Human Action Recognition using Transfer Learning with Deep Representations

Allah Bux Sargano [1,3,*], Xiaofeng Wang [2], Plamen Angelov [1], and Zulfiqar Habib [3]
[1] School of Computing and Communications Infolab21, Lancaster University, Lancaster LA1 4WA, UK;
[2] School of Computing and Communications, Northwest University, Xi'an, China;
[3] Department of Computer Science, COMSATS Institute of Information Technology, Lahore 54000, Pakistan;
[*] Correspondence: a.bux@lancaster.ac.uk

Abstract— Human action recognition is an important research area which has captured lot of attention from the research community due to its significant applications. Recently, due to the popularity and successful implementation of deep learning-based methods for image analysis, object recognition, and speech recognition. Researchers are motivated to shift from traditional feature-based approach to deep learning. This research work presents an innovative method for human action recognition using pre-trained Convolutional Neural Networks (CNNs) model as a source architecture for extracting features from the target dataset, followed by a hybrid Support Vector Machines and K-Nearest Neighbor (SVM-KNN) classifier for action classification. It has been observed that already learnt CNN based representations on large-scale annotated dataset are successfully transferable to action recognition task with limited training dataset. The proposed method is evaluated on two well-known action datasets, i.e., UCF sports and KTH. The comparative analysis suggests that the proposed method is better than handcrafted feature-based methods in terms of accuracy.

Keywords—action recognition; deep learning; transfer learning; hybrid classifier

I. INTRODUCTION

Over the last decade, researchers have been paying much attention towards human action recognition because of its numerous applications. These applications include: Human-Computer Interaction (HCI), video surveillance, Ambient-Assisted Living (AAL), entertainment, and intelligent driving [1, 2]. There are two major approaches for activity recognition; these include the traditional handcrafted feature-based representation, and learning-based representation. The learning-based representation, and in particular, the deep learning, introduced the concept of end-to-end learning by using the trainable feature extractor followed by a trainable classifier [3, 4]. The deep leaning based approaches have revealed the remarkable progress for action recognition in videos. The deep learning model introduced in [5] for reducing the dimensionality of the data, CNN [6] and Deep Belief Networks (DBNs) [7] have been widely used for image classification, object recognition, and action recognition.

However, training a new deep learning model from scratch requires huge amount of data, high computational resources, and hours, in some cases, days of training. In real-world applications, collecting and annotating huge amount of domain-specific data is time consuming and expensive. Hence, collecting the sufficient amount of domain-specific data may not be a viable option in many cases [8, 9], which makes it a quite challenging to apply deep learning models. For combating this challenge, researchers revisited their strategies for visual categorization to make them in-line with the working of the human vision system. Humans have capability to learn thousands of categories in their lives from just from few samples. It is believed that humans achieve this capability by accumulating the knowledge over the time period and transfer it for learning the new objects [10]. Researchers are convinced that, the knowledge of previous objects, assist in learning the new objects through their similarity and connection with the new objects. Based on this idea, some studies suggest that the deep learning models trained for a classification task, can be employed for new classification task [11-13]. Thus, the CNN models trained on a specific dataset or task can be fine-tuned for a new task even in a different domain [14-16]. This concept is known as transfer learning or domain adaptation.

The transfer learning has been studied as a machine learning technique since long time, for solving the different visual categorization problems. In recent years, due to explosion of information such as images, audios, and videos over the internet, demands for high accuracies, and computational efficiencies are increased. Due to these reasons, the transfer learning has attracted a lot of interests in the areas of machine learning and computer vision. When the traditional machine learning techniques have reached their limits, the transfer learning unlocks new flow of streams for visual categorization. It has primarily changed the approach, the way machines used to learn and treat the classification tasks. It has been applied successfully for visual categorization tasks in the domains of object recognition, image classification and human action recognition [17].

The transfer learning mainly employs two approaches: 1) preserving the original pre-trained network and updating the weights based on the new training dataset. 2) using pre-trained network for feature extraction, and representation followed by a generic classifier such as SVM for classification [18]. The second approach has been successfully applied for many recognition and classification tasks [11, 19]. Our proposed technique for human action recognition also falls under the second category. We investigated the recently proposed benchmark deep models such as AlexNet [20], and GoogleNet [21]. Based on the experimentations, we selected the AlexNet as source model for building a target model for the action recognition task. The source model has been used for feature extraction and representation followed by a hybrid (SVM-KNN) classifier for action recognition. The arrangement of the remaining sections is as follows: related work is presented in section II, methodology is elaborated in section III, and experimentation results, and conclusion are presented in section IV and V respectively.

II. RELATED WORK

This section discusses the literature review on existing methods for action recognition using handcrafted based representations and deep learning. The action recognition using handcrafted features descriptors such as extended SURF [22], HOG-3D [23], and some other shape and motion based features descriptors [24-28] have achieved remarkable performance for human action recognition. However, these approaches have several limitations: Handcrafted feature-based techniques require expert designed feature detectors, descriptors, and vocabulary building methods for feature extraction and representation. This feature engineering process is labor-intensive and requires expertise of the subject matter.

Due to these limitations, more research is directed to deep learning-based approach. This approach has been used in several domains such as image classification, speech recognition, and object recognition, just to name few [29]. These models have also been explored for human activity recognition. Some prominent contributions like 3D ConvNets [30], Convolutional RBMs [31], learning spatio-temporal with 3D ConvNets [32], Deep ConvNets [33], and Two-stream ConvNets [34] have achieved remarkable results. On-line deep learning is also getting more attention and some researchers have proposed action recognition using on-line deep learning approach [35]. In [36], a human action recognition method was proposed using unsupervised on-line deep learning technique. This method achieved accuracy of 89.86%, and 88.5% on KTH and UCF sports dataset respectively.

The handcrafted feature-based techniques, in particular, trajectory based methods have less discriminative power. Conversely, deep network architectures are inefficient in capturing the salient motion. For addressing this issue, [37] combined the deep convolutional networks with trajectory for action recognition. However, deep learning-based methods also have some limitations, these models require huge dataset for training, and collecting huge amount of domain-specific data is time consuming and expensive. Therefore, training the deep learning model from scratch is not feasible for domain-specific problems. This problem can be solved using pre-trained network as a source architecture for training the target model with small dataset, known as using transfer learning [18].

Fortunately, the winner models of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) such as AlexNet [20], GoogleNet [21], and ResNet [38] are publicly available as pre-trained networks. These networks can be used for transfer learning. One of the important ways to employ the existing models for new task is to use pre-trained models as feature extraction machine and combine this deep representation with off-the-shelf classifiers for action recognition [11].

Some researchers have also used cross-domain knowledge transfer for action recognition. In [39], the cross-domain knowledge transfer was performed between the KTH, TRECVID [40] and Microsoft research action dataset. The TRECVID and Microsoft research action datasets were used as a source domain while KTH was used a target domain. In addition to this, some researchers have used cross-view knowledge transfer, which is a special form of cross-domain knowledge transfer for multi-view action recognition.

III. METHODLOGY

In machine learning, utilizing the previously learnt knowledge for solving a new task is known as transfer learning or knowledge transfer [41]. The transfer learning using deep CNNs is very helpful for training the model with limited size dataset, because CNNs are prone to overfitting with small dataset. However, the overfitting can be avoided by increasing the size of the training data, but it is very difficult and expensive to provide the large amount of annotated data. In this situation, the transfer learning comes handy and solves this problem by using the pre-trained deep representation as a source architecture for building the new architecture [42]. In this work, we have employed the AlexNet [20] as a source architecture for solving human action recognition problem. The AlexNet was trained on ImageNet dataset and takes as input 224 x 224 pixels RGB image and categories it into the respected class. This architecture consists of five convolutional layers from C1-C5 and three fully connected layers Fc6-Fc8 as shown in the top row of the Fig. 1.

However, this architecture contains 60 million parameters, learning this much parameters for small training dataset of the new task is problematic and time consuming. Therefore, we have used source architecture as a feature extractor followed by an off-the-shelf hybrid SVM-KNN classifier for action recognition. The value of *'K'* in the nearest neighbor algorithm is selected through cross validation. The proposed work is innovative and presents an interesting combination of deep learning and hybrid classifier, which results in boosting the performance of the human recognition method. The experimentation results confirm the efficiency of the proposed work. Moreover, our experiments confirm that, a hybrid classifier has advantage over single classifier in boosting the accuracy of the classification system. The block diagram of the proposed methodology is shown in Fig. 1, and hybrid classification model based on SVM-KNN is presented in Fig. 2

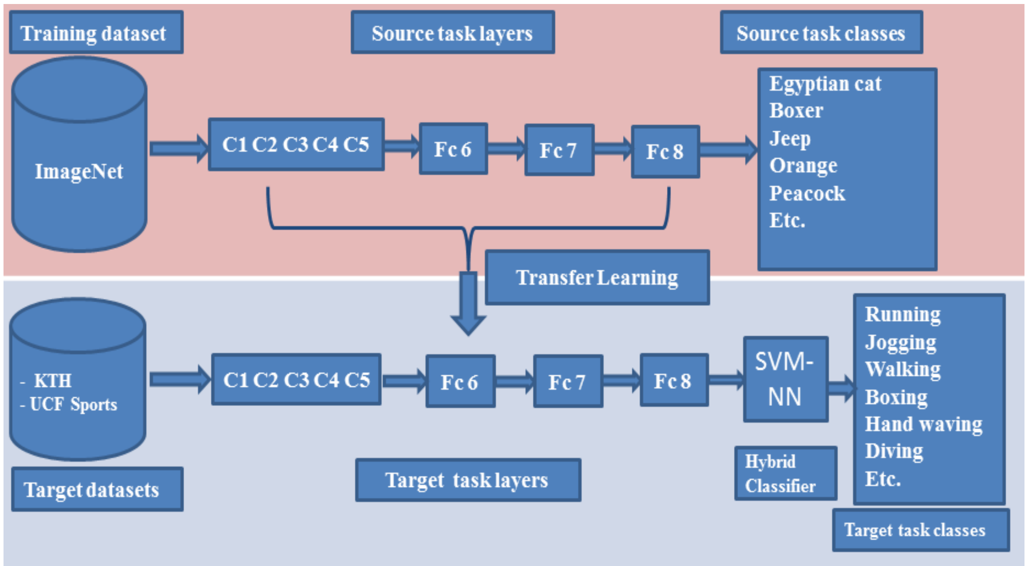

Fig. 1. Overview of the proposed system, first row indicates the source architecture and second row shows the target architecture.

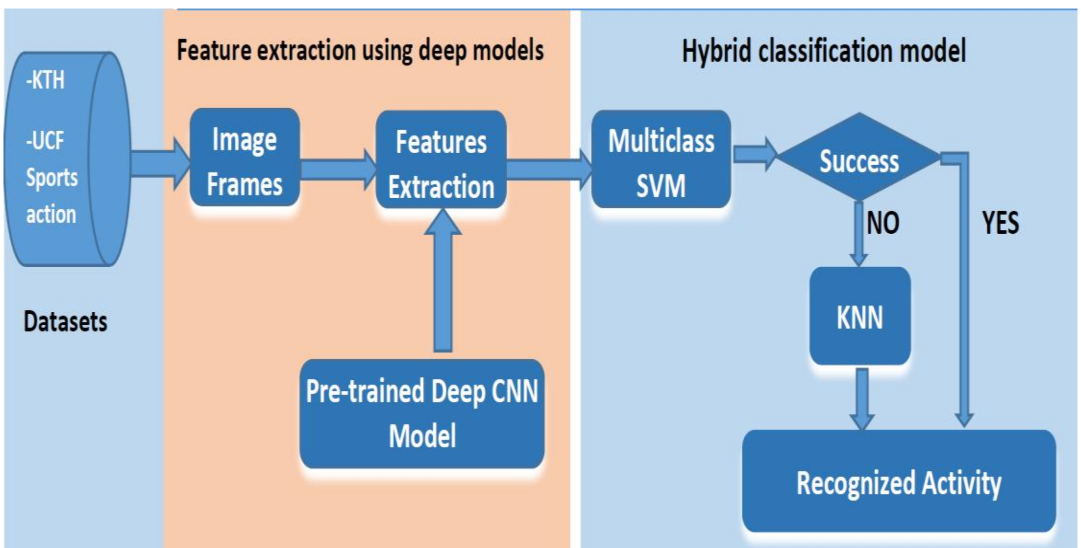

Fig. 2. Feature extraction and hybrid classification model

IV. EXPERIMENTATIONS AND RESULTS

This section discusses the experimental setup, training process and experimental results of the proposed technique. The proposed technique is tested on two well-known action datasets i.e., KTH [43], and UCF Sports [44]. The description of these datasets and comparative analysis are presented in the subsequent sections.

A. Evaluation on KTH dataset

The KTH [43] is well-known public dataset comprised of 6 actions, including waking, running, jogging, hand waving, boxing, and hand clapping. There were 25 actors involved in performing these actions in different setups including: outdoor, outdoor with variation in scale, outdoor with different clothes, and outdoor with illumination variations. The sample frames for each action from four different scenarios are shown in Fig. 3. This is a single view dataset with uniform background and recorded with fixed camera at the frame rate of 25fps.

During experimentation, the dataset is divided into two parts, One part is used for training while other one is used for testing the correctness of the proposed method same as [36]. The proposed method achieves 98.15% accuracy on KTH dataset, which is higher than the similar methods such as [26, 30, 36, 45-48], as shown in Table 1. The confusion matrix indicating the accuracy of each action and correspondence between the target classes along x-axis and output classes along y-axis is shown in Fig. 4.

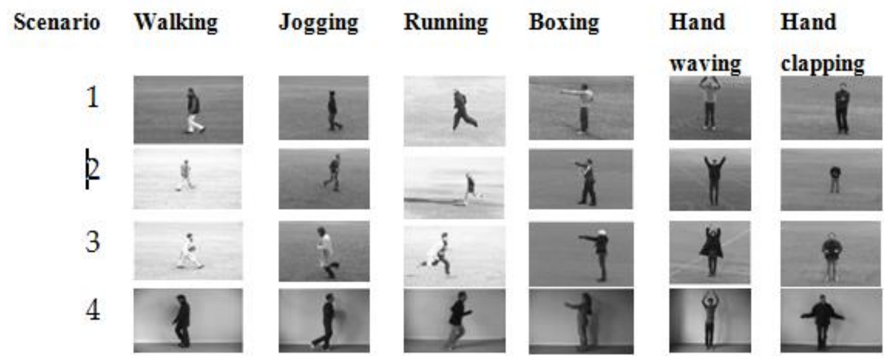

Fig. 3. Sample frames for each action from four scenarios in KTH dataset.

Table 1. Comparison of classification results on KTH dataset

Year	Method	Accuracy (%)
-	Proposed method (SVM-KNN)	98.15
-	Proposed method (KNN)	94.83
-	Proposed method (SVM)	89.91
2016	Charalampous and Gasteratos [36]	91.99
2016	Ahad et al. [45]	86.7
2016	Ding and Qu [46]	95.58
2013	Wang et al. [49]	94.2
2013	Ji et al. [30]	90.2
2013	Chaaraoui et al. [47]	89.86
2011	Le et al. [48]	93.9

	Boxing	Hand Clapping	Hand Waving	Jogging	Running	Walking
Boxing	1.0000	0	0	0	0	0
Hand Clapping	0	0.9963	0	0	0	0.0037
Hand Waving	0	0.0258	0.9705	0	0	0.0037
Jogging	0	0	0	0.9742	0.0111	0.0148
Running	0	0	0	0.0221	0.9668	0.0111
Walking	0	0	0	0.0185	0	0.9815

Fig. 4. Confusion matrix of KTH dataset with 6 human actions

B. *Evaluation on UCF sports action dataset*

The UCF dataset [44] encompasses 10 sports actions collected from videos broadcasted on television channels such as ESPN and BBC. These actions include: golf swing, diving, lifting, kicking, running, riding horse, swing-bench, skateboarding, swing-side, and walking. These actions were recorded in real sport environment exhibiting the variations in background, illumination conditions, and occlusions, which make it a challenging dataset. The sample frames for each action are shown in Fig. 5.

The proposed study uses a popular Leave-One-Out (LOO) cross validation scheme. Some other methods have also used Leave-One-Sequence-Out (LOSO), and Leave-One-Person-Out (LOPO) cross validation, which are quite similar to LOO validation [50]. In LOO cross validation, all video sequences are used for training except one, which is used for testing the performance of the classifier. This method is repeated for all available video sequences. Finally, the results of these sequences are summed up and average result is considered as a final result. This validation scheme has been employed by many similar research method such as [49, 51] for assessing the performance of their methods. Since, the proposed method uses the same validation scheme, it provides the fair comparison with similar methods. The proposed transfer learning method achieved an accuracy of 91.47% on UCF sports dataset which is higher than other similar methods as shown in Table 2. The detail confusion matrix indicating the accuracy of each action, and correspondence between the target classes along x-axis and output classes along y-axis and is shown in Fig. 6.

Fig. 5. Sample frames for each action from UCF sports dataset.

Table 2. Comparison of classification results on UCF sports action dataset

Year	Method	Testing scheme	Accuracy (%)
-	Proposed method (SVM-KNN)	LOO	91.47
-	Proposed method (SVM)	LOO	89.60
-	Proposed method (KNN)	LOO	82.75
2016	Tian et al. [51]	LOO	90.0
2016	Charalampous and Gasteratos [36]	-	88.55
2015	Atmosukarto et al. [52]	LOO	82.6
2014	Yuan et al. [28]	LOO	87.33
2013	Wang et al. [49]	LOO	88.0
2011	Le et al. [48]	-	86.5
2011	Wang et al. [53]	LOO	88.2
2010	Kovashka et al. [54]	LOO	87.27
2009	Wang et al. [55]	LOO	85.6

	Diving	Golf swing	Kicking	Lifting	Riding horse	Running	Skateboarding	Swing-bench	Swing-side	Walking
Diving	1.00000	0	0	0	0	0	0	0	0	0
Golf swing	0	0.9928	0.0036	0	0	0.0018	0.0018	0	0	0
Kicking	0	0.0109	0.7464	0.0036	0.0145	0.0978	0.0399	0.0036	0.0072	0.0761
Lifting	0	0	0	1.0000	0	0	0	0	0	0
Riding horse	0	0	0	0	0.9094	0.0036	0	0	0.0833	0.0036
Running	0	0.0634	0.1178	0	0.0054	0.7482	0.0018	0.0127	0.0072	0.0435
Skateboarding	0	0	0.0326	0	0.0236	0.0272	0.9112	0	0.0054	0
Swing-bench	0	0	0	0	0	0	0	1.0000	0	0
Swing-side	0	0	0	0	0	0	0	0	1.0000	0
Walking	0	0.0091	0.0598	0.0036	0.0254	0	0.0562	0.0054	0.0018	0.8388

Fig. 6. Confusion matrix of UCF sports action dataset.

V. CONCLUSION

This paper presents human action recognition method based on transfer learning using a pre-trained deep CNN architecture and a hybrid SVM-KNN classifier. The source architecture is used as a feature extractor machine for the new task and hybrid SVM-KNN classifier is trained on the target datasets. It was demonstrated that with the help transfer learning we can successfully utilize the already learnt knowledge for learning the new task with limited training dataset. Transfer learning is very useful when the dataset is not sufficient for training the deep learning model from scratch. Moreover, training a deep learning model from scratch requires much time and computational resources which can be saved using transfer learning. In addition to this, it was confirmed that a hybrid classifier has an advantage over the single classifier in boosting the accuracy of the recognition system. Moreover, unlike handcrafted representation based methods, the proposed approach is simpler and directly works with RGB images thus eliminating the need of preprocessing and manual feature extraction. The effectiveness of the proposed method was checked on two well-known KTH, and UCF sports action datasets, and achieved 98.15%, and 91.47% accuracies respectively. The comparative analysis confirms that the proposed methods outperforms the similar state-of-the-art methods for human action recognition using transfer learning. In future, we would like to extend this method for more complex datasets such as IXMAS, UCF-50, UCF-101, and HMDB-51.

ACKNOWLEDGMENT

The second author would like to thank China Scholarship and the National Natural Science Foundation of China for financial support (Grant No. 61602380 and 61673319).

REFERENCES

1. Aggarwal, J.K. and M.S. Ryoo, Human activity analysis: A review. ACM Computing Surveys (CSUR), 2011. **43**(3): p. 16.
2. Weinland, D., R. Ronfard, and E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding, 2011. **115**(2): p. 224-241.
3. Zhu, F., Sha, L., Xie, J., and Fang, Y., From handcrafted to learned representations for human action recognition: A survey. Image and Vision Computing, 2016.
4. Sargano, A.B., P. Angelov, and Z. Habib, A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Applied Sciences, 2017. **7**(1): p. 110.
5. Hinton, G.E. and R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science, 2006. **313**(5786): p. 504-507.
6. LeCun, Y., Bottou, L., Bengio, Y., Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.
7. Hinton, G.E., S. Osindero, and Y.-W. Teh, A fast learning algorithm for deep belief nets. Neural computation, 2006. **18**(7): p. 1527-1554.
8. Cao, X., Wang, Z., Yan, P., Li, X., Transfer learning for pedestrian detection. Neurocomputing, 2013. **100**: p. 51-57.
9. Wu, D., F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012. IEEE.
10. Fei-Fei, L. Knowledge transfer in learning to recognize visual objects classes. in Proceedings of the International Conference on Development and Learning (ICDL). 2006.
11. Azizpour, H., Razavian, S.A., Sullivan, J., From generic to specific deep representations for visual recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015.
12. Chatfield, K., Simonyan, K., Vedaldi, A., Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014.
13. Oquab, M., Bottou, L., Laptev, I., Sivic, J., Learning and transferring mid-level image representations using convolutional neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
14. Girshick, R., Donahue, J., Darrell, T., Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
15. Nam, H. and B. Han, Learning multi-domain convolutional neural networks for visual tracking. arXiv preprint arXiv:1510.07945, 2015.
16. Noh, H., S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. in Proceedings of the IEEE International Conference on Computer Vision. 2015.
17. Shao, L., F. Zhu, and X. Li, Transfer learning for visual categorization: A survey. IEEE transactions on neural networks and learning systems, 2015. **26**(5): p. 1019-1034.
18. Zeiler, M.D. and R. Fergus. Visualizing and understanding convolutional networks. in European Conference on Computer Vision. 2014. Springer.
19. Razavian, S.A., Azizpour, H., Sullivan, J., CNN features off-the-shelf: an astounding baseline for recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014.
20. Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Going deeper with convolutions. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
22. Willems, G., T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. in European conference on computer vision. 2008. Springer.
23. Klaser, A., M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. in BMVC 2008-19th British Machine Vision Conference. 2008. British Machine Vision Association.
24. Jain, M., H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
25. Fathi, A., Y. Li, and J.M. Rehg. Learning to recognize daily actions using gaze. in European Conference on Computer Vision. 2012. Springer.
26. Wang, H. and C. Schmid. Action recognition with improved trajectories. in Proceedings of the IEEE International Conference on Computer Vision. 2013.
27. Sargano, A.B., P. Angelov, and Z. Habib, Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines. Applied Sciences, 2016. **6**(10): p. 309.
28. Yuan, C., Li, X., Hu, W., Ling, H., Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition. IEEE Transactions on Image Processing, 2014. **23**(2): p. 658-672.
29. Angelov, P. and A. Sperduti. Challenges in Deep Learning. in Proc. European Symp. on Artificial NNs.
30. Ji, S., Xu, W., Yang, M., Yu, K., 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(1): p. 221-231.
31. Taylor, G.W., Fergus, R., LeCun, Y., Convolutional learning of spatio-temporal features. in European conference on computer vision. 2010. Springer.
32. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Learning spatiotemporal features with 3d convolutional networks. in 2015 IEEE International Conference on Computer Vision (ICCV). 2015. IEEE.

33. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Large-scale video classification with convolutional neural networks. in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.
34. Simonyan, K. and A. Zisserman. Two-stream convolutional networks for action recognition in videos. in Advances in Neural Information Processing Systems. 2014.
35. Bux, A., P. Angelov, and Z. Habib, Vision based human activity recognition: a review, in Advances in Computational Intelligence Systems. 2017, Springer. p. 341-371.
36. Charalampous, K. and A. Gasteratos, On-line deep learning method for action recognition. Pattern Analysis and Applications, 2016. **19**(2): p. 337-354.
37. Wang, L., Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
38. He, K., Zhang, X., Ren, S., Sun, J., Deep residual learning for image recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
39. Cao, L., Z. Liu, and T.S. Huang. Cross-dataset action detection. in Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. 2010. IEEE.
40. Smeaton, A.F., P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. in Proceedings of the 8th ACM international workshop on Multimedia information retrieval. 2006. ACM.
41. Aytar, Y., Transfer learning for object category detection. 2014, University of Oxford.
42. Su, Y.C., Chiu, T.H., Yeh, C.Y., Huang, H.F., Transfer Learning for Video Recognition with Scarce Training Data for Deep Convolutional Neural Network. arXiv preprint arXiv:1409.4127, 2014.
43. Schuldt, C., I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. 2004. IEEE.
44. Rodriguez, M., Spatio-temporal maximum average correlation height templates in action recognition and video summarization. 2010.
45. Rahman Ahad, M.A., M.N. Islam, and I. Jahan, Action recognition based on binary patterns of action-history and histogram of oriented gradient. Journal on Multimodal User Interfaces, 2016.
46. Ding, S. and S. Qu. An improved interest point detector for human action recognition. in Control and Decision Conference (CCDC), 2016 Chinese. 2016. IEEE.
47. Chaaraoui, A.A., P. Climent-Pérez, and F. Flórez-Revuelta, Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters, 2013. **34**(15): p. 1799-1807.
48. Le, Q.V., Zou, W.Y., Yeung, S.Y., Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. 2011. IEEE.
49. Wang, H., Klaser, A., Schmid, C., Liu, C.L., Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision, 2013. **103**(1): p. 60-79.
50. Vishwakarma, D. and R. Kapoor, Hybrid classifier based human activity recognition using the silhouette and cells. Expert Systems with Applications, 2015. **42**(20): p. 6957-6965.
51. Tian, Y., Ruan, Q., An, G., Fu, Y., Action Recognition Using Local Consistent Group Sparse Coding with Spatio-Temporal Structure. in Proceedings of the 2016 ACM on Multimedia Conference. 2016. ACM.
52. Atmosukarto, I., N. Ahuja, and B. Ghanem. Action recognition using discriminative structured trajectory groups. in 2015 IEEE Winter Conference on Applications of Computer Vision. 2015. IEEE.
53. Wang, H., Klaser, A., Schmid, C., Action recognition by dense trajectories. in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. 2011. IEEE.
54. Kovashka, A. and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010. IEEE.
55. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Evaluation of local spatio-temporal features for action recognition. in BMVC 2009-British Machine Vision Conference. 2009. BMVA Press.