have proposed the CBVR scheme using different encoding profiles. The scheme has analyzed the effects of the encoding scheme during the retrieval process. The scheme provided improved tolerance and robustness towards the noise and different encoding types.

Challenges: Challenges in developing the retrieval system for lecture videos are as follows:

The critical challenge is the recognition of the teaching topic by the retrieval system. Unavailability of the teaching topic may increase the complexity of the retrieval process[12].

The lecture video files have low level correlation among the features of different videos and hence, it is more challenging to retrieve the lecture video compared to other video files[12].

Some works have adopted the CNN based global descriptors for video retrieval but they face many challenges. The initial challenge in adopting the CNN based method for the video retrieval is the absence of the invariance in CNN method while geometric transformations occur in the input image. The geometric variations in the image include rotation of the image in the consecutive frame[24].

Another challenge confronting the CNN based method for the video retrieval is the performance degradation for rotated image query due to the global descriptors[24]. Use of the conventional hand crafted descriptors for the video retrieval has reduced the performance of the technique[24] as they are robust towards the scale and rotation changes occurring in the 2D plane.

MATERIALS AND METHODS

Proposed deep learning based video retrieval scheme: This section explains the proposed video retrieval scheme with the deep learning strategy. Lecture video retrieval designed in this work includes feature extraction, clustering and deep learning. Figure 1 shows the architecture of the proposed lecture video retrieval using deep learning technique.

Key frames are extracted from the videos present in the database. After extracting the key frames from the video, the features such as words, semantic words, context words and LDP features are extracted and the feature database is created. Then, the database is subjected to the FCM clustering and finally, DBN that gets trained with the cluster centroids is used for the retrieval of the video. While the user gives a search query arrives in to the video retrieval system, the above mentioned features are extracted from the query and given to the DBN for testing. The DBN classifier identifies the optimal cluster belonging to the query and the videos related to the optimal cluster are retrieved by the proposed deep learning based video retrieval scheme.

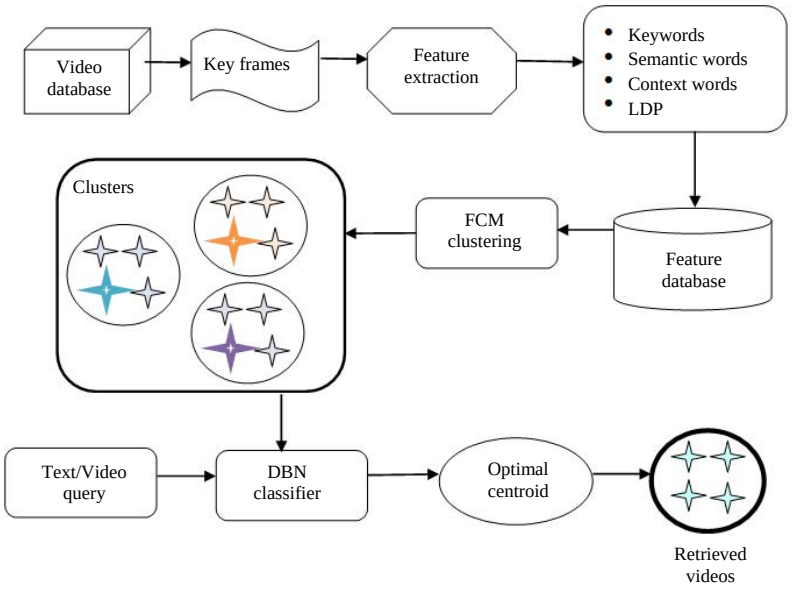

Fig. 1: Architecture of the proposed deep learning based video retrieval strategy

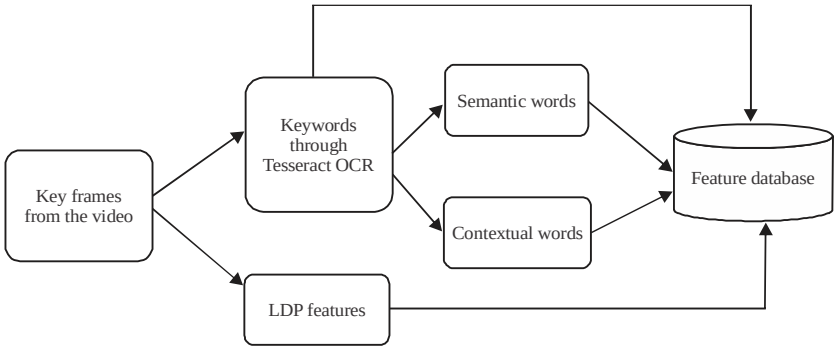

Fig. 2: Extraction of the features from the keyframes

Extraction of the key frame from input video: The initial stage in the proposed deep learning based video retrieval system is the extraction of the keyframes from the video. For the experimentation, this work considers the video archive with number of lecture video contents, expressed as:

$$D = \{R_1, R_2, ..., R_i, ..., R_V\} \quad (1)$$

where R_i, represents the ith lecture video in the video archieve and V refers to the total lecture videos in the database. The lecture videos have substantial information for processing. Use of all the data in the video for the processing makes the video retrieval more complex. Hence, this work extracts some of the keyframes in the video for the analysis. Consider the video R_i having K number of keyframes and then after the keyframe extraction, the video R_i is represented as:

$$R_i = \{R_i^1, R_i^2, .., R_i^k, ..., R_i^K\} \quad (2)$$

where, R^k_i represents the kth keyframe in the video R_i. Keyframe extraction improves the video retrieval process as most of the important image textures and features are present in the keyframe.

Feature extraction: The next major step in the proposed deep learning based video retrierval process is the extraction of the features from the keyframes of the videos. Various features extracted from the keyframes are shown in Fig. 2.

Extracting the keywords using the OCR technique: The keyframes in the video contain several keywords useful for video retrieval. Here, the OCR[14] is used for retrieving the keywords from the keyframes. Consider there are w numbers of keyword on a keyframe and the extracted keywords through the OCR technique are represented as:

$$s = \{s_1, s_2, , ..., s_z, ..., s_w\} \quad (3)$$

where, S_z indicates the Zth keyword in the keyframe.

Semantic words: Semantic words are extracted from the keywords after the identification of this keywords. The semantic words are the keywords having the similar synonym and it they are expressed as:

$$L = \{L_1^1, L_2^2, ..., L_w^v\} \quad (4)$$

where, L^1_1 indicates the semantic words for the keyword and the semantic words are expressed as follows:

$$L_1^1 = \{s_{w1}^1, s_{w2}^1, ..., s_{wn}^1\} \quad (5)$$

Contextual words: Contextual words are a collection of the frequently occurring keywords in the keyframe. Consider the keyword S_z occurs d number of times in the keyframe and the contextual words collect the more frequently occurring keywords in the particular keyframe. The contextual words are expressed by the following equation:

$$U = \{u_1, u_2, \ldots u_q\} \quad (6)$$

Where:
u_q = The contextual words in the keyframes
q = The total contextual words in the keyframe

Local directional pattern: The LDP features signify the direction of the pixels of the individual video frames. The LDP features are extracted by applying 8 masks to the keyframe. The LDP features[23] are extracted from the keyframe R^k_i. The masks are applied in reference to the centre point of the pixel. Consider the keyframe R^k_i has the centre pixel as (h_c, k_c) and the LDP features are obtained from the keyframe as:

$$LDP(h_c, k_c) = \sum_{f=0}^{7} o(r_f - r_g) 2^f \quad (7)$$

where, r_f indicates the fth mask used for obtaining the LDP feature.

Clustering the features; FCM algorithm: The features extracted from the keyframes are formulated as the feature database F_D. The next major task in the video retrieval process is the clustering of the features into G number of clusters, is carried out with the use of FCM algorithm. The mathematical formulation of the FCM algorithm[25] is as follows.

The initial step in the FCM clustering is the formulation of the fuzzy matrix, by computing the Euclidean distance measure. The fuzzy matrix is expressed as follows:

$$V = \sum_{p=1}^{s} \sum_{e=1}^{l} J_{pe}^{r} O_{pe}; \quad 1 \le r \le \infty \quad (8)$$

Where:
r = The fuzziness variable
O_{pe} = The Euclidean distance measure

and it is measured as:

$$O_{pe} = \left\| y_p - X_e \right\| \quad (9)$$

where, X_e refers to the cluster center and it is expressed as:

$$X_e = \frac{\sum_{p=1}^{s} J_{pe}^{r} y_p}{\sum_{p=1}^{s} J_{pe}^{r}} \quad (10)$$

The cluster center modifies the fuzzy matrix and it is expressed as:

$$J_{je} = \frac{1}{\sum_{l=1}^{j} \left(\frac{O_{pe}}{O_{le}} \right)^{\frac{2}{r-1}}} \quad (11)$$

FCM executes for the finite interval of time and finds the optimal centroid for the clustering. The centroids calculated through the FCM algorithm are expressed as follows:

$$C = \{C_1, C_2, \ldots, C_j \ldots C_G\} \quad (12)$$

Where:
C_j = The jth cluster centroid
G = Total number of clusters

Video retrieval using deep learning: Finally, after clustering the features using the FCM approach, the clustered features are fed to the DBN classifier for video retrieval. DBN[26] gets the clustered features from the FCM and tries to find suitable cluster matching with the query from the user. The proposed deep learning scheme performs the video retrieval using the training and the testing steps. Figure 3 presents the architecture of the proposed deep learning based video retrieval scheme.

As depicted in Fig. 3, the proposed deep learning system has three layers, namely RBM layer 1, 2 and MLP layer. The centroids of the clusters are fed as training input to the proposed deep learning scheme, the DBN finds the optimal centroid related to the query. The three layers of the DBN are interconnected with each other, with the output of one layer fed to the consecutive layer. The expression for the various layers is briefly described below.

The RBM layer 1 has visible neurons and hidden neurons for the processing. Both the input and the hidden layers of the RBM layer 1 are represented as follows:

$$A^1 = \{A_1^1, A_2^1, \ldots, A_j^1, \ldots, A_G^1\}; 1 \le j \le G \quad (13)$$

$$S^1 = \{S_1^1, S_2^1, \ldots, S_x^1, \ldots, S_y^1\}; 1 \le x \le y \quad (14)$$

where, A^1_j and S^1_k represent the jth visible neuron and xth hidden neuron of the RBM layer 1. The terns G and y refer to the total number of input and hidden neurons in the RBM layer 1. The input layer of the RBM 1 is fed with the features of the centroid of each cluster. The visible and the hidden layer of the RBM contains the biases which are represented as:

$$a^1 = \{a_1^1, a_2^1, \ldots, a_j^1, \ldots, u_G^1\} \quad (15)$$

$$b^1 = \{b_1^1, b_2^1, \ldots, b_x^1, \ldots, b_y^1\} \quad (16)$$

Where:
a^1_j = The bias corresponding to the jth visible layer
b^1_x = Corresponds to the bias of the xth hidden layer of RBM 1

The weights between the visible and the hidden neurons are given by the following expression:

$$Z^1 = \{Z_{jx}^1\}; 1 \le j \le G; 1 \le x \le y \quad (17)$$

where, Z^1_{jx} refers to the weight present amidst the jth visible and xth hidden neuron. For computing the output of the RBM layer 1, the bias present in the hidden units, feature inputs and the weights are used. The following expression indicates the output for the RBM layer 1 obtained through the hidden units:

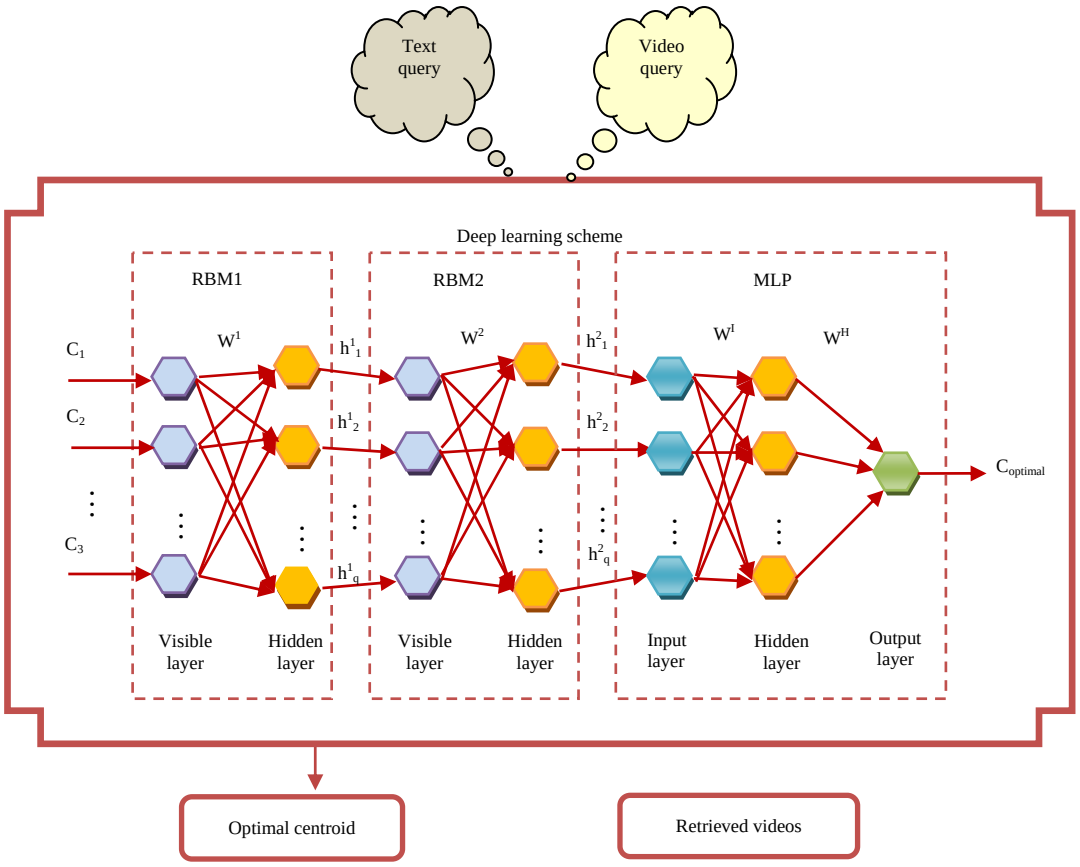

Fig. 3: Proposed deep learning scheme for the video retrieval

$$S^1_x = \sigma\left[b^1_1 + \sum_j A^1_j Z^1_{jx}\right] \quad (18)$$

where, σ indicates the activation function for computing the output of the RBM 1 which is expressed by the following expression:

$$S^1 = \left\{S^1_x\right\}; 1 \le x \le y \quad (19)$$

After computing the output at the RBM layer 1, the output is fed as the input to the visible units of the RBM layer 2. Thus, it is necessary to provide a similar number of visible units in RBM 2 as the output layer of RBM 1. The expression for the input of the RBM 2 is represented as follows:

$$A^2 = \left\{A^2_1, A^2_2, \ldots, A^2_G\right\} = \left\{S^1_x\right\}; 1 \le x \le y \quad (20)$$

where, S^1_x indicates the output of the xth layer in RBM 1. The hidden unit present in the RBM 2 is present as follows:

$$S^2 = \left\{S^2_1, S^2_2, \ldots, S^2_x, \ldots, S^2_y\right\}; 1 \le x \le y \quad (21)$$

Similar to the RBM layer 1, the RBM 2 has biases in both the input and the hidden units. The bias corresponding to the RBM layer 2 is represented as a^2 and b^2. The weight vector of the RBM layer 2 is given as follows:

$$Z^2 = \left\{Z^2_{xx}\right\}; 1 \le x \le y \quad (22)$$

where, Z^2_{xx} implies the weight between xth the hidden unit and the xth visible unit of the RBM layer 2. The expression for the output from the RBM layer 2 is represented as follows:

$$S^2_x = \sigma\left[b^2_x + \sum_j A^2_j W^2_{xx}\right] \forall A^2_j = S^1_x \quad (23)$$

where, b^2_x refers to the bias present in the xth hidden unit. The hidden unit of the RBM layer 2 is represented as follows:

$$S^2 = \left\{S_x^2\right\}; 1 \le x \le y \tag{24}$$

The output of the RBM layer 2 is directly fed to the MLP layer and thus, the visible units present in the MLP are represented by the following equation:

$$M = \left\{M_1, M_2, \ldots, M_x, \ldots, M_y\right\} = \left\{S_x^2\right\}; 1 \le x \le y \tag{25}$$

where, M_x indicates the xth input unit of the MLP layer. The hidden layer units of the MLP layer are expressed by the following equation:

$$N = \left\{N_1, N_2, \ldots, N_m, \ldots, N_n\right\}; 1 \le m \le n \tag{26}$$

where, n indicates the total number of units in the MLP layer. The MLP layer finds the optimal centroid among the input centroids and thus, the output of the MLP layer comprises of only one layer and it is represented as:

$$C = \left\{C_{optimal}\right\} \tag{27}$$

The output of the MLP layer depends on the weights in the input and the hidden units. The weight vector corresponding to the input unit of the MLP is expressed as:

$$Z^I = \left\{Z_{xm}^I\right\}; 1 \le x \le y; 1 \le m \le n \tag{28}$$

where, Z_{xm}^I indicates the weight between xth the input unit and mth hidden unit of the MLP layer. Now, the expression of the hidden unit in the MLP layer is expressed as:

$$N_m = \left[\sum_{x=1}^{y} Z_{xm}^I * M_x\right] Y_m \forall b_l = S_x^2 \tag{29}$$

where, Y_m indicates the bias present in the input unit. The expression for the weights present in the hidden unit of the MLP layer is expressed as:

$$Z^H = \left\{Z_m^H\right\}; 1 \le m \le n \tag{30}$$

The final output of the MLP layer depends on the hidden layer output and the weights present in the hidden unit. The expression for the MLP layer output is expressed as:

$$C_{optimal} = \sum_{m=1}^{n} Z_m^H * N_m \tag{31}$$

where, Z_m^H is the weight of the hidden unit of the MLP.

Training phase: This section presents the training procedure relating to the DBN. For the training, the features representing the centroids of the clusters are given as the training input to the RBM layer 1. As the proposed deep learning scheme involves the RBM and the MLP layer, each layer is trained using different algorithms. The training procedure identifies the suitable weights for the testing process.

Training of RBM layers: Initially, the RBM layer is fed with the centroid features for the training. The RBM layer 1 is consecutively connected with the RBM layer 2 and hence, the RBM 1's output serves as the input for the RBM 2. The training procedure aims to identify the optimal weights for both the RBM layers. The training of RBM layer is bound by the existing back propagation algorithm.

Training of MLP: After training the RBM layers, input of the MLP layers are taken from the output of RBM layers. In this work, the MLP layer is trained by the existing gradient descent algorithm which provides the expression for the weight update. After training, the optimal weight is identified for the input and the hidden layers for the MLP. The training procedure for the MLP layer is defined below:

In the initial step, the weights of the input layer Z^I and the hidden layer Z^H are chosen randomly based on expression (Eq. 29) and (31).

Input of the MLP layer depends on the output of the RBM 2 layer, the input sample $\{S_x^2\}$ is fed as the training input. In the next step, the output of the MLP layer, $C_{optimal}$ is found based on Eq. 27.

The training procedure finds the weight based on the minimal error such that the weight providing minimal output error is considered to be the optimal weight. Hence, the average error is computed based on following expression:

$$E_{avg} = \frac{1}{V}\sum_{i=1}^{V}\left(C_{optimal}^i - T^i\right)^2 \tag{32}$$

Where:
$C_{optimal}^i$ = The output of the MLP
T^i = The target response

In the next step, the partial derivative of the weights in the input and hidden units of the MLP is computed by the following expression:

$$\Delta Z_{xm}^I = -\eta\frac{\partial E_{avg}}{\partial Z_{xm}^I} \tag{33}$$

$$\Delta Z_m^H = -\eta\frac{\partial E_{avg}}{\partial Z_m^H} \tag{34}$$

where, η refers to the learning rate of the gradient descent algorithm. The weight update for the input and hidden units are calculated by the gradient descent algorithm is expressed below:

$$Z^{I}_{xm(G)}(t+1) = Z^{I}_{xm}(t) + \Delta Z^{I}_{xm} \quad (35)$$

$$Z^{H}_{m(G)}(t+1) = z^{H}_{m}(t) + \Delta Z^{H}_{m} \quad (36)$$

where $Z^{I}_{xm}(t)$ and $Z^{H}_{m}(t)$ indicate the weights in input and hidden unit of the MLP layer. From the newly computed weight, compute the MLP layer output and the corresponding average error using the expression (Eq. 32). The steps are repeated until the optimal weight with minimal average error is found.

Testing phase: Consider the user query Q provided to the system, the query from the user as a video query or the text query. After the query is given to the video retrieval system, the features are pull out from the query and provided as input to the RBM layer 1. The proposed deep learning based video retrieval system analyzes the feature input of the query and identifies the optimal centroid $C_{optimal}$ suitable for the query. After finding the suitable cluster for the query, all the video contents belonging to the optimal centroid $C_{optimal}$ are retrieved by the proposed deep learning based video retrieval system.

RESULTS AND DISCUSSION

This study briefly explains the simulation results achieved by the proposed lecture video retrieval using deep learning scheme. Simulation is done by choosing different query videos and the results are evaluated based on metrics namely recall, precision and F-measure.

Experimental setup: Experimentation of the proposed video retrieval using deep learning based strategy is done in the MATLAB tool. Further, the implementation of the entire proposed scheme is done in the PC with Windows 10 OS, 4 GB RAM and Intel I3 processor.

Database description: Experimentation of the proposed video retrieval using deep learning scheme is carried out by considering 60 videos of 6 categories. Each category include 10 video contents and hence most commonly used for the video retrieval. The categories include agriculture, India 2020, pollution, quantum optics, etc.

Performance metrics: For the evaluation of the proposed deep learning technique for the lecture video retrieval, three evaluation metrics, namely recall, precision and F-measure are considered. The expression for the evaluation metrics is stated below:

Precision: Precision refers to the ratio of the total number of relevant videos retrieved by the proposed classifier to the total count of relevant and irrelevant videos in the database.

F-measure: F-measure is defined as the measure of harmonic mean of the recall and the precision metrics.

Recall: Recall refers to the ratio of relevant videos retrieved by the classification model to the total number of relevant videos present in the database.

Comparative techniques: A comparative analysis is performed between various techniques namely k-NN+ OCR, k-NN,+WOCR, NB+OCR and CNB+OCR. The techniques are explained as follows:

k-NN+OCR: Here, the features are extracted using the OCR technique[27] and categorization is carried out using the k-NN technique[28].

k-NN+WOCR: Here, relevant features are extracted using the weighted OCR (WOCR) and the retrieval is carried out with the help of k-NN classifier.

NB+OCR: In this work, retrieval of the videos is done through the NB classifier.

CNB+OCR: Here, video retrieval is done using the CNB classifier[29].

The experimental results achieved through use of the proposed deep learning based video retrieval are presented in Fig 4. Both the video query and the text query have been used for the experimentation.

Figure 4a presents the text query provided by the user, the videos corresponding to the text query retrieved by the proposed scheme are presented in Fig. 4b. Figure 4c presents the video query provided by the user and Fig. 4d presents the retrieved video contents related to video query.

Comparative analysis: Here, a comparative analysis of the proposed methodology is done by considering both the text query and the video query. The analysis is done by varying the total number of retrieval and measured based on metrics namely recall, precision and F-measure.

Comparative analysis based for the video query: Figure 5 shows the comparative analysis of the proposed deep learning scheme for various video queries. Figure 5a presents the comparative analysis of the proposed deep learning based video retrieval scheme based on precision metric. The existing models, namely k-NN+OCR, k-NN +WOCR, NB+OCR and CNB+WOCR have achieved precision values of 0.5346, 0.6985, 0.695 and 0.9262, for the number of retrieval. Meanwhile, the proposed FCM+ DBN scheme has achieved a high precision with the value of 0.9503 for k = 16. Figure 5b shows the comparative analysis of the proposed deep learning based video retrieval scheme based on recall. The existing models,

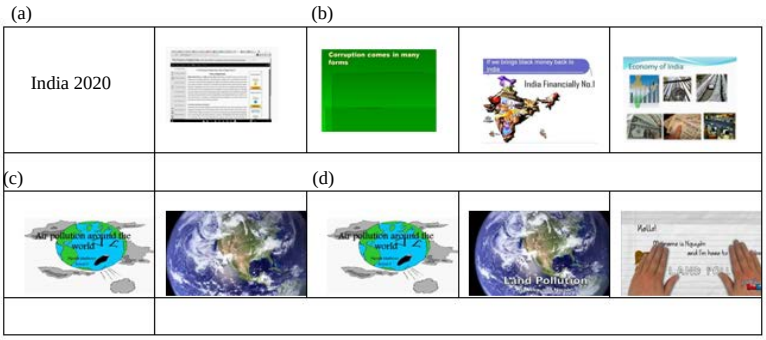

Fig. 4(a-d): Experimental results of the proposed lecture video retrieval using deep learning scheme for, (a) Text query, (b) Retrieved videos based on text query, (c) Video query and (d) Retrieved videos based on video query

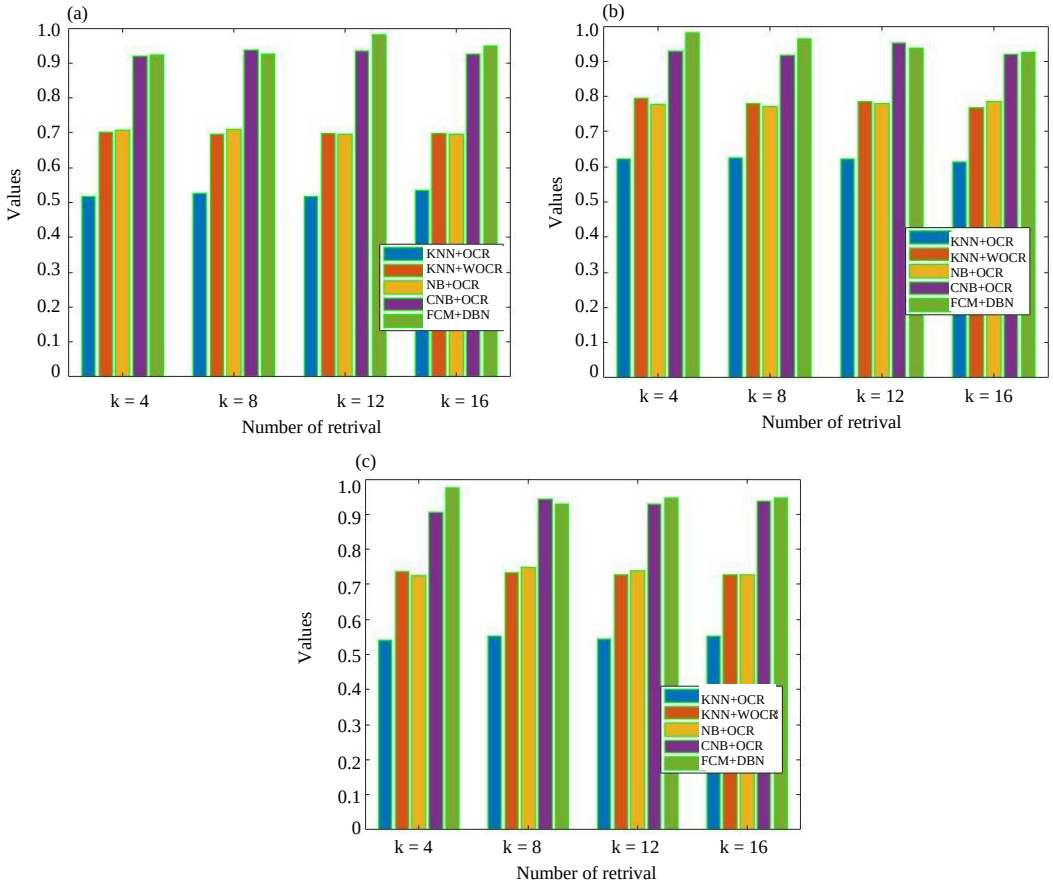

Fig. 5(a-c): Comparative analysis for the video query based on (a) Precision, (b) Recall and (c) F-measure

namely k-NN+OCR, k-NN+WOCR, NB+OCR and CNB+WOCR have achieved recall values of 0.6129, 0.7694, 0.7847 and 0.919, for the number of retrieval. Meanwhile, the proposed FCM+DBN scheme has achieved a high recall with the value of 0.925 for k = 16. Figure 5c presents the comparative analysis of the proposed deep learning based video retrieval scheme based on F measure for the cluster size as 6. The existing models, namely k-NN+OCR, k-NN+WOCR, NB+OCR and CNB+ WOCR have achieved the F-measure values of 0.5526, 0.7278 and 0.9385, for the number of retrieval. However, the proposed FCM+DBN scheme has achieved high F-measure with the value of 0.9463 for k = 16.

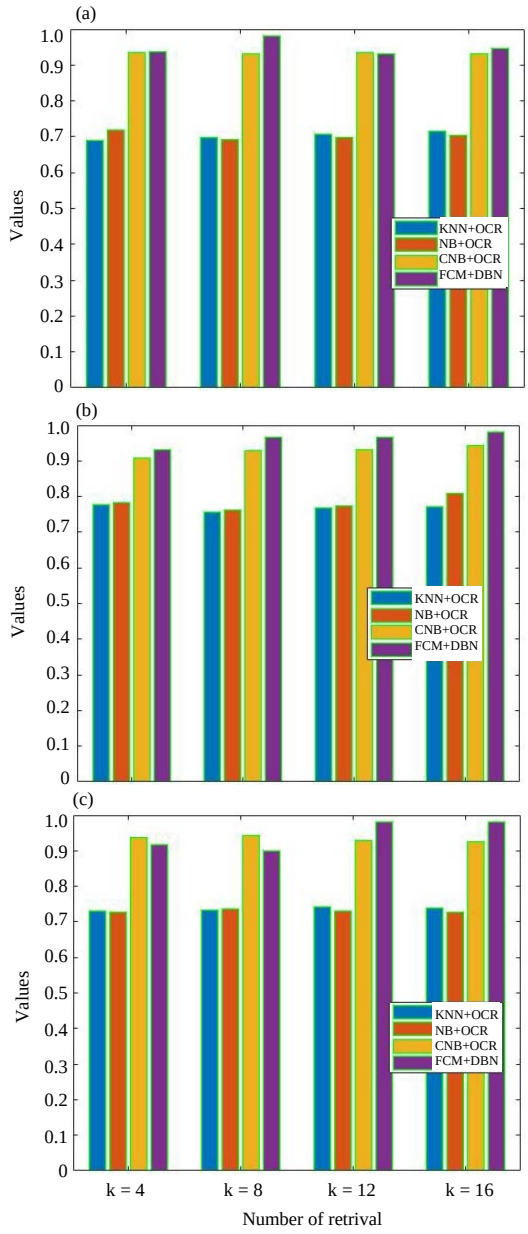

Fig. 6(a-c): Comparative analysis for the text query based on, (a) Precision, (b) Recall and (c) F measure

Comparative analysis based for the text query: Figure 6 shows the comparative analysis of the proposed deep learning scheme for various text queries. Figure 6a shows the comparative analysis of the proposed deep learning based video retrieval scheme based on precision metric. The existing models, k-NN+OCR, NB+OCR and CNB+WOCR have achieved precision values of 0.7156, 0.7029 and 0.9325, for the number of retrieval k = 16. Meanwhile, the proposed FCM+DBN scheme has achieved a high precision with the value of 0.9461 for k = 16. Figure 6b shows the comparative analysis of the proposed deep learning based video retrieval scheme based on recall metric. The existing models, k-NN+OCR, NB+OCR and CNB+WOCR have achieved recall values of 0.7711, 0.8096 and 0.942 for the number of retrieval. Meanwhile, the proposed FCM+DBN scheme has achieved a high recall with the value of 0.98 for k = 16. Figure 6c shows the comparative analysis of the proposed deep learning based video retrieval scheme based on F-measure metric. K-NN+OCR, NB+OCR and CNB+ WOCR have achieved F-measure values of 0.7391, 0.7266 and 0.9271 for the number of retrieval k = 16. Meanwhile, the proposed FCM+DBN scheme has achieved a high F-measure with the value of 0.98 for k = 16.

Table 1: Comparative discussion

	Performance metrics		
Comparative techniques	Precision	Recall	F-measure
k-NN+OCR	0.5188	0.6215	0.541500
k-NN+WOCR	0.6991	0.7936	0.737400
NB+OCR	0.6939	0.7778	0.723300
CNB+OCR	0.9353	0.9282	0.904600
Proposed FCM+DBN	0.9800	0.9800	0.974314

This study shows a comparative discussion of the proposed deep learning scheme as against other existing techniques for the retrieval. The performance of the proposed deep learning based video retrieval strategy is analyzed based on performance metrics namely recall, precision and F-measure. Table 1 presents the best performance of the proposed deep learning based video retrieval strategy.

As depicted in Table 1, the existing CNB+OCR has achieved values of 0.9282, 0.9353 and 0.9046, for recall, precision and F-measure, respectively. The performance of the existing models is not compactable enough for the video retrieval. The comparative analysis depicts the achievement of improved performance of the proposed FCM+DBN Model with values of 0.98 and 0.9743 for recall, precision and F-measure, respectively. The proposed deep learning scheme with FCM+DBN has improved the video retrieval process.

CONCLUSION

A video retrieval strategy using the deep learning scheme was developed. The videos in the video archive have been subjected to the key frame extraction process and the necessary keyframes have been extracted. Then, the features such as keywords, semantic words, contextual words and LDP features were extracted from the keyframes and formulated as the database. Then, using the FCM algorithm, the features were clustered into different groups. Then, features representing the cluster centroid were provided to the DBN for training and the DBN found the optimal centroid related to the query. For

the experimentation, the research has considered videos from different category and both the text query and video query were used for the retrieval. The performance of the proposed deep learning scheme for the video retrieval was compared with that of the various existing works and measured based on metrics such as precision, recall and F-measure. Results from the simulation depicts that the proposed deep learning strategy have proved to be efficient in video retrieval which has shown achievement of improved values of 0.98 and 0.9743 for recall, precision and F-measure, respectively.

REFERENCES

01. Rouhi, A.H. and J.A. Thom, 2018. Encoder settings impact on intra-prediction-based descriptors for video retrieval. J. Visual Commun. Image Represent., 50: 263-269.
02. Song, J., L. Gao, L. Liu, X. Zhu and N. Sebe, 2018. Quantization-based hashing: A general framework for scalable image and video retrieval. Pattern Recognit., 75: 175-187.
03. Hao, Y., T. Mu, R. Hong, M. Wang, N. An and J.Y. Goulermas, 2016. Stochastic multiview hashing for large-scale near-duplicate video retrieval. IEEE. Trans. Multimedia, 19: 1-14.
04. Fernandez-Beltran, R. and F. Pla, 2015. Incremental probabilistic latent semantic analysis for video retrieval. Image Vision Comput., 38: 1-12.
05. Yang, H., M. Siebert, P. Luhne, H. Sack and C. Meinel, 2011. Lecture video indexing and analysis using video OCR technology. Proceedings of the 2011 7th International Conference on Signal Image Technology & Internet-Based Systems, November 28-December 1, 2011, IEEE, Dijon, France, pp: 54-61.
06. Fernandez-Beltran, R. and F. Pla, 2016. Latent topics-based relevance feedback for video retrieval. Pattern Recogn., 51: 72-84.
07. Furini, M., 2018. On introducing timed tag-clouds in video lectures indexing. Multimedia Tools Appl., 77: 967-984.
08. Liang, B., W. Xiao and X. Liu, 2012. Design of video retrieval system using MPEG-7 descriptors. Procedia Eng., 29: 2578-2582.
09. Kanadje, M., Z. Miller, A. Agarwal, R. Gaborski, R. Zanibbi and S. Ludi, 2016. Assisted keyword indexing for lecture videos using unsupervised keyword spotting. Pattern Recogn. Lett., 71: 8-15.
10. Roy, S., P. Shivakumara, N. Jain, V. Khare, A. Dutta, U. Pal and T. Lu, 2018. Rough-fuzzy based scene categorization for text detection and recognition in video. Pattern Recogn., 80: 64-82.
11. Memar, S., L.S. Affendey, N. Mustapha, S.C. Doraisamy and M. Ektefa, 2013. An integrated semantic-based approach in concept based video retrieval. Multimedia Tools Applic., 64: 77-95.
12. Furini, M. and S. Mirri, 2017. On using on-the-fly students notes in video lecture indexing. Proceedings of the 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC), January 8-11, 2017, IEEE, Las Vegas, Nevada, pp: 1083-1088.
13. Fujii, A., K. Itou and T. Ishikawa, 2006. Lodem: A system for on-demand video lectures. Speech Commun., 48: 516-531.
14. Wu, S., H. Song, G. Cheng and X. Zhong, 2019. Civil engineering supervision video retrieval method optimization based on spectral clustering and R-tree. Neural Comput. Appl., 31: 4513-4525.
15. Chivadshetti, P., K. Sadafale and K. Thakare, 2015. Content based video retrieval using integrated feature extraction and personalization of results. Proceedings of the 2015 International Conference on Information Processing (ICIP), December 16-19, 2015, IEEE, Pune, India, pp: 170-175.
16. Chou, C.L., H.T. Chen and S.Y. Lee, 2015. Pattern-based near-duplicate video retrieval and localization on web-scale videos. IEEE. Trans. Multimedia, 17: 382-395.
17. Han, X., B. Singh, V.I. Morariu and L.S. Davis, 2017. VRFP: On-the-fly video retrieval using web images and fast fisher vector products. IEEE. Trans. Multimedia, 19: 1583-1595.
18. Yang, H. and C. Meinel, 2014. Content based lecture video retrieval using speech and video text information. IEEE. Trans. Learn. Technol., 7: 142-154.
19. Li, K., J. Wang, H. Wang and Q. Dai, 2014. Structuring lecture videos by automatic projection screen localization and analysis. IEEE. Trans. Pattern Anal. Mach. Intell., 37: 1233-1246.
20. Baidya, E. and S. Goel, 2014. LectureKhoj: Automatic tagging and semantic segmentation of online lecture videos. Proceedings of the 2014 7th International Conference on Contemporary Computing (IC3), August 7-9, 2014, IEEE, Noida, India, pp: 37-43.
21. Nguyen, N.V., M. Coustaty and J.M. Ogier, 2014. Multi-modal and cross-modal for lecture videos retrieval. Proceedings of the 2014 22nd International Conference on Pattern Recognition, August 24-28, 2014, IEEE, Stockholm, Sweden, pp: 2667-2672.
22. Araujo, A. and B. Girod, 2017. Large-scale video retrieval using image queries. IEEE. Trans. Circuits Syst. Video Technol., 28: 1406-1420.

23. Rahmani, F. and F. Zargari, 2018. Temporal feature vector for video analysis and retrieval in high efficiency video coding compressed domain. Electron. Lett., 54: 294-295.
24. Lin, J., L.Y. Duan, S. Wang, Y. Bai and Y. Lou *et al*., 2017. Hnip: Compact deep invariant representations for video matching, localization and retrieval. IEEE. Trans. Multimedia, 19: 1968-1983.
25. Bezdek, J.C., R. Ehrlich and W. Full, 1984. FCM: The fuzzy *c*-means clustering algorithm. Comput. Geosci., 10: 191-203.
26. Hinton, G.E., 2009. Deep belief networks. Scholarpedia, Vol. 4, No. 5.
27. Patel, C., A. Patel and D. Patel, 2012. Optical character recognition by open source OCR tool tesseract: A case study. Int. J. Comput. Appl., 55: 50-56.
28. Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman and A.Y. Wu, 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE. Trans. Pattern Anal. Mach. Intell., 24: 881-892.
29. Poornima, N. and B. Saleena, 2018. Multi-modal features and correlation incorporated naive bayes classifier for a semantic-enriched lecture video retrieval system. Imaging Sci. J., 66: 263-277.