Video Retrieval System Using Parallel Multi-Class Recurrent Neural Network Based on Video Description

Saira Jabeen, Gulraiz Khan, Humza Naveed, Zeeshan Khan and Usman Ghani Khan
{saira.jabeen, gulraiz.khan, humza.naveed, zeeshan.khan, usman.ghani}@kics.edu.pk

Abstract—**In recent times, there has been continuous interest in the area of content based information retrieval (CBIR) for images and video sequences. Exponential increase of multimedia data has triggered a cause for managing, storing and retrieving multimedia contents in convenient and efficient ways. Visual features from static images and dynamic videos are extracted to perform retrieval task. Once visual features are extracted, there is a need to search and retrieve relevant videos in efficient amount of time. This paper makes use of seven visual features; human detection, emotion, age, gender, activity, scene and object detection followed by sentence generation. Furthermore, generated sentence is used in multi-class recurrent neural network (RNN) to find genre of a video for retrieval task. Accuracy, precision and recall are used for evaluation of this framework on self generated dataset. Experiments show that our system is able to achieve high accuracy of 88.13%.**

I. Introduction

Multimedia information systems have widely been used in industrial and research based standards. It is becoming hard to obtain concerned video data and manage by individual human effort. Such information needs to be saved in an efficient way such to reduce the storage cost and maintain the maximum knowledge that a video contains. Usually a database is designed to keep the visual information of multimedia data with proper indexing for compression and searching in high power workstations and broadband networks. Videos have become mandatory for multimedia computing in everyday routine, from a television commercial (TVC) to a closed-circuit television (CCTV) photage, broadcasting, movies, education and military intelligence. For such a common and frequent usage, video data needs to be treated as simple as some textual data.

Content based image retrieval system has been of much interest for last decades [1–3]. Spatial and visual information of images is indexed in a database and on query, similar images to the query image are retrieved. For videos, the description is different, its subjective, variant, biased and ambiguous. So same information as in an image cannot be considered enough. Image based retrieval systems prohibit some features that are necessary to be considered in video retrieval systems such as temporal information. As temporally, a video may contain drastic changed visual content in the next moment as compared to previous. So the whole video should be treated as a structured sequence of frames that contains some meaningful information rather than a sequence of images.

To handle temporal information in a video our system benefits from RNN. Simple convolution network considers spatial information of frames without taking into account the temporal information.

In our paper, we present the content based video retrieval system that uses genre extracted from convolution and recurrent network. A set of hundred videos with nearly 300 to 500 frames per video are used to extract visual features. Based on these visual features a video is titled as multiple genres using multi-class RNN. For a query video visual features that are face, emotion, gender, age, activity, scene and objects are computed. Genres of a video is decided based on the high level features (HLFs). Initially, HLFs are used to generate a sentence followed by two parallel stacks of Convolutional long short-term memory(LSTM) to maintain spatial as well as temporal relation of features. Videos from database are retrieved simply by matching the genres of training videos.

In rest of the document, section II presents literature survey, section III and IV present methodology and genre prediction, respectively. Section V discusses the process of retrieving videos from database. In section VI experimental results of the video retrieval system are shared, section VII describes discussions and future work and section VIII discusses conclusion.

II. Literature Survey

Challenges in design and implementation of video retrieval systems have provoked many researchers in this area [1–3]. Many retrieval systems have been designed and implemented in image processing domain, speech recognition domain and databases discipline.

Shweta Ghodeswar and B.B. Meshram [1] in their contribution to video retrieval domain, converted a video into shots, extracted the key frames from the shot, obtained low level features of key frame and the relationship between features to represent these frames. These features were stored in database. Dynamic programming technique was used to check the similarity of input video features to the stored videos and similar videos were retrieved. Madhav Gitte et al. [2] used the same steps followed by the classification and clustering of video features by SVM and K-Means clustering respectively and similarity was found by euclidean distance formula. The work done by Simon Jones and Ling Shao is described as, video features, that is activity or motion saliency in a video, were populated in databases by visual codebook and represented as model of bag-of-words or spatio-temporal pyramid [3]. Top X videos similar to the query video were obtained based on the codebook. Gabriel, Mathias and Xavier [4] extended the existing LIRE CBIR open source tool to CBVR. LIRE's CBIR tool provides many local and global image features. Global significant features are joint color descriptor (JCD), pyramid histogram of oriented gradients (PHOG), MPEG-7 descriptors edge histogram, color layout and scalable color. While local features in LIRE include the bag of visual words approach and vector of locally aggregated descriptors (VLAD).

Scale invarient feature transform (SIFT) is computed and similarity is matched using feature vector of histograms for SIFT [5]. The second proposed methodology [5] trains these features via SVM, assigns a class for video and retrieve the videos from database based on that class. Alexander G. Hauptmann, Michael G. Christel and Rong Yan use the similar approach of bridging the gap between low level features and high level features, yielding semantic concepts. Semantic concepts such as cars, planes, roads, people, animals, and different types of scenes (outdoor, night time, etc.) are automatically detected from a video using particular algorithms [6]. Methodology introduced by Ja-HwungSu et al. generates temporal patterns of videos and apply the indexing technique via FPI-tree and AFPI-tree. Sequence is matched with the query clip to result similar videos [7].

For a video and textual sentence temporal features are necessary, Chiu JP et al. [8] used named entity recognition (NER) that automatically detects the words and character level features using the hybrid bidirectional LSTM and CNN architecture. They used the CoNLL-2003 and OntoNotes 5.0 dataset to achieve the F1 measure of 91.62% on CoNLL-2003 and 86.28% on OntoNotes 5.0. Santos CN and Guimaraes V employed CNNs to extract character-level features for using in NER and POS-tagging [9]. The model is language independent and learn the word features automatically. HAREM I corpus, and the SPA CoNLL-2002 corpus [10] had been used for training. Proposed system uses HLFs computed through computer vision approach to generate sentence followed by RNN network for genre prediction. Our system focused on both spatial as well as temporal features of a video to predict genre. After genres prediction videos are retrieved simply by similarity index.

III. Methodology

Proposed video retrieval system follows steps as shown in Fig. 1. For offline database managing, convert the training videos into frames and then extract HLFs (face, emotion, age, activity, gender, scene, objects) for each frame. After HLFs extraction, decide genres (funny, sad, kids, mature, female group, male group, male and female group, indoor , outdoor, sports, meeting, traffic, human video) for a video using RNN on sentence generated from HLFs. Insert video's meta-data in the database. For retrieving videos from databases, initially convert query video into frames and then detect genres just like training process. Finally, compare the genres with all videos' genres in database and retrieve all matched videos.

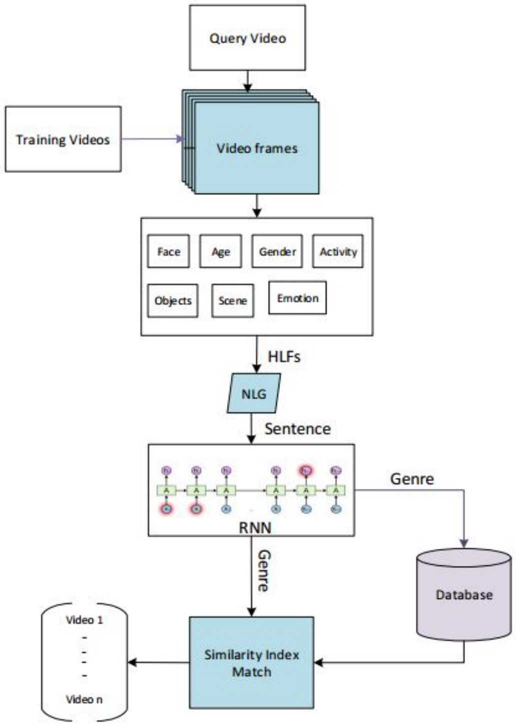


Fig. 1: Video Retrieval System Architecture

A. High Level Features (HLFs) Extraction Methods

Video retrieval has been accomplished by considering different contents of a video. A video has multiple visual features and all those features together deliver a description. In our system, we extracted seven visual features of a video frame i.e. human face existence, emotion, age, activity and gender of the detected person, scene of the particular frame and the objects present in the frame. Our pre-built system returns all HLFs for further processing. We have used Viola and Jones et al. method for fast human face detection [11]. Emotions are detected based on FACS (Facial Action Coding System), world widely defined standards for visual emotion detection [13]. Sixteen distances between twenty-four facial points are employed to detect emotions [12]. Facial emotions to be recognized are anger, smile, sad, surprise, neutral, fear and disgust. Gender detection is carried out using bag of visual words of low level SURF features [14]. Scene is detected using convolution neural network(CNN) followed by SVM classifier on CNN features and detects 16 classes [15]. Age of a person is estimated in a group. We incorporate four age groups baby, child, adult and old. Feature vector of a facial region is obtained using bio-inspired features [16, 17]. Human activity recognition is carried out by using diverse features: local binary pattern (LBP), histogram of oriented gradient (HOG), haar wavelets, SIFT, velocity and displacement. Finally, sequential minimal optimization is employed for classification of frames into 15 classes [18][19]. In our system we employed the open source code of YOLO available in C language [20] for detecting 80 objects.

B. Sentence Generation

Extraction of high level features (HLFs) is followed by sentence generation using natural language generation (NLG) techniques. High level features are passed through multiple steps of NLG. HLFs act as the input features to NLG module followed by missing words generation to makes a sentence. Initially, HLFs are divided into subject, verb and object triplet to position the words in generated sentence. After generation of sentence from subject, verb, object (SVO) triplet, intermediate words are introduced in sentence to make more realistic sentence.

1) Subject Verb Object creation: The basic structure of English sentence require proper placement of subject, verb and object. The extracted HLFs act as the input to SVO classification step. Age, gender and emotion contribute in subject creation. These three HLFs are concatenated in sequence as emotion, age and gender (Sad young female). Detected activity and detected object is considered as verb and object in the generated sentence, respectively. Multiple objects detected in the visual stream can be combined with respective activity, like sitting can be combined with chair. SVO creation can be defined by eq 1.

$$\begin{pmatrix} S \\ V \\ O \end{pmatrix} = \begin{cases} HLF & \left\{ \quad \text{if } HLF \,\epsilon\, (Emotion, Age, Gender) \right. \\ HLF & \left\{ \quad \text{if } HLF \,\epsilon\, (Activity) \right. \\ HLF & \left\{ \quad \text{if } HLF \,\epsilon\, (Objects) \right. \end{cases} \tag{1}$$

2) Missing words generation: After SVO triplet creation missing words are created to complete the sentence. We have used trigram for finding missing words in the sentence that can increase the readability of sentence. Need of missing words filling arise when we have more objects detected with less number of activities, for example we have only one activity "sitting" detected with two objects chair and remote, in this case we can not combine sitting with remote that produce a need to create missing verb. To generate suitable candidates for missing words, proposed system uses predefined Google corpus [21]. We have used words from this corpus to generate bigrams with our current generated SVOs. Probability of these bigrams (missing words concatinated with HLFs) is calculated using Google corpus. Concatenated word is considered as missing word provided the probability is above 0.5 threshold. Missing word between two words w_1 and w_2 can be defined as eq. 2.

$$MissingWord = \left\{ X \quad P(X, w_2) > 0.5 \; or \; P(w_1, X) > 0.5 \right. \tag{2}$$

Where, X is missing word between w_1 and w_2. P represents probability calculation function from given Google corpus. Multiple missing words can have probability greater than 0.5 and to select single word among all candidates missing words $argmax$ function on probability is employed. After finding required missing words system generate complete understandable sentence. The structure of complete sentence can be considered as eq. 3 with x as missing word.

$$S\,\{x_1\}\,V\,\{x_2\}\,O\; in\; \{scene\} \tag{3}$$

IV. Recurrent Neural Network for Multi-Class Genre calculation

For genre classification above described seven high level visual features are used in the form of generated sentence. In this section we introduce the architecture of sentence based genre prediction. Convolutional long short term memory (ConvLSTM) [22] is employed to build our genre prediction architecture. Initially, our system computes word embeddings using freely accessible 50-dimensional word embeddings trained on Wikipedia text and Reuters RCV-1 corpus [23]. Fifty dimensional features vectors for each word are combined in the form of two dimensional ($l \times d$) matrix with l as length of sentence and d as the size of embeddings for a single word. To make length l consistent along all sentences, we have considered l as 10 (maximum words in a sentence). Shorter sentences are made suitable to feed in our architecture by appending zeros on the last remaining rows of matrix. ConvLSTM consists of two basic constituents: two dimesional convolution and a successive LSTM unit. Convolution processes extract features from text embeddings and LSTM determine temporal relations in all words of a sentence. ConvLSTM can be defined by eq. 4 to 8.

$$i_t = \sigma[(w_{xi} * x_t) + (w_{hi} * h_{t-1}) + (w_{yi} \circ y_{t-1}) + b_i] \tag{4}$$

$$f_t = \sigma[(w_{xf} * x_t) + (w_{hf} * h_{t-1}) + (w_{yf} \circ y_{t-1}) + b_c] \tag{5}$$

$$o_t = \sigma[(w_{x\circ} * x_t) + (w_{h\circ} * h_{t-1}) + (w_{y\circ} \circ y_t) + b_\circ] \tag{6}$$

$$y_t = f_t \circ y_{t-1} + i_t \circ tanh(w_{xy} * x_t + w_{hy} * h_{t-1} + b_y) \tag{7}$$

$$h_t = o_t \circ tanh(y_t) \tag{8}$$

where, $\circ$ is Hadamard product, x_i, h_i and y_i denotes input, output and yields respectively, with three gates i_t, f_t and o_t as input, forget and output gate, respectively.

We have employed two different pipes of stacked ConvLSTM to find genre. Extracted embeddings are forward to convolution layers of each pipe followed by forward and backward LSTM network in respective pipe. Features from each Convolution processes are passed to LSTM layer that are normalised by batch-normalization processes in each pipe. Basic structure of each pipe is shown in Fig. 2. After flattening the features, these two pipes are then simply merged by concatenation of both features vectors followed by two fully-connected layers. To produce multiclass labels we have used sparse categorical cross-entropy as a function for calculating loss at the last layer of our prediction network. Training of our system for predicting multiple classes at the same time is accomplished using one hot vector for representing multiple labels for a single video with 1 having the respective class and 0 otherwise. Following are the genres into which a video lies.

- Funny Video
- Sad Video
- Kids Group Video
- Mature Group Video
- Female Group Video
- Male Group Video
- Male and Female Video
- Indoor Video
- Outdoor Video
- Sports Video
- Meeting Video
- Traffic Video
- Human Video

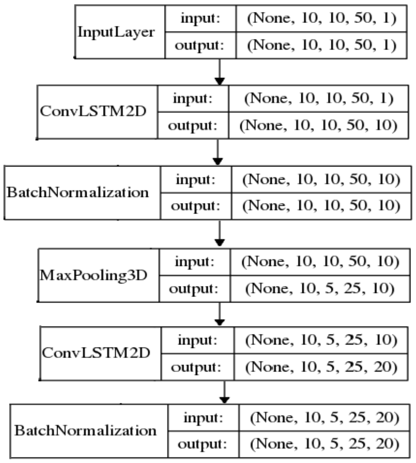

InputLayer
input: (None, 10, 10, 50, 1)
output: (None, 10, 10, 50, 1)
ConvLSTM2D
input: (None, 10, 10, 50, 1)
output: (None, 10, 10, 50, 10)
BatchNormalization
input: (None, 10, 10, 50, 10)
output: (None, 10, 10, 50, 10)
MaxPooling3D
input: (None, 10, 10, 50, 10)
output: (None, 10, 5, 25, 10)
ConvLSTM2D
input: (None, 10, 5, 25, 10)
output: (None, 10, 5, 25, 20)
BatchNormalization
input: (None, 10, 5, 25, 20)
output: (None, 10, 5, 25, 20)

Fig. 2: Convolution and LSTM based pipe

A. *Network Architecture*

This section describes the complete architecture of our genre calculation module. Input sentence is initially converted into embeddings matrix to feed into convolution process. After creating $(l \times d)$ matrix, convolution operator is applied on features set. Layers of our model are described as follow (Fig. 2).

- First layer of our network is ConvLSTM2D with input matrix of 10×50. we have used 10 filters of kernel size $[3 \times 3]$, stride $[1 \times 1]$ and padding $[1 \times 1]$ to produce output of $[10 \times 50 \times 10]$. Batch normalization is applied on output features to normalize output features within normal range.
- To reduce number of features, second layer apply 3D Maxpooling with pool size of $[1 \times 2 \times 2]$ that produce output of $[5 \times 25 \times 10]$.
- Third layer again apply ConvLSTM2D with 20 filters followed by batch normalization to produce output of shape $[5 \times 25 \times 20]$.
- Finally system apply flatten operation to further merge two pipes. Both pipes have single difference of forward and backward LSTM used in ConvLSTM2D layers.
- After merging forward and backward pipes system simply apply two fully connected layer with 1024 and 13 units in each fully connected layer.
- FC(13) with Softmax activation function produces a 13 dimensional vector containing probabilities for each class. A hot encoded output is generated by applying threshold of probability grater than 0.5.

Architecture of one pipe is shown in Fig. 2 and complete architecture of genre prediction using RNN is shown in Fig. 3.

V. Retrieve Videos from Databases

A video can have multiple genres. Extracted visual features are fed into NLG module then network predicts genres and decided genres are stored in a database structure to manage visual information. Similar videos are obtained on query, following an algorithm (1). A set of videos are processed one by one, offline. For each frame of one video visual features described in previous section are obtained and genres are determined as described in section IV. A database system is designed and populated with the metadata and genres a video contains.

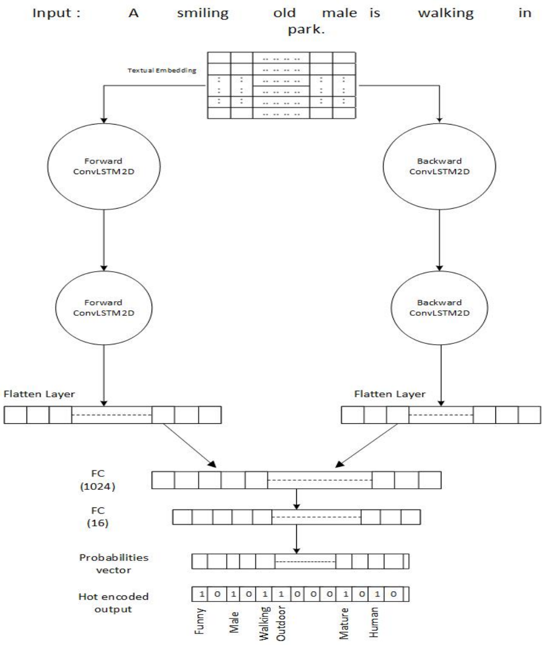


Fig. 3: Merged Multi-class RNN based framework for genre detection

Algorithm 1: Algorithm for Retrieving Videos

```
function RetrieveVideos $(a, b)$;
Input : Genres of Query Video $a[\,]$ and DB Videos
        with Genres $b[\,][\,]$
Result: Similar Videos Path $videos$
$j = 1$ ;
while $j < Length(b)$ do
    foreach $genre \in a$ do
        if $b[j].Contains(genre)$ then
            $videos$.Add($b[j]$);
        end
        $j = j + 1$ ;
    end
end
$videos$.Distinct();
return $videos$;
```

A. *Retrieving videos*

When a user chooses a query video, it is converted into frames and high level features are computed. Based on those visual features, genres are extracted. For each genre of a query video, videos are retrieved by matching it with all the genres of videos in databases. Overlapping is restricted by selecting distinct videos. Fig. 4. shows four frames from a funny retrieved video.

TABLE I: Categories of generated dataset

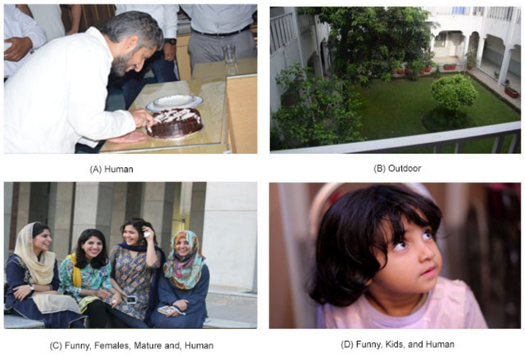
Videos	Labels	Videos	Labels
Video 1-5	Funny, Kids, Male & Female, Indoor, Human	Video 51-55	Mature, Male, Outdoor, Traffic, Human
Video 6-10	Funny, Mature, Male & Female, Indoor, Human	Video 56-60	Funny, Mature, Male & Female, Indoor, Meeting Room, Human
Video 11-15	Funny, Mature, Male & Female, Indoor, Human	Video 61-65	Funny, Kids, Female, Outdoor, Human
Video 16-20	Funny, Mature, Male & Female, Outdoor, Human	Video 66-70	Funny, Kids, Male, Indoor, Human
Video 21-25	Funny, Mature, Female, Indoor, Human	Video 71-75	Funny, Kids, Male, Outdoor, Sports, Human
Video 26-30	Sad, Kids, Male & Female, Outdoor, Sports, Human	Video 76-80	Kids, Male, Outdoor, Human
Video 31-35	Funny, Mature , Male ,Outdoor, Traffic, Human	Video 81-85	Outdoor, Traffic
Video 36-40	Sad, Mature, Male, Indoor, Meeting Room, Human	Video 86-90	Indoor, Meeting Room
Video 41-45	Sad, Mature, Male, Indoor, Meeting Room, Human	Video 91-95	Outdoor
Video 46-50	Sad, Mature, Male & Female, Indoor, Meeting Room, Human	Video 96-100	Outdoor

Fig. 4: Frames from a funny retrieved video

VI. Experimental Results

A. Dataset

Dataset used for training and testing purpose is self generated having hundred videos of multiple genres. Number of subjects in each video vary from 1 to 5 and frames per video ranges from 300 to 500. Dataset is generated in CVML lab[1] considering the categories described above and defined in Table I. From the generated datasets few frames of videos are shown in Fig. 5.



(A) Human

(B) Outdoor

(C) Funny, Females, Mature and, Human

(D) Funny, Kids, and Human

Fig. 5: Few frames from four videos

B. Evaluation

We have used Keras [24] library for Tensorflow to train our network on 200 epochs with 0.0001 learning rate. Model is trained using Geforce Nvidia 1080 Ti GPU having 11GB of memory size.

Performance of this information retrieval system is measured with binary classification i.e. either the results are relevant or not relevant. Accuracy, precision and recall matrix are used for evaluation of our system. We have achieved 88.13% accuracy for retrieving relevant videos as depicted by Table II. Video clips are evaluated against all categories. Precision and recall results are presented in the graphic charts (Fig. 6). A confusion matrix in Table II explains the results of retrieval system. Proposed system nearly takes half time of the video duration to processes and retrieve relevant videos. Comparison with other available methodologies has been provided in Table III.

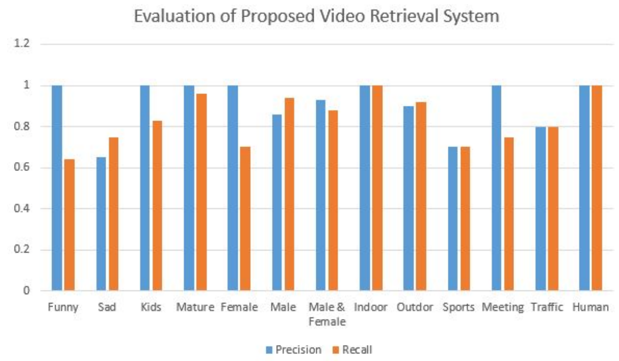


Fig. 6: Precision and recall measurements for self-generated dataset of 100 videos

VII. Discussion and Future Work

Features relevant to all the retrieval methods have already been extracted for the a set of videos while other computations and genre is decided at the time query is made. This retrieval can further be extended by tracking some specific label. Such as, based on face recognition one can retrieve all the videos a particular person appears in. Similarly, tracking some particular object like a car can be implemented.

VIII. Conclusion

Video dataset is increasing day by day with the increasing use of video recording devices such as mobile phones and handy cameras. Here we need a system which is able to retrieve the required content only by a query video. We present a system that can retrieve any type of videos using the maximum visual features a video can have. In this document, system that is presented extracts multiple visual features from every frame of a video followed by extraction of multiple categories for that video. Upon user request it decides the genres of query video and returns the similar genres videos

[1] http://www.kics.edu.pk/labs/about/cvml

TABLE II: Confusion matrix for evaluation measure of Video Retrieval System

	Funny	Sad	Kids	Mature	Females	Males	Male & Female	Indoor	Outdoor	Sports	Meeting	Traffic	Human
Retrieved Funny	32												
Retrieved Sad	8	15											
Retrieved Kids			25										
Retrieved Mature			3	48									
Retrieved Female					7								
Retrieved Male					3	33	2						
Retrieved Male & Female						2	31						
Retrieved Indoor								50					
Retrieved Outdoor									46		5		
Retrieved Sports										7		3	
Retrieved Meeting											15		
Retrieved Traffic										3		12	
Retrieved Human													80
Not Retrieved	10	5	2	2			2		4				

TABLE III: Comparison of our methodology with available methodologies

Paper	**Dataset**	**Precision**
B. V. Patel et al.[1]	-	0.629
Gitte et al. [2]	-	0.55
de Oliveira et al. [4]	Stanford I2V Dataset	0.450
Bajestani et al. [5]	KTH	0.257
Jos Timanta et al.[25]	-	0.42
Ours	**Self-generated**	**0.89**

as of the query video. Face, emotions, gender, scene, age, activity and objects are the visual features that are extracted from video frames. For training videos, all the genres are obtained using RNN algorithms described in the section IV and stored in database corresponding to each video. Videos with the closest genres are the retrieved videos from database. This methodology not only applicable for some particular type of videos, in fact it covers many aspects a video may contain human and the other similar videos that we need to obtain must have human being in it. Similarly, videos without human that may have scenes or objects can also be tracked.

IX. Acknowledgement

We would like to express our earnest gratitude to National ICT R&D Fund for full supporting our research work, full team (colleagues & management) and organization (KICS) for their support.

References

[1] Patel, B. V., & Meshram, B. B. (2012). Content based video retrieval. arXiv preprint arXiv:1211.4683.

[2] Gitte, M., Bawaskar, H., Sethi, S., Shinde, A.: Content based video retrieval system. Int. J. Res. Eng. Technol. 3(6), 1 (2014)

[3] S. Jones, L. Shao, Content-based retrieval of human actions from realistic video databases, Inform. Sci. 236 (2013) 56–65

[4] de Oliveira-Barra G, Lux M, Giró-i-Nieto X. Large Scale Content-Based Video Retrieval with LIvRE. In 14th International Workshop on Content-based Multimedia Indexing (CBMI). Bucharest, Romania: IEEE; 2016.

[5] Bajestani, Faride Jamali, Am F. Aminian, and Am M. Aminian. "Human actions retrieval from video databases according to the temporal feature by using multiple SVM and SIFT descriptor." Technology, Communication and Knowledge (ICTCK), 2015 International Congress on. IEEE, 2015.

[6] Hauptmann, Alexander G., Michael G. Christel, and Rong Yan. "Video retrieval based on semantic concepts." Proceedings of the IEEE 96.4 (2008): 602-622.

[7] Su, Ja-Hwung, et al. "Effective content-based video retrieval using pattern-indexing and matching techniques." Expert Systems with Applications 37.7 (2010): 5068-5085.

[8] Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." arXiv preprint arXiv:1511.08308 (2015).

[9] Santos, Cicero Nogueira dos, and Victor Guimaraes. "Boosting named entity recognition with neural character embeddings." arXiv preprint arXiv:1505.05008 (2015).

[10] Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2002, pages 155–158. Taipei, Taiwan.

[11] Viola, P. and Jones, M. "Rapid object detection using boosted cascade of simple features." IEEE Conference on Computer Vision and Pattern Recognition, 2001

[12] Mahapatra, D., Routray, A., & Mishra, C. (2006, December). An active snake model for classification of extreme emotions. In Industrial Technology, 2006. ICIT 2006. IEEE International Conference on (pp. 2195-2199). IEEE.

[13] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 19, pp. 757-763, 1997.

[14] Demirkus, Meltem, et al. "Gender classification from unconstrained video sequences." Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010.

[15] Zhou, Bolei, et al. "Learning deep features for scene recognition using places database." Advances in neural information processing systems. 2014.

[16] Levi, Gil, and Tal Hassner. "Age and gender classification using convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015.

[17] Guo, Guodong, et al. "Human age estimation using bio-inspired features." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[18] Wang, Heng, et al. "Dense trajectories and motion boundary descriptors for action recognition." International journal of computer vision 103.1 (2013): 60-79.

[19] Naveed, H., Khan, G., Khan, A. U., Siddiqi, A., & Khan, M. U. G. (2018). Human activity recognition using mixture of heterogeneous features and sequential minimal optimization. International Journal of Machine Learning and Cybernetics, 1-12.

[20] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR (2016)

[21] Evert, Stefan. "Google web 1t 5-grams made easy (but not for the computer)." Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop. Association for Computational Linguistics, 2010.

[22] Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems. 2015.

[23] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." Journal of Machine Learning Research 12.Aug (2011): 2493-2537.

[24] F. Chollet. Keras. https://github.com/fchollet/keras, Accessed on June, 2018.

[25] Tarigan, J. T., & Marpaung, E. P. (2018). Implementing Content Based Video Retrieval Using Speeded-Up Robust Features. International Journal of Simulation–Systems, Science & Technology, 19(3).