



Predicting Episodic Video Memorability Using Deep Features Fusion Strategy

Hasnain Ali, Syed Omer Gilani, Muhammad Jawad Khan,
Mohsin Jamil and Muazzam Khattak

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

February 24, 2022

PREDICTING EPISODIC VIDEO MEMORABILITY USING DEEP FEATURES FUSION STRATEGY

Abstract:

Video memorability prediction has become an important research topic in computer vision in recent years. The movie's input is highly remembered that gains much attention with unbounded time constraints. Episodic memory is a fascinating research area that needs much attention using video processing tools and techniques. Episodic memories are long-lasting with complete detail. Movies are one of the best instances of episodic memory. This paper proposes a novel framework to fuse deep features to predict the probability of recalling episodic events. Memories are reproducible and sensitive to sophisticated set of properties rather than low-level properties—the proposed framework pin up the fusion of text, visual and motion features. A fuzzy-based FastText model, a supervised text extraction module, is designed to extract the annotations with their relevant classes. The colour histogram analysis is done to determine the dominant colour region that performs as a connected fragment to form episodic video sequences. A novel Faster R-CNN is designed to discover the scene objects using an informative regional proposal network formation. Here, the modified loss function sorts out the lowest overlapping regions yielding the best proposals. The 'high-level' properties are collected using Principal Component Analysis (PCA) to form episodic shots. These are fused to estimate the memorability score. The proposed framework is implemented in Mediaeval 2018 datasets. A superior spearman's rank correlation result is achieved as 0.6428 short-term and 0.4285 long-term memorabilities than the latest comparable methods.

Keywords: Video memorability prediction; episodic memory; Faster R-CNN; Fuzzy based FastText; Regional Proposal Network & MediaEval 2018 datasets.

1. Introduction

Digital users are more likely to share videos online with the advent of information processing tools and technologies. In the era of information overload, it isn't effortless to get in the information they are interested in. People tend to recall fascinating events in the short/long term. Creating unique video content is crucial for engaging digital users with profitable marketing campaigns [1-3]. Henceforth, predicting and perceiving the video's memorability is one of the significant parts of the video analysis task [4,5]. Episodic memory is a late-developing approach that remembers and recalls the interesting events relevant to past experiences. It is described as the memory system that takes charge of preserving, encoding, and retrieval experienced events interacted with the accurate information of spatial and temporal context of encoded features. Episodic memory is not a typical memory system which probably unique to humans. It takes mental time to travel to recollect and re-experience the events from the present to the past through auto noetic awareness. Episodic learning-based video memorability prediction is the minor focused research area that motivates us to delve into this study [6].

Episodic learning systems portray as the essential foundation of our recollections. Human brains incessantly bombard the rich set of collective events. However, only a tiny fraction of events are recalled and crystallized into episodic memories. Many studies have stated that all events are not transformed into episodic memories. A human brain selects and interprets certain events and transforms them into narrative forms, represented as 'episodic memories'. It is significant to characterize each event and its learning modules to evaluate

those memories by learning the structural pattern of episodic memories. The memory recalls analysis relies on multiple factors such as

- Intrinsic information of subjects towards oneself, i.e. age, domain construction
- Relatable content evaluation, i.e. single item analysis, the similarity between formed events.
- Probing the memories concerning time, i.e. validation of encoding and testing
- Evaluation of recollection using objective and subjective metrics.

A movie is one of the stimuli that help investigate the memories of real-life events. Several aspects of episodic information collect from a movie [7, 8]. However, the information deduction from a single item is a complex process because it involves temporal sequences, spatial and temporal context, affective components and an underlying narrative. Though the subject is quite interesting, the collection of comprehensive memories for movie is assessed by an effective recall rate with précised information. Likewise, specific scenes create a more significant impact on human minds in the short term and the long term. Object detection is one of the well-known areas widely explored by deep learning algorithms in terms of the better scenes detection process. The objects can't be detected properly with the limited set of feature information [9, 10]. Due to their fixed dimensional locations, generating the fittest anchors in object detection is still a challenging task. A considerable number of anchors is required to estimate a recall rate that holds many irrelevant samples enclosed during the anchor extraction module. The general anchor generation method will have such problems, resulting in poor detection performance that lowers the memorability score predictions.

This paper proposes a novel video memorability framework using multi-modal features by improvising the combinational feature sets. This method can provide better short-term and long-term memorability score prediction comparable to the latest methods. The enrichments done in this paper are:

- a) A novel FastText-FIS model is designed to compute the short-term and long-term annotations with their text classes. In order to maintain the strength of semantic tasks of the sub-words, the proposed module has improved the weight estimation using fuzzy logic.
- b) A visual attention feature is extracted using a colour histogram that leverages the colour of the frames to create better visual information.
- c) A novel FR-CNN model is designed to detect the objects scene with the classified text. The modified loss estimation removes the irrelevant regions of a frame with the required proposals.
- d) The best combinations of annotations, dominant colours and scenes objects are trained and tested on the Mediaeval 2018 datasets.
- e) Experimental results have proved the efficacy of the proposed framework in terms of spearman's rank correlation, recall, mAP and F-measure metrics.

This paper arranges into sections:

“Literature Survey” that portrays the merits and demerits of the existing studies is given in Section 2.

“Proposed framework” that portrays working steps of each module in Section 3.

“Experimental Results and Discussion” that portrays the simulation setup, performance metrics and the achieved results in Section 4.

“Conclusion” that portrays the findings of this study is given in Section 5.

2. Literature Survey

The previous works related to features engineering, i.e. text, audio and motion for a better video memorability prediction model, are discussed. The review study conducts in two aspects, namely, text and visual feature analysis and motion feature analysis.

2.1 Text and visual feature analysis:

In [11], some deeply learned visual and textual features were extracted by Glove, C3D and I3D. These extracted features were trained and tested using regression models. Finally, the fusions of all features were passed through different regressor lines to estimate the memorability score. Though label information helps, it is not proper to use one-hot vectors. A practical set of features play a crucial role to address the continuous emotion in movies. Thus, the features such as Video Compressibility and Histogram of Facial Area (HFA) [12] were extracted and fed into the Mixture of Experts (MoE) that adaptively merged the emotional information. Finally, the Expectation-Maximization (EM) algorithm has estimated the memorability score. Different emotions are available in movies. However, the familiar emotions are considered by limiting the high-level semantic information. Convolutional Neural Network (CNN) was employed to extract the visual features, and then, the score was predicted using Long Short Term Memory (LSTM) network [13]. Inception V3-CNN and C3D features were also considered. In [17], the AMNet was an end-to-end architecture with soft attention and LSTM network. The transfer learning module and LaMem datasets were used to estimate the memorability score. It was observed that the highest attention maps of a video frame depict the highly memorable visual contents. Regardless, concept and semantic representations of the video frame are not focused.

A combination of visual approaches has been studied to predict the memorability score using the weighted average method [19]. I3D, ResNet -152 and ResNet-101 were concatenated on the video frames. Each frame is represented with captions to derive the local embedding's. Likewise, the captions were analysed using the combination of textual features such as self-attention, BERT and bag of words. The memorability decay of a scene is still not addressed. A hybrid approach was designed to compute the media memorability using fusion strategy [20]. The caption data were fed into the different embedding and recurrent layers of ResNet and AMNet. Then, data features were reduced to the required dimension levels until estimating a memorability score. The semantics information of video has described the efficiency for short-term videos. GloVE [21] is a Global Vector for Word Representation that depicts the fine-grained semantic and syntactic information. Creation of word vectors comprises two stages, viz, global matrix factorization and local context window methods. Finally, the weighted least square model trained the global word-word co-occurrences counts. However, the semantic features of each sub-word are not appropriately trained.

2.2 Motion features analysis

Feature extraction of informative regions using Local Binary Pattern (LBP) [14] was studied to determine the context's significance in the particular area. Compared to the SIFT feature descriptors, the LBP towards the region presented more intrinsic frame information. Human detection using a histogram of gradients [15] was studied. The collection of motion features from a video is a challenging task that was scrutinized by the histogram of edge pixels. However, the resolution of a frame distorts the specified region detection. Densenet CNN [16] was introduced to develop deep connection layers in a feed-forward fashion.

Feature maps helped gain collective knowledge of the network that kept the feature maps unchanged. This quality of CNN has minimized the loss function on deep network architectures. It has resolved the overfitting reduction in the region discovery process. Convolutional networks are widely adopted in various benchmarks. The computational efficiency and the low parameters optimization are the factors that deprive the object detection. Thus, the factorized convolutions and aggressive regularization [18] were computed on the ILSVRC 2012 classification datasets. It has achieved substantial gains with a lower error rate.

AMNet with the attention memorability estimation [22] was designed to generate an attentive mechanism on image regions. The incremental recall score via recurrent network has increased the performance gain of image regions and then classified using transfer learning. Finally, the classified images were grouped to estimate the memorability score. The attention map generation creates a sparsity issue in the more significant regions. In [23], spatiotemporal features' using 3D convolutional networks was studied on the large supervised video dataset. The compactness of the video descriptor is not achieved using C3D features. Action classification is one of the subsets of the object detection approach. Two-Stream Inflated 3D ConvNet (I3D) [24] was designed to detect the objects with their respective action classes. The kinetics information helped in the pre-training module for the faster classification module. The transferability of I3D models has increased the processing time on the test sets, and longer videos are not appropriately studied. In [25], the temporal segment networks were formulated by the idea of long-range temporal structure modelling. The longer videos were supervised using ConvNets to achieve faster action recognition. However, it yielded better outcomes for the limited training samples.

The conducted reviews show that the best combination of feature sets has a profound contribution to memorability score estimation. The features like text, visuals and motion have been limited according to the study's requirements. The collective set of events will impact the cognition ability, and thus, interpretations of the attentive features necessitate a better feature space transformation process. The memorability score prediction of episodic memory formation on movie events develops a tedious task using text and motion features.

3. Proposed Framework:

The proposed framework resolves the following research questions:

RQ-1: What are the key features that impact the cognitive ability for long-lasting?

RQ-2: How does the fused set of features impact the recall of recollecting the events?

RQ-3: How to effectively fuse and achieve the best combinational sets from the extracted features?

The proposed framework is explained as follows:

3.1 Video into Frame conversion:

The input videos are collected from the public repository, MediaEval 2018, which contains the set of Hollywood movies. As an initial step, the video converts into a set of frames. A rich set of information frames is considered to estimate the memorability score. Thus, the theme of the proposed study is to select the frames with rich data using multi-modal in-depth features selection process.

3.2 Text features:

FastText method is a word embedding technique that embeds the words with the available set of captions. New words affect the estimation of semantic information loss, which is resolved by designing a novel FastText method. The conventional FastText method performs on continuous skip-grams without eliminating the morphology of words. It takes each word as character n-gram. This model provides a vector for unknown words during the training process. Though it preserves sub-word information, it ignores the internal structure words. Hence, the novel FastText method improves the word embeddings during the representation of unknown words. Firstly, it trains the word vectors by the FastText model and fills in unknown word vectors by combining n-gram models. Secondly, the Fuzzy Inference System (FIS) idea is used to discover the multiple word vectors similar to the unfamiliar words. Finally, the numerous word vectors are fused to form a new representation of word vectors by the gate mechanism. The proposed steps of fuzzy-based FastText model are:

i) Embedding layer:

FastText embedding by way of 500 dimensions signifies video captions. A fundamental preprocessing step, i.e. stopwords and space, are removed. And every unique word is taken to create a dictionary. Likewise, a unique index is created for each individual word. Therefore, the input captions will combine uncommon words with their respective index. It is collectively represented as word embedding. For instance, let C be the caption holding complete sentences; a word embedding $word_i$ means each word in the captions C . Here, $word_i \in T^d$ and d are the dimension of the word embedding. In the proposed implementation, 500-dimensional FastText embedding is used and thus, it is stacked together to form a word embedding matrix WE_{mat} where $WE_{mat} \in T^{length \times dimension}$. Here, length denotes the maximum length of the sentence.

ii) Fuzzification:

Here, the crisp input values, WE_{mat} are transformed into a membership degree in the company of membership function. The input and output variables are represented in linguistic terms. The membership functions of the linguistic terms are obtained from the video captions data. The triangular membership functions are used to model the membership degrees of all variables under FIS. IMDB dataset [27] is used to design fuzzy rules. The range of the input variable, WE_{mat} under FIS are represented as, very high (VH), high (H), medium (M), low (L) and very low (VL). Finally, the range of $weight_{WE_{mat}}$ is represented as high (H), medium (M) and low (L) membership functions.

iii) Fuzzy rules:

Relied on the final weight of each sentence on the input variable, WE_{mat} , the if-then rules are expressed as follows:

If (WE_{mat} is VH and label is positive) then ($weight_{WE_{mat}}=H$).

iv) Defuzzification:

Standard centroid estimation estimates the weight for every sentence index with the baseline from aggregated output fuzzy set.

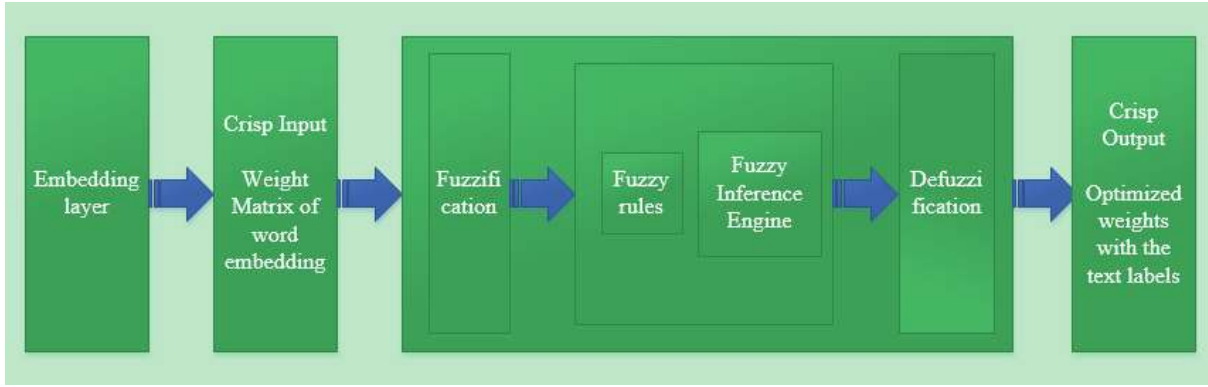


Fig. 1. Proposed Fuzzy based FastText model

3.3 Visual features:

The visual features play a significant role in generating episodic memories. Especially, colours have a unique ability to recollect interesting events. The colour histogram method is widely used in this study to derive the colour of the frames. Colour is a stimulus of human vision about Rd (red), Gr (Green) and Bl (Blue), which forms a colour space. Every region pixel is quantized from RGB colour space to HSV colour space. The HSV colour space [26] is formulated as:

$$H = \frac{\cos\left(\frac{Rd-Gr}{2} + \frac{Rd-Bl}{2}\right)}{\sqrt{(Rd-Gr)^2 + (Rd-Bl)(Gr-Bl)}}$$

$$S = 1 - \left(\frac{3}{Rd+Gr+Bl}\right) \times \min(Rd, Gr, Bl)$$

$$V = \frac{1}{3} (Rd + Gr + Bl)$$

Here, greys' hue value is defined as '0' for easy computation. Each colour component is quantified as H=20 bins, S=4 bins and V= 4 bins. At last, 16 * 4 * 4 histograms are concatenated to get a 256-dimensional vector. The colour histogram value is computed once the frame is converted into the HSV.

3.4 Motion features:

The motion features of the input video are extracted using a novel Faster Region Convolutional Neural Networks (FR-CNN). This step aims to segment the objects for an effective recall rate on memorable objects. The proposed FR-CNN is divided into four parts: Convolutional layers, Region proposal networks, Region of Interest (RoI) pooling, and classification layers.

3.4.1 Convolutional layers:

The traditional CNN contains 13 Conv layers, 13 ReLu layers, and 4 pooling layers. Here, the input matrix's length and width will change to half of the original length and width in the pooling layers.

3.4.2 Region Proposal Networks (RPN):

The frames with the objects are identified using RPN, which increases the identification speed of the frame generation process. It contains two steps: one is to estimate the softmax function using positive and negative anchor classification, and the other is to find the anchor points using bounding box regressions. Finally, the proposal layer synthesizes the positive anchor point with the relevant bounding box regression values. The proposals with small and out of boundary points are eliminated.

a) Anchors:

Let p be the maximum number of candidate boxes at each sliding window position. Each candidate box represents the class background, including a $4p$ regression layer a $2p$ classification layer. Then, the p candidate frames are parameterized, known as anchors. In general, the anchor is the middle point of the sliding window aspiring to the scale and aspect ratio. It is concluded that the convolution feature map size $W * H$ including an aggregate of $W * H * p$ anchors.

b) Estimating the anchors:

After estimating class background using feature map extraction, the convolution $1 * 1$ is employed to check the nine anchors at each pixel belonging to positive or negative classes from $H * W * 18$ matrix. The use of the softmax function is to determine whether objects are presented.

c) Bounding box regression:

The role of bounding box regression is to screw up the detection box spot. As we have taken the movie dataset, the anchor may contain multiple objects and thus, it is to be fine-tuning accurately. Assume a sliding window use 4-dimensional vector (m,n,h,w) representing the centre point coordinates (m, n) , height (h) , and width (w) . Let, $Anc = (Anc_m, Anc_n, Anc_h, Anc_w)$ be the anchor and $GT = (GT_m, GT_n, GT_h, GT_w)$ be the ground truth, then the regression of identifying the frame is done by transformation $GT' = F(Anc)$ such that $GT' \approx GT$. This is proposed in the FR-CNN. The proposed transformation network consists of four functions representing, $F = \{d_m(Anc), d_n(Anc), d_h(Anc), d_w(Anc)\}$ that translates the center points, and scales the height and width as:

$$GT'_m = Anc_w d_m(Anc) + Anc_m$$

$$GT'_n = Anc_h d_n(Anc) + Anc_n$$

$$GT'_w = Anc_w \exp(d_w(Anc_w))$$

$$GT'_h = Anc_h \exp(d_h(Anc_h))$$

By taking the ground truth and anchors on the regression function, the output is given as:

$$d * (Anc), * \in \{m, n, w, h\}$$

$$d_*(Anc) = W_*^T \phi(Anc)$$

Where,

$\phi(Anc)$ - The feature map input.

The modified loss function is:

$$Modified_Loss = t_{ab} \ln (O_b W_*^T \phi(Anc))$$

Where,

t_{ab} represents the offset between ground truth and anchor

$O_b(x_a, \theta)$ represents the output for the given input.

\ln is the natural log.

$$t_m = \frac{GT_m - Anc_m}{Anc_w}$$

$$t_n = \frac{GT_n - Anc_n}{Anc_h}$$

$$t_w = \ln\left(\frac{GT_w}{Anc_w}\right)$$

$$t_h = \ln\left(\frac{GT_h}{Anc_h}\right)$$

The function W is optimized as:

$$\hat{W}_* = \arg \min_{W_*} \sum_i^S |t_*^i - W_*^T \phi(Anc)| + \gamma |W_*|$$

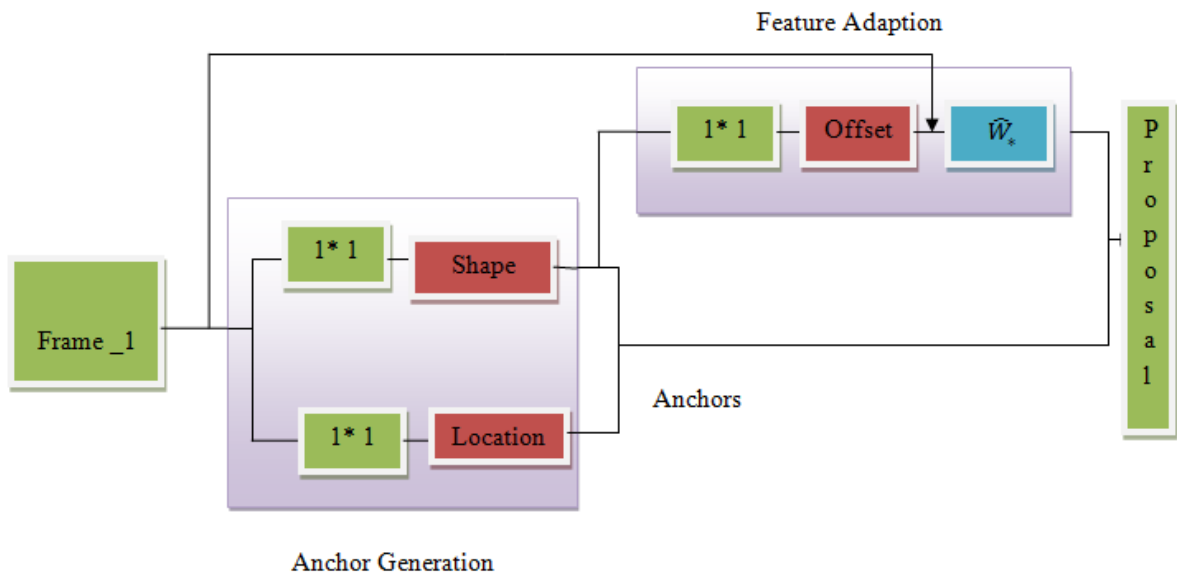


Fig. 2. Proposed Anchor generation

d) Proposal layer:

The role of the proposal layer is to estimate an accurate proposal from the coordinates of regression matrix $H \times W \times 4p$. The steps are:

- Finding the proposal from the above estimated RPN predicted offset and positive bounding box.
- Processing the bounding box that goes above the frame boundary
- Eliminating the bounding box more minor than the threshold

- Sort all (proposal, objectiveness score) pairs from the highest to the lowest value.
- Take the top (N) of all proposals.
- Estimate the non-maximum suppression on the positive bounding box.
- Again take the top (N) of all proposals.
- Exit the remaining proposal.

e) Region of Interest (RoI) pooling layers:

The role of RoI pooling is to modify each proposal concerning the size of $W^* H$ through pooling layers. The maximum pooling is analyzed on each part of $W^* H$ grid.

f) Classification layers:

Finally, this layer estimates the object class of each proposal by using proposal feature maps, fully connected layers and softmax layers. The class probability vector is achieved using bounding box regression until the required target frame is detected.

3.5 Feature reduction process:

The extracted features' dimensions are reduced using Principal Component Analysis (PCA). Since the proposed study intends to develop an episodic memory, the dimensions of the extracted features are reduced. PCA determines the eigenvectors of a covariance matrix with the highest Eigenvalues by following certain steps. The steps in PCA are:

- Let $X = \{x_1, x_2, \dots, x_n\}$ be the set of features for each video frame.
- Finding the mean for text, visual and motion features for each video stream
- Estimating the correlation matrix for all features.
- Finding the Eigen vectors and Eigen values from the correlation matrix.
- Finding the principal components (PC)
- Representing each frame as a linear combination of basis vectors.

To improve the descriptiveness, the PCs are reduced to 256 dimensions. Each video sequence has motion information, and thus, it is transformed into a single feature vector. The temporal dimensional issue is also handled here by generating a vector of fixed dimensions. After this process, two memorability scores are computed for each video sequence.

3.6 Late fusion

The final memorability score is calculated. The best combination of all features and their respective memorability score is optimized for the v^{th} video is given as:

$$\arg \min_{q_n} (q_n y_n - \text{label}_v)$$

Where,

$y_n \rightarrow$ Memorability score given by n^{th} stream of a video;

$\text{label}_v \rightarrow$ Memorability label of the v^{th} video;

$q_n \rightarrow$ Weight assigned to y_n

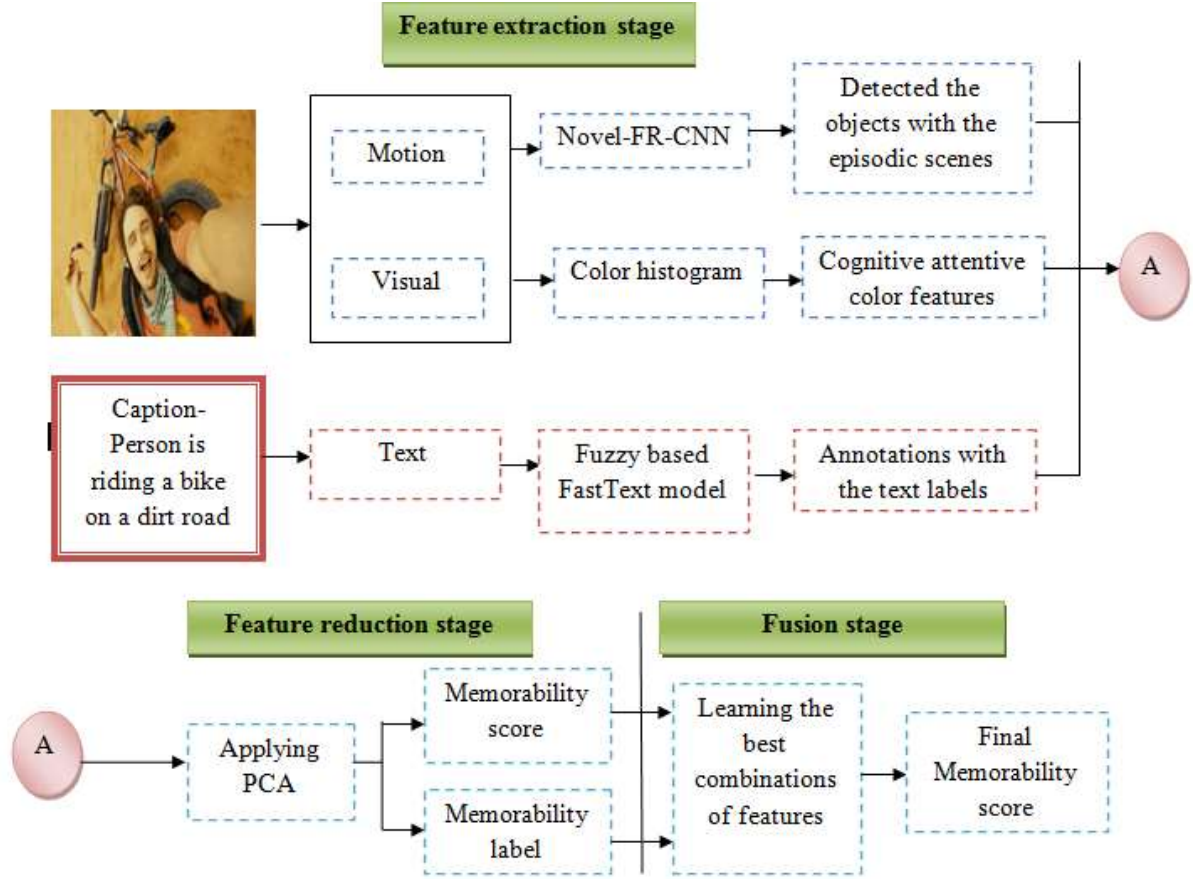


Fig. 3. Proposed episodic memorability framework- 3 deep features

4. Experimental settings:

The MediaEval 2018 dataset is analyzed to certify the efficacy of the proposed framework. It contains 6000 labelled videos with the text description and varied types of scene. The videos are labelled with the two memorability scores ranging from 0-1 representing the short-term and long-term memorability. The movie videos are divided into two sets, namely, 5000 training set and 1000 testing set. Since the video memorability prediction is viewed as a regression problem; the performance of the proposed framework is measure using spearman’s rank correlation, $\rho \in [-1, 1]$. The proposed framework is trained separately for short-term and long-term memorability score estimation. Then, both the scores are fused together according to the generation of episodic memory. The proposed framework is implemented in MATLAB 2019A.

4.1 Evaluation metrics:

The proposed framework is compared with the state-of-the-methods to show the efficiency of the proposed framework in terms of estimating the spearman rank coefficient, recall and F-measure for both the short-term and long-term memorability.

a) Spearman’s rank coefficient:

Spearman’s rank coefficient is to discover and test the linear relationship between ground truth and the predicted truth sets of data. It is expressed as:

$$r = 1 - \frac{6 \sum d^2}{n^3 - n}$$

Where,

$d \rightarrow$ difference between ground truth and the predicted truth data.

$n \rightarrow$ number of data samples.

b) Recall:

Recall defines the ability of the predictor system. It is a statistical measure that aligns with the ground truth values and the predicted values to the test video. It is expressed as:

$$Recall = \frac{\sum_{j=1}^T (a_j - \bar{a})(g_j - \bar{g})}{\sqrt{\sum_{j=1}^T (a_j - \bar{a})^2} \sqrt{\sum_{j=1}^T (g_j - \bar{g})^2}}$$

Where,

$T \rightarrow$ the aggregate count of test frames;

$g_j \rightarrow$ the ground truth value of j^{th} frame;

$\bar{g} \rightarrow$ the mean ground truth value;

$a_j \rightarrow$ the predicted value of the j^{th} frame;

$\bar{a} \rightarrow$ the average predicted value

c) Mean Average Precision (mAP):

Mean Average Precision (mAP) is estimated from the detection of objects using FR-CNN. It is estimated from the ratio of proposal regions between predicted bounding box and ground truth bounding box. It is expressed as follows:

$$mAP = \frac{1}{N} \sum_{k=1}^N AP_k$$

Where,

$N \rightarrow$ Number of proposal regions

d) F-Measure:

It is computed by harmonic mean analysis between the achieved mean average precision (mAP) and recall values. It is expressed as,

$$F - measure = \frac{2 * mAP * Recall}{mAP + Recall}$$

4.2 Results and Discussion:

Here, 5000 labelled videos are taken to build the training predictor system. The training predictor module is done separately for short-term memorability and long-term memorability.



Fig. 4. Sample frames for short-term collected from 127-hours movie.



Fig. 5. Sample frames for long-term collected from 127-hours movie.

To begin the experiments, the captions are analyzed using the fuzzy based FastText model. IMDB dataset is used for labelling purpose.

Table 1: Sample captions and their word embeddings

Sequence Name	No. of captions	1	2	3	4	5
127_hours_2000_2010_1	10	0.4	0.52	0.7	0.4	0.2
127_hours_2182_2192_5	10	0.75	0.46	0.43	0.7	0.8
127_hours_271_281_1	10	0.94	0.67	0.45	0.43	0.75
127_hours_285_295_2	10	0.5	0.56	0.45	0.78	0.2
2001_A_Space_Odyssy_1110_1120_4	10	0.23	0.32	0.47	0.68	0.2
2001_A_Space_Odyssy_1205_1215_5	10	0.46	0.81	0.49	0.67	0.93

Table 2: Fuzzy rule formation

Rule	Word embeddings matrix	Label	Weight of the word embedding matrix
R1	Low	Negative	Low
R2	High	Negative	Medium
R3	Low	Positive	Medium
R4	High	Positive	High

Table 3: Performance of the Fuzzy based FastText model:

Sequence Name	Precision	Recall	F-measure	No. of rules
127_hours_2000_2010_1	89.42%	64.23%	75.34%	12
127_hours_2182_2192_5	79.18%	76.92%	74.89%	56
127_hours_271_281_1	59.23%	71.95%	68.23%	34
127_hours_285_295_2	67.02%	43.23%	81.40%	87
2001_A_Space_Odyssey_1110_1120_4	84.32%	80%	66.34%	15
2001_A_Space_Odyssey_1205_1215_5	73.91%	89.19%	58.18%	34

The table 3 presents the performance analysis of the fuzzy based FastText model. The word embedding matrix is used as input to leverage the weight of the word embeddings. The fuzzy values are small and the numbers of captions are different for each video sequence.

Pertaining to this, the visual features are extracted using colour histogram values. The colour look-up table is referred to extract the histogram values of the captured frames in short-term and long-term. The main use of color histogram values is to find out the dominant and non-dominant regions for better visual perceptions. Finally, novel FR-CNN is employed to extract motion features by forming transformative regional proposal networks. It is quite common issue that, a frame can have multiple objects which is effectively resolved in this study. The impacts of the number of groups on multiple regions are considered. The proposed framework increases the detection performance by adjusting the intersected group of regions. The size of the object is limited in this study.

Table 4: Proposal network training – Parameter settings

Network parameters	Values settings
Initial learning rate	0.01
Gamma	0.1
Momentum	0.9
Weight decay	0.005
Max. number of iterations	60,000
Learning rate for first 10,000 iterations	0.01
Learning rate for next 50,000 iterations	0.001
Number of intersected regions	1 to 5

The detection performance of novel FR-CNN is evaluated using mean Average Precision (mAP). The modified loss function minimizes the computational complexity of proposal joining network during training process. It is fine-tuned according to the network parameters. The mAP value increases significantly when the number of intersected regions increases from

1 to 3. Therefore, the detection efficiency is achieved by considering the number of intersected regions to be 3.

Table 5. Number of intersected regions with respect to the text feature extraction module.
















Number of intersected regions					Time (sec)	mAP (%)
1	2	3	4	5		
					1.24	72.34%
					1.56	77.01%
					1.67	79.23%
					1.89	74.34%
					1.64	77.09%

Table 6: Performance of Recall and F-Measure for episodic video sequences.

Metrics	Short-term memorability	Long-term memorability
Recall	90.4762	90.4762
F-measure	90.6706	94.0325

Table 7: Spearman’s rank correlation – Comparative values for different methods

Existing Method	Short-term memorability	Long-term memorability
DCU-Ensembles	0.553	0.272
EFDF	0.518	0.261
TVM	0.522	0.277
LM-VSF	0.508	0.278
DFAN	0.496	0.249
ARN	0.494	0.265
VT-CRF	0.472	0.216
HF	0.470	0.266
Late fusion	0.5577	0.3443
Proposed	0.6428	0.4285

After the training process is completed, the testing set of 1000 videos is validated using the spearman’s rank coefficient, recall and F-measure metrics. The table 6 & 7 presents the samples spearman’s rank coefficient value for short-term and long –term memorability. It is clearly understood from the results that the proposed yields better memorability score than the existing methods. Each scene in episodic video sequence creates the impact on human cognitive ability. The main features like annotations, dominant colors and the relevant scenes objects creates a long-lasting impact on the human minds according to the time frame. The best combination of feature set is achieved by the PCA method. The sequence of the frames is splitted into shots using the proposed framework to detect the cuts between two consecutive frames. The ‘high-level’ properties describing the relevant text, visual and scene data are considered in the episodic memory experiment to compute the short-term and long-term memorability score. The success of the proposed framework is quite remarkable, considering

that, the proposed technique uses an only a single frame with the rich set of information that formulates the entire narrative of the movie.

5. Conclusion

This paper develops a new video memorability prediction model from multi-modal features using deep learning techniques. The fittest combinational features sets are fused to estimate the memorability score. Therefore, the feature extraction and reduction process are given more importance in this proposed study. Initially, the text features are extracted using a fuzzy-based FastText model that portrays the annotations of a frame with the classified texts. The dominant colours of a frame are performed using the colour histogram method. The novel Faster R-CNN improves the region proposal network with an enhanced anchor generation module that minimizes information loss. The 'high-level' properties are collected using Principal Component Analysis (PCA) to form episodic shots. These are fused to estimate the memorability score. The proposed framework is implemented in Mediaeval 2018 datasets. The achieved results convey the importance of the best combinational features selection strategy, which is not exploited by existing methods.

References

1. A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2390–2398.
2. M.-G. Constantin, B. Ionescu, C.-H. Demarty, N. Duong, X. Alameda-Pineda, and M. Sjöberg, "The predicting media memorability task at mediaeval 2019," Oct. 2019.
3. P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 145–152.
4. R. Gupta and K. Motwani, "Linear models for video memorability prediction using visual and semantic features.," in MediaEval, 2018.
5. D. Azcona, E. Moreu, F. Hu, T. E. Ward, and A. F. Smeaton, "Predicting media memorability using ensemble models," MediaEval, 2019.
6. Ferguson, R., Homa, D. & Ellis, D. Memory for temporally dynamic scenes. *Q J Exp Psychol (Hove)*, 2016. Pp. 1–14.
7. Lofus, E. F. Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning and Memory* 12, 2005. pp.361–366
8. Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A. & Rosenbaum, R. S. Te cognitive neuroscience of remote episodic, semantic and spatial memory. *Current opinion in neurobiology*, 16, 2006. pp. 179–190.
9. Andermane, N. & Bowers, J. S. Detailed and gist-like visual memories are forgotten at similar rates over the course of a week. *Psychon Bull Rev*, 22, 2015. pp. 1358–1363.
10. Furman, O., Dorfman, N., Hasson, U., Davachi, L. & Dudai, Y. Tey saw a movie: Long-term memory for an extended audiovisual narrative. *Learning & memory*, 14, 2007. pp. 457–467.
11. S. Wang, W. Wang, S. Chen, and Q. Jin, "Ruc at mediaeval 2018: Visual and textual features exploration for predicting media memorability.," in MediaEval, 2018.
12. A. Goyal, N. Kumar, T. Guha, and S. S. Narayanan, "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," in 2016 IEEE

- International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2822–2826.
13. L.-V. Tran, V.-L. Huynh, and M. Tran, “Predicting media memorability using deep features with attention and recurrent network,” in *MediaEval*, 2019.
 14. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
 15. A. Reboud, I. Harrando, J. Laaksonen, D. Francis, R. Troncy, and H. L. Mantecon, “Combining textual and visual modeling for predicting media memorability,” *MediaEval*, 2019.
 16. A. Viola and S. Yoon, “A hybrid approach for video memorability prediction,” in *MediaEval*, 2019.
 17. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
 18. S. Wang, L. Yao, J.-T. Chen, and Q. Jin, “Ruc at mediaeval 2019: Video memorability prediction based on visual textual and concept related features,” 2019.
 19. M. Heikkilä, M. Pietikainen, and C. Schmid, “Description of interest regions with local binary patterns,” *Pattern recognition*, vol. 42, no. 3, pp. 425–436, 2009.
 20. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” 2005.
 21. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
 22. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, “Amnet: Memorability estimation with attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6363–6372.
 23. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
 24. J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
 25. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*, Springer, 2016, pp. 20–36.
 26. Gonzalez, R.C. and Woods, R.E. *Digital Image Processing*. 2nd Edition, Prentice Hall, Upper Saddle River. 2002.
 27. IMDB dataset downloaded from: <https://www.imdb.com/interfaces/>
 28. MediaEval 2018 dataset downloaded from: <http://www.multimediaeval.org/mediaeval2018/>
 29. Roberto Leyva and Victor Sanchez, Video memorability prediction via late fusion of deep multi-modal features. Department of Computer Science, University of Warwick, Coventry, UK. 2021.