ISSN:0975-9646

A Review on Large-scale Video Classification with Recurrent Neural Network (RNN)

Bhagyashri P. Lokhande, Sanjay S. Gharde

Dept. of Computer Engineering
SSBT.s College of Engineering and Technology
Jalgaon, Maharashtra, India

***Abstract*— People have emphasis on retrieve videos on internet with specific category and it is infeasible to find video of interest. It becomes difficult to classify video with users demand due to limited research in video classification area. Further it affects the interest level of the users. It may become down, minimize or diverted. There is need to have easier method for users to access video of interest. To classify the videos, research has begun on large scale video classification. In order to review all the techniques related to video classification are compared to show the most suitable technique for video classification.**

***Keywords*— Video Classification, Large Scale Dataset, Convolutional Neural Network, Recurrent Neural Network, LSTM technique, Extraction of Features**

I. INTRODUCTION

Now days, people have access to a tremendous amount of video on internet. The amount of videos that viewer has to choose present globally is so large. So, it is infeasible for viewer to go through tremendous amount of videos and find video of interest. Images as well as videos have become global on the internet and it encourages people for the development of algorithm which include semantic content for various applications. It also includes search and optimization of videos for better classification result. One method that viewers use to deficient their choices is to look video for video within specific categories. Because of the huge amount of the videos to categories, research has begun on Large-Scale Video Classification [1].

Recently, Recurrent Neural Network (RNN) [2] has been express as an effective class of models for understanding image content. It gives state of the art which results on image recognition, segmentation, detection and retrieval. The literature review of the various techniques used for video classification are discussed here and compare all the techniques with each other on the basis of factors like video classification algorithm used in each approach, model for classification, features extracted for classification in each techniques. Lastly, it is analyzed that whether each technique is suitable for handling large scale dataset or not and whether all techniques are capable for deals with time as well as space complexity. Encourage by positive result in image and speech recognition, we study the performance of RNNs in large scale video classification and proposed a model to solve the video classification with visual feature extraction like color and texture.

A. Video Classification

Video classification differs from video indexing and retrieval, since in video classification, all videos are sorted by their categories, and each video is assigned a meaningful label. While in video indexing and retrieval, the aim is to accurately retrieve videos that match a users query. Many automatic video classification algorithms have been proposed, most of them can be categorized into four groups: text-based approaches, audio-based approaches, visual-based approaches, and combination of text, audio and visual features. Many standard classifiers, such as Gaussian Mixture Models (GMM), Bayesian, Support Vector Machines (SVM), Neural Networks and Hidden Markov Models (HMMs) have been applied in video classification and recognition [2].

Visual based approaches in video classification are categories in major five types such as Color Based, Shot Based, Object Based, MPEG and Motion Based approach. Color Based approach for visual feature extraction had some benefit such as it is simple to implement and process and it has crude representation. MPEG based visual feature is somewhat easy to extract from video clips but the required video must be in MPEG format then and then only these feature extraction technique works. In Shot Based visual feature, it is difficult to identify shot automatically and it may be result in non accurate prediction. Object Based visual feature if difficult to implement and it is limited on number of objects used to recognize video. It is also very expensive to build computationally. In Motion Based visual feature, it is difficult to distinguish between types of motion, and also the computational requirements for motion feature range varied from low to high. So, the best approach for video classification is color based visual feature extraction techniques. In color based visual feature, color space is calculated using RGB values. The RGB values are calculated using

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B} \qquad (1)$$

Then the distance is calculated for color space as,

$$KL(p \parallel q) = -\sum_{i=1}^{N} p(x_i) \log \frac{q(x_i)}{p(x_i)} \qquad (2)$$

Where,

N = number of bins in histogram,
$p(x_i)$ = probability of color x_i for one frame,
$q(x_i)$ = probability of color x_i for other frame

The texture visual feature is determined and video should be recognized and categorized as per viewers demand.

B. Convolutional Neural Network

A convolutional neural network [7] is a type of feed-forward artificial neural network where the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field.

When used for image recognition, convolutional neural networks consist of multiple layers of small neuron collections which look at small portions of the input image, known as receptive fields. The results of these collections are then tiled so that they overlap to obtain a better representation of the original image; this is repeated for every such layer [7].

C. Recurrent Neural Network

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. This cycle creates an internal state of the network which allows it to exhibit dynamic temporal behavior. RNNs can used their internal memory for processing arbitrary sequences of inputs.

This makes them applicable to tasks such as unsegmented connected handwriting recognition, which achieved the best known results [1].

II. RELATED WORK

The standard approach to video classification involves three major stages: First, local visual features that describe a region of the video are extracted either densely or at a sparse set of interest points. Next, the features get combined into a fixed-sized level of video description. One of the popular approach is to quantize all features using a learned k-means dictionary and accumulate the visual words over the duration of the video into histograms of varying spatio- temporal positions and extents. Lastly, a classifier (such as SVM) is trained on the resulting bag of words representation to distinguish among the visual classes of interest [2].

Numerous techniques exist at the moment for video classification. Many of such techniques are quantitative, and they offer a methodical approach for classifying of data and features extracted to different factors related to requirements for computing a priority.

Further techniques depend on carrying out for recognizing actions and some on video indexing. Features extracted with various methods and models are different and more complicated to evaluate result. Text based and audio based features are not given accurate result but visual based features like color, texture help to recognize video in a short time and computational complexity become increased with these features. Yet, this may forfeit a little consistency [1].

Rather some video classification techniques are recognized as Generalized Maximum Clique Problem (GMCP), Hidden Markov Model (HMM), Convolutional Neural Network (CNN) to recognize actions, Hidden Markov Model for video Indexing, and Evaluation of CNNs on large-scale video classification with spatio-temporal method etc [2].

A. Hidden Markov Model for video Indexing

The model was proposed by Stefan Eickeler et al. in 1999 [3]. It describes a new approach to content-based video indexing using Hidden Markov Models (HMMs). One feature vector is calculated for each image of the video sequence. These feature vectors classified by HMMs. The main advantage of these approach is the system has automatic learning capabilities. The motion based features liken average absolute deviation of the motion, centre of motion delta features are extracted. The presented approach works three times faster than real-time. But still this indexing approach is not working for more complicated tasks with more content classes. It is not applied for sport video dataset.

B. Hidden Markov Model (HMM)

Josh Hanna et al., in 2012, proposed a Hidden Markov Model (HMM) based classification technique for sports videos. As shown in above HMM model focuses on indexing of videos depends on contents. Here, they observe the speed of color changes which is computed for each video frame and used as observation sequences in HMM for classification. Experiments on 3 predefined genres (golf, hockey and football) give very satisfactory classification accuracy. Each video is considered to be a sequence of images with each image being represented with a color feature. Color data from each pixel in RGB color space is gathered and averaged for each frame. The speed of color change is calculated by red, green and blue saturations. It is done by subtracting each color saturation from the saturation of the previous frame. But still the problem arises when it comes to large scale dataset for classification purpose [4].

C. Convolutional Neural Network (CNN) to recognize actions

The model was proposed by Shuiwang Ji et al. in 2013. 3D CNN model for action recognition with novel approach is described here. The features like spatial i.e. related to space and temporal i.e. related to time extracted here and performing 3D convolutions. So, capture the motion information encoded in multiple adjacent frames.

A simple approach in this direction is to treat video frames as still images and apply CNNs to recognize actions at the individual frame level. It compare the 3D CNN model with two other baseline methods, which follow the state-of-theart bag-of-words (BoW) paradigm in which complex handcrafted features are computed. The 3D CNN model requires a large number of labeled sample strictly to categorized video [5].

D. Generalized Maximum Clique Problem (GMCP)

Shayan Assari et al. in 2014, propose a contextual approach to video classification based on Generalized Maximum Clique Problem (GMCP) which uses the co-occurrence of concepts as the context model. It represents a class based on the co-occurrence of its concepts and classifies a video based on matching its semantic co-occurrence pattern to each class representation. They

perform the matching using GMCP which finds the strongest clique of co-occurring concepts in a video.

The extracted Motion Boundary Histogram (MBH) features from the annotated clips and computed a histogram of visual words for each. Additionally, they propose a novel optimal solution to GMCP based on Mixed Binary Integer Programming (MBIP). But this method is unsuitable for solving binary detection problem. The GMCP problem is solved here by using Mixed Binary Integer Programming (MBIP) [6].

E. CNN evaluation on large-scale video classification

Andrej Karpathy et al., in 2014, proposed a Convolutional Neural Networks (CNNs) as a powerful class of models for text recognition and image recognition problems. It represent multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information. The multiresolution foveated architecture suggested as a promising way of speeding up the training.

The best spatio-temporal networks display significant performance improvements compared to strong feature-based baselines, but only a surprisingly modest improvement compared to single-frame models. Color based and objects based features are extracted from the video to recognize the category of video. Learned features for the first convolutional layer can be inspected. Interestingly, the context stream learns more color features while the high- resolution fovea stream learns high frequency grayscale filters [7].

Nowadays, it has become a growing concern that video classification are of diverse importance. Besides this, theoretically or practically, there has been a modest progress on the mechanisms for classifying videos with different techniques till date.

As shown in Table 1, it clearly mention that comparison of different approach shows that there is very less work done in video classification area with large scale dataset. The features extracted in each approach are differently distinguished but all the approaches are not worked for large scale dataset except CNN model. The capability of all approaches in case of time and space complexity are distinguishes here and gives as a result that each technique is capable for either time or space complexity.
So, here the comparison of all the approaches with each other is shown and proposed for the Recurrent Neural Network approach to get better performance and better result than previous techniques. The 3D CNN and CNN model also performs better in result as compared to other approaches.

III. PROPOSED WORK

Nowadays, it has become a growing concern that video classification are of diverse importance. Besides this, theoretically or practically, there has been a modest progress on the mechanisms for classifying videos with different techniques till date.

Review of the state of practice in video classification indicated that many organizations consider it crucial to classify videos on large scale; and to fix their decisions according to rational or quantitative data. Yet, hardly any organization actually knew about classifying videos with LSTM approach efficiently.

So, in order to make the classification of videos more efficient and its use in many different organizations much simplified, we have attempted to present a new approach towards Video Classification. It can deal with complexity, ambiguity, uncertainty and easily target the situations where complex service behavior can be deviated from user's expectations. This technique is a combination of Recurrent

Neural Network and LSTM technique; and can be recognized as 'Long Short Term memory Recurrent Neural Network'. It may lessen the requisite efforts and may enable to generate high-quality outcomes which are considered reliable by its users. This new technique will certainly prove helpful for many different organizations in their process of video classification.

Author	**Classification Parameters**				
	Approach	***Model for Classification***	***Features Extracted***	***Suitable for different size Large Scale Dataset***	***Capability to deal with time and space factor***
Stefan Eickeler and Stefan Muller 1999	Video Indexing for Automatic Video Classification	HMM Model	Motion Based Features	No	Only Time
Josh Hanna et al. 2012	Video Content Based Classification	HMM Model	Color Based Features	No	Only Time
Shuiwang Ji et al. 2013	Video Frame Based Classification for Multiple Contagious Frames	3D CNN Model	Motion Based Feature	Yes (Good in Performance)	Both Time and Space
Shayan Assari et al. 2014	Contextual Approach for Event Classification	GMCP Model and SVM Model	Motion Based Feature	No	No One
Andrej Karpathy et al. 2014	Context Stream and Fovea Stream for Speeding Up Runtime Performance	CNN Model, Multilayer Neural Network	Color Based and Object Based Feature	Yes (Better in Performance Than Other)	Both Spatial and Temporal with Best Result

Table 1. Comparison of Video Classification Techniques

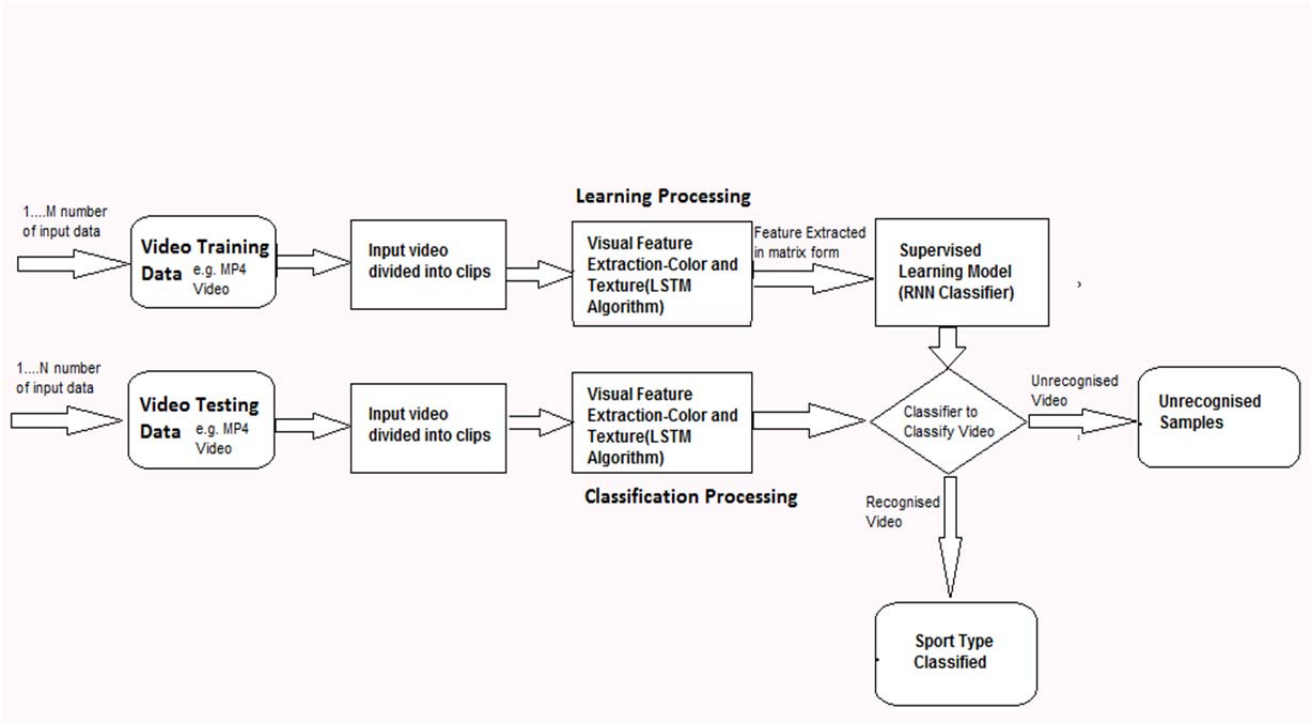

Fig 1. Video Classification Process with RNN

The Fig.1 gives rough idea about how the proposed approach should work. The visual features used are the spatial correlation between pair of color and texture. The 1....M number of sample videos are given input to the Training Data for video Classification. Then, each video is divided into small frames for save the time to classify and recognize the video. The LSTM algorithm is applied to given frames to extract visual features like color and texture from video clip. The features are extracted in matrix form as an output gives to RNN classifier to classify video and recognize type of video. It evaluates the effectiveness towards the classification result.

The Supervised Learning Model i.e. RNN Classifier we used to classify video takes lesser time to computer and classify video as per users requirement. Traditional RNNs can learn complex temporal dynamics by mapping input sequences to a sequence of hidden states, and hidden states to outputs via the following recurrence equations

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$z_t = g(W_{hz}h_t + b_z) \quad (2)$$

Where, g is an element-wise non-linearity, such as a sigmoid or hyperbolic tangent, x_t is the input, $h_t \in R^N$ RN is the hidden state with N hidden units, and y_t is the output at time t. For a length T input sequence (x_1; x_2; :::; x_T)i, the updates above are computed sequentially as h_1 (letting $h_0 = 0$), y_1, h_1, y_2, ..., h_T , y_T .

A. Feature Extraction

The feature extracted is Group of Picture i.e. (GOP) based. It is a independent unit in the MPEG video format. A set of feature vectors for each GOP contain color and motion information. Color Feature extracted from video is based on DC image. The DC image formed by the coefficients of all 8 × 8 blocks over the whole frame. So, the size of DC image is 8 times smaller than the original frame. The DC image converted into RGB color space. Motion feature information is captured in motion vector.

LSTMs provide a solution by incorporating memory units that allow the network to learn when to forget previous hidden states and when to update hidden states given new information. The advantages of LSTMs for modeling sequential data in vision problems are twofold. First, when integrated with current vision systems, LSTM model is straightforward for fine-tune end-to-end. Second, LSTMs are not confined to fixed length inputs or outputs allowing simple modeling for sequential data of varying lengths, like text or video.

IV. CONCLUSION

Review on video classification indicates that while the performance is not particularly sensitive to the architectural details of the connectivity at that time, all the models consistently perform slower than Convolutional Neural Network. The comparison of video classification techniques with different feature extraction shows some limitations like time complexity, computational complexity. It may increase with Recurrent Neural Network to get better result in short period of time. RNN may perform better than all other approach for large scale video classification. RNN classifier with LSTM algorithm used for feature extraction may increase performance of our model. The simplest feature extraction proposed in model results better in performance and percentage of accurate video classification may increase with RNN. Proposed RNN model with LSTM approach increase performance of video classification with better time and space complexity than others models specified in table.

ACKNOWLEDGMENT

Authors duly acknowledge the timely help and support received in the department of Computer Science, SSBT's College of Engineering and Technology, Jalgaon, Maharashtra, India. The authors are especially grateful to Prof. Dr. Girish Kumar Patnaik, Head of the Department of Computer Engineering for his valuable suggestions. Authors also like to thank Prof. Dr. Sanjay P. Shekhawat, Dean of Academics and Prof. Dr. Kishor S. Wani, Principal for supporting the research work.

AUTHOR'S CONTRIBUTION

Mrs. Bhagyashri Lokhande is a post graduate student under the guidance of Mr. Sanjay S. Gharde in the Department of Computer Engineering. Miss. Bhagyashri Lokhande has contributed to conceptualizing the research work Mr. Sanjay S. Gharde has reviewed the research work and assisted in writing the manuscript of the article.

REFERENCE

[1] J. Schmidhuber, "Deep learning in neural networks: An overview,\ The Swiss AI Lab IDSIA, Tech. Rep. IDSIA-03-14 / arXiv:1404.7828v1 [cs.NE], 2014.

[2] D. Brezeale, D. J. Cook, and S. Member, \Automatic video Classification: A survey of the literature," IEEE Transactions on Systems, Man, and Cybernetics, Part C.

[3] S. Eickeler and S. Mller, \Content-based video indexing of tv broadcast news using hidden markov models," 1999, pp. 2997-3000.

[4] J. Hanna, F. Patlar, A. Akbulut, E. Mendi, and C. Bayrak, \Hmm based classification of sports videos using color feature." in IEEE Conf. of Intelligent Systems. IEEE, 2012, pp. 388{390. [Online]. Available: http://dblp.uni-trier.de/db/conf/is/is2012.html#HannaPAMB12

[5] S. Ji, W. Xu, M. Yang, and K. Yu, \3d convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell.,vol. 35, no. 1, pp. 221{231, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.59

[6] S. Modiri Assari, A. Roshan Zamir, and M. Shah, \Video classification using semantic concept co-occurrences," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, \Large-scale video classification with convolutional neural networks," in Proceedings of International Computer Vision and Pattern Recognition (CVPR 2014), 2014.