Human Action Recognition using Multi-Kernel Learning for Temporal Residual Network

Saima Nazir[1,2,4], Yu Qian[4], Muhammad Haroon Yousaf[1], Sergio A. Velastin[2,3,4], Ebroul Izquierdo[2] and Eduard Vazquez[4]

[1]*University of Engineering and Technology Taxila, Pakistan*

[2]*Queen Mary University of London, U.K.*

[3]*Universidad Carlos III de Madrid, Spain*

[4]*Cortexica Vision Systems Ltd., U.K.*

Keywords: Deep Learning, Residual Network, Spatio-Temporal Network, Temporal Residual Network, Human Action Recognition.

Abstract: Deep learning has led to a series of breakthrough in the human action recognition field. Given the powerful representational ability of residual networks (ResNet), performance in many computer vision tasks including human action recognition has improved. Motivated by the success of ResNet, we use the residual network and its variations to obtain feature representation. Bearing in mind the importance of appearance and motion information for action representation, our network utilizes both for feature extraction. Appearance and motion features are further fused for action classification using a multi-kernel support vector machine (SVM). We also investigate the fusion of dense trajectories with the proposed network to boost up the network performance. We evaluate our proposed methods on a benchmark dataset (HMDB-51) and results shows the multi-kernel learning shows the better performance than the fusion of classification score from deep network SoftMax layer. Our proposed method also shows good performance as compared to the recent state-of-the-art methods.

1 INTRODUCTION

Human action recognition is an important yet challenging task in computer vision field (Nazir et al., 2018). For the last decade, many researchers are focusing on deep convolutional networks for human action recognition. Since the introduction of a large scale image repository (ImageNet) (Deng et al., 2009) and high performance computational machines (GPUs), image and action recognition using deep convolutional neural network (CNN) is enjoying a remarkable triumph (Tran et al., 2018; Pham et al., 2018; Herath et al., 2017). However to achieve accurate recognition at the level of human understanding of ongoing actions in real environment is still strenuous.

After the introduction of AlexNet (Krizhevsky et al., 2012), researchers are now focusing on increasing the depth of state-of-the-art CNN architectures. Networks such as VGG network (Simonyan and Zisserman, 2014b) are going deeper and deeper having 8 layers (VGG-M) to 16 (VGG-16) and 19 (VGG-19) layers respectively. However simply stacking layers to go deeper does not guarantee increase in performance. He et al., (He et al., 2016) state that a 56-layer plain network shows higher training and test error as compared to a 22-layer plain network. Consequently, deeper networks are harder to train, and their performance may get saturated or keep on decreasing by adding more layers.

To overcome this problem, He et al., (He et al., 2016) proposed a novel residual network, ResNet, by introducing identity shortcut connections to fit the input from the previous layer to the next layer without any modification of the input. To get this identity mapping, they introduced a novel residual learning framework. The residual function is introduced to let the layers fit the residual mapping.

After the introduction of such deeper network, ResNet (He et al., 2016), many researchers have proposed variations of ResNet to improve performance in different computer vision applications such as image classification, object and, action recognition. Initially ResNet (He et al., 2016) was introduced for image recognition task.

Many researchers are endeavoring for the extension of 2D convolutional neural network architecture to the temporal domain. As shown in fig. 1 (b), Feichtenhofer et al. (Feichtenhofer et al., 2016a) integrated the space-time information by injecting a residual connection between the appearance and motion

Nazir, S., Qian, Y., Yousaf, M., Velastin, S., Izquierdo, E. and Vazquez, E.
Human Action Recognition using Multi-Kernel Learning for Temporal Residual Network.
DOI: 10.5220/0007371104200426
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 420-426
ISBN: 978-989-758-354-4
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

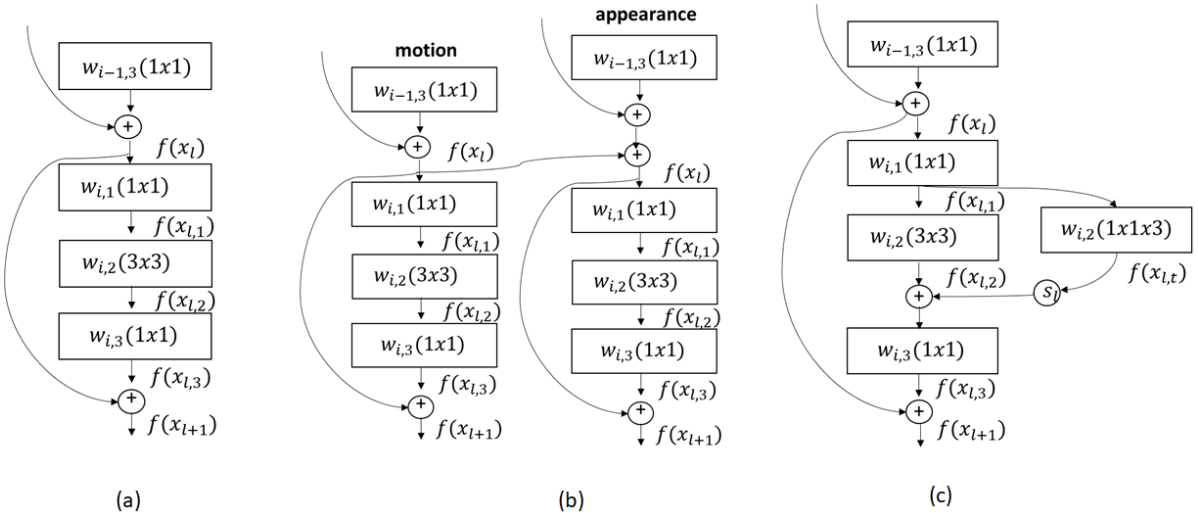

Figure 1: (a) Original residual unit [ResNet] (b) Spatio-temporal residual unit [ST-ResNet] (c) spatio-temporal residual unit [T-ResNet].

pathways of a two stream architecture. To operate on adjacent feature maps in time domain they applied spatiotemporal filter kernel instead of spatial filter kernel.

To transform spatial ConvNet into spatiotemporal ConvNet, Feichtenhofer et al. (Feichtenhofer et al., 2017b) modify the basic residual block (He et al., 2016) and inject the temporal information using 1D temporal filter. As shown in fig. 1 (c) they introduced an additional convolutional block in the original residual unit along with affine scaling layer S_l.

The closest work to ours is temporal ResNet (Feichtenhofer et al., 2017b), where they augment the temporal information in the spatial residual network with convolution across time. In contrast to their work, we extracted the features from temporal ResNet for both appearance and flow streams and fused both network features for the classification using a multi-kernel SVM.

2 RELATED WORK

Human action recognition has remained one of the important and challenging tasks in computer vision field. Many researchers have contributed by proposing hand-crafted feature representation for action recognition. Spatio-temporal interest point (3D Harris) (Laptev, 2005), 3D-scale invariant feature transform (3D SIFT) (Scovanner et al., 2007) and dense trajectories (Wang et al., 2011) are a few of the state-of-the-art hand-crafted feature representation approaches. Since the introduction of CNN, many researchers are now concentrating on the use of deep learning approaches for the task of action recognition. Zeiler and Fergus (Zeiler and Fergus, 2014) stated that convolutional neural network can serve the purpose of feature extraction as the first layer learns low level features and high-level features are learned by top layers in CNNs.

Many researchers have extended the traditional 2D convolutional neural network (CNN) to integrate temporal information. 3D CNN is proposed by Hara et al., (Hara et al., 2017) to assimilate the temporal information by using a 3D kernel to extract information from both spatial and temporal domain. Limiting spatial space looks reasonable whereas, for the extraction of motion information, limiting time domain to a few frames seems deficient. Ng et al. (Yue-Hei Ng et al., 2015) stated that instead of learning temporal information from few frames, temporal pooling can be applied to integrate the temporal domain information in network layers.

Simonyan et al. (Simonyan and Zisserman, 2014a) extended the convolutional neural network by proposing a two-stream architecture. Similar to the way human vision cortex learns appearance and motion information using ventral (object recognition) and dorsal (motion recognition) stream, they used two pathways to learn spatial and temporal information for appearance and motion respectively. Spatial stream is learned on the model pretrained on ImageNet while temporal stream is trained on an action recognition dataset considering the unavailability of pretrained model on flow information.

Following sucess of the of two-stream architecture, Feichtenhofer et al. (Feichtenhofer et al., 2016a) extended the two-stream convolutional neural net-

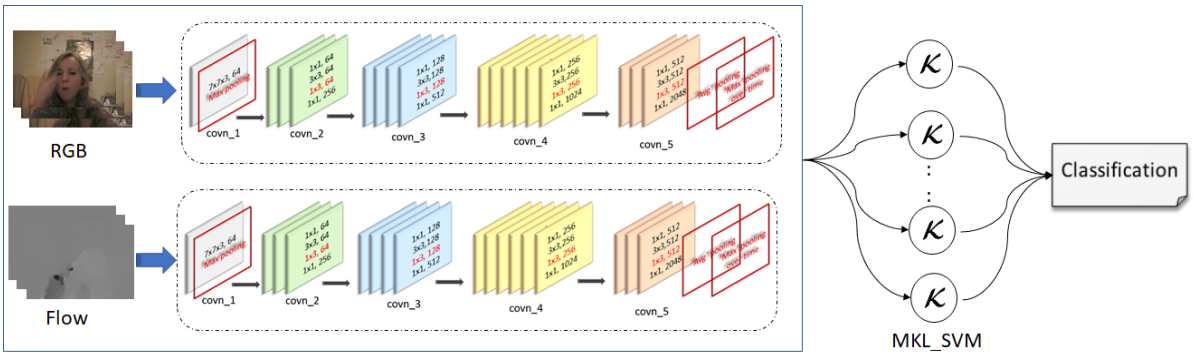

Figure 2: Proposed framework: Multi-kernel SVM learning using temporal residual network features for human action recognition.

work and proposed a spatio-temporal residual network. Driven by the success of ResNet in ILSVRC 2015, they used ResNet which is pretrained on ImageNet for action recognition. They embed the residual connection between appearance and motion streams to allow the hierarchical learning of spatiotemporal features for both streams.

Hara et al. (Hara et al., 2017) also utilized the ResNet model for human action classification task. They make use of 3D convolutional kernels to integrate temporal information in residual network. They trained and evaluated their proposed 3D ResNet model on the relatively larger datasets (Kinetics and Activity Net) available for human action recognition task. Temporal ResNet (Feichtenhofer et al., 2017b) also integrates the temporal information in spatial residual network by modifying the basic residual function by adding an additional temporal conv block.

3 PROPOSED METHOD

In this paper, we evaluate the performance of residual network (ResNet) and its variations for human action recognition. We adopt two-stream framework, the networks are trained on appearance (RGB frames) and motion (optical flow) respectively. We utilize the residual network (ResNet) (He et al., 2016), Spatio-temporal residual network (ST-ResNet) (Feichtenhofer et al., 2016a) and temporal residual network (T-ResNet) (Feichtenhofer et al., 2017b) for the feature extraction task. We extract features for both appearance and motion representation. Later these features are passed to Multi-kernel support vector machine (SVM) for the action classification task. Fig. 2 shows the general framework for action classification using a multi-kernel SVM where features are extracted from RGB and flow stream supplied to temporal residual networks. Using the approach shown in fig. 2 we have also extracted the features for the other two networks, i.e. residual network and spatio-temporal residual network. Details of features extraction from these three different residual networks are provided in section 3.1, 3.2 and 3.3.

3.1 Feature Extraction using Residual Layers (ResNet)

As the base representation, we utilize the ResNet50 model provided by (Feichtenhofer et al., 2016a) for feature extraction. They used a ResNet50 (He et al., 2016) model pretrained on ImageNet and replace the last classification layer as per the requirement of the HMDB51 dataset. We truncate the network and obtain 2048 features from the pooling layer after convolution block 5 for both appearance and motion stream separately. For further information on network training, model construction please refer to (Feichtenhofer et al., 2016a).

3.2 Feature Extraction using Spatiotemporal Residual Layers (ST-ResNet)

We also make use of spatio-temporal residual network (ST-ResNet) (Feichtenhofer et al., 2016a) to extract the appearance and motion features. This network architecture integrates the motion information in spatial stream by adding a residual connection between both streams. To learn spacetime features they also included temporal convolution in residual units. We truncate the network at pooling layer after last convolutional block to obtain 2048 for both streams. For further information on network training and model construction please refer to (Feichtenhofer et al., 2016a).

Table 1: Accuracy (%) for HMDB-51 dataset using residual network (ResNet).

Split	RGB		Flow		RGB + Flow		RGB + Flow + iDT	
	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM
1	49.0%	50.3%	57.7%	63.5%	63.3%	67.5%	67.5%	70.1%
2	48.7%	48.3%	54.6%	59.3%	61.8%	65.0%	67.7%	67.1%
3	49.6%	48.0%	56.5%	61.8%	61.8%	65.8%	65.6%	67.9%
Average	49.1%	48.9%	56.3%	61.5%	62.3%	66.1%	66.9%	68.4%

Table 2: Accuracy (%) for HMDB-51 dataset using spatio-temporal residual network (ST-ResNet).

Split	RGB		Flow		RGB + Flow		RGB + Flow + iDT	
	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM
1	56.2%	56.6%	58.8%	61.7%	68.0%	67.6%	69.1%	68.3%
2	56.8%	57.1%	59.8%	62.1%	66.0%	65.8%	66.2%	66.3%
3	55.5%	56.4%	59.8%	65.2%	66.3%	67.0%	66.7%	67.1%
Average	56.2%	56.7%	59.5%	63.0%	66.8%	66.8%	67.3%	67.2%

3.3 Feature Extraction using Temporal Residual Layers (T-ResNet)

We also use the temporal residual network architecture (T-ResNet) for the extraction of appearance and motion features. For appearance, we utilized the architecture proposed in (Feichtenhofer et al., 2017b) and build the spatiotemporal residual network on ResNet50 (He et al., 2016) which is pre-trained on ImageNet. In the next stage, we inject the temporal information in this spatiotemporal residual unit. For the flow extraction, we utilized the model provided in (Feichtenhofer et al., 2016a) and inject the temporal connection in spatiotemporal residual network which is pre-trained on the HMDB-51 dataset. In contrast to (Feichtenhofer et al., 2017b), where they fused the prediction of both networks, we truncate both networks at pooling layer after conv_5 to get the appearance and flow features representation separately. For further information on network training and model construction please refer to (Feichtenhofer et al., 2017b).

3.4 Support Vector Machine with Multi-kernel Learning

Given training tuples $(x_i; y_i)$ and weights w, under a Hinge loss, an SVM solves the primal problem [19],

$$\begin{aligned} \min_{w,b,\vartheta} \frac{1}{2} w^T w^+ C \sum_{i=1}^{n} \vartheta_i \\ s.t. y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \vartheta_i \\ \vartheta \geq 0, i = 1, \ldots, n \end{aligned} \tag{1}$$

As is customary in kernel methods, computations involving ϕ are handled using kernel functions $k(x_i; x_j) = \phi(x_i).\phi(x_j)$. In all our experiments, a Radial Basis Function (RBF) based kernel has been used. *C (fixed at 1)* is the penalty parameter and ϑ is the slack variable. For multiple kernel learning (we follow the recipe by (Sonnenburg et al., 2006), and formulate a convex combination of sub-kernels as,

$$k(x_i; x_j) = \sum_{k=1}^{K} \beta_k K_k(x_i; x_j) \tag{2}$$

In contrast to (Sonnenburg et al., 2006), we use *L2* regularized $\beta_k \geq 0$ and $\sum_{k=1}^{K} {\beta_k}^2 \leq 1$. *L2* regularised multiple kernel is learnt by formulating eqn. 2 as a semi-infinite linear programming (SILP) problem. During each iteration, an SVM solver is first instantiated to obtain the weighted support vectors; subsequently, a linear programming (LP) problem is solved using Mosek (mos, 2012).

4 EVALUATION

4.1 Dataset

We used the HMDB-51 (Kuehne et al., 2011) human action recognition dataset for the evaluation of proposed method. HMDB-51 is a large-scale benchmark dataset containing 7k videos. It is divided into training, testing and validation sets containing 3570, 1530 and 1749 videos respectively. These videos are taken from digital movies and different public video databases like YouTube. This dataset is captured in realistic environment and comparatively more challenging than the other benchmark datasets like UCF101. For evaluation we used the standard evaluation measures and report the performance as average accuracy over three splits. All experiments were done on Intel Xeon E5-2687W 3 GHz 128 GB workstation with two 12GB nVIDIA TITAN Xp GPUs.

Table 3: Accuracy (%) for HMDB-51 dataset using temporal residual network (T-ResNet).

Split	RGB		Flow		RGB + Flow		RGB + Flow + iDT	
	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM	SoftMax	MKL_SVM
1	51.3%	53.9%	64.2%	64.9%	67.2%	70.1%	67.9%	71.1%
2	51.5%	53.2%	64.6%	65.0%	68.6%	69.1%	69.4%	69.8%
3	49.9%	51.9%	65.8%	65.2%	68.6%	68.9%	68.8%	69.9%
Average	50.9%	53.0%	64.9%	65.0%	68.1%	69.4%	68.7%	70.3%

Table 4: Comparison with the state-of-the-art methods for HMDB-51 dataset.

Method	Results (Accuracy %)	Reference
Two-stream	59.4	(Simonyan and Zisserman, 2014a)
Rank Pooling (ALL)+ HRP (CNN)	65.0	(Fernando and Gould, 2017)
Convolutional Two-stream (without/with iDT)	65.4/69.2	(Feichtenhofer et al., 2016b)
Temporal-Inception	67.5	(Ma et al., 2017)
Temporal Segment Network (2/3 modalities)	68.5/69.4	(Wang et al., 2016)
TS-LSTM	69	(Ma et al., 2017)
ST-ResNet(without/with iDT)	66.4/70.3	(Ma et al., 2017)
ST-multiplier network(without/with iDT)	68.9/72.2	(Feichtenhofer et al., 2017a)
Two-Stream I3D on split-1	66.4	(Carreira and Zisserman, 2017)
Two-Stream I3D with Kinetics pre-training	80.7	(Carreira and Zisserman, 2017)
Temporal ResNet	67.2	(Feichtenhofer et al., 2017b)
MKL_SVM ResNet (RGB+Flow+iDT)	68.4	Our
MKL_SVM ST-ResNet (RGB+Flow+iDT)	67.2	Our
MKL_SVM T-ResNet (RGB+Flow+iDT)	70.3	Our

4.2 Results and Discussion

In this section we present the results on three different residual networks by comparing the performance for action classification using SoftMax and multi-kernel support vector machine. For multiple network scores (the output of SoftMax layer) fusion, we used a 1:1 weighting scheme. These scores were taken from the respective models. For MKL_SVM, we tuned the parameters on validation dataset and chose *L2-norm* as it always performed better than *L1-norm*. We fixed penalty parameter *C=1* as it shows better performance than *C=10 or 100*. Along with appearance and motion features, we have also considered the iDT features which are computed by following the method described in (Wang et al., 2013).

As the base representation of our proposed methods, we used ResNet50 (Feichtenhofer et al., 2016a) trained on HMDB-51 dataset. Table 1 reports the performance of ResNet50 over 3 splits. ResNet trained on flow information shows better performance as compared to the ResNet trained on RGB information. By combing both RGB and flow information we get a better result as compared to individual network. We compared the performance of Multi-kernel SVM and SoftMax using RGB+flow and RGB+flow+iDT information and results shows the MKL_SVM has 4% and 2% increase in performance as compared to SoftMax respectively.

Similiarly, Table 2 reports the spatio-temporal residual network performance for HMDB-51 dataset. We used the ST-ResNet50 (Feichtenhofer et al., 2016a) model trained for HMDB-51 dataset for both appearance and flow extraction. Multiple kernel learning provides better performance as compared to SoftMax specially when network is trained using flow information. However in case of learning three kernels for RGB, flow and iDT features, MKL couldn't outperform SoftMax.

Table 3 shows the performance of temporal residual network for RGB, flow and iDT features. We trained the T-ResNet (Feichtenhofer et al., 2017b) using RGB and flow information. MKL_SVM always show better performance than SoftMax when multi-kernel learning is applied on features obtained from temporal residual network.

By comparing the performance of three residual networks i.e. ResNet, ST-ResNet and T-ResNet, we can conclude that using temporal residual network (T-ResNet) that integrates the temporal domain information in basic residual unit of ResNet, it is beneficial for action recognition where the importance of temporal information along with spatial information in non-negligible.

We conclude the evaluation section with the comparison of our proposed methods with state-of-the-art deep learning methods. As shown in Table 4. multi-kernel learning SVM approach boost the re-

sults and provide state-of-the-art performance for HMDB-51 dataset. Feichtenhofer et al. (Feichtenhofer et al., 2017a) used ResNet-50 for appearance and ResNet-152 for motion in their proposed architecture for HMDB-51 dataset. They concluded that for the HMDB-51 dataset deeper appearance network degrades performance while for deeper motion network sizable gain is observed. The performance of our proposed network can also be improved by using ResNet-152 for motion stream. The highest accuracy shown in Table 4. was released by DeepMind in July 2017 which uses 240K Kinetics dataset pre-training with end-to-end finetuning on HMDB-51 (Carreira and Zisserman, 2017), without Kinetics pretraining model, their proposed two-stream I3D shows 66.4% accuracy on the split-1 of HMDB-51. The better performance is explainable by the better quality of Kinetics dataset that is used for pre-training the network.

5 CONCLUSION

In this paper we proposed a new method to improve the performance for human action recognition. The proposed method is based on an existing state-of-the-art deep learning network, T-ResNet, which recently shows noticeable performance for the task of human action recognition task. We utilized the prevalent variations of residual network for the extraction of motion and appearance features for benchmark dataset. We extracted features from popular ResNet, spatio-temporal ResNet and temporal ResNet. We evaluated multi-kernel learning approach in comparison with SoftMax and experiments shows that MKL_SVM always outperforms SoftMax. In future, we can train the state-of-the-art residual networks on larger datasets like Kinetics600 to improve the performance for action recognition.

ACKNOWLEDGEMENTS

Sergio A Velastin has received funding from the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander. Authors also acknowledge support from the Higher Education Commission, Pakistan.

REFERENCES

(2012). Mosek toolbox. http://docs.mosek.com/6.0/toolbox/]. Last accessed on: 2018-06-30.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee.

Feichtenhofer, C., Pinz, A., and Wildes, R. (2016a). Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*, pages 3468–3476.

Feichtenhofer, C., Pinz, A., and Wildes, R. P. (2017a). Spatiotemporal multiplier networks for video action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7445–7454. IEEE.

Feichtenhofer, C., Pinz, A., and Wildes, R. P. (2017b). Temporal residual networks for dynamic scene recognition. In *CVPR*, volume 1, page 2.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016b). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941.

Fernando, B. and Gould, S. (2017). Discriminatively learned hierarchical rank pooling networks. *International Journal of Computer Vision*, 124(3):335–355.

Hara, K., Kataoka, H., and Satoh, Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*, volume 2, page 4.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Herath, S., Harandi, M., and Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE.

Laptev, I. (2005). On space-time interest points. *International journal of computer vision*, 64(2-3):107–123.

Ma, C.-Y., Chen, M.-H., Kira, Z., and AlRegib, G. (2017). Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *arXiv preprint arXiv:1703.10667*.

Nazir, S., Yousaf, M. H., Nebel, J.-C., and Velastin, S. A. (2018). A bag of expression framework for improved human action recognition. *Pattern Recognition Letters*, 103:39–45.

Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., and Velastin, S. A. (2018). Exploiting deep residual networks for human action recognition from skeletal data. *Computer Vision and Image Understanding*.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM.

Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(Jul):1531–1565.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer.

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*, pages 4694–4702.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS