Still Image-based Human Activity Recognition with Deep Representations and Residual Learning

Ahsan Raza Siyal[1], Zuhaibuddin Bhutto*[2], Syed Muhammad Shehram Shah[3], Azhar Iqbal[4], Faraz Mehmood[5], Ayaz Hussain[6], Saleem Ahmed[7]

Department of Electronic Engineering, Dawood University of Engg. & Technology, Pakistan[1]
Department of Computer System Engineering, Balochistan University of Engg. & Technology, Pakistan[2]
Department of Software Engineering, Mehran University of Engineering & Technology, Pakistan[3]
Deptartment of Basic Sciences, Dawood University of Engineering & Technology, Pakistan[4,5]
College of Information and Communication Engineering, SungkyunKwan University, Suwon, Republic of Korea[6]
Department of Computer System Engineering, Dawood University of Engg. & Technology, Pakistan[7]

***Abstract*—Iterative Recognizing human activity in a scene is still a challenging and an important research area in the field of computer vision due to its various possible implementations on many fields including autonomous driving, bio medical, machine intelligent vision etc. Recently deep learning techniques have emerged and successfully deployed models for image recognition and classification, object detection, and speech recognition. Due to promising results the state of art deep learning techniques have replaced the traditional techniques. In this paper, a novel method is presented for human activity recognition based on pre-trained Convolutional Neural Network (CNN) model utilized as feature extractor and deep representations are followed by Support Vector Machine (SVM) classifier for action recognition. It has been observed that previously learnt CNN knowledge from large scale data-set could be transferred to activity recognition task with limited training data. The proposed method is evaluated on publicly available stanford40 human action data-set, which includes 40 classes of actions and 9532 images. The comparative experiment results show that proposed method achieves better performance over conventional methods in term of accuracy and computational power.**

Keywords*—Human activity recognition; action recognition; deep learning; transfer learning; residual learning

I. INTRODUCTION

Over the recent decade, the human activity recognition has been a highlighted topic for researchers because of its various applications which include video surveillance, Human-machine interaction, ambient-assisted living, smart system design and autonomous driving. Automatic classification of an action in a given scene is a challenging and critical task. There are two main approaches for human activity recognition; which includes traditional handcrafted features representations and a deep learning approach or deep representations. The learning-based approach introduce concept of classification by trained feature extractor followed by a state-of-the-art classifier. The deep learning approach have made remarkable growth in activity recognition task.

The deep learning model in [1] has introduced multi-stream 3D CNN for limited learning data. The author in [2] utilizes integrating body pose, part shape, and motion data for activity recognition task. The process of training the deep neural network from the scratch involves huge amount of trainable data and learning this much of parameters require high computational resources and hours or days for training. Training the model for real-world applications, collecting annotated hug amount of specific task related data is very time consuming and costly [3]. Therefore, accumulation of sufficient task related learning data may not be feasible choice in many cases [4]. It is challenging to produce adequate results by applying deep learning methods. For mitigating this issue researchers reviewed their approach for implementation of deep learning models on smaller data-sets and this makes them relating the problem to human vision system. We humans learn several categories in our lives just from few samples and this capability is achieved by accumulating previously learnt knowledge over the period of time and transferring it for learning the new task [5]. Researchers came to the conclusion that previously learnt knowledge contribution in learning new tasks through connection and similarity between new task and the old one can produce significant improvement in efficiency of methods. By this idea studies suggest pre-trained models for classification can be utilized to classify new classification task [6]. Hence, the CNN models trained to classify certain object can be fine-tuned for new task even in different domain [7].

In this paper, human activity recognition method is proposed which is based on pre-trained Convolutional Neural Network (CNN) model utilized as feature extractor and deep representations are followed by Support Vector Machine (SVM) classifier for action recognition. The results show that proposed method can produce significant performance results for human activity recognition.

The rest of the research paper is outlined as follows. The Learning methods are explained in Section II. Then Section III details about the related work. The Section IV methodology applied for the proposed approach. The experimental results are illustrated and explained in Section V. Finally, paper is concluded in conclusion section.

*Coresponding Author

II. Related Work

This section of the paper discusses existing state-of-the-art methods for human activity recognition using both approaches of handcrafted techniques and deep representation learning. The methods that use handcrafted techniques such as extended SURF [14], HOG-3D [15]. The techniques which utilizes motion-based feature descriptors such as exploiting motion in [16], using gaze [17], by improved trajectories [18], from multiple views based on view-invariant feature descriptor [19], have achieved notable performance for classifying human activities. However, these methods have some limitations such as, requirement of proficient intended feature detector and descriptors for feature extraction. This process requires skilled manpower, consumes time and it is a cost inefficient method.

For all these reasons, researchers prefer deep learning approaches for human activity recognition. This approach has been used for various domains in the recent past such as image classification, object detection, speech recognition. Same approach has also been explored for human activity recognition. Some of the contributions are as using Global spatial-temporal attention [20], based on skeleton data [21], 3D ConvNets [22], learning Spatio-Temporal with 3DConvNets [23].

Human activity recognition based on videos has always been highlighted research topic for researchers and they have contributed a lot in the last decade. Image-based action recognition models have been developed and evaluated for efficient action recognition. Researchers have contributed to resolve issues related to accuracy improvement and less computational power requirements.

Some researchers have used transfer learning in cross-domain for human action recognition to improve accuracy and performance of the model such as cross-domain knowledge transfer was carried out in [24]. Based on human poses human part-detectors can be employed to detect different human parts in the scene and then these parts are encoded into poses for human activity classification [25]. In [26], author employs trained neural network to perform pose estimation. Additionally, it transfers previously learnt knowledge to target model in order to perform training of new task has shown improved accuracy of the model and it saves time and money. Furthermore, research work has been reported reported in [28-31].

III. Learning Methods

A. Residual Learning

Residual learning is a machine learning technique in which the network has stack of layers. Let us assume $H(x)$ contained by a neural network block, where x denotes the input parameter. The difference between the true distribution $H(x)$ and input sample x is given as $x = H(x) - x$. When we rearrange it we will get $H(x) = F(x) + x$. Both equations will be approaching a value or curve which has certain limit called asymptote line but the way of understanding for both equations could be different. If all the added layers are similar or they return the identical value, then error can occur which will not be greater than shallower part. If the system has nonlinear layers, then it can create issue in order to find the approximate value in identity mapping layers. Another solution for this might be that if identical mappings are good to choose then the weights of nonlinear layers will assume towards value zero to get the result of identical mappings by multilayer mappings.

But in actual scenario the identical mappings are not most favorable, so we have to reformulate the process to get the accurate result. It can also be solved when identical mapping has the value closer to zero than it will be easy to solve. Experiments show that residual functions have small and simple responses to provide reasonable approach to solution.

B. Identity Maping using Shortcuts

As we know that residual learning has number of layers. It can be written in equation as:

$$y = F\left(x, \{w_i\}\right) + x \tag{1}$$

here x and y are the input and output, respectively, and $F\left(x, \{w_i\}\right)$ represents residual learning. In Fig. 2, it has been shown that the two layers in which $F = W2\sigma\left(w_i x\right)$ here $\sigma =$ rectified linear unit. $F + x$ perform element wise addition. This is also a short cut connection which will neither consider as an extra element in the equation nor it will create a difficulty in calculation process. Plain or residual networks have same number of parameters used for example computational cost, width of depth etc. so we can easily compare them. In above equation x (input) and F must have same dimensions if the dimensions are not same then we are able to calculate another kind of projection called as liner a projection

$$y = F\left(x, \{w_i\}\right) + W_{sx} \tag{2}$$

Experiments shows that to label the degradation problem the identical mapping is enough, but we can also use square matrix. F (residual function) has two layers and its value is adaptable but more layers are also possible. If F (residual function) has single layers, then it's called as linear layer. We have seen that we have fully occupied layers to keep it simple that is applicable to the first layer also called as convolution layer. Multiple convolutional layers can be shown by the function $\left(x, \{w_i\}\right)$. This function will also use to perform element wise addition which includes two featured maps and work as channel by channel one by one.

C. Transfer Learning

The transfer learning approach tries to utilize the previously learnt knowledge to solve another problem which may have different domain. To initialize the process notation is used which is introduced by Pan and Yang (2010). Notation

has two major components, domain D which is learnable data-set having given probability distributions, data may be images with different resolutions and pixel values, and other component is task T which may be defined as target function or labels. Labels are the marking of the image which provides the class or category which assist to learn them accordingly.

Transfer learning is the deep learning technique to set a deep neural network using features learnt for a source problem (TS, DS) and the same network can be fine-tuned and employed for target task (TT, DT). This approach has been introduced to generate more accurate model than training a model from scratch. Another utilization of the pre-trained neural network for TS is as a feature extractor for TT and the extracted features can be utilized on training another machine learning method to optimize accuracy. By utilizing this approach, one will represent DT data learnt for TS task by use of pre-trained network representations on small data which is inadequate to train these methods.

Transfer learning is one of the approaches that utilizes pretrained models as feature extractor by replacing fully connected layer and extracting feature data from last pooling layer of the deep neural network which may be followed by a generic state-of-the-art SVM classifier [8], this approach has been employed on many classification and recognition tasks [9]. In this paper our proposed model also falls under the same approach. We evaluate recently benchmark models like GoogleNet [10], VGG-16 , VGG-19 [11] and ResNet-18 [12] on stanford40 [13] data-set, based on performance in terms of accuracy and learning validation we selected ResNet-18 as a source model for generating a target model to classify action in a given image. ResNet-18 is used to extract features from input image and deep representations are followed by SVM classifier for action class recognition.

IV. Methodology

Transfer learning in a machine learning approaches utilizes previously learnt knowledge to train a model for new task with less computational requirements as compare to train the model from scratch. Transfer learning based on pre-trained CNN are useful in training the model with smaller data-sets. However, CNN are prone to overfitting with limited data-set but it can be avoided by increasing the training data on the expenses of increased cost and time-consuming process. For these reasons transfer learning is very convenient way to train a model with help of pre-trained deep representations as source architecture for creating new architecture in order to perform a new task. In the proposed method, we have evaluated publicly available popular pre-trained networks including GoogleNet, VGG16, VGG19 and ResNet-18 and selected the pre-trained model ResNet-18 on bases on performance on the stanford40 data-set for human action recognition problem. The ResNet-18 has been trained on a million of images of 1000 different categories of ImageNet data-set [27]. ResNet-18 architecture consists of 18 layers and having input layer of size 224x224x3. This pre-trained CNN has the ability to categorize 1000 different classes like pencil, mouse, cat, dog, keyboard and many more. For the reason of extensive learning of deep representations of various classes of images, it is very useful to transfer this learnt knowledge to classify human action recognition.

In this paper, pre-trained CNN Reset-18 model is used as a feature extractor, input image is augmented with a size of ResNet-18 first layer (224x224x3). The architecture consists of 17 convolutional layers C1 to C17 and one fully connected layer (fc1000). Features are collected from the last pooling layer in the case of ResNet-18 ‘Pool_5’ layer, and with these deep representations state-of-the-art classifier is trained to classify action in the given image. Fig. 1 shows the block diagram of proposed method.

A. CNN Architecture for Feature Extraction

Deep CNN (DCNN) is one of the general classes of deep neural networks (DNN). In past years, DCNN has contributed in a large scale to the computer vision field by surpassing the performance of machine learning algorithms. However, problems come up whenever network become widen. Many researchers have quoted that the accuracy becomes compromised turning into saturated. There is also a case of test errors at the time of training deep networks consisting several layers. The DCNNs are also vulnerable to the problems of vanishing gradient. In that case, a minor gradient prevents the layer weights to be updated.

Whenever this problem comes, the training of the deeper layers of the network is inefficient due to slow training of deeper layers. Therefore, residual learning can be used to counter this problem.

In case of residual learning, there is a training of the network of features are conducted instead of features. ResNet is a DCNN model which is considered state-of-the-art; it carries a concept of residual learning. In its architecture, ResNet uses an alternative connection through connecting a n_{th} layer to the input with a $n+i$ layer. The architecture of ResNet consists of many residual building blocks. Let’s consider the input of residual block to be $x_i - 1$ and the output is x_i, after attempting several operations like, convolution, ReLU activation and batch normalization on input, the output is $f(x_i - 1)$. looking residual learning, it can be defining $f(x_i - 1) = x_i - x_i - 1$. However, $x_i = f(x_i - 1) + x_i + 1$ is obtained. Through this method, the information of previous layer is enabled to be added is the current layer. In Fig. 2, a basic building block of ResNet is shown. Many variations of ResNet can be found. The evaluation of knowing accuracy using ResNet-18 is presented in this work, which consist of 17 layers of convolutional and one fully connected layer. Through residual learning, the layers have been stacked up one after another. Every convolutional layer in the block of residual is followed by its connected ReLU layer and batch normalization layer. Max pooling layer is used at the end of first residual block. An average pooling layer is used after the fifth residual block which is later on connected with the FC layer. The size of the FC layer is equivalent to the class number.

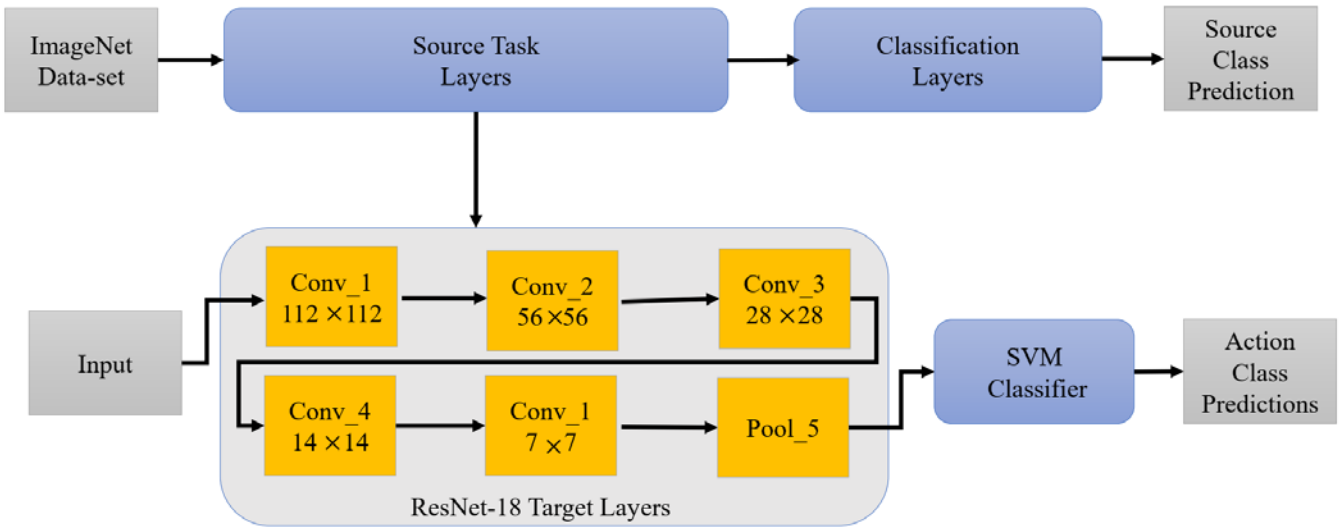

Fig 1. Block Diagram of Proposed Method.

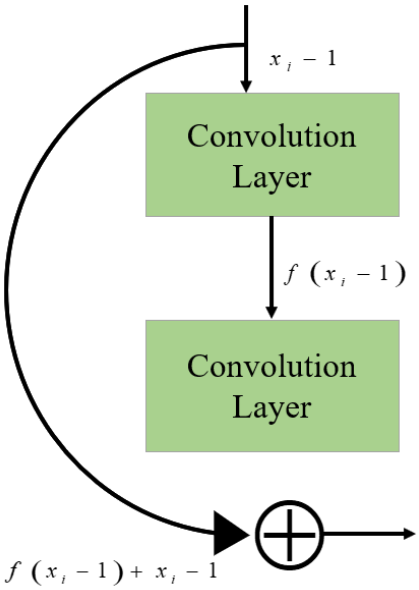

Fig 2. Resnet Architecture: Basic Building Block of Residual Learning.

B. Classification

Classification is a method of designating an unknown sample to already defined class or it is based on the data which is used to help in understanding the program. When there are different classifiers Support Vector Machine (SVM) is one of the classifiers that uses classification algorithm with two groups of data. Tasks related to pattern recognition SVM is considered as the popular one. Basically, SVM operates on machine learning algorithm. In this method there is a n-dimensional space. The data items are represented by x_i and y_i, each one has its own representation that is x_i represents the attribute of sample and y_i denotes the class usually it has positive or negative value. In SVM the classification is found by using a line called as hyper plane which divides the plane in to two parts and the classes on lie on the either sides. The cost function $\left(\frac{1}{2}W^T W\right)$ is kept maximum to find out the hyper plane. This can be limited by an equation as:

$$y_i = \begin{cases} +1 & if\ w \text{ x } x_i + b \geq 1., \\ -1 & \text{otherwise,} \end{cases} \quad (3)$$

In this equation w represents the weight factor and b is the distance between the hyper planes to the origin which is called as bias. Since we are interacting with 10 classes of human actions this is definitely a problem which has not only one class but has more than one class. The multi class problems can be used to solve the problems occur in binary class. The equation shown above is having the contents which are made by combining different binary class SVMs to get a multi class SVM. The two other kinds of programming which can be used are one-vs-one and one-vs-all methods.

Now we talk about One-vs–all method. In this method there are n numbers of binary classifiers each of them recognizing a particular class. The c is using for c-class problems and ith classifier is used to create a boundary between other classifiers. There is another method called as winner-takes-all which is used to assign value to a class the value is unknown sample usually a negative value is also be accepted.

On the other hand one-vs-one method is also called as binary classifier which contains two samples i and j they can be assigned when we pick a positive sample of class i and negative sample of class j.In this method an approach for classification is used called as max-wins voting in which each classifier gives the value of one or two classes the vote will be in the favor of assigned class and increased by one and an unknown sample is also assigned to a class which has largest vote.one-vs-one method provides better accuracy that's why

SVM is using this problem. However, the initial of one-vs-one method has greater overhead than multi class classifier.

V. Experimentation and Results

In this section we discuss the experiment setup in terms of preprocessing, learning process and evaluated proposed method observations and results. The proposed method is tested on publicly available Stanford 40 human action data-set. It includes 40 different classes of daily life human actions like phoning, walking, jumping and more. Each action class have approximately 180 to 300 images of bounding box of the action performing person, some sample images from the data-set with their corresponding labels are shown in Fig. 3. For experimental purpose only 10 classes are used to evaluate the performance of four pre-trained models on the data-set , Images in the data-set are different in size to make these images compatible with input layer of pre-trained models, images are gone through augmentation process as a preprocessing followed by learning implementation and based on the performance of Resnet-18 it is selected as source architecture for feature extraction for the proposed method, Table I shows class-wise accuracy of the pre-trained models and Fig. 4 shows overall accuracy on 10 classes of the data-set.

In the proposed method, based on accuracy on the data-set we selected Resnet-18 as source architecture for feature extraction, Resnet-18 has input layer of size 224 by 224 and images in the data-set have different sizes, images are augmented to the size of input layer of pre-trained Resnet-18. The architecture comprises of 17 convolutional layers and only one fully connected layer. The proposed method utilizes the architecture for feature extraction there for deep representations are extracted from last pooling layer which is 'pool5' in case of Resnet-18. These deep representations and then followed by state-of-the-art SVM classifier for predicting action class in a given image, the proposed method achieves 87.22% accuracy on the data-set. Fig. 5 shows the classified actions from the test data-set with predicted labels and confidence score and Fig. 6 shows the confusion matrix.

TABLE I. Comparison of Classification Results on Stanford Data-Set

S. No	Action	ResNet-18 [12]	VGG-16 [11]	VGG-19 [11]	GoogLeNet [10]
1	Applauding	71.19	64.41	64.41	64.41
2	Brushing	70	70	43.33	56.67
3	Cleaning	85.71	93.65	96.83	89.83
4	Climbing	98.31	94.92	91.53	91.53
5	Cutting trees	88.33	95	95	93.33
6	Cooking	92.94	92.94	91.76	96.47
7	Jumping	96.55	96.55	93.1	90.8
8	Phoning	72.73	77.92	79.22	70.13
9	Playing guitar	91.86	80.23	84.88	94.19
10	Riding bike	96.51	91.86	96.51	96.51

Fig 3. Shows Sample Images from Stanford40 Data-Set with Labels.

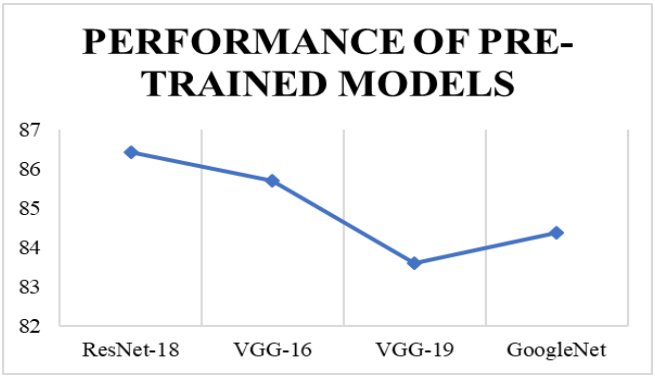

Fig 4. Overall Accuracy of Pre-Trained Models on Stanford40 Data-Set.

Fig 5. Classified Actions on Test Data-Set with Predicted Labels and Confidence Score.

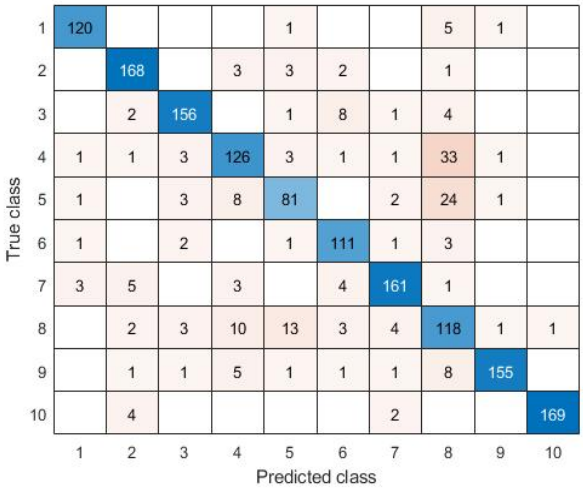

Fig 6. Confusion Matrix of Predicted Classes.

VI. Conclusion and Future Work

In this paper a novel method has evaluated based on transfer learning of deep representations by using pre-trained Resnet-18 convolutional neural network as source architecture for feature extraction and state-of-the-art SVM classifier is trained to classify target data-set. It has been established that by using transfer learning technique previously learnt knowledge can be utilized to learn new task with limited data size. Transfer learning comes very handy when the data-set is not adequate for training the deep model from the scratch and also training the deep model from the scratch on very large amount of data requires computational resources and it is very time costly way which can be avoided using transfer learning. In addition to that it has been noted that SVM as a classifier performs better than a convolutional neural network and moreover, handcrafted representation-based methods require preprocessing and manual feature extraction, the proposed method eliminate these requirements as it directly accepts RGB images as input and extract features from them. The performance of the proposed method is evaluated on open source stanford40 human action data-set, it achieves 87.22% accuracy.

Some future directions in the human activity recognition and classification research are given as:

A. Utilizing Image-based Models for Other Area of Research

Deep learning network has been emerged and proven its superiority over other traditional methods in many areas of research and similarly in computer vision field. However, video-based models are training with the complexity and difficult implementation, thus benefiting from pre-trained models on images would be better solution to explore. In addition, image-based models have done a better job on capturing spatial relationships of objects which might be utilized in action recognition. These image-based models can be explored for medical image processing, disease detection and classification.

B. Interpretability on Temporal Extent

All the frames in the video are not equally significant for activity recognition, few of the frames are critical and required to learn deep representations of temporal interpretability of video-based models. Above all else, activities, particularly long-length activities can be considered as a sequence of primitives. It is intriguing to have an interpretability of these primitives for example, how are these primitives sorted out in the temporal area in activities, how would they add to the arrangement task, can we just utilize not many of them without sacrificing recognition performance in order to achieve fast training and less computation.

C. Complexity Reduction Techniques

Learning deep representations is very complex and specially dealing with dataset having high dimension would require high computational power and time. It is commonly helpful to reduce data dimension not only for reason of computational efficiency but also improve accuracy of analysis. These dimensions reduction technique can be apportioned into two significant manners, they can be separated as techniques that can be utilized for supervised or non-supervised learning and into techniques that either entail feature selection or feature extraction.

References

[1] V. A. Chenarlogh, F. Razzazi “Multi-stream 3D CNN structure for humanaction recognition trained by limited data”, IET Computer Vision, Vol. 13 Issue 3, pp. 338-344, 2019.

[2] H. El-Ghaish, M. Hossain, A. Shoukry, And R. Onai, ”Human Action Recognition Based onIntegrating Body Pose, Part Shape, andMotion”,IEEE Access, vol.6, pp. 49040 – 49055, 2018

[3] Hossein Rahmani, Ajmal Mian and Mubarak Shah, “Learning a DeepModel for Human ActionRecognition fromNovel Viewpoints”, IEEE Transactions on Pattern Analysis and Machine Intelligence ,Vol. 40 , Issue: 3 , pp. 667 – 681, 2017.

[4] Cao, X., Wang, Z., Yan, P., Li, X., “Transfer learning forpedestrian detection”, Neurocomputing100, p. 51-57, 2013.

[5] F. Fei, L. “Knowledge transfer in learning to recognize visual objects classes”, International Conferenceon Development and Learning (ICDL). 2006.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, “Return of the devil inthe details: Delving deep into convolutional nets”, arXiv preprintarXiv:1405.3531, 2014.

[7] H. Nam, and B. Han, “Learning multi-domain convolutional neural networks for visual tracking”, arXiv preprintarXiv:1510.07945, 2015.

[8] M. D. Zeiler, and R. Fergus. “Visualizing and understanding convolutional networks”, European Conference on Computer Vision, Springer, 2014.

[9] H. Azizpour, S. A. Razavian, J. Sullivan, “From generic to specific deep representations for visual recognition”, IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015.

[10] S. Christian, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions", IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015

[11] S. Karen, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556 (2014).

[12] H. Kaiming, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition", IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[13] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and

Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain, November 6-13, 2011.

[14] G. Willems, T. Tuytelaars, and L. V. Gool. "An efficient dense and scale-invariant spatio-temporal interest point detector" European conference on computer vision. , Springer 2008.

[15] A. M. Klaser, Marsza, and C. Schmid. "A spatio-temporal descriptor based on 3d-gradients", BMVC-19th British Machine Vision Conference, 2008.

[16] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition", IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[17] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize dailyactions using gaze", European Conference on Computer Vision, Springer, 2012.

[18] H. Wang, and C. Schmid, "Action recognition with improved trajectories", IEEE International Conference on Computer Vision. 2013.

[19] A. B. Sargano, P. Angelov, and Z. Habib, "Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines", Applied Sciences, vol. 6, issue 10, pp. 309, 2016.

[20] Y. Han , S. L. Chung , A. M. Ambikapathi ,W.Y. Lin, S.Su, "Robust Human Action Recognition Using Global Spatial-Temporal Attention for Human Skeleton Data", International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8-13 July 2018.

[21] W. Cho, T. Z. Win, M. T. A. Win. "Human Action Recognition System based on Skeleton Data", IEEE International Conference on Agents (ICA), Singapore, 28-31 July 2018.

[22] S. Ji, W., Yang, M. Yu, "3D convolutional neural networks for human action recognition", IEEE transactions on pattern analysis and machine intelligence, vol.35 issue 1: p. 221-231.2013

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, "Learning spatio temporal features with 3D convolutional networks", IEEE International Conference on Computer Vision (ICCV).2015.

[24] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection", IEEE conference on Computer vision and pattern recognition (CVPR), San Francisco USA, June 2010.

[25] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance", IEEE Conference on Computer Vision and Pattern Recognition, pp. 3177–3184. 2011.

[26] J. Tompson, R. Goroshin, A. Jain, Y. L. Cun, and C. Bregler, "Efficient object localization using convolutional networks", IEEE Conference. on Computer Vision and Pattern Recognition, pp. 648–656, 2015.

[27] J. Deng, W. Dong, *et at*., "ImageNet: A Large-Scale Hierarchical Image Database", IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 2009.

[28] Z Bhutto, *et at.*, "Scaling of Color Fusion in Stitching Image", International Journal of Computer Science and Network Security, Vol. 19, No. 4, pp. 61-64. April 2019.

[29] A. S. Chan, K. Saleem, Z. Bhutto, *et at.*, "Feature Fusion Based Human Action Recognition in Still Images", International Journal of Computer Science and Network Security, Vol. 19, No. 11, pp. 151-155, November 2019.

[30] N. K. Baloch, Z. Bhutto, *et at.*, "Finger-vein Image Dual Contrast Enhancement and Edge Detection", International Journal of Computer Science and Network Security, Vol. 19, No. 11, pp. 184-192, November 2019.

[31] A. Siyal, Z. Bhutto, *et at*., "Ship Detection in Satellite Imagery by Multiple Classifier Network", International Journal of Computer Science and Network Security, Vol. 19, No. 8, pp. 142-148. August 2019.