Human activity recognition using combinatorial Deep Belief Networks

Shreyank N Gowda
Indian Institute Of Science
Bangalore, India
kini5gowda@gmail.com

Abstract

Human activity recognition is a topic undergoing a great amount of research. The main reason for that is the number of practical applications that are developed using activity recognition as the base. This paper proposes an approach to human activity recognition using a combination of deep belief networks. One network is used to obtain features from motion and to do this we propose a modified Weber descriptor. Another network is used to obtain features from images and to do this we propose the modification of the standard local binary patterns descriptor to obtain a concatenated histogram of lower dimensions. This helps to encode spatial and temporal information of various actions happening in a frame. This further helps to overcome the dimensionality problem that occurs with LBP. The features extracted are then passed onto a CNN that classifies the activity. Few standard activities are considered such as walking, sprinting, hugging etc. Results showed that the proposed algorithm gave a high level of accuracy for classification.

1. Introduction

An action is a set of movements of the human body and these movements tend to be sequential. Action recognition intuitively means recognizing a set of actions. From an outlook of computer vision, action recognition refers to learning a set of video sequences to identify the sequence of movements associated with a particular action and using the knowledge obtained to predict a future action based on the movements associated. Action recognition is a key component in many applications, namely, human-computer interaction, surveillance, video analysis etc [1-3]. Recognizing an action helps the system to summarize an entire event. For instance, action recognition is becoming an integral part of sports systems for summaries as seen in Han et al [4].

Presently, most action recognition methods can be divided into two general categories. Older methods employed handcrafted feature descriptors to represent an action. These efficient descriptors would have texture and appearance information encoded in them. Some examples of these methods are the Histogram of Optical Flow proposed by Laptev et al [5], Histogram of Oriented Gradients proposed by Dalal et al[6], Motion Boundary Histogram proposed by Dalal et al [7] etc.

Most of the recent work involves techniques trying to adapt learning-based feature representations to area of interests in human recognition. Some examples of these are the Bag of Words method proposed by Fei-Fei et al [8], model and sparse coding proposed by Lee et al [9]. Also, recently information from multi-view depth images has been under focus as shown by Yang et al [10], data-driven convolutional network has been used as an example of a deep learning method by Krizhevsky et al [11].

Some methods combine the hand crafted features method and deep learning algorithms, an example is proposed by Karpathy et al [12]. Learning based feature descriptors tend to be more advantageous than handcrafted feature descriptors. The reason is that learning based descriptors have the capability to learn additional features that cannot be encoded by handcrafted feature descriptors. Most frameworks obtain information from the video under consideration rather than the images in each frame.

We consider the possibility of utilizing features from the frames and the difference in the frames (video). On this consideration, we propose the use of a dual deep network, one for extracting features from the frame and another for extracting features from the video. We use handcrafted feature descriptors to locate interest points of the person in the video and to track the movements of those interest points. We then use deep learning to predict the action by learning features from the video without the use of handcrafted features. We use the output from each network as input to a regression network which predicts the action.

Chapter 2 talks about the related work. Chapter 3 talks about the various steps involved in the proposed algorithm. Chapter 4 shows us the experimental results and Chapter 5 gives us a brief conclusion.

2. Related Work

Since the proposed work is an improved framework of deep learning based feature representation, we discuss feature extraction by deep learning techniques as the fundamental part of our related work.

Bag of Words [8] is a model that quantifies various local features into a combined feature space. Basically, the local image features are considered as words and this is used to aid in the generation of a sparse histogram about the complete feature space by checking the frequency of each feature. The Bag of words model gives us a way to generate a descriptor by using the sparse histogram as a vector. The main limitation of the bag of words model is that spatial relationships between the patches in the image are disregarded.

Another unsupervised method that helps to find representations of higher-level features when given input data that is unlabeled is the sparse coding algorithm proposed by Lee et al [9]. During the training phase, the method learns a relatively smaller quantity of bases to correlate the input data and further minimize the object function. These bases are used for encoding of the data and then providing a feature representation. Using a sparse matrix was shown to be able to save features and reduce the space and cost needed to lose permissible information in adequate quantity.

Principal Component Analysis [13] is a method that reduces and simplifies the number of features extracted for feature representation. Principal component analysis procures the main or principal components and their respective feature scores from input data by performing characteristic decomposition of the covariance matrix. In the end, all features that contribute to a great extent are considered and the rest are removed. This leads to a huge reduction in the dimensions of the feature matrix and hence reduces the memory required and enhances the speed of the operations. We look at a non-linear method of PCA for dimensionality reduction in our proposed method.

Liu et al [14-15] proposed methods to human action recognition and grouping. First, they used the hierarchical part-wise bag-of-words representation. This encoded the local as well as the global visual prominence in the context of the body structure cue. Next, they formulated the multiple/single-view human action recognition as a MTSL (Multi-tasked Structural Learning) problem. This gave them two advantages, firstly, maintaining the consistency between action classification based on the body and action classification based on parts and secondly, discovering action-specific and action-shared feature sub-spaces that strengthened the learning process.

Karpathy et al [12] proposed the use of Convolutional Neural Networks with the purpose of processing large-scale video classification. Video classification was also shown possible by experimental evaluation of multiple approaches. Input was processed at two spatial resolutions by the algorithm and this improved the speed of execution of the Convolutional Neural Networks without a change in accuracy.

Ryoo et al [16] proposed a pooled-time representation which captures ego motion information in first-person videos. The reason to do so is to track the variation of element in each frame descriptor vectors over time. The proposed algorithm generated many pooling operators to the time series and this helped to obtain an efficient feature representation for a video.

A video feature descriptor was proposed by Wang et al [17]. They used two-stream convolutional neural networks to obtain the deeply learned feature maps and next, they extracted the feature descriptors by using feature map normalization and a trajectory based pooling method. The trajectory pooling method helps to link the feature map and the handcrafted features closely.

Recently, the development of accurate depth sensors has made it feasible to obtain 3D information in real-time. Based on the availability of depth data, an algorithm was proposed to detect objects by Shotton et al [18]. They designed an intermediate body parts representation that would map the pose estimation problem into a pixel classification problem. The algorithm was invariant to pose, clothing, body shape etc. 3D proposals of the joints were made based on confidence scores. The results of these algorithms inspired research into developing skeleton-based approaches using HMMs or Linear Dynamical Systems.

These methods however, represented motion with a set of parameters that had no physical meaning. A method was developed by Ofli et al [19] that considered an action as an ordered sequence of most informative joints (SMIJ). The disadvantage of this method was that it could not differentiate different motion about the same joint because of the coarse nature of the representation.

Depth images help to provide additional body information to differentiate activities. This has helped to motivate recent research in activity recognition. Oreifej et al [20] proposed an algorithm utilizing a histogram of an oriented 4D surface normal to capture motion and geometry cues together. To build the histogram, they created 4D projectors that quantize the 4D space and represent the various directions for the 4D normal.

In most research, the computational complexity and speed have not been considered and hence there is a difficulty in executing these for real-time practical applications. To increase this efficiency, motion features need to be obtained directly from video sequences and this will help to represent an activity in an intuitive and efficient manner.

3. Proposed Approach

3.1. Novelty in contribution

The novelty in the contribution is the design of a modified Weber descriptor that is used in the proposed approach. Further, we propose a new modification of the LBP to overcome the dimensionality problem that occurs with it. Also, the use of two networks for classifying motion-related features and static features is a concept that has not been used in the field of action recognition before, to the best of my knowledge.

3.2. Dataset

We use two datasets for training our proposed model and testing it with other recently developed algorithms.

Firstly, we use the HMDB dataset [21] and secondly, we use the Hollywood2 dataset [22]. The HMDB dataset has 51 types of actions and each action has more than 100 videos. There is a huge variation within each action video due to changes in scale, brightness, background etc.

We consider the following actions: running, walking, kissing, standing, sitting, handshakes and hugs.

Figure 1 shows a snapshot of some of the clips used from both datasets.

Figure 1. Snapshot of (a) Running (b) Hugging

3.3. Interest point selection

Interest points need to be detected on a person in order to obtain information relevant to their body movement. In that regard, we consider the use of interest points on features such as the head, neck, shoulders, elbows, wrists, spine, hips, ankles, knees and feet. An example skeletal figure with interest points marked is shown in Figure 2.

3.4. Feature Selection from motion

We used two deep belief networks, one for motion and one for the image. The combination score obtained was then used as input to a Deep regression neural network. From the motion point of view, features are extracted by using a motion based Weber Local Descriptor. These features include stride length and speed of stride etc. The original Weber Local descriptor was proposed by Chen et al [23]. This descriptor consisted of two components: the magnitude and the orientation.

The magnitude is defined in (1) and the orientation in (2).

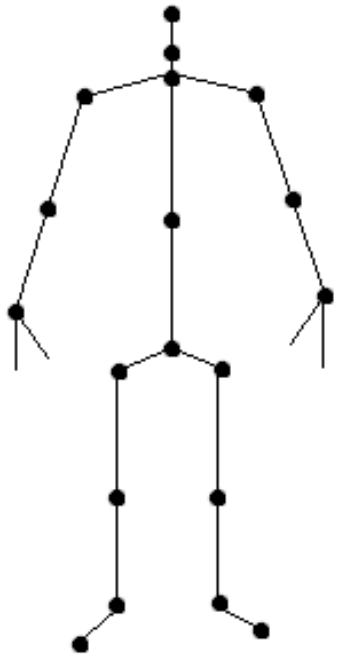

Figure 2. Skeletal image with interest points marked

$$M(x_c) = tan^{-1}(A\sum_{i=0}^{p-1}\frac{x_i - x_c}{x_c}) \quad (1)$$

The inverse of tan function is utilized to prohibit the output from being too large. This also reduces the side effect of noise. Here, x_c represents the center pixel, x_i represents the neighboring pixels and ranges in value from 0 to p-1, p is the total number of neighbors and A is a parameter used to regulate the intensity differences between neighboring pixels.

$$O(x_c) = tan^{-1}(\frac{x_1 - x_5}{x_3 - x_7}) \quad (2)$$

Here, x_1-x_5 and x_3-x_7 helps to obtain the pixel differences between neighboring pixels in horizontal and vertical directions.

According to [23] M and O are then linearly quantized into T dominant differential magnitudes and orientations respectively. We have set T as 10 for our experimental evaluations.

The WLD descriptor is illumination and invariant and also efficient in terms of computation. The WLD feature is not rotation invariant. To overcome this we rebuild the WLD histogram by considering the histograms of the neighboring regions and aggregating them. We then align the histograms to their dominant direction. The orientation is quantified into 12 dominant bins, each bin used to cover 30 degrees.

The histogram of Weber gradients from surrounding regions is taken as the local feature. This helps to deal with occlusions. The samples in each neighboring window is weighted by the Weber magnitude and the rotation is performed using the dominant orientation. Pixels are then normalized into 144 elements that are grouped as 4x4 grids. Each grid has its own histogram. This results in 16 x 12 dimensions in a feature vector.

To extract features from the frame or from a static point of view, we use a supervised deep learning method. We consider the static positions of skeletal joints or in specific the interest points from each frame occurring before time t0+v. This is stored as shown in (3).

$$x_0^{t0} = [j^{t0^T}, j^{t0+1^T}, ..., j^{t0+k-1^T}] \tag{3}$$

To obtain dynamic information such as speed and displacement variations between frames, we consider various frames and calculate the difference in positions as shown in (4) to (7).

$$x_1^{t0} = [\Delta j^{t0^T}, \Delta j^{t0+1^T}, ..., \Delta j^{t0+k-1^T}] \tag{4}$$

$$\Delta j^{t0+i} = j^{t0+i} - j^{t0} \tag{5}$$

The range of i is from 0 to k-1.

$$x_2^{t0} = [\Delta^2 j^{t0^T}, \Delta^2 j^{t0+1^T}, ..., \Delta^2 j^{t0+k-1^T}] \tag{6}$$

$$\Delta^2 j^{t0+i} = \Delta j^{t0+i} - \Delta j^{t0+i-1} \tag{7}$$

(4) and (5) help to obtain displacement related information and (6) and (7) help to obtain speed or velocity related information.

3.5. Feature selection from images

LBP has been used as a descriptor for videos for some time now. However, for instance, the VS-LBP descriptor implemented by Yeffet et al in [23] suffers from a problem with regards to memory mainly due to its dimensionality. To overcome this, we can use less temporal and spatial information. However, this will lead to a decrease in the ability to make accurate predictions.

We propose an extension of the LBP operator called the STLBP. Here, first, a video sequence is analyzed frame by frame. The people are objects of interest in the video and hence first the people are detected using a cascade of HOGs as proposed by Zhu et al in [27]. We then compute the local features of the person by training a RBF-SVM. Further, these features are used to calculate a histogram, H-STLBP. This global spatial information is further represented by spatiograms of STLBP as S-STLBP. Finally, the video is analyzed in the time domain by applying TPM to generate a set of multitemporal histograms. These histograms contain both spatial and temporal information from different subsequences. The concatenation of all these histograms forms the final depiction of the video sequence.

The LBP descriptor can be described by (8) where gc corresponds to gray level value of center pixel and gp for the local neighborhood, where p ranges from 0 to P-1.

$$LBP = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{8}$$

's(y)' is equal to 1 for every y greater than or equal to 0 and 0 otherwise.

If we consider a 8 neighborhood boundary for instance and use LBP we obtain 256 labels for a histogram. We need to reduce this. To aid us with this reduction we divide the LBP obtained into 4 parts that contain only 2 binary digits each. An example can be seen in Figure 3.

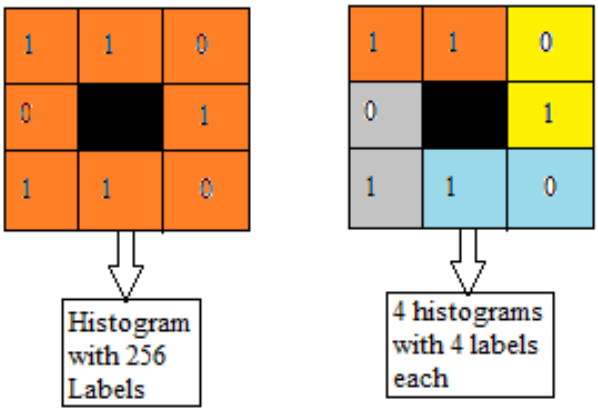

Figure 3. Comparison of histograms generated by LBP and STLBP

(9) shows us how the STLBP ends up being represented.

$$STLBP_i = \sum_{p=\frac{P}{4}(i-1)}^{\frac{P}{4}-i-1} s(g_p - g_c)2^{pmod\frac{P}{4}} \tag{9}$$

3.6. Kernel PCA and sparse encoding for dimensionality reduction

Dimensionality reduction is an important phase in any machine learning algorithm, especially in terms of increasing the speed of the algorithm by reducing the number of random variables. To aid with this, we utilise the kernel principal component analysis method. This is a non linear dimensionality reduction technique and was chosen for convenience. We use kernel PCA based on experimental results obtained. The sparse encoding further reduces features and selects only the most important ones.

3.7. Prediction Phase

Figure 4 shows the prediction outline for the proposed methodology.

The output from the two networks is passed as input to a standard CNN which classifies the output to one of the actions under consideration.

4. Experimental Results and Analysis

The proposed method was tested and compared using first the HMDB51 and then the Hollywood2 database. The proposed method was compared against the method proposed by Yue-Hei Ng et al [24], Lan et al [25], Simonyan et al [26] and Wang et al [17]. Table 1 corresponds to the accuracy comparison for the algorithms tested using HMDB51 database. The proposed method and the method in [26] have similar methodologies. Both methods use a combination of networks. However, the proposed method performs early fusion of spatial and temporal information and the [26] uses late fusion.

Method	Accuracy
CNN, optical flow, LSTM [24]	79.64
IDT, FV, temporal scale invariance [25]	77.48
CNN, IDT, FV, trajectory, SVM [17]	80.46
Two-Stream CNNs [26]	78.43
Proposed	**80.48**

Table 1. Accuracy comparison using HMDB51

Table 2 corresponds to the accuracy comparison using Hollywood 2 database.

Method	Accuracy
CNN, optical flow, LSTM [24]	84.19
IDT, FV, temporal scale invariance [25]	86.46
CNN, IDT, FV, trajectory, SVM [17]	91.18
Two-Stream CNNs [26]	89.78
Proposed	**91.21**

Table 2. Accuracy comparison using Hollywood2

Figure 3 shows the confusion matrix for the proposed method using the HMDB51.

5. Conclusion

A novel method was developed for action recognition. Two deep belief nets were developed. The first one was used to obtain features from motion and to do these we developed a modified WLD. The second one was used to obtain features from static images and for this, we developed a modified LBP that helped to eliminate the dimensionality problem associated with LBP.

The output from both networks was used as input to a CNN that classified the action into one of the seven actions selected.

The novelty in the proposed approach was the modification of the WLD and using separate deep networks to extract features from video and frames along with the modification of the LBP.

The output showed the credibility of the proposed approach and also the stronger efficiency of the algorithm in comparison to other recently developed algorithms. In terms of the accuracy, the algorithm showed results as good as any of the state of the art algorithms.

6. References

[1] Lao, W., Han, J. and De With, P.H., 2009. Automatic video-based human motion analyzer for consumer surveillance system. IEEE Transactions on Consumer Electronics, 55(2).

[2] Zhang, B., Perina, A., Li, Z., Murino, V., Liu, J. and Ji, R., 2016. Bounding multiple gaussians uncertainty with application to object tracking. International Journal of Computer Vision, 118(3), pp.364-379.

[3] Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J. and Liu, H., 2016. 3D action recognition using multi-temporal depth motion maps and fisher vector. In Proceedings of International Joint Conference on Artificial Intelligence (pp. 3331-3337).

[4] Han, J., Farin, D. and de With, P.H., 2008. Broadcast court-net sports video analysis using fast 3-D camera modeling. IEEE Transactions on Circuits and Systems for Video Technology, 18(11), pp.1628-1638.

[5] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B., 2008, June. Learning realistic human actions from movies. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.

[6] Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[7] Dalal, N., Triggs, B. and Schmid, C., 2006, May. Human detection using oriented histograms of flow and appearance. In European conference on computer vision (pp. 428-441). Springer Berlin Heidelberg.

[8] Fei-Fei, L. and Perona, P., 2005, June. A bayesian hierarchical model for learning natural scene categories. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 524-531). IEEE.

[9] Lee, H., Battle, A., Raina, R. and Ng, A.Y., 2007. Efficient sparse coding algorithms. Advances in neural information processing systems, 19, p.801.

[10] Yang, Y., Wang, X., Liu, Q., Xu, M. and Yu, L., 2015. A bundled-optimization model of multiview dense depth map synthesis for dynamic scene reconstruction. Information Sciences, 320, pp.306-319.

[11] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

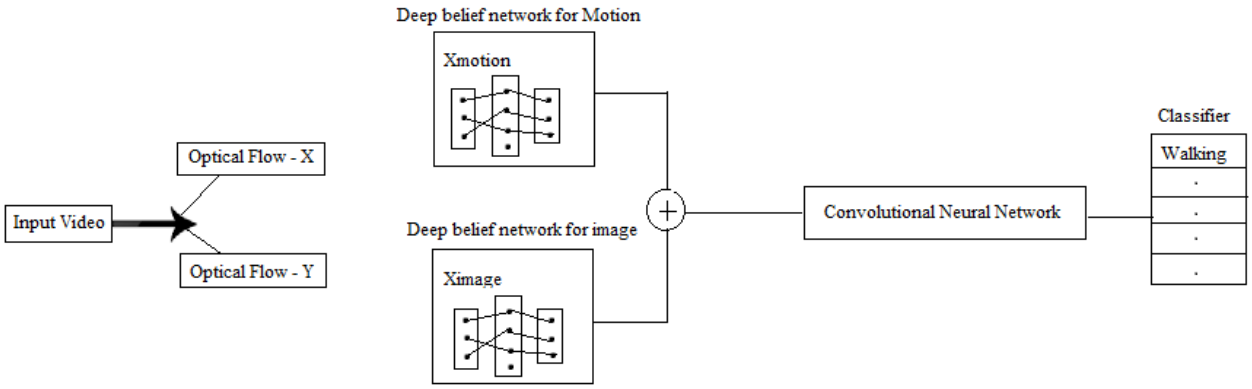

Figure 4. Pipeline of classification using Deep Belief Network

	Walking	Kissing	Hugging	Running	Standing	Sitting	Handshake
Walking	61.4	0	3.4	34.8	0	0	0.4
Kissing	0	80.1	11.4	0	0	0	8.5
Hugging	0	10.6	82.2	0	0	0	7.2
Running	36.2	0	0	63.8	0	0	0
Standing	1.6	0	0	4.8	93.6	0	0
Sitting	0	2.4	0	0	0	92.4	5.2
Handshake	0	3.4	2.3	0	1.7	2.7	89.9

Figure 5. Confusion Matrix for proposed method tested against HMDB51

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

[13] Jolliffe, I., 2002. Principal component analysis. John Wiley and Sons, Ltd.

[14] Liu, A.A., Su, Y.T., Jia, P.P., Gao, Z., Hao, T. and Yang, Z.X., 2015. Multiple/single-view human action recognition via part-induced multitask structural learning. IEEE transactions on cybernetics, 45(6), pp.1194-1208.

[15] Liu, A.A., Xu, N., Su, Y.T., Lin, H., Hao, T. and Yang, Z.X., 2015. Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing, 151, pp.544-553.

[16] Ryoo, M.S., Rothrock, B. and Matthies, L., 2015. Pooled motion features for first-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 896-904).

[17] Wang, L., Qiao, Y. and Tang, X., 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4305-4314).

[18] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. and Moore, R., 2013. Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), pp.116-124.

[19] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R. and Bajcsy, R., 2014. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. Journal of Visual Communication and Image Representation, 25(1), pp.24-38.

[20] Oreifej, O. and Liu, Z., 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 716-723).

[21] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., 2011, November. HMDB: a large video database for human motion recognition. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE.

[22] Marszalek, M., Laptev, I. and Schmid, C., 2009, June. Actions in context. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2929-2936). IEEE.

[23] Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X. and Gao, W., 2010. WLD: A robust local image descriptor. IEEE transactions on pattern analysis and machine intelligence, 32(9), pp.1705-1720.

[24] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G., 2015. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).

[25] Lan, Z., Lin, M., Li, X., Hauptmann, A.G. and Raj, B., 2015. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 204-212).

[26] Simonyan, K. and Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).