ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Video Classification with Recurrent Neural Network

Bhagyashri P. Lokhande, Sanjay S. Gharde

M.E Student, Dept. of CSE, SSBT's College of Engineering and Technology, Jalgaon, Maharashtra, India

Assistant Professor, Dept. of CSE, SSBT's College of Engineering and Technology, Jalgaon, Maharashtra, India

ABSTRACT: People have emphasis on retrieval of 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. There is need to have easier method for users to access video of interest. Due to the problems such as misdetection, increased computational loads on system and poor video quality number of methods becomes unsuccessful, computationally expensive and hard to implement. The proposed system gives the solution to the current problem using Recurrent Neural Network. Recently all video classification benchmarks performed for clip level prediction but the proposed system worked for clip level prediction to global video level prediction using Recurrent Neural Network. Different patterns are generated for each class for classification. Hue, Saturation, Value color model is used to extract color features from each frame. Recurrent Multilayer Perceptron Neural Network is used to classify videos with the color features model and pattern generated for classes. Implementation results show that the proposed system increases the performance of the proposed system by increasing peak signal-to-noise and reduces the Mean Absolute Error, Mean Percentage Error and Relative Standard Error and hence perform better.

KEYWORDS: Video Classification, Recurrent Neural Network, Recurrent Multilayer Perceptron, Heu-Saturation-Value (HSV) Color Model, Peak Signal to Noise Ratio (PSNR), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Relative Standard Error (RSE)

I. INTRODUCTION

Today people have access to a huge amount of videos on internet. To choose the video with user interest from a large dataset is infeasible as a very few benchmarks are proposed till today to classify video. One solution to the current problem is to categories videos of user's interest. To categories the video, the research has begun on video classification. A series of images contains multimedia sequences is used to refer here as a video. Video classification differs from video indexing and retrieval. In video classification, all videos are put into categories and each video is assigned a meaningful label. In other hand, video indexing and retrieval, the main aim is to accurately retrieve videos that match a users query. Video classification is a part of pattern recognition which is a sub domain of Machine Learning (ML). Pattern recognition [1] [2] [3] is one of the most significant application. Every instance in any dataset used by machine learning is represented using the same set of features. The features may be continuous, categorical or binary types. The learning is called supervised when the instances in it are given with known labels. Exactly opposite to unsupervised learning, where instances are unlabeled. Many ML applications involve tasks that can be set up as supervised. Data classification is a one of the major category of pattern recognition for several applications like [4] speech recognition, hand writing recognition, video classification and so on. In particular, the work is concerned with video classification in which the output of instances is categorized into distinct classes.

Video classification [5] is a branch of pattern recognition which belongs to statistical pattern recognition category. Video classification is differ from text classification and audio classification which are also a branches of pattern recognition. Video is nothing but collection of frames and a collection of continuous frames form a video. Process of video classification requires preprocessing on video, feature extraction from frames and classify video according to particular class they belong. Video classification refers to the problem of multimedia database retrieval. Internet search is an application of video classification. The input patterns given to system are video clips and the pattern classes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

generated are video genre such as sport video, action etc. Video classification can be done with different approaches and applying different classifiers. Many applications of video classification like Satellite video, Broadcast video had problems like tracking, human face recognition, activity recognition and activity based person identification, scene understanding on internet and Military area also. Research and development efforts in video classification based on content and features extraction techniques to efficiently classify videos is started now to get better result for real time applications in pattern recognition domain. It is observed that many techniques are implemented for classifying videos. These techniques are used under various scenarios such as text based image retrieval, content based image retrieval, scene labeling, Nighttime Surveillance, and using Multiple Time-Spatial Images. But in all cases some limitations are seen. A new technique can be implemented to solve the limitations of existing system. This new technique can be used for video genre detection and classification.

Recurrent neural networks (RNN) [6] [7] are a widely used tool for the prediction of time series, context dependent pattern classification tasks such as speech recognition. Generally contribute to integrating the context of the input feature vector to be classified is the feedback connection in RNN networks. It described that the contribution of the feedback connections in RNN is primarily a smoothing mechanism. It is achieved by moving the class boundary of an equivalent feed forward network classifier. Recurrent neural network architectures can have many different forms. One common type of a standard Multi-Layer Perceptron (MLP) with added loops. It can exploit the powerful non-linear mapping capabilities of the MLP, and also it have some form of memory.

The visual feature extracted from each frame is useful in sport video classification system to distinguish the sport class. The visual features are described by Hue, Saturation, and Value (HSV) color model as shown in Figure 1.1. In visual features various types of features are used to classify the multiple classes. The color features are the most widely used visual features in classification they are easier to extract compared with shape feature. A video frame is composed of a set of dots known as pixels and the color of each pixel is represented by a set of values from a color space. Many color spaces exist for representing the colors in a frame. Two of the most popular are the red-green-blue (RGB) and hue saturation-value (HSV) [8] color spaces. In the RGB color space, the color of each pixel is represented by some combination of the individual colors red, green and blue.

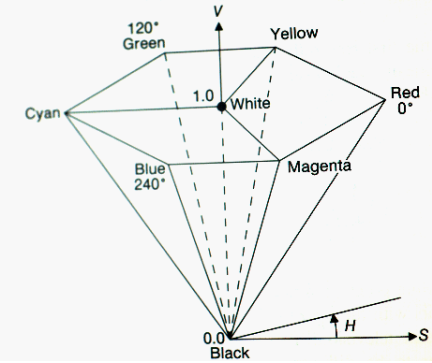

Figure 1.1 HSV Color Model

In the HSV color space as shown in Figure 1.1, colors are represented by hue (i.e., the wavelength of the color percept), saturation (i.e., the amount of white light present in the color), and value (also known as the brightness, value is the intensity of the color). Color moments are measures that can be used differentiate images and it is based on their features of color. Once calculated the color moments, it provides a measurement for color similarity between images. The basis of color moments produces in the assumption that the distribution of color in an image can be interpreted as a probability distribution. A number of unique moments are characterized by probability distribution for e.g. Normal distributions are differentiated by their mean and variance. Hence it follows that if the color in an image follows a certain probability distribution, the moments of given distribution can be used further as features to identify that image based on color.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

II. RELATED WORK

David Gibson et al., in [9], presents preprocessing on each video frame by transformed frame into the eigenspace via principal component analysis (PCA) and kernel PCA, respectively. The purpose of this transformation is to perform dimensionality reduction so that high-dimensional image space can be represented by a much lower dimension space, whilst retaining the significant variations of the original data. Daniilidis et al., in [10], presents naive strategy to categorization foundered by bottom up temporal segmentation. The difficulty of partitioning a video into actions purely based on low-level cues. Andrej Karpathy et al.,in [4], present a preprocessing by cropping the centre of region of each frame resizing them to 200 × 200 pixels, randomly sampling a 170 × 170 region, and finally randomly flipping the images horizontally with 50% probability. These preprocessing steps are applied consistently to all frames that are part of the same clip. As a last step of preprocessing we subtract a constant value of 117 from raw pixel values, which is the approximate value of the mean of all pixels in our images. Karpathy et al., in [4], proposed a Convolutional Neural Networks (CNNs) as a powerful class of models and it is used for text recognition and image recognition problems. It represents multiple approaches for increasing and expanding the connectivity of a CNN in time domain to take advantage of local spatio-temporal information. The multiresolution foveated architecture demonstrated as a favorable and hopeful way of speeding up the training. The best spatio-temporal networks display notable performance improvements as 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. Ji et al., in [11], proposed 3D CNN model for action recognition with novel 3D CNN model used for action recognition. This model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generates multiple channels of information from the input frames, and the final feature representation combines information from all channels. To further boost the performance, the regularizing the outputs with high-level features and combining the predictions of a variety of different models. Then apply the developed models to recognize human actions in the real-world environment of airport surveillance videos. Assari et al., in [12], 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 represent a class based on the co-occurrence of its concepts and classify a video based on matching its semantic co-occurrence pattern to each class representation. The matching was performed using GMCP which finds the strongest clique of co-occurring concepts in a video. Additionally, propose a novel optimal solution to GMCP based on Mixed Binary Integer Programming (MBIP). Hanna et al., in [13], proposed a Hidden Markov Model (HMM) based classification technique for sports videos. Speed of color changes is computed for each video frame and used as observation sequences in HMM for classification. Eickeler et al., in [14], proposed a new approach to content-based video indexing using Hidden Markov Models (HMMs). In this approach one feature vector is calculated for each image of the video sequence. These feature vectors are modeled and classified using HMMs. This approach has many advantages compared to other video indexing approaches. The system has automatic learning capabilities. It is trained by presenting manually indexed video sequences. To improve the system we use a video model that allows the classification of complex video sequences.

From the review it is summarized that, instead of using text based and audio based feature extraction method for video classification, visual based feature extraction method perform better. In feature extraction process the visual based feature HSV color model are better than color histogram and other method. The HSV feature includes both local and global properties, while in color histogram it take only global properties. HSV color feature robustly tolerates large change in appearance and shape. They are suitable for large databases. Next various classifiers are used for classification such as ANN, SVM and HMM. The ANN are much faster in larger datasets compared to SVM. The ANN outperforms than HMM in video classification, where HMM classifier may lead to classification error especially when there is small subset of features.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

III. PROPOSED ALGORITHM

The main purpose of video classification is to classify video with users demand by reducing misdetection in pattern with a proposed classifier and decrease computational load of a system. In proposed system, the HSV features are calculated from each frame to preprocess the frame for classification and the classification is performed with RMLP neural network for categories video belongs to each class.

A. *Architecture of Proposed System*: The basic foundation of the proposed video classification system is in the way of classifying data into classes belonging to each category. Instead of making clip level classification only, the proposed system worked for global video level classification of videos. The Figure 3.1 shows overall structure of the proposed system.

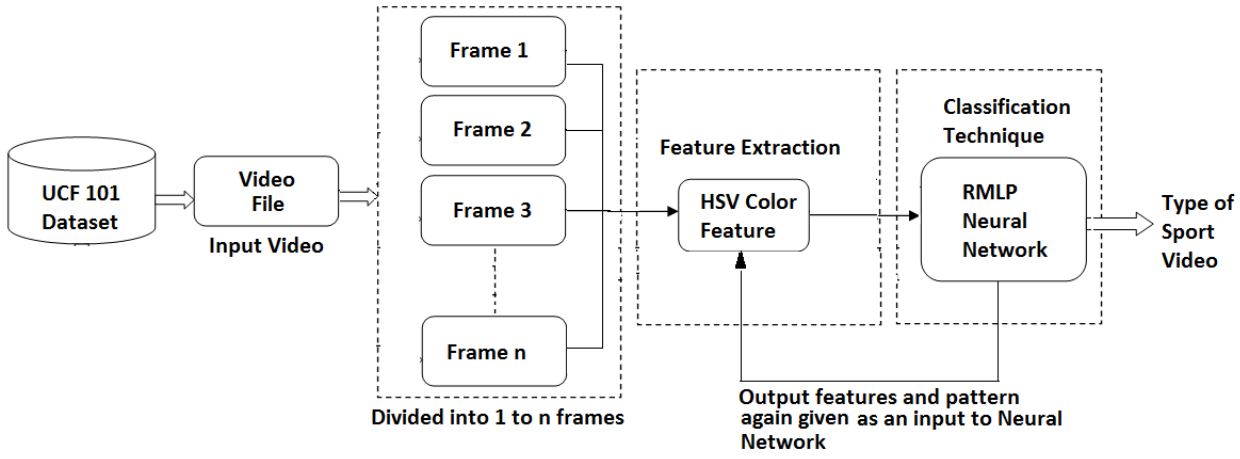

Figure 3.1: Architecture of Proposed System

The architecture for video classification shows the overall process of the video classification. First of all, the video is taken from UCF 101 dataset for classification. The input video file is then divided into number of frames for further process. Then features are extracted from each frame for classification. HSV visual color features are used to calculate features from frames. Then the extracted features along with fixed patterns generated for each class of sport is given to the RMLP neural network for classification. The RMLP neural network classify each video with its category by calculating error term and weighted sum and generate class for sport. If the patterns and features given for test video are match with train file then the sport class classify accurately. The output is again given to the input of neural network with recurrent connection to generate better and accurate result. The number of output is depending on the number of recurrence given to the network to classify video.

Algorithm 1 shows the training of a proposed system with required parameter and set values of learning rate and momentum to train neural network. It takes the video as an input to the system and generate frames of video to process for classification. The features are extracted from each frame and stored in a _le with respect to pattern generated for each class. The train file is stored in database for further testing process which is given in Algorithm 2

B. *Algorithm 1: Training RMLP*

Require: Vt (Training Video), t(Time for Video), S(Sigmoidal value), M(momemtum), L(Learning Rate), W(UpdatedWeight)

1: $V_t = \{V_1; V_2; V_3 \ldots V_n\}$
2: t=0, M = 0.6, L = 0.4, F = Features Of Frames
3: for i=1 to n do

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

```
 4:    MomentumBackpropagation(F,P)
 5:    extract F = {F_i1; F_i2; F_i3….. F_ik}
 6:    F_ik is set of frames from video
 7:    for j=3 to k do
 8:       D = |F_ij − F_ij − 1|
 9:       if D > t then
10:       key[t + +] = F_ij
11:      end if
12:   end for
13:   for j=1 to t do
14:      X(j) = Feature(key[j])
15:    Tag(j) = Tag given for video
16:    key
17:  end for
18: end for
```

Algorithm 2 shows the testing of a proposed system with taking testing samples and comparing with train sample to test. The classification is performed on the basis of recurrent connection which is given by the user and depends on the number of output values generated for system. The output neuron calculated in proposed system is j = 3 so the recurrent loop for classification is set to 3. It gives better result than previous one technique by filtering and testing the unknown samples 3 times to generate pattern and match with train file. Once the pattern is match, the output count decides the class of the input sample and generates output with result.

C. *Algorithm 2: Testing RMLP*

```
Require: Vt (Training Video), t(Time for Video), S(Sigmoidal value), M(momemtum),
    L(Learning Rate), W(UpdatedWeight)
 1: V_t = {V_1; V_2; V_3….V_n}
 2: t=0, M = 0.6, L = 0.4, F = Features Of Frames
 3: for i=1 to n do
 4:    MomentumBackpropagation(S,M,L)
 5:    extract F = {F_i1; F_i2; F_i3….. F_ik}
 6:    F_ik is set of frames from video
 7:    for j=3 to k do
 8:       D = |F_ij − F_ij − 1|
 9:       if D > t then
10:       key[t + +] = F_ij
11:      end if
12:   end for
13:   for j=1 to t do
14:      X(j) = Feature(key[j])
15:    Tag(j) = CLASSIFY(X(j), KNOWLEDGEBASE)
16:   end for
17: end for
```

IV. SIMULATION RESULTS

The performance metrics shows the performance of the proposed system. In the proposed system the result depends upon the parameter accuracy. The accuracy of classification is evaluated by using following Equation 1 for color feature.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

$$Accuracy = 100 - \left[\frac{100 * Number\ of\ Misclassified\ samples}{Total\ Number\ of\ Samples}\right] \quad \ldots (1)$$

The term PSNR is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. Because many signals have a very wide dynamic range, (ratio between the largest and smallest possible values of a changeable quantity) the PSNR is usually expressed in terms of the logarithmic decibel scale. The mathematical representation of the PSNR Equation 2 is as follows:

$$PSNR = 20log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \quad \ldots (2)$$

where the MSE (Mean Squared Error) Equation 3 is:

$$MSE = \left(\frac{1}{(m*n)}\right)^{*} sum(sum(f-g)^2) \quad \ldots (3)$$

Where f: Represents the matrix data of our original image
g: Matrix data of our degraded image in question
m: Numbers of rows of pixels of images and i represents the index of that row
n: Number of columns of pixels of image and j represents the index of that column
MAX: Maximum signal value that exists in our original known to be good image

The MAE, MPE and RSE are described below to minimize the error rate of the proposed system. The MAE is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The RSE of a sample mean is the standard error divided by the mean and expressed as a percentage or as a fractional value. The mean percentage error (MPE) is the computed average of percentage errors by which forecasts of a model differ from actual values of the quantity being forecast.

$$MAE = \frac{1}{n}\sum_{i}^{b}|f_i - y_i| = \frac{1}{n}\sum_{i}^{n}|e_i| \quad \ldots (4)$$

Where f_i: The prediction
y_i: The true value
e_i: The average of the absolute errors

$$RSE = \frac{s}{\sqrt{n}} \quad \ldots (5)$$

Where s: Sample standard deviation
n: Size (number of observations) of the sample.

$$MPE = \frac{100per}{n}\sum_{i}^{n}\frac{a_t - f_t}{a_t} \quad \ldots (6)$$

Where a_t: Actual value of the quantity being forecast
f_t: A forecast
n: Number of different times for which the variable is forecast

In experimental results the result analysis is performed on proposed work. Table 1 shows the result analysis of each sport class. It shows total number of training samples and total number of testing samples. Total classified samples and misclassified samples are observed. The total no of training sample is 63 and total number of testing sample is 51. From the testing samples 30 are classified samples and 21 are misclassified sample. The accuracy of each class is calculated and the values are obtained as shown in Table 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

In Table 2 row defines the actual sport class and column defines sport class. Confusion matrix shows how many misclassifications occurred and also which sport class is predicted. In Table 2 BT- Baseball Pitch, BS- Basketball Shooting, GS- Golf Swing, SJ- Soccer Juggling, TS- Tennis Swing, CB- Cricket Bowling, VS- Volleyball Spiking.

Sr. No.	Types of Sports	Total Training Samples	Total Testing Samples	Classified Samples	Misclassified Samples	Accuracy
01	Baseball Pitch	02	01	01	00	100%
02	Basketball Shooting	05	04	03	01	75%
03	Golf Swing	06	06	03	03	50%
04	Soccer Juggling	10	10	05	05	50%
05	Tennis Swing	15	10	06	04	60%
06	Cricket Bowling	11	10	07	03	70%
07	Volleyball Spiking	14	10	05	05	50%

Table 1: Result Analysis for Classification of Various Sports Classes

	BT	BS	GS	SJ	TS	CB	VS
BT	1	0	0	0	0	0	0
BS	0	3	0	0	0	1	1
GS	0	1	3	1	1	0	0
SJ	0	1	0	5	1	1	1
TS	0	0	0	1	6	1	1
CB	0	0	0	0	2	7	0
VS	0	0	0	0	0	0	5

Table 2: Confusion Matrix for Classification of Sport Video

The result generated for each class is shown in Table 3. The values get by taking average of the testing samples for each class. From result it shows that, the PSNR values generated for each class is greater and the mean square error rate is minimum means the system performs better in accuracy and generate better result than the previous system. In some classes the PSNR value is less and the MSE rate is very high indicate that the system perform better for the particular classes not for the all classes belongs to sport category.

Sport Class	Total Training Samples	Total Testing Samples	Average PSNR	Average MPE	Average MAE	Average RSE
BT	02	01	12.08789345	-4.231266778	0.042312667	$1.2929180855E^{-1}$
BS	05	04	15.07546678	-0.830588765	0.008305887	$6.73003866858E^{-7}$
GS	06	06	13.54434567	-50.5345009477	0.505345009	0.06764567781^{-7}
SJ	10	10	11.98723456	-23.987453455	0.239874534	$0.0342235565E^{-3}$
TS	15	10	8.76893456	-5.435678990	0.054356789	$1.675839455E^{-5}$
CB	11	10	11.67963456	-0.763682934	0.000763682	3.865398024^{-2}
VS	14	10	12.56743589	-4.326782367	0.043267823	$1.745367857E^{-4}$

Table 3: Result Parameter Analysis for Classification of Various Sports Classes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

The Table 3 shows the MPE, MAE and RSE values of each class and it find out that MPE and MAE for some classes are generate very low and shows the accuracy of result parameter is better than other one. The RSE i.e. relative standard error shows the error rate relative to the MPE and MAE. It result in a better performance than other system for RMLP neural network. The above result parameter related to each class shows that the system perform better than previous system and the performance of the system is increased by using RMLP neural network.

V. CONCLUSION AND FUTURE WORK

The proposed system is developed to analyze the sports videos on small scale dataset. The Video Classification with Recurrent Neural Network system will helps to recognize videos from specific class accurately. In proposed system, Visual feature extraction is used for extracting the features. In feature extraction HSV color space features are evaluated. The results are analyzed for HSV features by applying RMLP neural network. Based on these feature and result parameters generated for PSNR, MPE, MAE and RSE shows system performance is increased. The PSNR ratio generated is increased to 15.054. The MPE rate for proposed system is decreased to -12.873. The MAE rate is also decreased to 0.01287. The RSE of the system is shows the less standard rate for the each class and it is decreased to 13.9131. An RMLP classifier is used to accurately categorize input sport video into different categories like Baseball Pitch, Basketball Shooting, Golf Swing, Soccer Juggling, Tennis Swing, Cricket Bowling and Volleyball Spiking. The proposed system gives increased classification accuracy. The proposed system definitely reduces the video classification problems, which is very essential in various applications such as Real Time system like YouTube and Military Area. In future work, incorporate the more broader categories in the dataset to obtain more powerful and generic feature. Investigate the more powerful approaches like motion based feature extraction and explore hybrid approach for classification as a more powerful technique for video classification.

REFERENCES

1. T. M. Mitchell, “Machine Learning”, 3rd ed., ser. 5. McGraw-Hill Science/ Engineering/Math, vol. 4, ch. 8, pp. 121-125, March 1997.
2. E. Alpaydin,” Introduction to Machine Learning”, 2nd ed., ser. 4. The MIT Press, vol. 3, ch. 1, pp. 5-15, 2010.
3. N. J. Nilsson, “Introduction to Machine Learning: An Early Draft of a Proposed Textbook”, 3rd ed., ser. 3. The MIT Press, vol. 2, pp. 124-130, 1996.
4. 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 the 2014 IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, pp. 1725-1732, 2014.
5. D. Brezeale and D. J. Cook, “Automatic video classification: A survey of the literature.” IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 38, no. 3, pp. 416-430, 2008.
6. M. Hsken and P. Stagge, “Recurrent neural networks for time series classification.” Neurocomputing, vol. 50, pp. 223-235, 2003.
7. T. Burrows and M. Niranjan, “The use of recurrent neural networks for classification”, in Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Work-shop, pp. 117-125, Sep 1994.
8. C. A. Poynton, A Technical Introduction to Digital Video. New York, NY, USA: John Wiley & Sons, Inc., 1996.
9. D. P. Gibson, N. W. Campbell, and B. T. Thomas, “Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion”, in International Conference on Pattern Recognition (2). IEEE Computer Society, pp. 814-817, 2002.
10. S. Satkin and M. Hebert, “Modeling the temporal extent of actions”, in Computer Vision ECCV 2010, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, vol. 6311, pp. 536-548, 2010.
11. 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.
12. 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), pp. 2529-2536, June 2014.
13. 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, pp. 388-390, 2012.
14. S. Eickeler and S. Mller, “Content-based video indexing of tv broadcast news using hidden markov models”, pp. 2997-3000, 1999.