International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 2, (July-December 2018), pp 329-336

CLASSIFICATION OF COMPRESSED VIDEOS

Dipti Singh[1] and Priyanka Sharma[2]

[1]*Computer Science and Engineering, Nirma University, Ahmedabad, Gujarat-382481. Email: 15mcec11@nirmauni.ac.in*
[2]*Professor of Computer Science, Nirma University, Ahmedabad, Gujarat-382481. Email: priyanka.sharma@nirmauni.ac.in*

Abstract: Classification Of Compressed Videos is one of the challenging aspect in today's date. Analyzing and classifying the contents of videos is an important factor in retrieval of Data. In today's time, around 72, 000 videos are uploaded on YouTube per minute so it becomes very difficult to classify all the videos manually or using certain low level features only. In order to properly classify the data content of a video, we need to have a proper knowledge of the various features of a video. Previously, many classification tasks has been performed based on the text, audio, video or many a times these features can be fused as in case of HMM which uses both text and audio. In this paper, we will actually use features like motion vectors, gradient descent and optical flow. These features will be compared and based on the comparison result further classification will be done accordingly.

Keywords: Compressed Videos, bag of visual words, mac-roblock, HEVC.

1. INTRODUCTION

Video classification includes huge video management, Web video Search, Video surveillance. In many of the video compression techniques, motion based features plays an important role as it reflects how dynamically the features changes with respect to time and space. Previously, Optical flow[1] was being used as an important aspect to recognize humans and detect their activities in the given video. Detection of shot changes, Object Detection and segmentation, motion detection, gradient descent all the aspect were used individually or in an integrated form to analyze the content of videos. video processing application in which maximum motion related information is extracted. One of the motion related feature is Macroblock and Histogram of Motion gradients[2] which is further used for classification of these videos. Likewise we have feature extraction method which includes optical flow, histogram of gradients, Motion boundary histograms. When we consider the Human Visual System (HVS), the focus of object detection and identification is the salient region whereas the background region is not of that much importance and is generally ignored. In order to achieve discriminative feature sets, mainly more importance is given to salient region so that much of the robustness is achieved. Generally, we extract local features from the spatial gradient, temporal difference, optical flow and trajectory which reflect the temporal structure and temporal motion of the video content. Videos are everywhere and we need to properly classify them based on their content. Videos are required and use in every field be it in research, sports, education. Hence, classifying, analysing and retrieving the video is very important for gaining proper knowledge. Need for working on compressed videos:

1. Compressed videos require very less disk space for storage.
2. Bandwidth requirement for transferring the video content is also very less.
3. Reading and writing on the disk becomes very easy due to compression.
4. Order of bytes is independent so a compression does not affect the content of the video.

Bag of words which has proved to be successful in the domain of pixel. In a video an object while moving have varying shape and size as well as the location of the object keeps on changing in both the foreground

as well as the background region. This method includes the following important steps: First all the coding units in the quad tree is defined using syntax features of HEVC . With the help of a clustering method like *k* means we generate a codebook. After the second step all the moving objects in the video are coded with the help of code words from the codebook. Finally, we make use of classifier which can classify between the person and vehicle. Its very challenging to identify all the striking features which can discriminate properly between the two objects.

Along with the visual features, temporal features also plays an important role in describing about the content of the video. Many researches have been done on the video classification based on the motion information. Generally, the information regarding the motion is extracted from the optical flow which is computed through the vertical and horizontal components of the frames in the case of temporal evolution. But the computation of optical flow is very cost effective. To extract the motion information from the videos is highly cost effective. Among all the feature extraction method, optical flow consumes more than percent of the total time. Histogram of Optical gradient[3] provides an effective method to extract the motion information using the spatial temporal derivation. After the extraction of descriptor the next important step is feature encoding. There are different methods proposed for this step which has different level of accuracy and efficiency. This steps forms the crucial step in order to achieve an efficient output.

2. LITERATURE REVIEW

Sovan Biswas et. al., [4] make use of motion vectors as a part of feature extraction. From the motion vectors, orientation information is extracted. From the hierarchical space time cubes, we calculate histogram of motion oriented vectors. This paper is based on three evaluation steps: (1) Preprocessing (2) Feature Extraction based on Histogram of Motion Vectors (3) Extraction of Video Feature. All the motion vectors which are computed are not of use, some of them are noisy. In Order to remove the redundancies and lower down the computation cost, we need to prune them. Camera parameter also needs to be estimated. Finding the region of interest in a video is the most crucial part. Most of the noisy vectors have very high magnitude sometimes it is almost the size of the frame so in this case we can truncate those noisy vectors which have size more than 10 percent of the frame. In order to have an effective output, we need and Region of Interest which can be found through spatial motion orientation gradient and motion magnitude gradient. Feature Extraction: Feature is extracted through space time cube generation. From the motion vectors, temporal cubes are generated from overlapping *b* frames. It is further divided into three levels coarser, medium and finer each having its own division coarser (1x1x1), medium (3x1x1) and finer (5x1x1). Histogram of oriented motion vector is very efficient in separating the different motion vector but sometimes it fails to do so. For example, a person walking towards left is considered to be same as a person walking towards right. In this place it makes use of normalization factor to get rid of the anomalies.

Liang Zhao et. al., [2] reflects the importance on a video where it is first necessary to identify whether an image in the given video belongs to the foreground part or the background part. Usage of second classifier which actually gets trained to distinguish between person and vehicle. In many video indexing applications such as video surveillance and classification, Compressed domain moving object segmentation has a crucial role to play. Encoding features has been developing from the time of H. 264[4] but more advanced features have been provided in HEVC. Hence, applying HEVC[5] on the compressed domain for the purpose of classification and segmentation has become an interesting are for research. In a video, there are various objects which keep on moving. In this paper, mainly person and vehicles are the two objects considered here. Segmentation and classification are two methods that are applied on these two objects for proper classification. In a video, it is first necessary to identify whether an image in the given video belongs to the foreground part or the background part. The important scenes take place in the foreground part and most of the scenes in the background region can

be discarded. For this purpose, it is required to train a classifier to distinguish between the two regions. The second step includes the usage of second classifier which actually gets trained to distinguish between person and vehicle.

Weiyao Lin et. al., [6] focuses on Video processing application in which maximum motion related information is extracted. One of the motion related feature is Macroblock which is further used for classification of these videos. First the Motion Vector field is analyzed in the compressed domain and then we classify each Macroblocks[6] obtained from several individual frames. The macro block thus obtained from the different frames is used to classify them into different classes and the information which is obtained through these classes is used to describe the content of the frame. Macroblock classification is both low computational in nature as well as very highly efficient method. In the practical scenario, generally videos after getting processed, they are stored in a compressed format. In this compressed form, Motion estimation is performed in order to get rid of all the temporal and spatial redundancies. Motion Estimation is a process which identifies the similarity between the various adjacent frames. Hence, it becomes very easy for us to determine the motion pattern and similarity between the various frames. Motion Vector information is extracted from the frames in the compressed domain which exists in the bitstream form. This paper also reflects the usage of MB class information in detecting the Shot Change. Shot is defined as the continuous length of video frames which is mainly captured by one camera action[7]. A single operation of camera captures the whole sequence of frames whereas the shot change is defined as the boundary where the difference between the two shots can be seen. Change in the shot can be of varying types like it can be gradual or it can be very smooth and this change is detected where there is two dissimilar videos.

Xiaolin Tang [3] explains how trajectory method can be used in the process of video classification then two approaches are used: Detection of Feature point and extraction of local feature. In this paper it is shown that it is first necessary to detect the salient region in

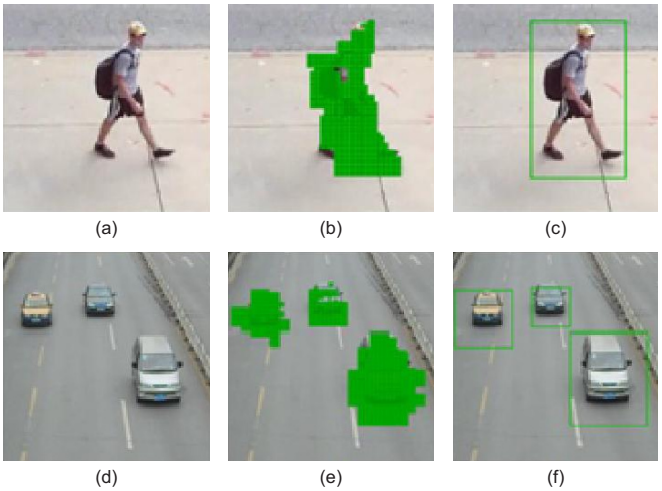

Figure 1: Object Segmentation and Classification

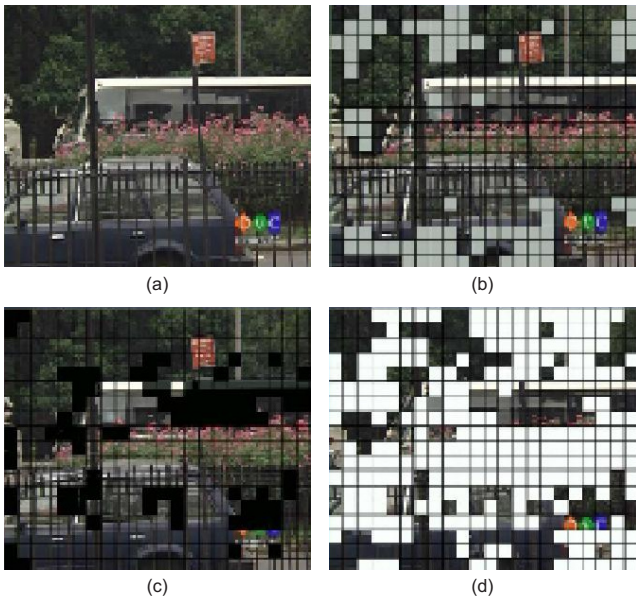

Figure 2: Macroblock representation of an image

the entire frame and from this salient region we identify speech-impaired feature. This method in general reduces greatly the feature points which are noisy in nature from the background region and hence it ultimately leads to low computational cost. Removal of noisy features leads to highly discriminative features. Both feature point detection as well as the local feature extraction helps in the recognition of action in various datasets like UCF50, HMDB. The whole process includes the calculation of saliency map[3] in which the very first step is generation of super pixel whose complexity is given as O(N). This complexity can be reduced by down sampling in which we reduce the number of input frames. If we downsample the original

frame by a factor $x = 1/W$ in both dimensions then by a factor of $1/n2$ it is reduced when it is down sampled. In the graph manifold ranking method, there are vertices and these vertices are connected through edges. Super pixel generated above is treated as the vertices and they are connected to the local neighbors. Hence these vertices and edges are used in the calculation of saliency map. We can define the object of interest in a video sequence by identifying their motion characteristics as well as their color contrast. Their boundary is also very different as they become the main focus during the shot capture.

3. PROPOSED FRAMEWORK

In the phase of implementation, we make use of Bag of Visual Words. This method is mainly used in the process of recognizing any action going on in the video. The very first step is to encode all the features with the help of a codebook. Fisher kernel encoding and sparse [8] are the few methods which are used as an encoding methods used in BoVW. In this paper, various pooling and normalization technique along with encoding method are experimented with to identify an object in the video and its action. Dataset UCF50[8] is used where 50 classes of different human actions like playing tabla, playing piano are depicted with each having around 50 different videos representing different actions. As we know, in the field of video surveillance techniques method for human action recognition plays an important role. In order to get BOVW[8] of action recognition we carry out the whole step in following three steps: (a) Extraction of features (b) Vector quantization of the extracted features. (c) Construction of histogram representation[1] with sum pooling with the help of L1 normalization technique. The last step includes the usage of nonlinear svm classifier which takes final histogram representation as an input. The very next step is feature encoding . We have five different encoding methods which needs to be compared with the help of different datasets for example, Video quantization encoding method, sparse encoding, fisher encoding. Different Encoding Methods[1].

(a) *Vector Quantization:* With the help of hard assignment, a histogram of visual word is obtained. Histogram of visual words can be obtained through the method of sum pooling and normalized l1 normalization.

(b) *Soft-assignment Encoding:* This method is done through two steps: all the visual features are assigned a codeword and at the same time k nearest neighbor is identified and they are assigned to the features. In order to control the softness, we make use of a single parameter i.e., x.

(c) *Linear Encoding:* This method is used as a fast encoding method. Features of k nearest neighbours is used as the base feature. K is given value of 5 and then we make use of pooling and normalization.

(d) *Sparse Encoding:* Here we use y as another parameter so that a balance can be maintained between loss term and regularization of sparse. Default value of y is given as 0. 15 and on this we use max pooling and normalization with norm l2.

(e) *Fisher kernel encoding:* This method requires us that we train GMM for all the input features. As we know, the features have very high dimensionality like histogram of gradients have a dimension of 162 and it is necessary to remove the redundancy. If we use directly EM algorithm then the singular covariance problem takes place. Other factors are also taken into considerations which are the size of the codebook as well as the computational cost for the whole process. After the above step of feature encoding is performed then comes the step for pooling and normalization. In this method both max pooling as well as the sum pooling is analyzed. After the process of pooling is carried out, normalization is done through three methods: l1, l2 and power normalization.

4. METHODOLOGY

Video Volume: Large amount of videos are given as input. Videos from various fields like research,

education, news, sports, entertainment etc. are provided as an input the classification technique. Various datasets like UCF 50, UCF 101, HMDB, KTH, and BOLLYWOOD 50 contain various videos of different classes. These Videos are subdivided into training and testing data as well as validation data. So many videos are taken into use so that training of the data is properly executed. These videos are given as an input to the classifier like SVM which classifies them accordingly to their respective classes.

Feature Extraction: [1] In the MPEG compressed videos, group of pictures is the independent unit. It actually comprises of the sequence of I, B, P frames. Compression can be mainly performed on the B and P frames with very low loss of data. From this group of pictures, we can extract feature vectors which actually comprises of notion and color information. For example, a video in MP4 can be used to extract frames mainly Intra frames with the help of ffmpeg and further from these frames, motion vectors can be extracted to obtain the orientation information. Preprocessing: When we extract motion vectors then most of the vectors extracted are not of our use. These motion vectors are noisy in nature. Removal of these leads to lower computational cost. In order to increase the efficiency of coding, prefiltering algorithm is applied for the calculation of frame difference . In video compression technique, a video in MP4 format is applied to ffmpeg to extract Intra frame and from these frames we can extract motion vectors. Hence it saves computational time from extracting vectors from B and P frames. Classification: In todays time we have huge lot of databases of videos from different fields and it is not possible to manage it efficiently through a single framework. So todays need is implementation of a proper algorithm which can efficiently classify all the videos in their respective domain. Previously, classification of videos has been carried out on different modalities like video, audio, text and sometimes they are integrated together to yield better results. Proper and efficient classification of videos leads to easy analysis and retrieval from their respective domain. In case of compressed domain classification, various features extraction method like motion vector, Gradient descent, Optical Flow, SIFT. Large amount of videos are given as input. Videos from various fields like research, education, news, sports, entertainment etc. are provided as an input the classification technique. Various datasets like UCF 50, UCF 101, HMDB, KTH, and BOLLYWOOD contain various videos of different classes. These Videos are subdivided into training and testing data as well as validation data. HEVC makes use of a new approach for object segmentation and classification which is a quad tree approach. In this approach, partitioning is based on the coding tree units. The root of the tree is formed by the Coding tree unit which is further subdivided into Coding units. Size of the various units in the tree is determined by the quad tree approach . Coding unit further forms the root for the prediction tree. Hence in this way, there are three levels in which coding tree unit becomes the main root of the tree then Coding unit becomes the root of the prediction tree. There is only one level in the prediction tree which only determines how the coding unit can be subdivided into prediction blocks. In case of inter partition mode we have only 8 partition mode where as in case of intra partition mode, we have only 2 Partition modes. The usage of this quad tree approach greatly increases the efficiency of the coding of compressed domain using HEVC method. Hence Hevc method can be effectively used in the video classification field. In this paper, we mainly aim to explore the different fusion methods so that we can produce a state-of-the-art action recognition system. If we want to use a single descriptor then only we can make use of Bag of Visual words approach as it ignores the usage of fusing multiple descriptors. We can have different levels of fusion descriptors which includes descriptor level, representation level as well as score level. When we talk about descriptor level fusion then we consider multiple descriptors from the same cuboid whereas in representation level, which is performed at video level in which each descriptor if first given as an input to the BoVW framework independently and then the overall representation is used to train the classifier. In case of score level, each descriptor is given as an input to the framework on the individual basis and then further it is used to train a recognition classifier.

In BoVW, we make use of various descriptors (HOG, MBH, HOF) and feature extractors like dense trajectories which is very helpful in depicting the visual pattern of a cuboid. For better results and performance improvement we can make use of fusion of various descriptors. There are various components which were made used in BOVW framework. They are as follows:

1. The very step in this framework is data pre-processing in which we remove the unwanted and noisy data so that a better final recognition is obtained.
2. For encoding methods, high dimensional super vector is used which yields very effective and efficient results as compared to other encoding methods.
3. Finally, we make use of pooling and normalization which is one of the major step which is used in this framework specially sum pooling along with power l2 normalization.

The overall sequential execution of the steps comprises of the following steps:

1. The very step includes feature extraction which includes detection and description of the low level features. We can make use of various local descriptors which can describe major features like motion boundary, static appearance as well as the various motion undergoing in the foreground or the background region. We have local features which includes space time interest points as well as the dense trajectory and they are used widely because of their ease of use as well as their robust performance. In case of STIP, we make use of HOG and HOF. We have another low level feature descriptor which includes iDTs which stands for improved dense trajectories which uses much more sophisticated engineering skills and can define more refined low level features.
2. *Feature pre-processing:* In this step we make use of principle component analysis. This process is a statistical procedure which is used to pre-process the features which makes use of the orthogonal transformation which can map features with principle components which is a set of linearly correlated variables.
3. *Codebook generation:* Third step includes the division of the various feature space into regions which is described by its center which is referred as codeword and generative model further explains the probability distribution of features.
4. *Pooling and normalization:* Inorder to have the global representation of a video we can make use of the code coefficients in pooling operation. There are two types of pooling operation: max pooling and sum pooling. In sum pooling we calculate the overall coefficients whereas in max pooling we consider the maximum value among all.

This pooling operation is further normalized by the normalization methods which includes l1, l2 and l3 normalization techniques.

5. PERFORMANCE ANALYSIS

As we know, in the field of human surveillance technique, method of action recognition plays an important role and it is highly efficient and effective in classifying the different actions of human.

Figure 3: Confusion table when the codebook size is 300

K means method is considered as a hard clustering algorithm whereas GMM can be considered as a soft clustering. GMM provides the additional feature of providing the shape of the probability distribution along with the mean of the visual words. . We have five different encoding methods which needs to be compared with the help of different datsets for

example, Video quantization encoding method, sparse encoding, fisher encoding. Other factors that are also taken into consideration which are the size of the codebook as well as the computational cost for the whole process. After the above step, feature encoding is performed then comes the step for pooling and normalization. In this method both max pooling as well as the sum pooling is analysed. After the process of pooling is carried out, normalization is done through three methods: l1, l2 and power normalization. Result below shown includes the accuracy level of different classes. Accuracy of a class reflects how efficiently the different actions belong to that class is efficiently classified. Few of the classes have efficiency less than 50 percent which means the actions were not properly identified and that is why it could not be classified in their respective classes whereas in most of the other classes accuracy level is above 80 percent which leads to a better classification of their respective actions. Results of classification of all 50 classes is shown as follows:

```
Mean per-class accuracy: ...
Accuracy of class 1: 0.831429
Accuracy of class 2: 0.707571
Accuracy of class 3: 0.978286
Accuracy of class 4: 0.925333
Accuracy of class 5: 1.000000
Accuracy of class 6: 0.980000
Accuracy of class 7: 1.000000
Accuracy of class 8: 0.994286
Accuracy of class 9: 0.783333
Accuracy of class 10: 0.972000
Accuracy of class 11: 0.950238
Accuracy of class 12: 0.827905
```

Figure 4: Result of classification for classes from 1-12

```
Accuracy of class 13: 0.948667
Accuracy of class 14: 0.906905
Accuracy of class 15: 0.890667
Accuracy of class 16: 0.766000
Accuracy of class 17: 0.778286
Accuracy of class 18: 0.909619
Accuracy of class 19: 0.910000
Accuracy of class 20: 0.860365
Accuracy of class 21: 0.716364
Accuracy of class 22: 0.981111
Accuracy of class 23: 0.840905
Accuracy of class 24: 0.462381
Accuracy of class 25: 0.688000
```

Figure 5: Result of classification for classes from 13-25

```
Accuracy of class 13: 0.948667
Accuracy of class 14: 0.906905
Accuracy of class 15: 0.890667
Accuracy of class 16: 0.766000
Accuracy of class 17: 0.778286
Accuracy of class 18: 0.909619
Accuracy of class 19: 0.910000
Accuracy of class 20: 0.860365
Accuracy of class 21: 0.716364
Accuracy of class 22: 0.981111
Accuracy of class 23: 0.840905
Accuracy of class 24: 0.462381
Accuracy of class 25: 0.688000
```

Figure 6: Result of classification for classes from 16-37

```
Accuracy of class 13: 0.948667
Accuracy of class 14: 0.906905
Accuracy of class 15: 0.890667
Accuracy of class 16: 0.766000
Accuracy of class 17: 0.778286
Accuracy of class 18: 0.909619
Accuracy of class 19: 0.910000
Accuracy of class 20: 0.860365
Accuracy of class 21: 0.716364
Accuracy of class 22: 0.981111
Accuracy of class 23: 0.840905
Accuracy of class 24: 0.462381
Accuracy of class 25: 0.688000
```

Figure 7: Result of classification for classes from 38-50

6. CONCLUSION AND FUTURE WORK

In this paper, it has been shown how we can make use of different encoding methods like fisher kernel then pooling methods like max pooling and sum pooling along with the different normalization methods for the creation of Bag of Visual words. Comparison between the different feature extraction method is done in order to find out which method yield out better results with lower computational cost and higher accuracy. Experiment has been carried out on the dataset UCF50 which comprises of 50 different classes and each class comprising of 25 videos. Results shows that new encoding method which has been used with pooling and different normalization method give better results in action recognition within a video. . In this paper, it has been shown how we can make use of different encoding methods like fisher kernel then pooling methods like max pooling and sum pooling along with the different normalization methods for the creation of Bag of Visual words. Comparison between

the different feature extraction method is done in order to find out which method yield out better results with lower computational cost and higher accuracy. Experiment has been carried out on the dataset UCF50 which comprises of 50 different classes and each class comprising of 25 videos. Results shows that new encoding method which has been used with pooling and different normalization method give better results in action recognition within a video.

Future Work includes many other feature extraction method can be used where the computation part is faster. It can be applied on more challenging datasets like HMDB and Youtube where millions of videos are there in order to improve the accuracy of the classifier. In the feature extraction phase, slight changes can be done where the features can be divided into two streams: spatial and temporal and this can be fed to the classifer as an input to the classifer for better accuracy.

References

[1] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in Asian Conference on Computer Vision, pp. 572–585, Springer, 2012.

[2] I. C. Duta, J. R. Uijlings, T. A. Nguyen, K. Aizawa, A. G. Hauptmann, B. Ionescu, and N. Sebe, "Histograms of motion gradients for real-time video classification," in Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on, pp. 1–6, IEEE, 2016.

[3] X. Tang, A. Bouzerdoum, and S. L. Phung, "Video classification based on spatial gradient and optical flow descriptors," in Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on, pp. 1–8, IEEE, 2015.

[4] S. Biswas and R. V. Babu, "H. 264 compressed video classification using histogram of oriented motion vectors (homv)," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 2040–2044, IEEE, 2013.

[5] L. Zhao, D. Zhao, X. Fan, and Z. He, "Hevc compressed domain moving object detection and classfication," in Circuits and Systems (ISCAS), 2016 IEEE International Symposium on, pp. 1990–1993, IEEE, 2016.

[6] W. Lin, M.-T. Sun, H. Li, Z. Chen, W. Li, and B. Zhou, "Macroblock classification method for video applications involving motions," IEEE Transactions on Broadcasting, Vol. 58, No. 1, pp. 34–46, 2012.

[7] H.-H. Chen, "video compression tutorial," Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC, 2014.

[8] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," Computer Vision and Image Understanding, Vol. 150, pp. 109– 125, 2016.