J. Vis. Commun. Image R. 104 (2024) 104309

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Full Length Article

Optimized deep learning enabled lecture audio video summarization☆

Preet Chandan Kaur[a,*], Dr. Leena Ragha[a,b]

[a] Computer Engineering/Faculty, Ramrao Adik Institute of Technology, D.Y PATIL Deemed to be a University, Nerul, Navi Mumbai 400706, Maharashtra, India
[b] Department of Computer Science and Engineering, BLDEA's V.P. Dr. P. G. Halakatti College of Engineering Technology, Ashram Road, Vijayapur 586103, Karnataka, India

ARTICLE INFO

Keywords:
Audio Video Summarization
Deep Residual Network
Video Shot Segmentation
YCbCr Space Colour Model
E-learning

ABSTRACT

Video summarization plays an important role in multiple applications by compressing lengthy video content into compressed representation. The purpose is to present a fine-tuned deep model for lecture audio video summarization. Initially, the input lecture audio-visual video is taken from the dataset. Then, the video shot segmentation (slide segmentation) is done using the YCbCr space colour model. From each video shot, the audio and video within the video shot are segmented using the Honey Badger-based Bald Eagle Algorithm (HBBEA). The HBBEA is obtained by combining the Bald Eagle Search (BES) and Honey Badger Algorithm (HBA). The DRN training is executed by HBBEA to select the finest DRN weights. The relevant video frames are merged with the audio. The proposed HBBEA-based DRN outperformed with a better F1-Score of 91.9 %, Negative predictive value (NPV) of 89.6 %, Positive predictive value (PPV) of 90.7 %, Accuracy of 91.8 %, precision of 91 %, and recall of 92.8 %.

1. Introduction

In today's era, with the advancement in camera, internet, storage and display technologies, video plays a vital role in everyday life and finds many applications that use video as a communication medium. The raw original records of the video are generally affected by various noises and unwanted audio and/or video frames. Sports analysts, medical professionals, and researchers also employ video summarization for focused analysis in their respective domains Furthermore, entertainment, traffic management, crisis response, and environmental monitoring all leverage video summarization to extract crucial insights from extensive video datasets [39], thereby optimizing decision-making and enhancing user experiences across diverse fields. In the field of education, video summarization can play a key role in enhancing the learning process [15]. Most of the time, we're exposed to raw education videos with different things that affect the learning process. These videos haven't been changed in any way, so they are called "raw education videos.".

Video summarization has attracted significant interest due to its valuable ability to enhance video browsing [40]. Video summarization is making shorter, more interesting versions of long video lessons, which makes it easier for students to understand difficult ideas and quickly review important points. This not only saves time and storage space, because they can find short summaries that fit their needs [17]. The quality of education on an e-learning platform can be improved by making better summaries from long videos. So, summarizing helps students learn better, save time, and understand complex topics more easily [18]. CPSP [32] refines audio-visual features, making them distinguishable from negatives enhancing the features benefit of AVE [33] classification.

There's a growing need for automated methods to create summaries of educational lecture videos. In our work, we're also looking at the audio part of the videos. In educational videos, it's crucial to sync the images with the audio [16,38] to find the important parts and create a summarized version of the lecture video. In online classes, the lecture videos are stored on servers, and students can watch them at their place. Static summarization makes a sequence of pictures that show the most important moments in a video. It's precise but doesn't use sound, just the pictures [19]. Dynamic summarization, on the other hand, creates short videos that give you a quick idea of what the video is about. It uses both the video and its sound. To make these summaries, we use different methods like breaking the video into smaller parts, understanding what's happening in the video, and how things move [20 21]. Classical techniques for video summarization work by analyzing short parts of

☆ This paper has been recommended for acceptance by Kuo-Liang Chung.

* Corresponding author.
E-mail addresses: preetchandan.kaur@rait.ac.in (P. Chandan Kaur), cse.leenaragha@bldeacet.ac.in (Dr.L. Ragha).

https://doi.org/10.1016/j.jvcir.2024.104309
Received 18 January 2024; Received in revised form 10 September 2024; Accepted 5 October 2024
Available online 9 October 2024

videos to find the most important parts with the least repetition. They use things like visual characteristics and how often something appears to figure this out. Some methods even use high-level features like objects and people, but creating detectors for all these things can be hard, which is why we use video summarization techniques [22].

1.1. Motivation

The explosion of video content has made it challenging to review and analyse all that video data. Especially when dealing with long lecture videos, using computer vision methods to detect and classify events can become slow and demanding because they have to process every single frame of the video. To tackle this problem, we need new ways to summarize videos effectively. These methods should be able to quickly capture the most important parts of videos, especially when dealing with educational or lecture videos. So, we're looking for ways to make video summarization work well for all kinds of videos, including those used for teaching and learning. Essentially, summarizing lecture videos aims to simplify the learning process and make educational content more digestible.

The main contributions of this work are:

- **Proposed Honey Badger-based bald eagle algorithm based Deep Residual Network (HBBEA_DRN):** The important audio and video segments are selected using Deep Residual Network (DRN) with Honey Badger-based bald eagle algorithm (HBBEA) training. Here, the HBBEA is developed by integrating both the Bald Eagle Search (BES) and Honey Badger Algorithm (HBA). From each video shot, the audio and video within the video shot are segmented using the HBBEA.

2. Literature survey

Davila et al. [1] introduced a technique to extract and condense handwritten content from videos. They used a fully convolutional network called FCN-LectureNet, treating videos as binary images. This network mined the handwritten content to create a summarized version with the relevant segments. However, this method didn't consider additional factors to improve how content flows over time. To enhance temporal segmentation, Urala Kota et al. [2] proposed a model to mine and simplify lecture videos. They employed a deep learning pipeline to detect handwritten text for indexing video collections. However, this approach struggled to accumulate and annotate a broader range of lecture data. Yuan et al. [3] developed CRSum, a neural network for video summarization. It combined feature extraction, temporal modelling, and summary creation, introducing a new loss function, the Sobolev loss. Yet, this method had limitations in handling complex tasks. For intricate tasks, Mussel Cirne, M.V., and Pedrini, H [4] introduced a video summarization approach named VISCOM, utilizing colour co-occurrence matrices. However, this method incurred significant computational time. Gharbi et al. [5] proposed an effective keyframe mining technique based on local interest points and repeatability graph clustering. This method, though successful, didn't involve the concept of a visual query. Basavarajaiah, M., and Sharma, P. [6] presented GVSUM, a generic video summarization method. It produced a summary by selecting keyframes during main scene changes. However, this approach didn't incorporate audio and metadata. To include audio and metadata, Jin, H et al. [7] developed super pixel segmentation based on image similarity, adapting it to summarize videos and mine keyframes while reducing redundancy. Rafiq et al. [8] introduced a novel technique that classifies video scenes using a summarization model. Creating concise video summaries is complex, requiring substantial manual effort and extensive computational resources. Nevertheless, frames that don't neatly belong to any specific class may be incorrectly classified into the closest suitable class. Zhao., *et al* [30] introduced the Audio Visual Recurrent Network (AVRN) to exploit the audio and visual information for the video summarization task. This method manipulates both audio and visual information to improve understanding of video content and structure. However high-quality audio and visual data are essential for effective performance. This classification challenge can be enhanced through the optimization techniques commonly employed across various domains. Optimization techniques [41 42 43] include metaheuristic optimization techniques, such as Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Simulated Annealing (SA), Ant Colony Optimization (ACO), and others, which are widely employed across diverse domains. These techniques find applications in engineering design, scheduling problems, financial modelling, bioinformatics, network design, evolutionary robotics, function optimization, image processing, neural network [31 36] training, and various combinatorial optimization problems. Their versatility allows researchers and practitioners to adapt these algorithms to address complex optimization challenges in fields ranging from logistics and telecommunications to structural design and machine learning [35 37 10].

In the previous discussion of the literature survey, we noticed that Bald Eagle Optimization (BEO) and Honey Badger Optimization (HBO) techniques are metaheuristic optimization algorithms inspired by the characteristics and behaviours of bald eagles and honey badgers, respectively. BEO is applied to solve optimization problems across diverse domains, including engineering design, scheduling, and bioinformatics. Its inspiration from the hunting behaviour of bald eagles makes it particularly suitable for problems requiring adaptability and efficient exploration of search spaces. On the other hand, HBO, inspired by the tenacity and fearless nature of honey badgers, is utilized in optimization tasks such as feature selection, image segmentation, and parameter tuning in machine learning. Both techniques leverage the unique traits of these animals to enhance their search mechanisms and find high-quality solutions in various application domains [11 34].

As per the literature survey performed previously, there isn't specific information available on the direct application of BEO or HBO techniques to video summarization tasks. These optimization algorithms are relatively new, and their applications may vary across different domains. The choice of optimization technique for such tasks often depends on the specific characteristics of the problem and the objectives of the summarization.

However, it's important to note that research in the field of optimization techniques and their applications is continually evolving [14]. Researchers are exploring and adapting these algorithms for video summarization. There are some challenges mentioned in section 2.1 as per the rigorous survey carried out in the previous section.

2.1. Challenges

Here are some Challenges that have been addressed by lecture video summarization methods:

Complexity of Deep Models: One method called FCN-LectureNet tried to classify lecture videos, but it showed that using deep models to directly figure out which frames are important can be risky. It's still challenging to accurately classify different types of content like text, graphics, and math within the video.

Content Overlapping: Some methods used detailed pixel masks of the lecturer's writing to piece together the handwritten content. However, this approach sometimes causes the content to overlap or get jumbled up [2].

Complex Analysis for Similarities: Another method, VISCOM, used colour co-occurrence matrices to define video frames. However, it required a deep analysis of distance functions to determine how similar pairs of frames were, which made it more complex [4].

Inaccurate Segmentation: Super pixel-based video summarization methods didn't always give precise results, especially when the video content changed quickly or shifted regularly. This led to issues with the similarity of keyframes [7].

Lack of Search Engine Indexing: Lecture videos are valuable for

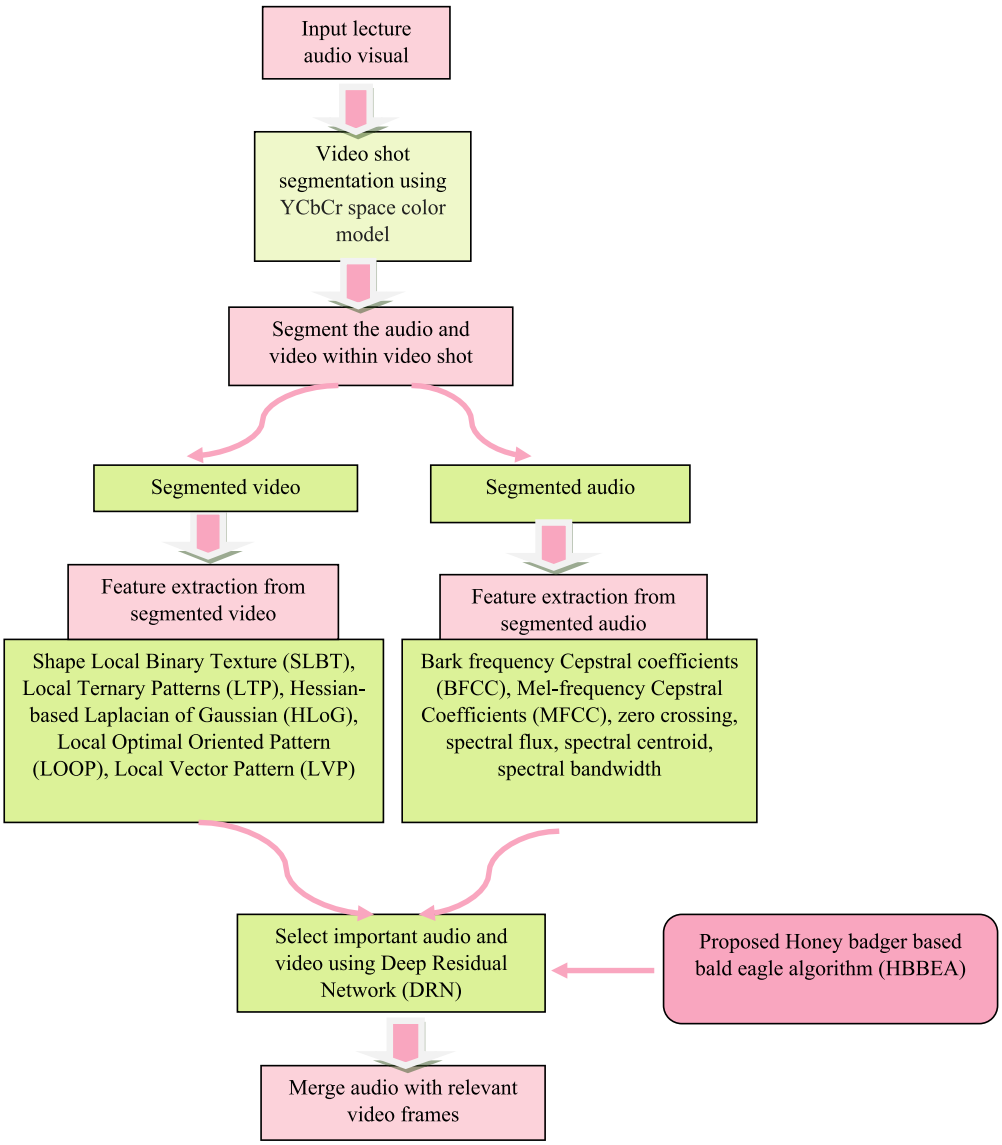


Fig. 1. Preview of lecture audio video summarization model using HBBEA-based DRN.

students and educators worldwide, but they often aren't searchable on search engines. Creating annotations (labels or descriptions) for these videos manually is a difficult and time-consuming task.

The aim here is to make a smart way of summarizing raw lecture videos. We start by breaking the video into different parts based on colours, and we use something called the YCbCr color model for this. Then, we separate the audio (sound) and video (pictures) parts of the videos using an algorithm called HBBEA. After that, we look at the information in the sound and the pictures separately, put it all together, and feed it into a model called DRN. DRN helps us figure out which parts of the video and audio are important. To make DRN work well, we use HBBEA to adjust its settings and weights. So, in simple terms, we're using a combination of colour, frame and sound to pick out the important bits from raw lecture videos.

3. Proposed Honey Badger-based bald eagle algorithm (HBBEA)-based Deep Residual Network (DRN) for lecture audio video summarization model

The purpose is to present a fine-tuned deep model for lecture audio video summarization. Initially, the input lecture audio-visual video is taken from the dataset. Then, the video shot segmentation (slide segmentation) is done using the YCbCr space colour model [9]. From each video shot, the audio and video within the video shot are segmented using the HBBEA. The Honey Badger-based Bald Eager Algorithm (HBBEA) is obtained by combining BES [10] and HBA [11]. After that, feature extraction for each segmented audio and video frame is done. Here, the audio features, like BFCC [12], MFCC [12], zero crossing, spectral flux [24], spectral centroid [13], and spectral bandwidth, whereas video features, such as SLBT [25], LTP [26], HLoG [27], LOOP [23], and LVP [29] are extracted. Once the segmented audio and video frames are extracted, the important audio and video segments are selected using DRN [14]. The DRN training is executed with HBBEA to tune the finest DRN weights. Finally, the relevant video frames are merged with the audio. Thus, the HBBEA-based DRN are utilized for lecture audio video summarization. Fig. 1 shows a preview of the lecture audio video summarization model using HBBEA-based DRN.

3.1. Congregate lecture audio-visual video

Audiovisual (AV) represents the processing of electronic media, which includes both visual and sound units, like films, and slide-tape presentations. Digital video is considered to be an emerging storage and exchange interface through the quick design in high-speed net-

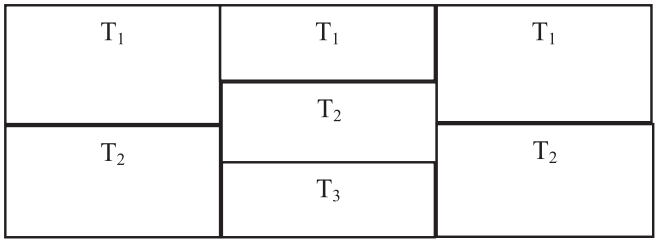


Fig. 2. Solution model.

works. There exist recordings of audiovisual content, which are utilized more frequently in e-learning. Consider a dataset F, which comprises input lecture audio-visual video content, and is represented by,

$$F = \{K_1, K_2, \cdots, K_l, \cdots, K_m\} \tag{1}$$

where, articulates total videos and K_l is l^{th} a lecture video.

3.2. Video shot segmentation using YCbCr space colour model

Here, K_l is subjected to segmentation of video shots. Here, the segmentation of video shots is done using the YCbCr space colour model [9]. It is modelled to YCbCr space colour in which the three colour values Y, Cb and Cr are considered. Here, the Y, Cb and Cr indicate the Luminance component, blue-difference and red-difference chroma components. Each channel of colour in a frame evaluates the moments. Thus, each video frame is adapted as nine moments wherein the three moments for each 3 color channel are considered. These colour moments include mean, standard deviation and skewness, which is devised next to this. The mean represents the value of the average colour in a frame. The square root using the distribution of variance is termed standard deviation. The skewness indicates the degree of asymmetry in a distribution. The mean, standard deviation and skewness are explicated in equations (2), (3) and (4). These three colour moments are described in the form of pixels. Each pixel in a frame is modelled as $G_{d,e}$, which represents the pixel of d^{th} the colour channel of e^{th} the frame.

$$E_d = \frac{1}{D}\sum_{D}^{e-1} G_{d,e} \tag{2}$$

$$C_d = \sqrt{\left(\frac{1}{D}\sum_{D}^{e-1}\left(G_{d,e} - E_d\right)^2\right)} \tag{3}$$

$$B_d = \sqrt[3]{\left(\frac{1}{D}\sum\nolimits_{D}^{e-1}\left(G_{d,e} - E_d\right)^3\right)} \tag{4}$$

where, E_d, C_d and B_d articulates mean, standard deviation and skewness of d^{th} colour channel. Enumerate the addition of weighted differences amidst the moments of two distributions considering expression (5), which is utilized for getting the dissimilarity amidst the colour distributions of two frames in a video. Thus, the inputted video is segmented into the number of frames based on the dissimilarity between the two frames. The segments are notated as,

$$R = \{R_1, R_2, \cdots, R_g\} \tag{5}$$

where, g refers to total segments.

3.3. Segment audio and video

Here, the audio and video segments are obtained separately considering the data present in memory. The audio segmentation technique discovers the correlations amidst the audio features while the video segmentation discovers the correlations amidst the keyframes. Here, the audio and video segmentation is performed with HBBEA. The solution model, fitness and HBBEA steps are defined below.

3.3.1. Solution encoding

The representation of the solution considering different videos is revealed in Fig. 2. Here, the segments of audio and video are separately obtained. Consider slide 1, which is divided into two, three and two segments as represented below.

3.3.2. Derive fitness function

The fitness is modelled as,

$$Fit = \frac{1}{2}*\left(\sum_{k=1}^{g}\frac{1}{|J_k|}\sum_{h=1}^{|J_k|}\|\widehat{L} - L_h^k\|\right) + \left(\sum_{k=1}^{g}\frac{1}{|N_k|}\sum_{h=1}^{|N_k|}\|\widehat{N} - N_h^k\|\right) \tag{6}$$

where $\widehat{L}$ is the average of frames in k^{th} the segment, $\widehat{N}$ is the average of signals in k^{th} the segment, J_k is the total number of frames in k^{th} segment and N_k is the total number of signals in k^{th} segment,g represents total segments.

3.3.3. Steps of HBBEA

The HBBEA is produced by the alliance of BES and HBA. BES [10] is motivated by the mimicking behaviours of bald eagles throughout the hunting to illustrate the consequences of each hunting phase. It is plotted into three parts, named selection of search space, search selected search space, and swooping. It improved its effectiveness by improving its efficacy and incorporating powerful operators. To produce a global best solution, HBA [11] is incorporated. HBA is motivated by the intellectual foraging strategy of honey badger to devise an effective search phase to solve the optimization issues having complicated search spaces. It is effective in balancing exploration–exploitation and is best in solving unsupervised and supervised issues. Thus, the alliance of BES and HBA provided the finest competence. The HBBEA steps are illustrated below.

Step 1) Initialization

The fundamental step is to define the populace that contains the production of bald eagles notated by,

$$T = \{T_1, T_2, \cdots, T_\mu, \cdots, T_\vartheta\} \tag{7}$$

where, ϑ is total bald eagles, T_μ signifies μ^{th} bald eagles.

Step 2) Enumerate the fitness function

The fitness is previously inspected in section 3.3.2.

Step 3) Choose the stage.

According to BEO [10], the select phase comprises bald eagles that determine and select the best area in the selected search space wherein it hunts prey and is provided by,

$$T(a+1) = T_{best} + \gamma^* n(T_{mean} - T(a)) \tag{8}$$

wherein, γ stands for attribute, which manages change of position, n signifies arbitrary number amongst 0–1,T_{best} symbolizes optimum position, T_{mean} signifies eagles which used up all the data with previous search, and $T(a)$ symbolizes the present position of the bald eagle, and $T(a+1)$ refers to subsequent bald eagle location.

$$T(a+1) = T_{best} + \gamma^* n^* T_{mean} - \gamma^* n^* T(a) \tag{9}$$

For escaping with local optimal, the HBA [11] is being utilized. From the HBA algorithm, the update expression is modelled by,

$$T(a+1) = T_{prey} + H^* p_7^* \eta^* q_a \tag{10}$$

$$q_a = T_{prey} - T(a) \tag{11}$$

$$T(a+1) = T_{prey} + H^* p_7^* \eta^* \left(T_{prey} - T(a)\right) \tag{12}$$

$$T(a+1) = T_{prey} + H^* p_7^* \eta^* T_{prey} - H^* p_7^* \eta^* T(a) \tag{13}$$

Table 1
Pseudo-code of HBBEA.

```
Input: Population T
Output: Finest solution T_best
Begin
Randomly initiate point T_c for c point.
Enumerate fitness
While (termination criterion not met)
//Select stage
For (each point c in population)
Update with expression (21)
If f(T_new) < f(T_c)
T_c = T_new
If f(T_new) < f(T_best)
T_best = T_new
        End if
        End if
End for
// Search in phase
   For (each point c in population)
      Update with expression (22)
If f(T_new) < f(T_c)
T_c = T_new
If f(T_new) < f(T_best)
T_best = T_new
          End if
        End if
End for
// Swoop phase
   For (each point c in population)
      Update with expression (29)
If f(T_new) < f(T_c)
T_c = T_new
If f(T_new) < f(T_best)
T_best = T_new
          End if
        End if
End for
   Fix a = a + 1
   End while
End
```

$$H^*p_7{}^*\eta^*T(a) = T_{prey} + H^*p_7{}^*\eta^*T_{prey} - T(a+1) \tag{14}$$

$$T(a) = \frac{T_{prey}(1 + H^*p_7{}^*\eta) - T(a+1)}{H^*p_7{}^*\eta} \tag{15}$$

Substitute expression (15) in expression (9),

$$T(a+1) = T_{best} + \gamma^*n^*T_{mean} - \gamma^*n\left(\frac{T_{prey}(1 + H^*p_7{}^*\eta) - T(a+1)}{H^*p_7{}^*\eta}\right) \tag{16}$$

$$T(a+1) = T_{best} + \gamma^*n^*T_{mean} - \gamma^*n\frac{T(a+1)}{H^*p_7{}^*\eta} - \gamma^*n\left(\frac{T_{prey}(1 + H^*p_7{}^*\eta)}{H^*p_7{}^*\eta}\right) \tag{17}$$

$$T(a+1) + \gamma^*n\frac{T(a+1)}{H^*p_7{}^*\eta} = T_{best} + \gamma^*n^*T_{mean} - \gamma^*n\left(\frac{T_{prey}(1 + H^*p_7{}^*\eta)}{H^*p_7{}^*\eta}\right) \tag{18}$$

$$T(a+1)\left[1 + \frac{\gamma^*n}{H^*p_7{}^*\eta}\right] = T_{best} + \gamma^*n^*T_{mean} - \gamma^*n\left(\frac{T_{prey}(1 + H^*p_7{}^*\eta)}{H^*p_7{}^*\eta}\right) \tag{19}$$

$$T(a+1)\left[\frac{H^*p_7{}^*\eta + \gamma^*n}{H^*p_7{}^*\eta}\right] = T_{best} + \gamma^*n^*T_{mean} - \gamma^*n\left(\frac{T_{prey}(1 + H^*p_7{}^*\eta)}{H^*p_7{}^*\eta}\right) \tag{20}$$

The final update expression of HBBEA is provided as,

$$T(a+1) = \frac{(T_{best} + \gamma^*n^*T_{mean})(H^*p_7{}^*\eta) - \gamma^*n\left(T_{prey}(1 + H^*p_7{}^*\eta)\right)}{H^*p_7{}^*\eta + \gamma^*n} \tag{21}$$

Step 4) Search stage

In the search stage, the updated position of bald eagles is modelled as,

$$T_{c,new} = T_c + A(c)^*(T_c - T_{c+1}) + I(c)^*(T_c - T_{mean}) \tag{22}$$

Here, *AI(c)*refers to the present location of c^{th} bald eagle in the x-axis, $I(c)$ signifies the present location of c^{th} bald eagle in the y-axis, T_c symbolizes c^{th} bald eagle's location and T_{c+1} is $(c+1)^{th}$ bald eagle's location.

The current location of c^{th} bald eagle in the x-axis is provided by,

$$A(c) = \frac{AI(c)}{\max(|AI|)} \tag{23}$$

The present location of c^{th} bald eagle in the y-axis is modelled by,

$$I(c) = \frac{QI(c)}{\max(|QI|)} \tag{24}$$

$$AI(c) = I(c)^*\sin(\theta(c)) \tag{25}$$

$$QI(c) = I(c)^*\cos(\theta(c)) \tag{26}$$

The angle provided with a bald eagle while updating the location is noted by,

$$\theta(c) = f^*\pi^*rand \tag{27}$$

$$I(c) = \theta(c) + O \times rand \tag{28}$$

where f refers to attribute with value 5–10, O signifies a random number, which relies on value amidst 0.5 to 2 and aids to identify search cycles count, *rand* is arbitrary, $\theta(c)$ denote angle.

Step 5) Swooping stage.

In the swooping stage, the updated position of bald eagles is modelled as,

$$T_{c,new} = rand^*T_{best} + AI(c)^*(T_c - i1^*T_{mean}) + QI(c)^*(T_c - i2 \times T_{best}) \tag{29}$$

Here,$i1, i2 \in [1, 2]$.

The present location of c^{th} bald eagle in the x-axis is modelled by,

$$AI(c) = \frac{AI(c)}{\max(|AI|)} \tag{30}$$

The present location of w^{th} bald eagle in the y-axis is modelled by,

$$QI(c) = \frac{QI(c)}{\max(|QI|)} \tag{31}$$

$$AI(c) = I(c)^*\sinh(\theta(c)) \tag{32}$$

$$QI(c) = I(c)^*\cosh(\theta(c)) \tag{33}$$

The angle attained by the bald eagle while updating the location is modelled by,

$$\theta(c) = f^*\pi^*rand \tag{34}$$

$$I(c) = \theta(c) \tag{35}$$

Step 6) Re-enumeration of fitness

Considering each population, fitness is re-enumerated to define the best solution.

Step 7) Termination

Imitate aforesaid steps until the crucial factor of termination is acquired. Table 1 explores the pseudo-code of HBBEA.

3.4. Acquisition features from segmented video

The features obtained through the segmented videos are briefly described below.

a) SLBT

SLBT feature [25] unifies both shape and texture data and it is similar to Active Appearance Model (AAM). Consider a 3×3 window having a centre pixel (U_r, V_r) intensity value represented as W_r wherein $W_r\,(r = 1, 2, 3, ..., 7)$ equal to the grey metric of eight adjoining pixels. These pixels are thresholds using the centre metric W_r as $\psi\,(X\,(W_0 - W_r), ..., X\,(W_7 - W_r))$ an operation of $X\,(Y)$ is notated by,

$$X(Y) = \begin{cases} 1\,, & z > 0 \\ 0, & z \le 0 \end{cases} \tag{36}$$

$$LBP\,(U_r,\, V_r) = \sum_{r=0}^{7} X\,(W_{rr} - W_r)\,2^{rr} \tag{37}$$

Also, texture modelling is executed with PCA and it is described by,

$$W_{tt} = x_{tt}\,(y - \overline{y}) \tag{38}$$

The unified texture and shape attribute vector is evaluated with the below equation. As values of shape and texture are in assorted units, then it is imperative to discover the weights of the diagonal matrix z_s. The unified feature vector is explicated as,

$$W_{st} = \begin{pmatrix} z_s\, W_s \\ W_{tt} \end{pmatrix} \tag{39}$$

$$v = x_{st}\,(W_{st} - \overline{W}_{st}) \tag{40}$$

The generated SLBT feature is notated by E_1.

b) LTP

Several segmented areas are more consistent and genuine to examine if the feature robustness is enhanced in the areas. LTP [26] is less responsive and more dominant to noise in consistent segmented areas. LTP grey levels are quantized for zero in a zone width of $\pm j$ around t_u and a 3-valued operation is utilized for replacing the indicator $b(o)$ and is modelled as,

$$E_2 = b(o, t_u, j) = \begin{cases} 1 & ; o \ge t_u + j \\ 0 & ; |o - t_u| < j \\ -1 & ; o \le t_u - j \end{cases} \tag{41}$$

Thus, the user-based threshold is explicated j which builds codes of LTP to be more defiant to noise and invariant to grey-level alteration,t_u represent the central pixel and o is notated by adjoining pixel. The LTP feature is notated as E_2.

c) HoG

HOG [27] is executed on the Scale-Invariant Features Transform (SIFT) descriptor and is enumerated by first, vertical L_φ and horizontal gradients L_λ to filter an image with subsequent kernels.

$$L_\lambda = (-1, 0, 1)\; and\; L_\phi = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \tag{42}$$

The magnitude (m) and orientation $(\tau)$of gradients (δ, θ) considering each pixel are evaluated as,

$$i(\lambda, \phi) = \sqrt{Q_\lambda^2 + Q_\phi^2}\; and\; \varsigma(\lambda, \phi) = \arctan\left(\frac{Q_\phi}{Q_\lambda}\right) \tag{43}$$

wherein, Q_λ and Q_φ signify horizontal and vertical gradient approximations. The HoG output generated through feature mining is notated as E_3.

d) LOOP

The Local Binary Pattern (LBP) and Local Directional Pattern (LDP) are unified considering a non-linear way to offer this feature. The equation of LOOP [23] value using pixel is provided as;

$$LOOP(aa_{cc}, bb_{cc}) = \sum_{ff}^{7} gg(ee_{ff} - ee_{cc})2^{\alpha} \tag{44}$$

$$dd(hh) = \begin{cases} 1 & ; If\; hh \ge 0 \\ 0 & ; Otherwise \end{cases} \tag{45}$$

wherein, ee_{ff} and ee_{cc} symbolize the intensity of the image in 3×3 the neighbourhood, α signify the pixel's exponential weight, and is discovered using rank and magnitude of Kirsch masks, which is determined using its eight pixels and noise-free image's centre pixel is expressed using coordinate aa_{cc}, bb_{cc}. The LOOP feature is denoted as E_4..

e) LVP

The LVP [29] characteristic is modelled by,

$$F_{\alpha,\varepsilon}(L_{c_1}) = (J(L_{\alpha,\varepsilon}) - J(L_{e_1})) \tag{46}$$

Where in, α symbolizes the count of neighbourhood pixels using targeted pixels L_{c_1} from the radius.

$$\begin{aligned} LVP_{X,Y,\alpha}(L_{c_1}) = \{&q_5(F_{\alpha,\varepsilon}(L_1, Y), F_{\alpha+45^\circ,\varepsilon}(L_1, Y), F_{\alpha,\varepsilon}(L_{c_1}), F_{\alpha+45^\circ,\varepsilon}(L_{c_1})), \\ &q_5(F_{\alpha,\varepsilon}(L_2, Y), F_{\alpha+45^\circ,\varepsilon}(L_2, Y), F_{\alpha,\varepsilon}(L_{c_1}), F_{\alpha+45^\circ,\varepsilon}(L_c)), ..., \\ &q_5(F_{\alpha,\varepsilon}(L_{l,Y}), F_{\alpha+45^\circ,\varepsilon}(L_{l,Y}), F_{\alpha,\varepsilon}(L_{c_1}), F_{\alpha+45^\circ,\varepsilon}(L_{c_1}))\}|_{l=1,2,...,X;Y=1} \end{aligned} \tag{47}$$

$$\left[\begin{matrix} q_5(F_{\alpha,\varepsilon}(L_{l,Y}), F_{\alpha+45^\circ,\varepsilon}(L_l, Y), \\ F_{\alpha,\varepsilon}(L_{c_1}), F_{\alpha+45^\circ,\varepsilon}(L_{c_1}) \end{matrix}\right] = \begin{cases} 1, & if\; F_{\alpha+45^\circ,\varepsilon}(L_{l,Y}) - \left(\dfrac{F_{\alpha+45^\circ,\varepsilon}(L_{c_1})}{F_{\alpha,\varepsilon}(L_{c_1})} \times F_{\alpha,\varepsilon}(L_{l,Y})\right) \\ 0, & else \end{cases} \ge 0 \tag{48}$$

wherein,$F_{\alpha,\varepsilon}(L_{c_1})$ symbolize the direction vector at the targeted pixel, and ε symbolize separation amidst the pixel in question and its neighbours, and the ratio of transformation is provided as $q_5(.,.)$. The LVP feature is notated by E_5..

Thus, the feature vector obtained with segmented video features is explicated as,

$$\kappa_1 = \{E_1, E_2, E_3, E_4, E_5\} \tag{49}$$

3.5. Acquisition features from segmented audio

The features obtained through the segmented audio are briefly described below.

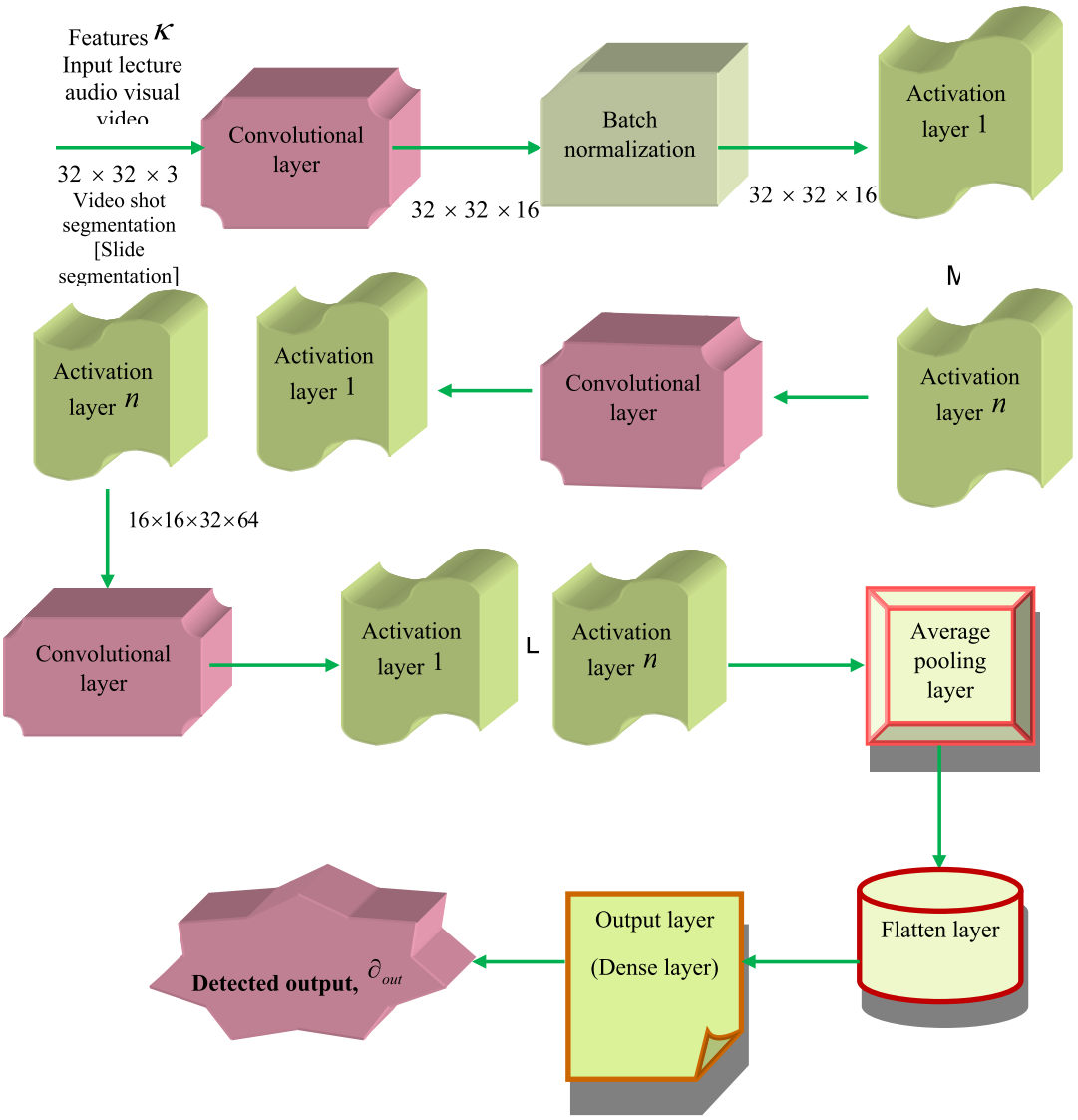


Fig. 3. Preview of DRN.

a. BFCC

BFCC [12] is a model utilized to mine audio segments and is described using logarithmic illustration. The BFCC is enumerated as,

$$Bark(\partial) = 13jtan(0.76\partial/100) + 3.5 \quad (50)$$

where, ∂ tends to frequency. The BFCC feature is denoted by T_1.

b. MFCC

MFCC [12] represents a technique utilized to excavate the features using the obtained audio segments and is enumerated considering dissimilarities amidst speaker voice and nearby platforms. Thus, the MFCC feature is indicated by P_3.

$$Mel(\partial) = 2595 \times \log(1 + \partial/700) \quad (51)$$

where, ∂ tends to frequency. The MFCC feature is denoted by T_2.

c. Zero-crossing

It represents a point wherein the sign of mathematical function alters in a signal. It helps to evaluate speech frequency. The zero-crossing features are denoted by T_3.

d. Spectral flux

The term spectral flux [28 24] is denoted as a 2-norm of frame-to-frame spectral amplitude variance vector and is notated by T_4.

e. Spectral centroid

Spectral centroid [28 13] is utilized to evaluate the brightness of the audio segment, and is expressed by,

$$T_5 = \left(\sum_{mm=nn_1}^{nn_2} kk_{mm} ll_{mm} / \sum_{mm=nn_1}^{nn_2} ll_{mm} \right) \quad (52)$$

where kk_{mm} signifies the frequency of bin mm, ll_{mm} represents a spectral value of bin mm and nn_1 nn_2 exposed band edges, and T_5 is the spectral centroid.

f. Spectral bandwidth

This feature is used to compute the smallest bandwidth voice using the higher frequency voice and is notated by T_6.

Thus, the feature vector obtained with segmented audio features is explicated as,

$$\kappa_2 = \{T_1, T_2, T_3, T_4, T_5, T_6\} \tag{53}$$

Hence, the final feature vector fed to DRN is denoted as κ, which represents the alliance of segmented video features κ_1 and segmented audio features κ_2.

3.6. Select important audio and video segments by DRN

The important audio and video segments are selected using the proposed HBBEA-based DRN. The DRN preview and steps of HBBEA are discussed below.

3.6.1. DRN model preview

The DRN [14] includes various layers, like residual blocks, linear classifier, pooling and convolutional layer. The DRN model acquires enhanced training and learning efficiency with a limited amount of training instances. Thus, DRN is utilized for selecting imperative audio and video segments. The feature vector κ is attained as DRN input and the structure is displayed in Fig. 3.

i. Convolutional layer

The conv layer expression is modelled by,

$$R(s) = \sum_{\ell=0}^{\eta-1} \sum_{\lambda=0}^{\eta-1} \upsilon_{\ell,\lambda} \bullet s(T + \ell)(\Upsilon + \lambda) \tag{54}$$

$$R_e(s) = \sum_{\varphi=0}^{\theta-1} \upsilon_{\aleph} * s \tag{55}$$

Here, s explicates the 2-dimensional outcome generated through the prior layer, υ is $\eta \times \eta$ kernel matrix, and learnable attribute T and Υ is adapted for accumulating coordinates ℓ, λ is the index of location in 2-dimensional kernel matrix, $\upsilon_{\aleph}$ explicates kernel size of $\aleph^{th}$ neuron, and $*$ stands for cross-correlation operator.

ii. Pooling layer

The Pool layer is linked with the convolutional layer, is adapted to reduce the spatial size of the feature map, and smoothly manages overfitting.

$$\ell_{out} = \frac{\ell_{in} - \aleph_{\ell}}{\chi} + 1 \tag{56}$$

$$\lambda_{out} = \frac{\lambda_{in} - \aleph_{\lambda}}{\chi} + 1 \tag{57}$$

Here, $\ell_{in}\lambda_{in}$ express width and height of input 2-dimensional matrix, $\aleph_{\ell}$ represent the height of kernel, $\aleph_{\lambda}$ is kernel width, ℓ_{out} and λ_{out} endows width and height of 2-dimensional matrix output.

iii. Activation function

The ReLU is notated as,

$$\mathrm{Re}LU(s) = \begin{cases} 0, & s < 0 \\ s, & s \geq 0 \end{cases} \tag{58}$$

Here, s signify the input feature.

iv. Batch normalization

The set of training instances is divided into disparate diminutive sets, known as mini-batches in the batch normalization process and these mini-batches are adapted to train the process.

v. Residual blocks

The blocks of residual allocate a correlation substitute amongst the convolutional layers. It involves shortcut correlation through input to output in contrast to CNN. In addition, the input is linked directly to the output layer that is modelled by,

$$\xi_{in} = \chi(s) + s \tag{59}$$

Furthermore, the element matching factor is notated by,

$$\xi_{out} = \chi(s) + \varepsilon_T s \tag{60}$$

where, s and $\xi_{in}\xi_{out}$ refers to residual blocks input and output, φ signifies mapping relation amongst output and input layer, ε_T signifies weight matrix.

vi. Linear classifier

The linear classifier is modelled by,

$$\varsigma = \varepsilon_T + o \tag{61}$$

where o refers to bias and DRN output is modelled by ∂_{out}.

3.6.2. Training using HBBEA

The steps of HBBEA are the same as in section 3.3.3. Here, the fitness considered is a mean square error which is modelled by,

$$\Re = \frac{1}{\ell} \sum_{\nu=1}^{\ell} [\lambda_{\nu} - \partial_{out}]^2 \tag{62}$$

where, $\Re$ articulates fitness function, ℓ states total samples such that $1 \leq \nu \leq \ell$, term λ_{ν} reveals target output, and ∂_{out} denotes DRN output.

4. Results and discussion

The propensity of HBBEA-DRN is examined by changing the training percentage in the x-axis. The proposed HBBEA-DRN is scripted in Python.

4.1. Dataset description

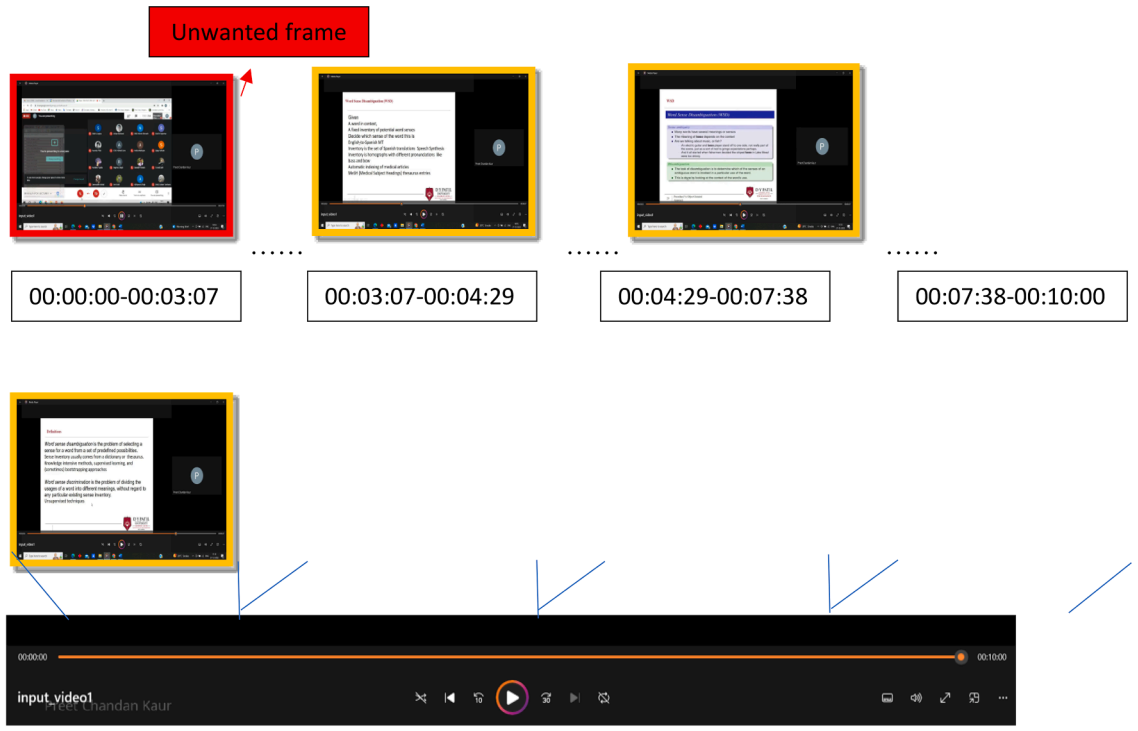
The lecture video dataset is employed for the analysis. Here, five types of lecture videos are considered. Each video comprises 1561 frames which provide visual content, audio content and data. Here, the audio feature size is (1561, 135). Here, the video feature size is (1561, 160). Here, the audio and video feature size is (1561, 295).

4.2. Experimental results

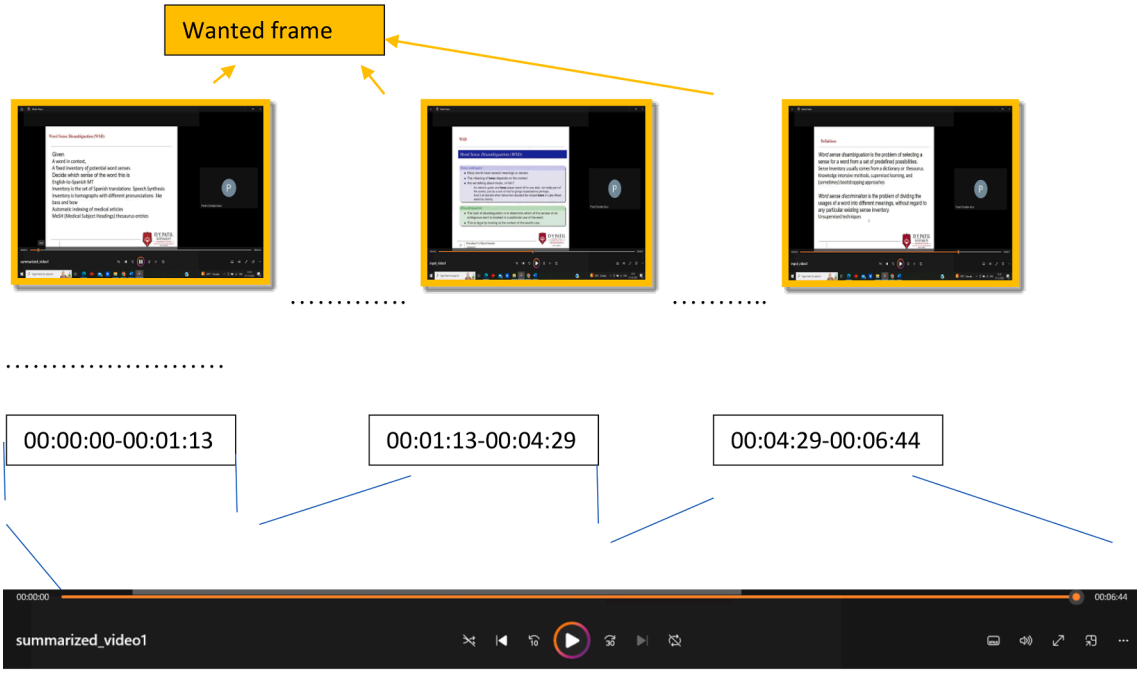
The experimental resultants of HBBEA-DRN are expressed in Figs. 4, 4a) specifies the Input video frames with the timeline, 4b) indicates summarized video frame with timeline, 4c) implies SLBT image, 4d) indicates HoG image and 4e) signifies LOOP image. Notably, in Fig. 4, one can observe the removal of undesired frames from the 00:03:07 timeline by HBBEA-DRN. The initial duration of the input lecture video is 10 min, and through the summarization process, the duration of the resulting video is reduced to 6.44 min. This indicates a successful reduction of 3.56 min in the summarized video duration, demonstrating the efficiency of the HBBEA-DRN approach in achieving more concise and focused lecture video summaries. The outcomes are obtained using five sets of videos which have 1561 frames. Some of the frames are described in the below figures.

4.3. Metrics used

Metrics like Accuracy, NPV, Precision, Recall and F-measure are the

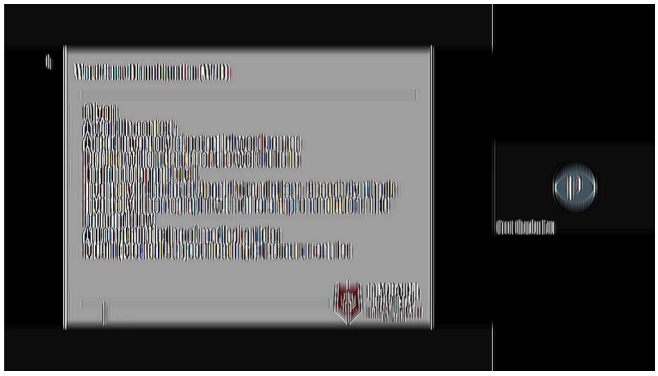


a) Input video frames with the timeline

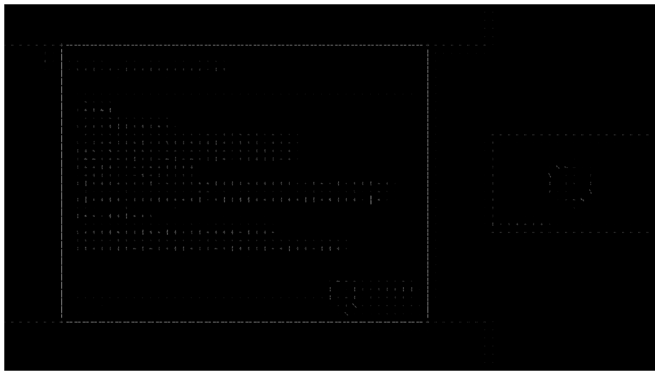


b) Summarised video frames with the timeline

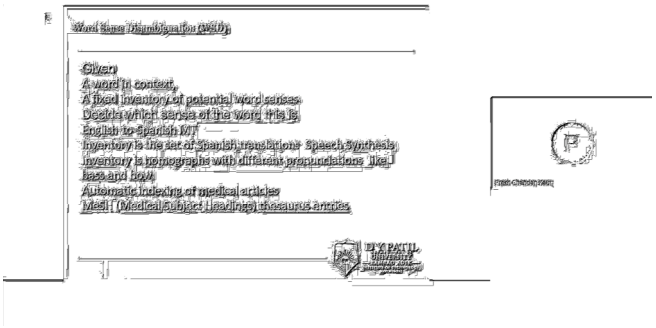
Fig. 4. Experimental input and outcomes of HBBEA-DRN for lecture video summarization.



(c) SLBT



(d) HoG



(e) LOOP

Fig. 4. (*continued*).

metrics which is used by HBBEA-DRN for audio–video summarization. The ability of HBBEA-DRN is inspected with different metrics and is examined below.

(a) Accuracy

It measures the correctness of representing the video's key aspects. It exposes the probability that underwent the correct outcome and is expressed as,

$$A = \frac{\mathrm{O} + Y}{\mathrm{O} + \mathrm{E} + T + Y} \tag{63}$$

where, O refers to truly positive, ϒ is truly negative, E states falsely positive, and T articulates falsely negative.

(b) NPV

NPV is a metric used to assess the ability of a summarization method to exclude irrelevant or redundant content. It explicates probability that overcomes negative test outcome in which each will truly not contain that precise result and is represented as,

$$N^{PV} = \frac{\Upsilon}{\Upsilon + \mathrm{T}} \tag{64}$$

(c) Precision

Precision measures the proportion of relevant content in the summary. It displays the closeness of various data instances amidst each other to model summarization and is represented as,

$$\rho = \frac{\mathrm{O}}{\mathrm{O} + \mathrm{E}} \tag{65}$$

(d) Recall

Recall evaluates the ability of the summarization method to include all relevant content from the original video. It provides computation of positive set classifications number, and is notated as,

$$\Re = \frac{\mathrm{O}}{\mathrm{O} + \mathrm{T}} \tag{66}$$

(e) F-measure

The F-measure combines precision and recall into a single metric and is modelled as,

$$\psi = 2 * \frac{\rho^{*}\Re}{\rho + \Re} \tag{67}$$

wherein, ρ and $\Re$ is precision and recall.

4.4. Comparative analysis

The analysis is implemented with video, audio and video-audio data using different metrics.

Certain previously developed schemes adapted for assessment involve FCN-LectureNet [1], DL [2], CRSum [3], VISCOM [4] and proposed HBBEA_DRN.

4.4.1. Video data evaluation

Fig. 5 explicates the video data evaluation using different measures

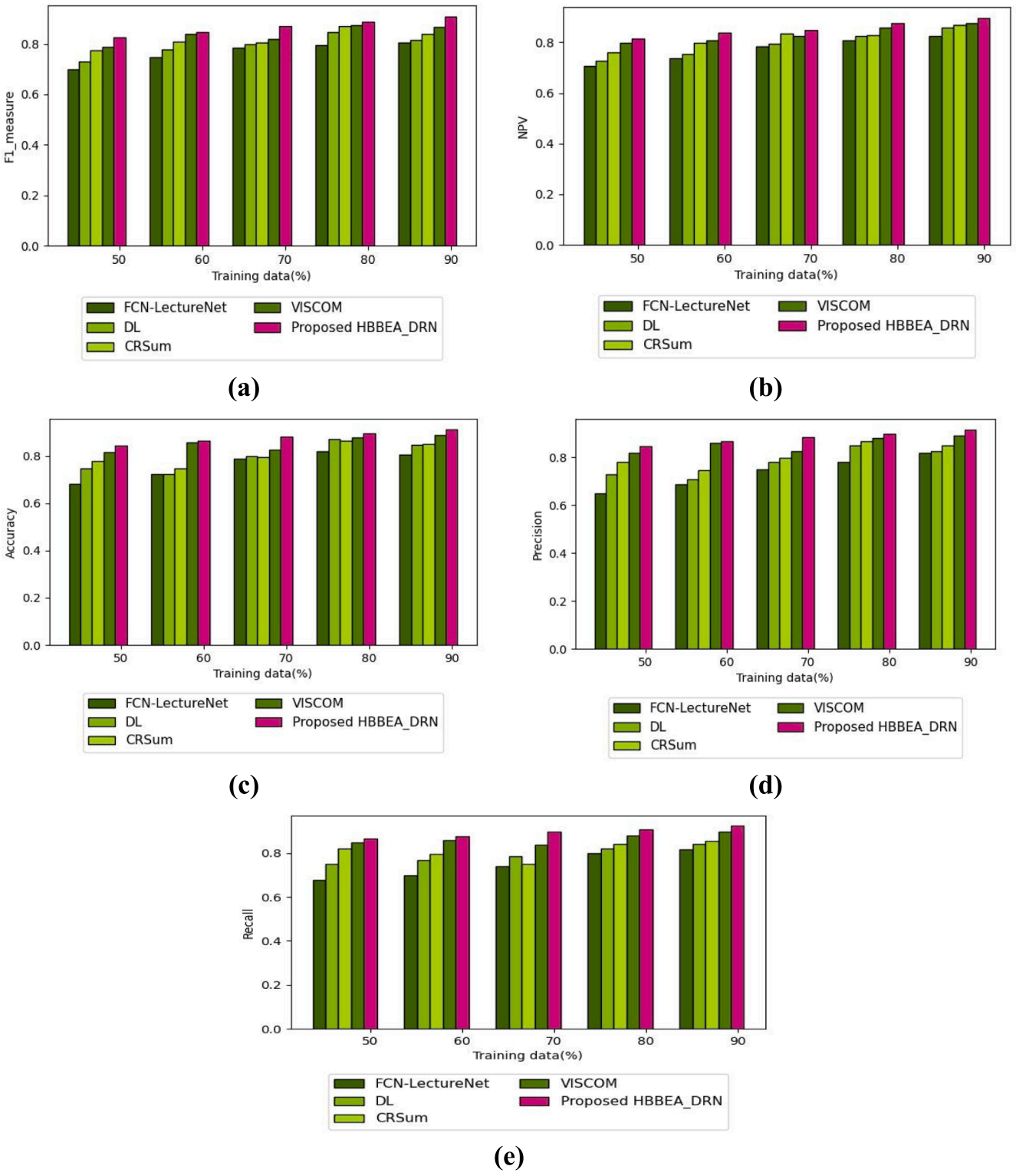


Fig. 5. Video data evaluation using a) F1-measure b) NPV c) Accuracy d) Precision e) Recall.

by changing the training percentage. The F1-Score-assisted evaluation is presented in Fig. 5a). Attaining 50 % training data, the F1-measure enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.699, 0.730, 0.775, 0.789, and 0.826. Besides attaining 90 % training data, the F1-measure enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.806, 0.817, 0.840, 0.869, and 0.909. The evaluation of existing HBBEA_DRN using F1-measure is 11.331 %, 10.121 %, 7.590 %, and 4.400 %. The NPV-assisted evaluation is presented in Fig. 5b). Attaining 50 % training data, the NPV enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.705, 0.725, 0.760, 0.799, and 0.816. Besides attaining 90 % training data, the NPV enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.827, 0.860, 0.870, 0.875, and 0.897. The evaluation of existing HBBEA_DRN using NPV is 7.803 %, 4.124 %, 3.010 %, and 2.452 %. The Accuracy-assisted evaluation is presented in Fig. 5c). Attaining 50 % training data, the Accuracy

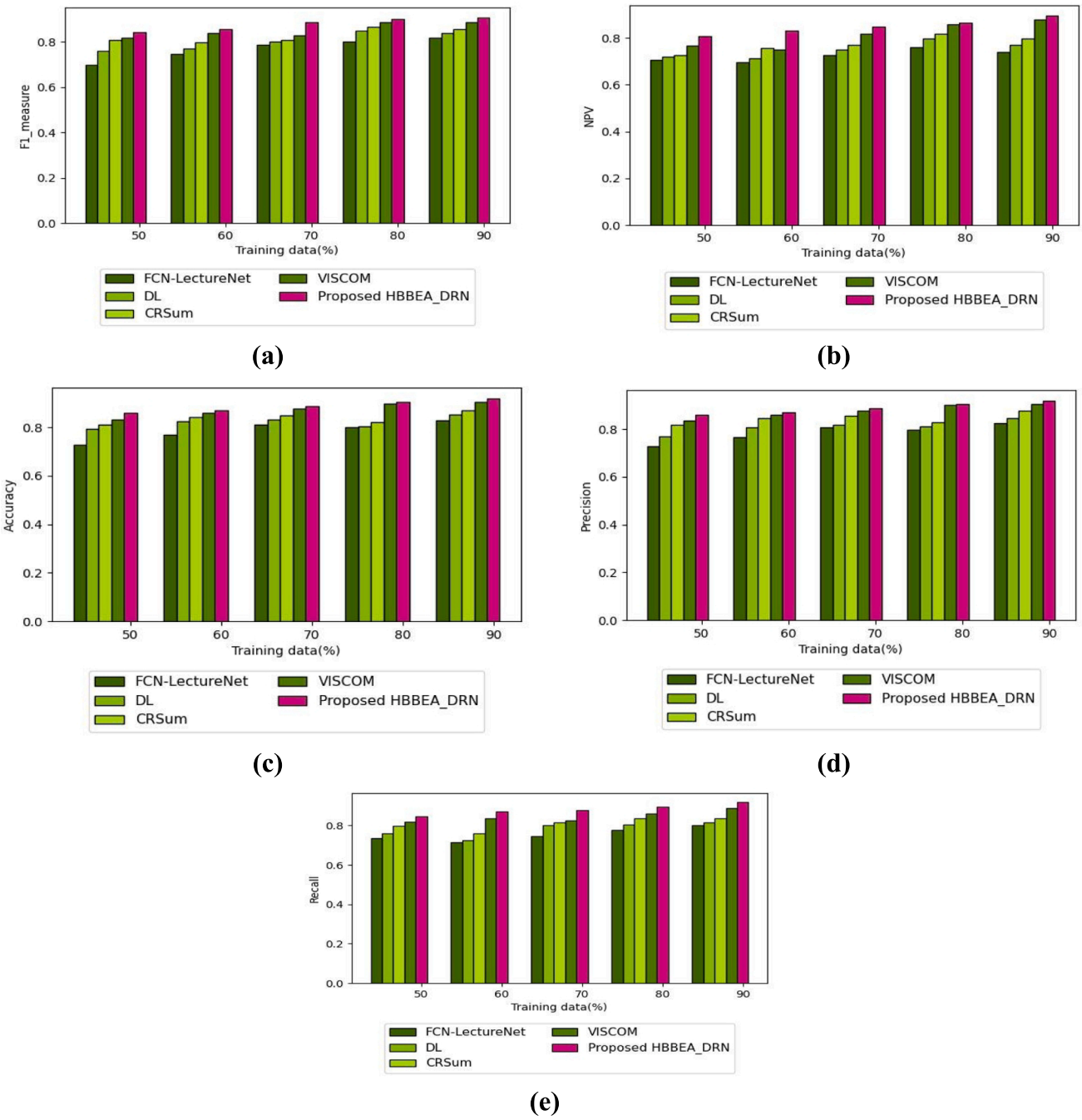


Fig. 6. Audio data evaluation using a) F1-Score b) NPV c) Accuracy d) Precision e) Recall.

enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.683, 0.748, 0.779, 0.815, and 0.844. Besides attaining 90 % training data, the Accuracy enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.806, 0.846, 0.850, 0.888, and 0.913. The precision-assisted evaluation is presented in Fig. 5d). Attaining 50 % training data, the precision enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.649, 0.730, 0.780, 0.817, and 0.847. Besides attaining 90 % training data, the precision enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.818, 0.826, 0.850, 0.890, and 0.915. The evaluation of existing HBBEA_DRN using precision is 10.601 %, 9.726 %, 7.103 %, and 2.732 %. The recall-assisted evaluation is presented in Fig. 5e). Attaining 50 % training data, the recall enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.677, 0.750, 0.819, 0.850, and 0.867. Besides attaining 90 % training data, the recall enumerated by FCN-Lecture Net, DL, CRSum, VISCOM and proposed HBBEA_DRN are 0.817, 0.840, 0.857, 0.897, and 0.926. The evaluation of existing HBBEA_DRN using recall is 11.771 %, 9.287 %, 7.451 %, and 3.131 %.

4.4.2. Audio data evaluation

Fig. 6 endows the audio data evaluation with different metrics by altering the training percentage. The F1-Score-based evaluation is presented in Fig. 6a). Acquiring 50 % training data, the F1-measure enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.696, 0.759, 0.807, 0.817, whereas for HBBEA_DRN is 0.840. Besides acquiring 90 % training data, the F1-measure enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.819, 0.839, 0.855, 0.887,

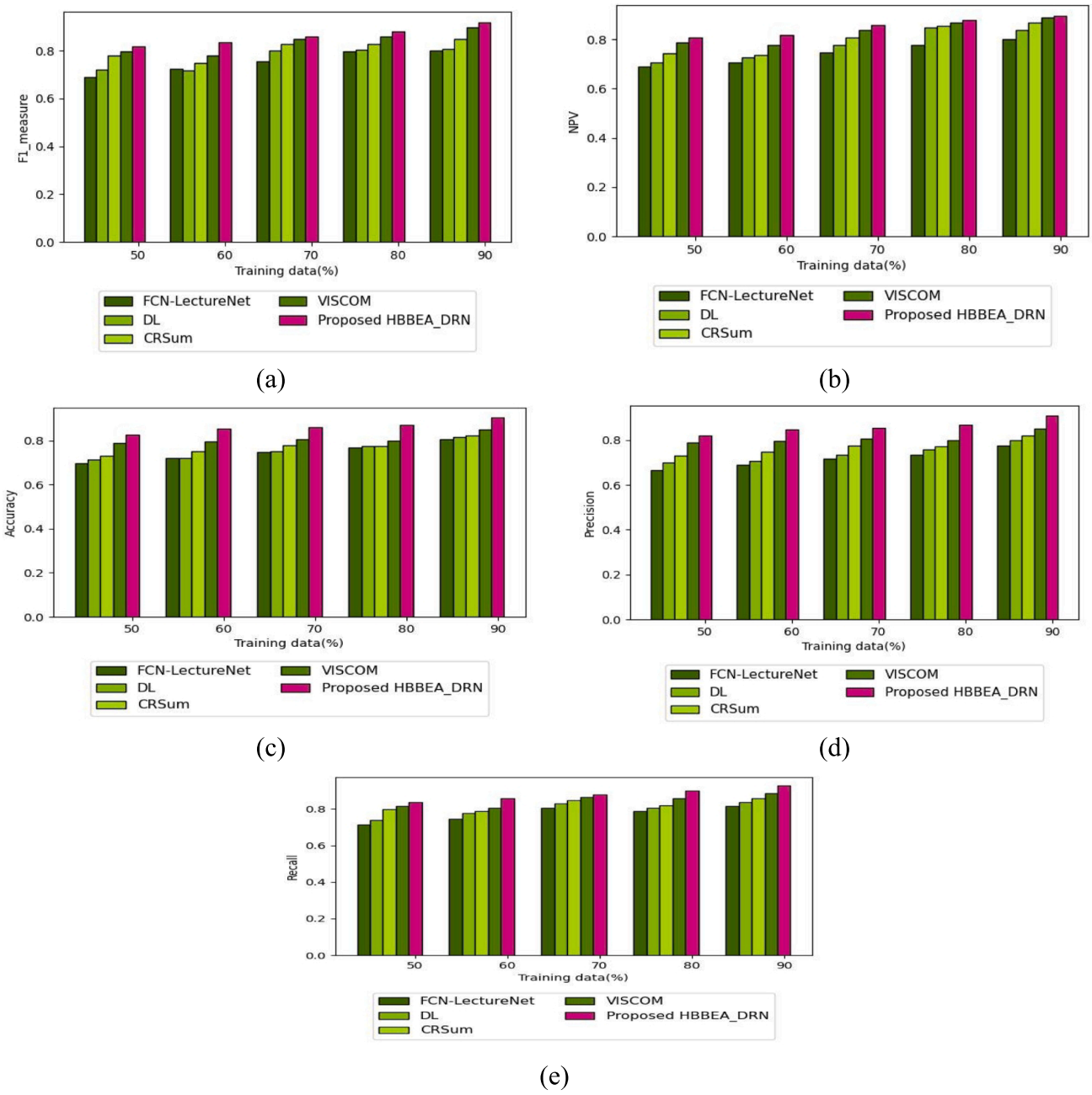


Fig. 7. Video-Audio data evaluation using a) F1-Score b) NPV c) Accuracy d) Precision e) Recall.

whereas for HBBEA_DRN is 0.907. The evaluation of existing HBBEA_DRN using F1-measure is 9.702 %, 7.497 %, 5.733 %, and 2.205 %. The NPV-assisted evaluation is presented in Fig. 6b). Acquiring 50 % training data, the NPV enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.707, 0.719, 0.726, 0.765, whereas for HBBEA_DRN is 0.807. Besides acquiring 90 % training data, the NPV enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.739, 0.769, 0.797, 0.880, whereas for HBBEA_DRN is 0.896. The evaluation of existing HBBEA_DRN using NPV is 17.522 %, 14.174 %, 11.049 %, and 1.785 %. The Accuracy-assisted evaluation is presented in Fig. 6c). Acquiring 50 % training data, the Accuracy enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.729, 0.794, 0.810, 0.831, whereas for HBBEA_DRN is 0.860. Besides acquiring 90 % training data, the Accuracy enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.828, 0.852, 0.869, 0.903, whereas for HBBEA_DRN is 0.918. The precision-assisted evaluation is presented in Fig. 6d). Acquiring 50 % training data, the precision enumerated by FCN- Lecture Net, DL, CRSum, and VISCOM are 0.727, 0.769, 0.817, 0.834, whereas HBBEA_DRN is 0.860. Besides acquiring 90 % training data, the precision enumerated by FCN-Lecture Net, DL, CRSum, and VISCOM are 0.825, 0.847, 0.876, 0.905, whereas for HBBEA_DRN is 0.918. The evaluation of existing HBBEA_DRN using precision is 10.130 %, 7.734 %, 4.575 %, and 1.416 %. The recall-assisted evaluation is presented in Fig. 6e). Acquiring 50 % training data, the recall enumerated by FCN-Lecture Net, DL, CRSum and VISCOM are 0.737, 0.760, 0.797, 0.820, whereas for HBBEA_DRN is 0.847. Besides acquiring 90 % training data, the recall enumerated by FCN-Lecture Net, DL, CRSum and VISCOM are 0.800, 0.817, 0.835, 0.889, whereas for HBBEA_DRN is 0.920. The evaluation of existing HBBEA_DRN using recall is 13.043 %, 11.195 %, 9.239 %, and 3.369 %.

Table 2
Comparative analysis.

Variation	Metrics	FCN-Lecture Net	DL	CRSum	VISCOM	HBBEA_DRN
Video data	**F1-Score (%)**	80.6	81.7	84.0	86.9	**90.9**
	NPV (%)	82.7	86.0	87.0	87.5	**89.7**
	Accuracy (%)	80.6	84.6	85.0	88.8	**91.3**
	Precision (%)	81.8	82.6	85.0	89.0	**91.5**
	Recall (%)	81.7	84.0	85.7	89.7	**92.6**
Audio data	**F1-Score (%)**	81.9	83.9	85.5	88.7	**90.7**
	NPV (%)	73.9	76.9	79.7	88.0	**89.6**
	Accuracy (%)	82.8	85.2	86.9	90.3	**91.8**
	Precision (%)	82.5	84.7	87.6	90.5	**91.8**
	Recall (%)	80	81.7	83.5	88.9	**92**
Video-Audio	**F1-Score(%)**	80	80.7	84.9	89.7	**91.9**
	NPV(%)	80	83.7	86.7	89	**89.6**
	Accuracy (%)	80.7	81.5	82.2	85.0	**90.5**
	Precision(%)	77.6	80	82.0	85	**91**
	Recall(%)	81.7	83.7	85.7	88.7	**92.8**

4.4.3. Video-audio data evaluation

Fig. 7 explicates the audio data evaluation using different measures by shifting training percentages. The F1-Score-assisted valuation is presented in Fig. 7a). Acquiring 50 % training data, the F1-measure explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.688, 0.720, 0.780, 0.799, and 0.816. Besides acquiring 90 % training data, the F1-measure explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.800, 0.807, 0.849, 0.897, and 0.919. The evaluation of existing HBBEA_DRN using Accuracy is 12.948 %, 12.187 %, 7.616 %, and 2.393 %. The NPV-assisted evaluation is presented in Fig. 7b). Acquiring 50 % training data, the NPV explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.690, 0.707, 0.744, 0.786, and 0.807. Besides acquiring 90 % training data, the NPV explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.800, 0.837, 0.867, 0.890, and 0.896. The evaluation of existing with HBBEA_DRN using NPV is 10.714 %, 6.584 %, 3.236 %, 0.669 %. The Accuracy-assisted evaluation is presented in Fig. 7c). Acquiring 50 % training data, the Accuracy explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.697, 0.712, 0.731, 0.790, and 0.825. Besides acquiring 90 % training data, the Accuracy explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.807, 0.815, 0.822, 0.850, and 0.905. The evaluation of existing HBBEA_DRN using ACCURACY is 11.135 %, 8.820 %, 5.402 %, and 0.882 %. The precision-assisted evaluation is presented in Fig. 7d). Acquiring 50 % training data, the precision explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.666, 0.699, 0.730, 0.790, and 0.820. Besides acquiring 90 % training data, the precision explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.776, 0.800, 0.820, 0.850, and 0.910. The evaluation of existing HBBEA_DRN using precision is 14.725 %, 12.087 %, 9.890 %, and 6.593 %. The recall-assisted evaluation is presented in Fig. 7e). Acquiring 50 % training data, the recall explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.715, 0.737, 0.797, 0.817, and 0.837. Besides acquiring 90 % training data, the recall explicated by FCN-Lecture Net, DL, CRSum, VISCOM and HBBEA_DRN are 0.817, 0.837, 0.857, 0.887, and 0.928. The evaluation of existing HBBEA_DRN using recall is 11.961 %, 9.806 %, 7.650 % and 4.418 %.

4.5. Comparative discussion

Table 2 explores techniques evaluation with video, audio and video-audio data with definite metrics. Using video data, the highest F1-score of 90.9 %, NPV of 89.7 %, Accuracy of 91.3 %, precision of 91.5 % and recall of 92.6 % is produced by HBBEA_DRN. Using audio data, the highest F1-score of 90.7 %, NPV of 89.6 %, Accuracy of 91.8 %, precision of 91.8 % and recall of 92 % is produced by HBBEA_DRN. Using video-audio data, the highest F1-score of 91.9 %, NPV of 89.6 %, Accuracy of 90.5 %, precision of 91 % and recall of 92.8 % is produced by HBBEA_DRN. The highest precision is due to DRN which helps to segment the audio precisely by taking imperative features. The highest recall is due to the proposed optimization algorithm whose combined benefits led to better recall. The highest F1 score is due to elevated precision and recall values. Better Accuracy and NPV are attained due to combined usage of DRN and optimization algorithms which helps to detect the positive predicted values and negative predicted values.

5. Conclusion

The goal is to create an optimized deep-learning model for summarizing lecture audio and video content. We start by using a dataset and segmenting the lecture video into shots using the YCbCr color model. Each video shot is then divided into audio and video segments using HBBEA. Next, we extract features from each segmented audio and video frame. For audio, we consider features like BFCC, MFCC, zero crossing, spectral flux, spectral centroid, and spectral bandwidth. Video features include SLBT, LTP, HoG, LOOP, and LVP. After extracting these features, we use a method called DRN to select the most important audio and video segments. We train DRN using HBBEA to fine-tune its weights for optimal performance. Finally, the selected video frames are combined with the audio. The proposed HBBEA-based DRN has demonstrated superior performance, achieving a high F1-Score of 91.9 %, NPV of 89.6 %, Accuracy of 91.8 %, precision of 91 %, and recall of 92.8 %. This approach outperforms others in efficiently summarizing lecture videos. Future work will involve testing the model on additional databases to validate its effectiveness. In conclusion, the HBBEA-DRN combination has proven to be the most effective method for summarizing lecture videos.

CRediT authorship contribution statement

Preet Chandan Kaur: Writing – original draft. **Leena Ragha:** Visualization, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] K. Davila, F. Xu, S. Setlur, V. Govindaraju, FCN-LectureNet: Extractive summarization of whiteboard and chalkboard lecture videos, IEEE Access 9 (2021) 104469–104484.

[2] B. Urala Kota, K. Davila, A. Stone, S. Setlur, V. Govindaraju, Generalized framework for summarization of fixed-camera lecture videos by detecting and binarizing handwritten content, Int. J. Document Anal. Recognition (IJDAR) 22 (3) (2019) 221–233.

[3] Y. Yuan, H. Li, Q. Wang, Spatiotemporal modelling for video summarization using convolutional recurrent neural network, IEEE Access 7 (2019) 64676–64685.

[4] M.V. Mussel Cirne, H. Pedrini, VISCOM: A robust video summarization approach using colour co-occurrence matrices, Multimed. Tools Appl. 77 (1) (2018) 857–875.

[5] H. Gharbi, S. Bahroun, E. Zagrouba, Key frame extraction for video summarization using local description and repeatability graph clustering, SIViP 13 (3) (2019) 507–515.

[6] M. Basavarajaiah, P. Sharma, GVSUM: Generic video summarization using deep visual features, Multimed. Tools Appl. 80 (9) (2021) 14459–14476.

[7] H. Jin, Y. Yu, Y. Li, Z. Xiao, Network video summarization based on key frame extraction via superpixel segmentation, Trans. Emerg. Telecommun. Technol. 33 (6) (2022) 3940.

[8] M. Rafiq, G. Rafiq, R. Agyeman, G.S. Choi, S.I. Jin, Scene classification for sports video summarization using transfer learning, Sensors 20 (6) (2020) 1702.

[9] S.K. Nayak, J. Majumdar, Hybrid method of video shot segmentation based on YCbCr space color model, J. Eng. Res. Rep. 20 (11) (2021) 8–17.

[10] H.A. Alsattar, A.A. Zaidan, B.B. Zaidan, Novel meta-heuristic bald eagle search optimisation algorithm, Artif. Intell. Rev. 53 (3) (2020) 2237–2264.

[11] F.A. Hashim, E.H. Houssein, K. Hussain, M.S. Mabrouk, W. Al-Atabany, Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems, Math. Comput. Simul 192 (2022) 84–110.

[12] Kumar, C., Rehman, F., Kumar, S. and Mehmood, A. and Shabir, G.,"Analysis of MFCC and BFCC in a Speaker Identification System", In proceedings of International Conference on Computing, Mathematics and Engineering Technologies, 2018.

[13] Hassan, A.R. and Haque, M.A., "Computer-aided sleep apnea diagnosis from single-lead electrocardiogram using dual-tree complex wavelet transform and spectral features," In Proceedings of International Conference on Electrical & Electronic Engineering (ICEEE), pp. 49-52, 2015.

[14] Z. Chen, Y. Chen, L. Wu, S. Cheng, P. Lin, Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions, Energ. Conver. Manage. 198 (2019) 111793.

[15] Liu, T. and Kender, J.R, "Lecture videos for e-learning: Current research and challenges," In IEEE Sixth International Symposium on Multimedia Software Engineering, pp. 574-578, 2004.

[16] S. Repp, A. Gross, C. Meinel, Browsing within lecture videos based on the chain index of speech transcription, IEEE Trans. Learn. Technol. 1 (3) (2008) 145–156.

[17] Ngo, C.W., Wang, F. and Pong, T.C, "Structuring lecture videos for distance learning applications," In Fifth International Symposium on Multimedia Software Engineering, pp. 215-222, 2003.

[18] Davila, K. and Zanibbi, R., "Whiteboard video summarization via spatio-temporal conflict minimization," In 14th IAPR International conference on document analysis and recognition (ICDAR), vol.1, pp. 355-362, 2017.

[19] G.C. Lee, F.H. Yeh, Y.J. Chen, T.K. Chang, Robust handwriting extraction and lecture video summarization, Multimed. Tools Appl. 76 (5) (2017) 7067–7085.

[20] H.B. UlHaq, M. Asif, M.B. Ahmad, R. Ashraf, T. Mahmood, An effective video summarization framework based on the object of interest using deep learning, Math. Probl. Eng. (2022).

[21] T. Sebastian, J.J. Puthiyidam, A survey on video summarization techniques, Int. J. Comput. Appl 132 (13) (2015) 30–32.

[22] Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J. and Yokoya, N., "Video summarization using deep semantic features," In Asian conference on computer vision, pp. 361-377, 2016.

[23] T. Chakraborti, B. McCane, S. Mills, U. Pal, LOOP descriptor: Local optimal-oriented pattern, IEEE Signal Process Lett. 25 (5) (May 2018).

[24] Li, Su, and Yang, Y., "Power-Scaled Spectral Flux and Peak-Valley Group-Delay Methods for RobustMusical Onset Detection", In ICMC,2014.

[25] Lakshmi Prabha, N. S. and Majumder, S. "Face Recognition System Invariant to Plastic Surgery", In Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, pp. 258-263, November 2012.

[26] X. Tan, B. Triggs, Enhanced local texture feature sets for FaceRecognition under difficult lighting conditions, IEEE Trans. Image Process. 19 (6) (2010) 1635–1650.

[27] M. Zhang, T. Wu, K.M. Bennett, Small blob identification in medical images using regional features from optimum scale, IEEE Trans. Biomed. Eng. 62 (4) (September 2014) 1051–1062.

[28] G. Sharma, K. Umapathy, S. Krishnan, Trends in audio signal feature extraction methods, Appl. Acoust. 158 (2020) 107020.

[29] K.C. Fan, T.Y. Hung, A novel local pattern descriptor—local vector pattern in high-order derivative space for face recognition, IEEE Trans. Image Process. 23 (7) (2014) 2877–2891.

[30] B. Zhao, M. Gong, X. Li, AudioVisual video summarization, IEEE Trans. Neural Networks Learn. Syst. 34 (8) (2021) 5181–5188.

[31] H. Li, J. Zhu, C. Ma, J. Zhang, C. Zong, Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video, IEEE Trans. Knowl. Data Eng. 31 (5) (2019) 996–1009.

[32] J. Zhou, D. Guo, M. Wang, Contrastive positive sample propagation along the audio-visual event line, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7239–7257.

[33] Zhou, J. et al., "Audio–Visual Segmentation," In Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13697, 2022.

[34] Zhou, J., Guo, D., Zhong, Y., and Wang, M., Improving Audio-Visual Video Parsing with Pseudo Visual Labels, ArXiv, 2023.

[35] Shen, Xuyang, Li, D., Zhou, J., Qin, Z., He, B., Han, X., Li, A., et al. "Fine-grained audible video description." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10585-10596, 2023.

[36] Tian, Yapeng, Guan, C., Goodman, J., Moore, M. and Xu, C., "Audio-visual interpretable and controllable video captioning," In IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops, 2019.

[37] Z. Li, D. Guo, J. Zhou, J. Zhang, M. Wang, Object-aware adaptive-positivity learning for audio-visual question answering, Proc. AAAI Conf. Artif. Intell. 38 (4) (2024) 3306–3314.

[38] Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J. and Wilson, K. "CNN architectures for large-scale audio classification," In 2017 IEEE international conference on acoustics, speech and signal processing (icassp), pp. 131-135, 2017.

[39] Carreira, J. and Zisserman, A. "Quo Vadis, action recognition? A new model and the kinetics dataset," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308, 2017.

[40] Z. Yunzuo, Y. Liu, C. Wu, Attention-guided multi-granularity fusion model for video summarization, Expert Syst. Appl. 249 (2024) 123568.

[41] Y. Zhang, T. Liu, P. Yu, S. Wang, R. Tao, SFSANet: Multiscale object detection in remote sensing image based on semantic fusion and scale adaptability, IEEE Trans. Geosci. Remote Sens. 62 (4406410) (2024) 1–10.

[42] Z. Yunzuo, C. Wu, W. Guo, T. Zhang, W. Li, CFANet: Efficient detection of UAV image based on cross-layer feature aggregation, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11.

[43] Y. Zhang, T. Zhang, C. Wu, R. Tao, Multi-scale spatiotemporal feature fusion network for video saliency prediction, IEEE Trans. Multimedia 26 (2024) 4183–4193.