International Journal of Engineering & Technology, 7 (2.31) (2018) 85-88

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Visual recognition and classification of videos using deep convolutional neural networks

N. Shobha Rani[1*], Pramod N. Rao[2], Paul Clinton[3]

[1]Dept of Computer Science, Amrita School of Arts & Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.
[2]Dept of Computer Science, Amrita School of Arts & Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.
[3]Dept of Computer Science, Amrita School of Arts & Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.
**Corresponding author E-mail:n_shobharani@asas.mysore.amrita.edu*

Abstract

Classification of videos based on its content is one of the challenging and significant research problems. In this paper, a simple and efficient model is proposed for classification of sports videos using deep learned convolution neural networks. In the proposed research, the gray scale variants of image frames are employed for classification process through convolution technique at varied levels of abstraction by adapting it through a sequence of hidden layers. The image frames considered for classification are obtained after the duplicate frame elimination and each frame is further rescaled to dimension 120x240. The sports videos categories used for experimentation include badminton, football, cricket and tennis which are downloaded from various sources of google and YouTube. The classification in the proposed method is performed with Deep Convolution Neural Networks (DCNN) with around 20 filters each of size 5x5 with around stride length of2 and its outcomes are compared with Local Binary Patterns (LBP), Bag of Words Features (BWF) technique. The SURF features are extracted from the BWF technique and further 80% of strongest feature points are employed for clustering the image frames using K-Means clustering technique with an average accuracy achieved of about 87% in classification. The LBF technique had produced an average accuracy of 73% in differentiating one image frame to other whereas the DCNN had shown a promising outcome with accuracy of about 91% in case of 40% training and 60% test datasets, 99% accuracy in case of 60% training an 40% test datasets. The results depict that the proposed method outperforms the image processing-based techniques LBP and BWF.

***Keywords**: Sports videos, convolutional neural networks, local binary patterns, bag of words features, SURF, K-Means clustering, video processing.*

1. Introduction

Simple classification tasks like visual recognition of objects can be achieved aptly with good efficiency through Machine Learning (ML) techniques. However, the implication of massive computational complexities onlarge scale image classification tasks had led to the motivation of using Deep Learning Systems (DLS) [18]. These systems are dominant in its usage due to its high efficiency and robustness in prediction of outcomes. ML based techniques consider the feature extraction and classification as two isolated protocols and usually limited by small scale data set classifications [19]. DLS are more powerful tools with capability of processing datasets in the volumes of millions and therefore enabling the machines to perform large scale image classification tasks.

In order to learn variety of object categories ranging interms of 1000's or even higher from millions of images, one needs a specialized image processing system with large scale learning capabilities [1]. DLS inheriting Convolutional Neural Network (CNN) architecture is one of such efficient model inhibiting learning at varying levels of abstraction on the same image. The CNN outperforms ML techniques is the process of visual recognition of image contents or classification of images at large scale [2]. CNN architectures originally investigated and proposed byKrizhevsky et al 2012 [1]; Zeiler and Fergus in 2013 [4]; Sermanet et al in 2013 [5] had provided significant improvement in accuracy and performance of many large-scale classification and recognitiontasks which are remained as irresolvable research challenges in the field of computer vision.

In the proposed investigation, CNN is applied on the classification of sports videos of multiple categories. The number of image frames generated from videos of longer durations is quite high even after elimination of duplicate frames from the sliced videos. This result in a large volume of image frames (distinct) generated from various sports video collections are considered here as sources for classification of video types. Hence, the problem of video classification turns out to be a large-scale image classification problem [6, 7, 8]. In this research, we focus on the problem of sports video classificationfrom a collection including various other sports video types which will be usually obtained after first round of user feedback cycle in the information retrieval process. The objective of this investigation is to mainly improvise the accuracy in retrieval of sports videos of particular type as specified and expected by user, rather than mixing up of various other sports video categories.

Until recently, many works are proposed in the area of image classification using deep Convolutional networks with varied architecture type, however the challenging aspect of proposed research lies classification sports videos in the local context rather globally which including various other image categories. For example, the classification two similar sports videos like Tennis and Badminton (local context) is challenging rather than entertainment and sports (global context). Some of the works reported in the literature include., large video classification using Convolutional neural networks performed on over 1 million videos including 487 classes by Karpathy et al [6], a supervised rule based classification approach is proposed by Zhou et al [9] by utilizing visual and motional characteristic descriptors in the visual contents of videos, Brezale et al had reviewed the literature of various automatic video classification techniques including supervised and unsupervised clustering strategies, Huang et al [10] had made an attempt on classification of video and audio data using hidden markov models by extracting the multi modal features which has shown a significant improvement in prediction of video content type. An investigation is carried out by Lin et al [11] for classification of news videos by using the floating text in video frames as well as the visual contents by using support vector machine classifier that had reflected a significant impact on improvement of precision and recall rates. A spatial temporal super feature vector is used for classification of video types into news, cartoon, sports, commercial and music using probabilistic modeling and principal component analysis which has shown an overall improvement in performance of system [12]. Face and text trajectories are employed by Dimitrova et al [13] for classification of video clips by using hidden markov models for classifying videos into commercial, news, sitcom and soap and achieved an accuracy of about 80%. Yang et al [14] had proposed an approach for classification of videos using bag of visual words features based on the key points extracted from salient image patterns, Zhao et al [15] had proposed a method for analysis of video contents using the features of local binary patterns and claimed that LBP features are efficient in recognition of changes incurred due to dynamic textures in videos.Lippmann et al [16] had examined the applicability of variety of neural network classifiers towards video type detection including probabilistic classifiers, kernel and exemplar classifiers etc.Geoffrey et al [17] had analyzed the applicability of the deep belief networks towards the classification of digits.

It is clear from the literature that, most of the video classification tasks are handled through the use of supervised and unsupervised classification techniques. Despite the limitations of computational, performance considerations and constraints on datasets limits, the use of ML techniques had provided the appreciable results. However, the use of deep learning networks for classification of videos had accelerated the performance of the video classification systems by extending the capability of specialized systems to handle millions of datasets through the additional computational power of Graphics Processing Unit(GPU). Hence, an attempt is made in the proposed investigation to analyze the performance of system towards classification of only sports videos. The rest of the paper is organized as follows, section 2 describes the proposed methodology, the analysis of methods applied is discussed with experimental evidences in section 3 and finally section 4 concludes the proposed work.

2. Proposed technique

The methodology applied to classify the videos include, convolutional neural networks and bag of words which as discussed subsequently.

Convolution neural network

The computational complexities involved in interpretation and recognition of visual contents of video frames reinforces the problem of video classification for further investigations. In the proposed methodology, the classification of sports videos is performed using the deep Convolutional neural networks. Figure 1 depicts the overall flow of video type classification using the Convolutional neural networks.

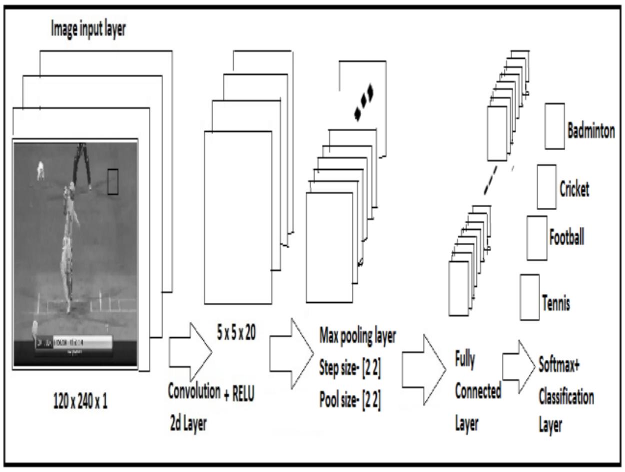
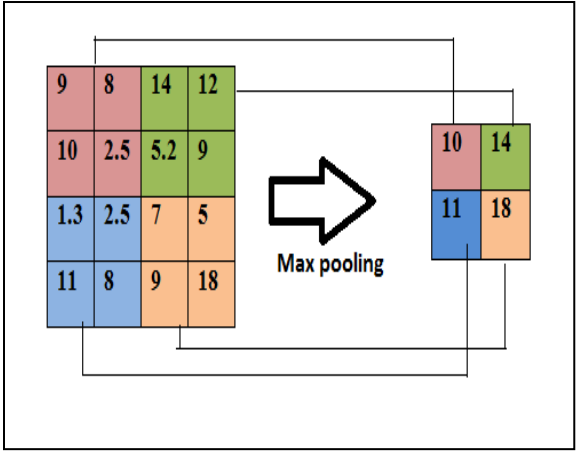


Figure 1: CNN architecture for sports video classification

The overall architecture of CNN includes Image input layer, convolution layer which employs 20 filters of size 5 x 5 to create the subsequent higher-level abstraction features of the image. Further, RELU layer acts like rectifier activation function which adds non-linearity to the network and it does not show any impact on the spatial dimensions changes in the image and also it helps in speeding up the training process. The output of the RELU layer is acquired to the max pooling layer which performs down sampling, thereby reducing the dimensionality of the input image representation. A max filter is convolved over the non-overlapping sub regions of the initial image to obtain the abstract representation of input image as depicted in figure 2. As a result of which number of learning parameters are reduced leading to low computational costs.



9	8	14	12
10	2.5	5.2	9
1.3	2.5	7	5
11	8	9	18

10	14
11	18

Figure 2: Max pooling from size of 4 x 4 to 2 x 2

Mapping from internal representation to generalized representation

Max pooling layer provides the generalized representation of the internal image representation which is invariant to various geometric transformations through a series of stacked hidden layers. In the proposed architecture, the pool size of the max pooling layer is chosen as width of ‘2’ and height of ‘2’ with a stride length of ‘2’. The size of the output OP_{sz} produced by pooling layer with input of size IP_{sz} is given by (1).

$$OP_{sz} = \frac{(IP_{sz} - P_{sz} + 2 * Pad_{sz})}{Str_{len}} + 1 \qquad (1)$$

Where pool size is indicated by P_{sz} , padding size is represented by Pad_{sz} and Str_{len} is the stride length used in max pooling layer.

Higher level abstraction of the features is achieved by assuming the input volumes as the output obtained max pooling layer preceding it and outputs an N dimensional vector where N is the number of classes into which the inputs are classified based on the probabilities returned to each class. The outputs from max pooling layer are directed to fully connected layers which connects each neuron in one layer to every other neuron in the other layer leading to formation of a multi-layer perceptron neural network. Output size is specified through the fully connected layer, in the sports video classification system as a total of 4 classes.

The output is ((120-5 + 2x0) / 1) +1 = 116 for one direction in the convolutional layer (feature mapping). The max pool in the proposed architecture has areas that do not overlap, it is down-sampled by 2 in each direction, i.e.,116/2 = 58. For one channel of the convolutional layer, the output of the max pool layer is 58 × 58 = 3364. The convolutional layer has 20 channels, so the output of the max pool layer is 3364x20 = 67280, which is the size of the fully connected layer input. The fully-connected layer returns a 4 cross-67280 matrix with a values from the Gaussian distribution with mean 0 and standard deviation 1 in the proposed architecture

Bag of features

The methodology for extraction of bag of features is depicted in figure 3. Initially, the process begins by identification of strong feature point locations in the input image I is accomplished using SURF which is popularly known as Speeded Up Robust Features. The process of identifying strong feature point locations involves feature extraction, feature description and feature matching.

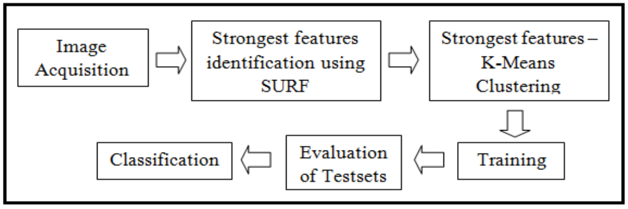


Image Acquisition
Strongest features identification using SURF
Strongest features – K-Means Clustering
Training
Evaluation of Testsets
Classification

Figure 3: Classification using bag of Features

SURF is a scale-invariant feature descriptor for object recognition and image classification. SURF can be seen as the difference between Gaussian and automatic size selection. First, construct a scale space where the input images are filtered in different proportions. The scale space can be assumed to be a pyramid in which two consecutive images are related by a scale change and the scales are then grouped by octaves, i.e., a large change in the size of a Gaussian filter. The calculations done by the Gaussian filter are replaced by fast approximations.

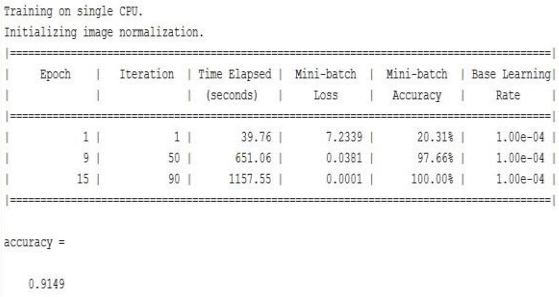
3. Experimental analysis

The objective of video classification is to assign input patterns of the image frames to one of the four classes. The input image frames extracted from the videos of multiple sport categories is comprised a total of 500 frames of each category leading to a total of 500 x 4 frames. The outcomes of our experimentations are discussed as follows. Figure 4 shows few instances of image frames employed for classification.

Figure 4: Instances of image frames used for classification

Convolution neural network

The classification of videos using convolutional neural networks is accomplished in training and testing phases. The total number of image frames includes a total of 2000 frames in which the experimentation is carried out by analyzing the training and testing datasets in various ratios. Initially the classification is carried out by dividing the entire datasets into 60% and 40% for training and testing. Further, it is tested by considering both training and testing in equal proportions. Figure 5 represents the outcomes of the experimentation using CNN for total number of epochs equal to 15, number of iterations as 90 and simulation on single CPU.

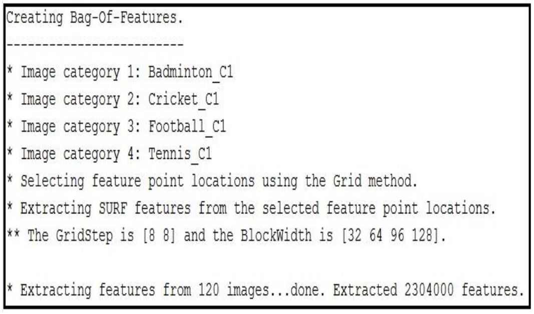


```
Training on single CPU.
Initializing image normalization.
|=======================================================================================|
|     Epoch     |   Iteration   | Time Elapsed |  Mini-batch  |  Mini-batch  | Base Learning|
|               |               |  (seconds)   |     Loss     |   Accuracy   |     Rate     |
|=======================================================================================|
|             1 |             1 |        39.76 |       7.2339 |       20.31% |     1.00e-04 |
|             9 |            50 |       651.06 |       0.0381 |       97.66% |     1.00e-04 |
|            15 |            90 |      1157.55 |       0.0001 |      100.00% |     1.00e-04 |
|=======================================================================================|

accuracy =

    0.9149
```

Figure 5: Performance metrics of experimentation using CNN

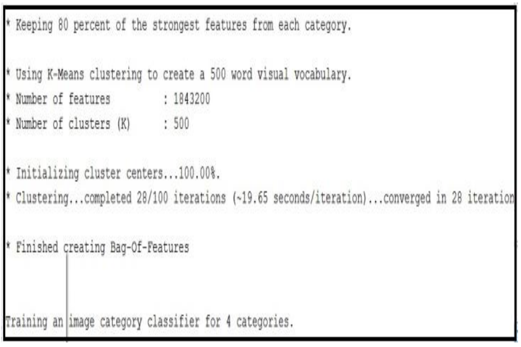
Bag of features

The training set is assumed to be around 80% for classification of videos using Bag of Features. For the reduced training samples still, the performance of SURF in terms of total elapsed time and accuracy is very poor. Figure 6, figure 7, figure 8 depicts the outcomes obtained with Bag of features technique.



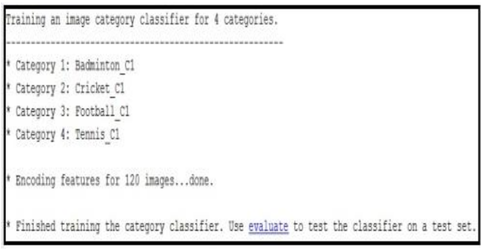
```
Creating Bag-Of-Features.
-------------------------
* Image category 1: Badminton_C1
* Image category 2: Cricket_C1
* Image category 3: Football_C1
* Image category 4: Tennis_C1
* Selecting feature point locations using the Grid method.
* Extracting SURF features from the selected feature point locations.
** The GridStep is [8 8] and the BlockWidth is [32 64 96 128].


* Extracting features from 120 images...done. Extracted 2304000 features.
```

Figure 6: Creation of bag of features using SURF



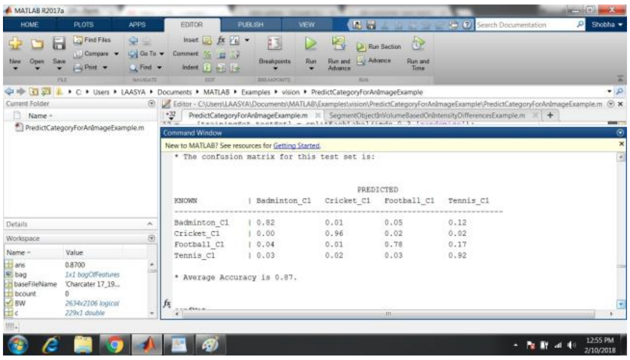
```
* Keeping 80 percent of the strongest features from each category.

* Using K-Means clustering to create a 500 word visual vocabulary.
* Number of features        : 1843200
* Number of clusters (K)    : 500

* Initializing cluster centers...100.00%.
* Clustering...completed 28/100 iterations (~19.65 seconds/iteration)...converged in 28 iteration

* Finished creating Bag-Of-Features


Training an image category classifier for 4 categories.
```

Figure 7: Identification of cluster centers - K-Means clustering



```
Training an image category classifier for 4 categories.
--------------------------------------------------------
* Category 1: Badminton_C1
* Category 2: Cricket_C1
* Category 3: Football_C1
* Category 4: Tennis_C1

* Encoding features for 120 images...done.

* Finished training the category classifier. Use evaluate to test the classifier on a test set.
```

Figure 8: Training of sport data-Bag of features



```
* The confusion matrix for this test set is:


                           PREDICTED
KNOWN          | Badminton_C1   Cricket_C1   Football_C1   Tennis_C1
--------------------------------------------------------------------
Badminton_C1   | 0.82           0.01         0.05          0.12
Cricket_C1     | 0.00           0.96         0.02          0.02
Football_C1    | 0.04           0.01         0.78          0.17
Tennis_C1      | 0.03           0.02         0.03          0.92

* Average Accuracy is 0.87.
```

Figure 9: Confusion matrix of classified image frames- Bag of Features

An average of about 87% is achieved with Bag of Features technique for classification of sports videos.

4. Conclusion

Performance of the convolutional neural networks and Bag of Features is analyzed in the proposed method for classification of sports videos. It is evident; the performance of convolutional neural network is appreciable compared to Bag of Features technique. The experimentation is performed on a single CPU and accuracy of about 92% is achieved with CNN for equal rate of training and testing data samples.

References

[1] Krizhevsky A, Sutskever I & Hinton GE, “Imagenet classification with deep convolutional neural networks”, *Advances in neural information processing systems*, (2012), pp.1097-1105

[2] Ciregan D, Meier U & Schmidhuber J, “Multi-column deep neural networks for image classification”, *IEEE conference on Computer vision and pattern recognition (CVPR)*, (2012), pp.3642-3649.

[3] Simonyan K & Zisserman A, “Very deep convolutional networks for large-scale image recognition”, *arXiv preprint arXiv:1409.1556*, (2014).

[4] Zeiler MD & Fergus R, “Visualizing and understanding convolutional networks”, *European conference on computer vision*, (2014), pp. 818-833.

[5] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R & LeCun Y, “Overfeat: Integrated recognition, localization and detection using convolutional networks”, *arXiv preprint arXiv:1312.6229*, (2013).

[6] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R. & Fei-Fei L, “Large-scale video classification with convolutional neural networks”, *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, (2014), pp.1725-1732.

[7] Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R & Toderici G, “Beyond short snippets: Deep networks for video classification”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), pp.4694-4702.

[8] Brezeale D & Cook DJ, “Automatic video classification: A survey of the literature”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol.38, No.3, (2008), pp.416-430.

[9] Zhou W, Vellaikal A & Kuo CC, “Rule-based video classification system for basketball video indexing”, *Proceedings of the ACM workshops on Multimedia*, (2000), pp.213-216.

[10] Huang J, Liu Z, Wang Y, Chen Y & Wong EK, “Integration of multimodal features for video scene classification based on HMM”, *IEEE 3rd Workshop on Multimedia Signal Processing*, (1999), pp. 53-58.

[11] Lin WH & Hauptmann A, “News video classification using SVM-based multimodal classifiers and combination strategies”, *Proceedings of the tenth ACM international conference on Multimedia*, (2002), pp.323-326.

[12] Xu LQ & Li Y, “Video classification using spatial-temporal features and PCA”, *International Conference on Multimedia and Expo*, (2003).

[13] Dimitrova N, Agnihotri L & Wei G, “Video classification based on HMM using text and faces”, *10th European Signal Processing Conference*, (2000), pp.1-4.

[14] Yang J, Jiang YG, Hauptmann AG & Ngo CW, “Evaluating bag-of-visual-words representations in scene classification”, *Proceedings of the international workshop on Workshop on multimedia information retrieval*, (2007), pp.197-206.

[15] Zhao G, Ahonen T, Matas J & Pietikainen M, “Rotation-invariant image and video description with local binary pattern features”, *IEEE Transactions on Image Processing*, Vol.21, No.4,(2012), pp.1465-1477.

[16] Lippmann RP, “Pattern classification using neural networks”, *IEEE communications magazine*, Vol.27, No.11,(1989), pp.47-50.

[17] Hinton GE, Osindero S & The YW, “A fast learning algorithm for deep belief nets”, *Neural computation*, Vol.18, No.7,(2006), pp.1527-1554.

[18] Rani NS & Ashwini PS, “A Standardized Frame work for Handwritten and Printed Kannada Numeral Recognition and Translation using Probabilistic Neural Networks”, *IJISET - International Journal of Innovative Science, Engineering & Technology*, Vol.1, No.4, (2014).

[19] Pushpa BR, Anand C & Mithun NP, “Ayurvedic Plant Species Recognition using StatisticalParameters on Leaf Images”, *International Journal of Applied Engineering Research*, Vol.11, No.7,(2016), pp.5142-5147.