Extreme Low Resolution Action Recognition with Spatial-Temporal Multi-Head Self-Attention and Knowledge Distillation

Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang
National Taiwan University of Science and Technology, Taiwan, R.O.C.
Email: {d10602806,d10702801,ytchen,whf}@mail.ntust.edu.tw

Abstract

This paper proposes a two-stream network with a novel spatial-temporal multi-head self-attention mechanism for action recognition in extreme low resolution (LR) videos. The new approach first utilizes a super resolution (SR) mechanism to provide better visual information to facilitate the network training. To provide more discriminative spatio-temporal features, a knowledge distillation scheme that consists of teacher and student models is employed to enhance the network model using the knowledge from a high resolution (HR) model. Moreover, the two-stream network is combined with a new spatial-temporal multi-head self-attention network to efficaciously learn the long-term temporal dependency. Simulations demonstrate that the proposed method surpasses the state-of-the-art works for extreme LR action recognition on two widespread HMDB-51 and IXMAS datasets.

1. Introduction

Action recognition in extreme low resolution (LR) videos has received growing interests in security and surveillance [1–5], where privacy-preserving issues are the main concern. However, analyzing LR videos is not a simple task due to their substantial loss of visual information, which can engender misleading cues. Simultaneously, action recognition in extreme LR videos, which suffers some common issues encountered in high resolution (HR) videos such as view point changes, background clutter, inter-class similarity, and occlusion, is a challenging issue.

Over the past few years, a number of convolutional neural-network (CNNs) has been addressed for action recognition in extreme low resolution (LR). For instance, Yu *et. al.* [1] introduced a low rank representation for videos and a data-driven learning to speed up the convergence of training for LR videos. Ryoo *et al.* [6] introduced an inverse super resolution paradigm to learn the image transformation for generating multiple low resolution videos from a single video, and then later further improved the action recognition accuracy by using a multi-Siamese network to learn the shared embedding space in [4]. Chen *et al.* [7] made use of a filter sharing scheme to jointly train both of the HR and LR networks. Also, Rahman *et al.* [5] combined textural features and classical shape and motion features to enhance the action recognition accuracy in LR videos. However, the aforementioned methods did not consider long-term temporal dependency information for action recognize, which is particular important when the visual qualities is severely degraded. Xu *et al.* [8] integrated a 3D ConvNet and recurrent neural network to exploit the temporal dependency information. However, RNN in general has a slow converge rate and requires a variety of large training data, so it is not effective in learning temporal dependency of distant temporal positions [9].

In this paper, we present a two-stream CNN for action recognition in extreme LR videos. It first utilizes a super resolution (SR) mechanism, which can provide better visual information than the distorted extreme LR images, to facilitate the two-stream network training. To provide more discriminative spatio-temporal features, a knowledge distillation scheme that consists of teacher and student models is employed to enhance the network model using the knowledge from a HR model. Moreover, the two-stream network is integrated with a new spatial-temporal multi-head self-attention network to efficaciously learn the long-term temporal dependency, which is essential when the action contains several sub-actions or the spatial information is severely impaired. Note that in contrast to [10], which relies on an object detector and thus is not suitable for extreme LR videos, our self-attention incorporates SR and knowledge distillation to guide the network to provide more discrimina-

2473-9944/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCVW.2019.00125

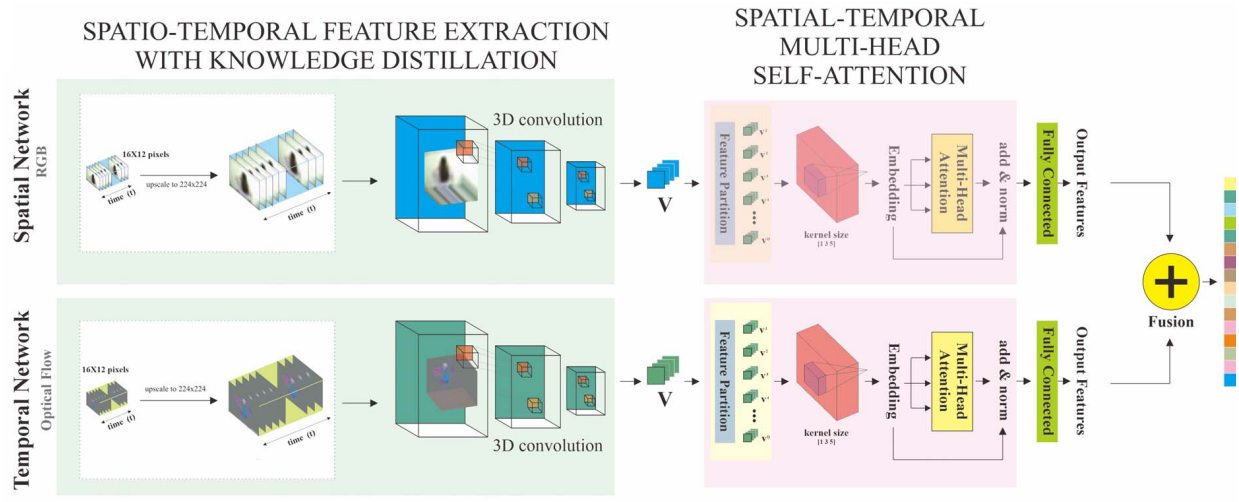

Figure 1: Overall architecture of the proposed method, which consists of a super resolution mechanism to enrich the visual information, knowledge distillation to assist the network training using high resolution videos, spatio-temporal feature extraction with two-stream I3D, and spatial-temporal multi-head self-attention to leverage the long-term temporal dependency.

tive features. Simulations reveal that the new method surpasses the state-of-the-art works on the widespread extreme LR HMDB-51 and IXMAS datasets.

The contributions of this paper include: i) a two-stream CNN together with a new multi-head self-attention is addressed to learn the temporal dependency across the frames in extreme LR videos; ii) a knowledge distillation mechanism, which possesses the advantage of teacher and student models, is utilized to obtain more discriminative features. To the best of authors's knowledge, it is the first time that knowledge transfer from HR to LR is considered by utilizing knowledge distillation in action recognition; iii) an SR mechanism is employed to provide richer visual information to expedite the two-stream network training for LR videos.

2. Related Work

Recently reported approaches for action recognition and detection in extreme LR videos rely on deep spatio-temporal CNN features [5, 11–13]. For example, Herrmann *et al.* [13] took advantage of large-scale datasets for effective training in face recognition. Wang *et al.* [14] developed a partially coupled network to enhance the robustness of the CNN features. Dimiccoli *et al.* [15] investigated the trade-off between recognition accuracy of daily activity using a combination of CNN features and a random forest classifier, and the privacy level captured by wearable cameras. Recently, Ren *et al.* [16] devised an adversarial training that renders human face anonymous while detecting actions to preserve privacy-sensitive information. Recently, Wu *et al.* [17] trained deep network directly on the compressed video to remove superfluous information and obtain more representative motion information.

Numerous methods have been proposed for restoring LR images using an SR mechanism. For instance, Dong *et al.* [18] used a deep convolutional network to learn a mapping between high- and low-resolution images to reconstruct HR images. Lim *et al.* [19] performed an optimization algorithm on residual networks by removing redundant modules from the conventional networks. Huang *et al.* [20] introduced a wavelet-based fully convolutional network that is able to learn wavelet coefficients from LR images. Ledig *et al.* [21] used a generative adversarial network (GAN) with perceptual loss function to preserve the texture of a single image. Recently, Haris *et al.* [22] developed a dense up-and-down projection unit that allows an efficient accumulation of multi-resolution features.

Integrating temporal dependency information into CNN has shown to be beneficial to recognize actions with several sub-actions and high inter-class similarity. In order to capture various temporal dependency from skeletal coordinates, Lee *et al.* [23] incorporated temporal sliding windows into a long short-term memory (LSTM) network. Martinez *et al.* [24] employed gated recurrent unit (GRU) to model human motion from motion capture (mo-cap) data. Tanfous *et al.* [25] applied a bidirectional LSTM to sparse coding to represent 3D skeletal sequences for action recognition. Zhu *et al.* [26]

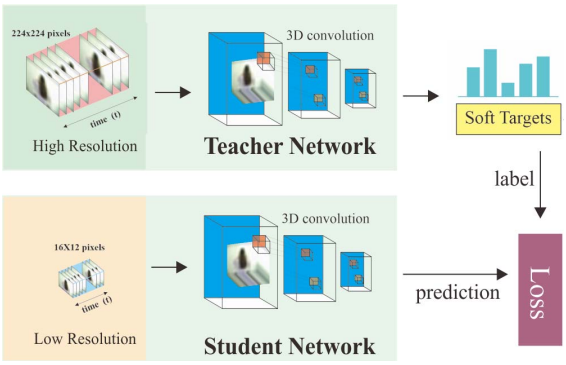

Figure 2: Illustration of knowledge distillation.

introduced a variant of convolutional LSTM, in which the spatial convolution map is passed on to the input-to-state transition with the same gates as the conventional fully connected LSTM. Different from the aforementioned approaches, our method combines a two-stream network with a self-attention mechanism to effectively model the long-term temporal dependency without encountering the vanishing gradient problem.

Recent progress in knowledge distillation has demonstrated that it is possible to transfer knowledge from a complex CNN model into a more compact one with a small performance gap. To construct more compact filters, Iandola *et al.* [27] reduced the number of network parameters and squeezed the number of the input channels. Also, Shang *et al.* considered a concatenated ReLU, which uses pairing filters in the lower layers, to reduce the number of network parameters. Zhang *et al.* [28] proposed an extension of MobileNet [29] based on Interleaved Group Convolution (IGC), in which two complementary group convolutions are performed alternatingly. In this paper, aside from reducing the network complexity, our objective is to use the knowledge distillation to take the advantage of the knowledge from the HR model to enhance the quality of the LR network.

3. Proposed Method

In this section, we introduce the proposed action recognition method for extreme LR videos. We begin with knowledge distillation, which is employed to invigorate the two-stream network by transferring knowledge from the HR to LR models in Sec. 3.1. Subsequently, spatio-temporal feature extraction with the two-stream network is discussed in Sec. 3.2. Finally, a self attention network, which is designed to learn the temporal dependency across the frames, is described in Sec. 3.3. For easy reference, the overall procedures of the proposed method are illustrated in Fig. 1.

3.1. Knowledge Distillation

Training a deep network for extreme LR videos is arduous because even a slight change of view points can cause an object to be misidentified, leading to unstable decision boundary [6]. Inspired by the fact that HR models can be generated using publicly available HR videos, in contrast to previous methods that resort to data augmentation with various HR transformations to ease the two-stream network learning, here, we consider a different approach to resolve the training issue by transferring knowledge from the HR into LR network models with knowledge distillation [30].

Before conducting knowledge distillation, we employ an SR mechanism as a pre-processing step, which can enhance the quality of LR images, to bolster the learning capability of the LR networks. This mechanism is beneficial to make the network learn more discriminative features extracted from LR videos. In light of the success of Deep Back-Projection Network (DBPN) [22], which consists of several downsample and upsample layers representing the image degradation and key components of the images, respectively, we upsample the LR frames to obtain better image representation.

DBPN comprises of three main stages. The first stage is the initial feature extraction by two convolutional layers. Subsequently, the initial features are broadcast through a sequence of projection units, which in turn adjust the LR and targeted SR feature maps. Lastly, the SR images are reconstructed using the concatenated features across all upsampling projection units. Hence, the input LR images, which propagate through DBPN, can be reconstructed as SR images by a large scaling factor.

The HR network is regarded as a teacher network and the softmax value from this network is used as the ground truth of the LR network. This strategy allows the use of soft targets instead of hard targets. Class probabilities $\mathbf{Y} = \{y_1, \cdots, y_K\}$ are obtained from logits $\mathbf{X} = \{x_1, \cdots, x_K\}$ by a softmax function that can be expressed as [30]

$$y_i = \frac{e^{x_i/T}}{\sum_j e^{x_j/T}}, \quad i = 1, \cdots, K, \tag{1}$$

where T is a temperature variable that controls the softness of the probabilities over classes. The knowledge is thus transferred to the LR network regarded as a student network by training it on a transfer set using a soft target distribution generated by the teacher network with a high temperature. The student network is also trained to generate correct labels by using a weighted average of two objective functions [30]. The first objective function

Figure 3: Visualization of some images with different resolutions in the HMDB-51 and the IXMAS datasets.

is the cross entropy [31] with the soft targets and a high temperature while the other one uses the ground truth labels. For easy reference, this scheme is illustrated in Fig. 2.

3.2. Spatio-Temporal Feature Extraction

Making inference based on a single LR frame can lead to inaccurate action recognition caused by distorted spatial information. In light of this, we consider I3D [32], which processes a sequence of frames simultaneously to generate spatio-temporal features. I3D adopts inception mechanism and expands 2D filters into their 3D counterparts. Two-stream architecture is considered in I3D, where stacks of RGB and optical flow images are utilized as inputs. Here, we train the two-stream I3D using SR videos with knowledge distillation that leverages the knowledge from HR videos.

For every video, we partition the total number of frames, N, into P non-overlapping sequences of N_p consecutive frames. Next, for each sequence, we generate the spatio-temporal features from the last convolutional layer of I3D that is a tensor with a dimension of $N_t \times W \times H \times F$, where $N_t < N_p$ is the temporal dimension after applying temporal pooling within I3D layers, F is the dimension of the feature map channel, and W and H are the width and the height of the feature map, respectively. Thereafter, we concatenate the features in the time domain to obtain the final feature representation for every video, $\mathbf{V} \in N_s \times W \times H \times F$, where $N_s = P \times N_t$.

3.3. Spatial-Temporal Multi-Head Self-Attention Network

Self-attention mechanisms have been shown to be effective for learning long-term temporal dependency in various natural language processing tasks, see, *e.g.*, [9, 33, 34]. In contrast to RNN, each temporal position within self-attention can attend other distant positions directly so that it is easier to learn long-term temporal dependency. However, there are only limited studies on the benefits of self-attention in video understanding, where the spatio-temporal features are considered. Moreover, since the spatial information degrades in extreme LR, the capability to learn long-term temporal dependency is reduced as well. To account of this, a new spatial-temporal multi-head self-attention network is addressed to encapsulate long-term temporal dependency in the spatio-temporal features.

Our spatio-temporal multi-head self-attention, as illustrated in Fig. 1, is based on multi-head self-attention [9], where several self-attention layers are computed simultaneously. It is noteworthy that the original multi-head self-attention only processes a sequence of vectors

Table 1: Parameter settings for training I3D and the convolutional self-attention networks.

	I3D with Knowledge Distillation	Spatial-Temporal Self-Attentions
Pre-trained model	Kinetics+ImageNet [32]	-
T	5	-
Optimizer	Adam	Adam
Learning rate	0.0001	0.0002
Activation Function	ReLU	ReLU
Epoch	20	100

Table 2: Performance comparison of the proposed method with various mechanisms.

Two-stream	Super Resolution	Knowledge Distilation	Spatial-Temporal Self-Attention	Accuracy HMDB-51	Accuracy IXMAS
✓	-	-	-	52.61	93.89
✓	✓	-	-	53.92	94.44
✓	-	✓	-	54.26	94.87
✓	-	-	✓	54.31	94.44
✓	-	✓	✓	56.12	95.12
✓	✓	✓	-	56.67	95.56
✓	✓	✓	✓	**57.84**	**97.22**

and the relationships among the parallel self-attention heads are not considered. To circumvent these setbacks in videos, we utilize a 3D convolutional layer [32] to learn local spatio-temporal dependency. First, we partition the feature representation $\mathbf{V}$ and obtain a set of new feature representation $\mathbf{V}_A = \{\mathbf{V}^1, \cdots, \mathbf{V}^D\} \in \mathbb{R}^{D \times N_s \times (W \times H) \times (F/D)}$, where D is the number of parallel heads. The new feature representation $\mathbf{V}_A$ is then forwarded to the 3D convolutional layer that still preserves the original feature size. Afterward, the dimension of the feature map of the 3D convolutional layer is reshaped into $D \times (N_s \times W \times H) \times (F/D)$ and passed on to the parallel self-attention layers. The self-attention function applied to each head is defined as [9]:

$$S(\mathbf{V}^h) = \frac{B_1(\mathbf{V}^h)B_2(\mathbf{V}^h)^T}{\sqrt{dim(\mathbf{V}^h)}} B_3(\mathbf{V}^h) + B_3(\mathbf{V}^h), \tag{2}$$

where $h = 1, \cdots, D$, and B_1, B_2, and B_3 are linear projection layers that map every spatio-temporal positions into the same embedding space. Based on (2), we can simultaneously model the compatibility of every pair of spatio-temporal positions. The output of the self-attention for every parallel head is then concatenated and reshaped into $\mathbf{V}_B \in \mathbb{R}^{N_s \times (W \times H \times F)}$. Finally, $\mathbf{V}_B$ is passed on to the fully-connected and softmax layers to obtain the class probability. We use a late fusion strategy [35] to aggregate the features from the two streams.

4. Experimental Results and Discussions

4.1. Low Resolution Datasets

To create LR videos, the original videos in the HMDB-51 [36] and IXMAS [37] datasets are downsampled into 16×12 resolution using the average downsampling [4]. The HMDB-51 dataset consists of 6,765 videos and 51 action classes that encompasses several viewpoints, high inter-class similarity, background clutter, *etc*. Meanwhile, The IXMAS dataset is comprised of 1,800 videos and 11 action classes that are incurred by occlusion, inter-class variation problem, and strong viewpoint changes. From Fig. 3, we can observe the differences between HR and LR images, where the latter are generally more difficult to analyze because of a significant loss of the visual information.

4.2. Experimental Setup

The learning process consists of training the two-stream network using the knowledge distillation and the self-attention network, which are conducted separately. For reference, the hyper-parameters used in these two learning procedures are summarized in Table 1. Following [32], we set N_p and F as 64 and 1,024, respectively, and the kernel size of the spatial-temporal multi-head self-attention as $[1, 3, 5]$ with a stride of 1. As [9], we use D = 8 and a drop out rate of 0.9. We generate the LR testing images using the same procedure as [8] and use the evaluation metrics provided by [36, 37] for HMDB-51 and IXMAS, respectively.

4.3. Ablation Studies

Impact of SR: First, we inspect the performance improvement with the SR mechanism, as shown in Table 2, from which we can note that by utilizing the SR mechanism to augment the training data, the accuracy of the two-stream network is improved by about 1.3% and 0.5% on HMDB-51 and IXMAS, respectively. This is because this mechanism can provide richer visual information that can facilitate the training on LR videos.

Impact of Knowledge Distillation: Next, we examine the impact of knowledge distillation on the performance of the proposed approach. As presented in Table 2, we can note that the knowledge distillation can enhance the performance of the two-stream network by about 1.9% and 0.3% on HMDB-51 and IXMAS. This is because this scheme can help the network learn more discriminative features by transferring the knowledge from HR into LR models. We can also observe that the knowledge distillation can further improve the performance of the SR assisted two-stream network by about 3% and 1% on HMDB-51 and IXMAS, respectively. This is because the SR mechanism can provide training data with richer details close to HR models to facilitate knowledge transfer.

Impact of Spatial-Temporal Multi-Head Self-Attention: We evaluate the impact of the spatial-temporal multi-head self-attention network on the performance of the proposed method, as shown in Table 2, from which we can see that even without the knowledge transfer from HR model and the training data augmentation from SR, the spatio-temporal multi-head self-attention network can still improve the action recognition results of the two-stream network by about 1.7% and 0.5% on HMDB-51 and IXMAS, respectively. This performance gain indicates that temporal dependency information is indeed essential to recognize actions from LR videos, in which the spatial information is extremely impaired. The performance is further improved by about 1.8% and 0.2% on HMDB-51 and IXMAS, respectively, when the spatio-temporal multi-head self-attention is trained with the data generated by SR. This is because the training data is richer in visual

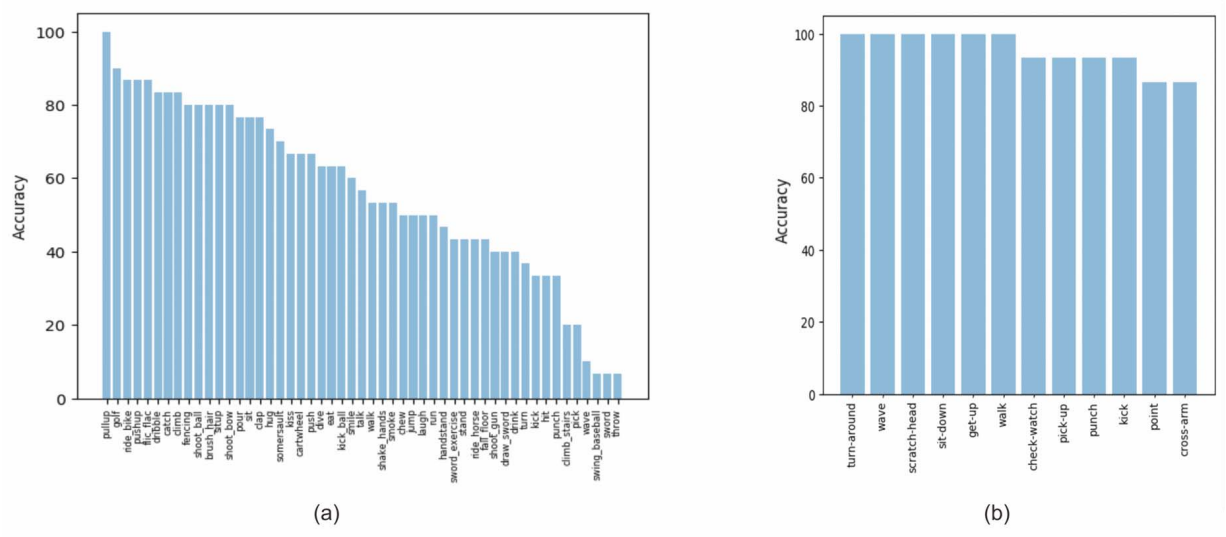

Figure 4: The accuracy of the proposed method for each action on HMDB-51 and IXMAS.

Figure 5: Some snapshots of the (a) successful and (b) failure cases.

information so the network model can work better on LR videos. The most significant improvement by the spatio-temporal multi-head self-attention is achieved by the integration with all other schemes. This is because the spatio-temporal multi-head self-attention can learn the long-term temporal dependency information with more discriminative features obtained from the student network.

Based on the above observations, in the following simulations, the proposed method is equipped with the SR, knowledge distillation, and spatial-temporal multi-head self-attention mechanisms.

Assessment of the Proposed Method: We also scrutinize the accuracy of the proposed method for each class. The action recognition accuracy for each action class in HMDB-51 and IXMAS is depicted in Figs. 4 (a) and 4 (b), respectively. We can notice from Fig. 4 (a) that the proposed method is able to provide accurate recognition for periodic actions such as 'dribble' and 'clap' and non-periodic actions including 'shoot ball' and 'brush hair.' Also, our approach performs well on facial actions that contain small movement, which is difficult to recognize in many LR videos, such as 'chew', 'kiss', and 'eat'. However, the actions such as 'throw', 'sword', 'swing baseball' can not be precisely recognized, as these actions have similar sub-action movements that resemble other classes, *e.g.* 'sword' and 'fencing'. Meanwhile, as shown in Fig. 4 (b), the proposed method is doing

Table 3: Comparison of the action recognition accuracy on the low resolution HMDB-51 and IXMAS datasets.

Methods	Modalities	Accuracy	
		HMDB-51	IXMAS
pLRN+Tennet [1]	RGB	21.7	-
ISR [6]	RGB	28.68	-
Dai et al. [12]	RGB	-	80
Semi-Coupled [7]	RGB and Optical Flow	29.2	93.7
Rahman et al. [5]	RGB and Texture	34.57	-
Multi-Siamese [4]	RGB and Optical Flow	37.7	-
Fully-Coupled [8]	RGB and Optical Flow	44.96	-
I3D [31]	RGB and Optical Flow	52.61	93.89
Ours	RGB and Optical Flow	57.84	97.22

well on most of the action classes in IXMAS, except for 'point' and 'cross arm,' which have hand movement similar to other classes. As an illustration, we also provide snapshots of some successful and failure recognition results, as shown in Fig. 5 (a), from which we can see that 'pull up' and 'ride bike' can be well recognized because these type of videos still provide decent spatial information. On the other hand, failure cases happen when the videos contain group actions as in 'sword' and the images are severely distorted as in 'throw,' as depicted in Fig. 5 (b).

4.4. Comparison with the State-of-the-Art Works

In this subsection, we compare the proposed method with some state-of-the-art works, including pLRN+Tennet [1], ISR [6], Dai *et al.* [12] Semi-Coupled [7], Rahmat *et al.* [5], Multi-Siamese [4], Fully-Coupled [8], and I3D [32] on the LR HMDB-51 dataset, where [32] is trained on the LR dataset. From Table 3 we can see that pLRN+Tennet [1] exhibits inferior performance, as this approach trades speed for accuracy. ISR [6] attains better accuracy by using different types of subpixel transformations from HR frames. By implementing joint training with the same filters for both of the HR and LR networks, Semi-Coupled [7] yields slightly better performance than ISR. We can also find that [5] is superior to the previous methods by combining the textural information with the shape-motion features. Multi-Siamese [4] provides even higher accuracy by integrating the two-stream Siamese network with the pyramid pooling. As Fully-Coupled [8] is focused on modelling temporal dependency information by applying GRU directly on C3D, it further improves the action recognition accuracy. I3D [32], which adopts the inception mechanism and a longer sequence of frames than C3D, substantially outperforms the aforementioned methods. Our proposed method surpasses the state-of-the-art works by employing the potent I3D features along with efficacious knowledge transfer by knowledge distillation and the long-term temporal dependency acquired by spatial-temporal multi-head self-attention.

Next, we make a comparison on the LR IXMAS dataset, as shown in Table 3, from which we can find that [12] yields the worst action recognition accuracy because it only utilizes a pixel-wise time-series algorithm on gray-scale video sequences. Also, the performance of [7] exceeds that of [12] by using semi-coupled networks that jointly optimize both of the HR and LR networks. Again, our approach that benefits from the knowledge transfer from HR and learns long-term dependency information provides the best performance.

5. Conclusions

This paper has developed an effective framework for action recognition in extreme LR videos, which integrates a two-stream network with a new spatial-temporal multi-head self-attention mechanism. Also, an SR mechanism is considered to enhance the degenerated visual information in LR videos. Additionally, the two-stream network is trained by taking advantage of the HR model with a knowledge distillation scheme. With such a combination, the new two-stream network can effectively learn long-term temporal dependency to achieve better recognition accuracy. Simulation shows that the proposed approach excels the state-of-the-art methods on the common HMDB-51 and IXMAS datasets.

Acknowledgment

This work was supported by the Ministry of Science and Technology, R.O.C. under contracts MOST 107-2221-E-011-124 and MOST 107-2221-E-011-078-MY2.

References

[1] Tingzhao Yu, Lingfeng Wang, Chaoxu Guo, Huxiang Gu, Shiming Xiang, and Chunhong Pan. Pseudo low rank video representation. *Pattern Recognition*, 85:pages 50–59, 2019.

[2] Hong-Kai Chen, Xiao-Guang Zhao, Shi-Ying Sun, and Min Tan. PLS-CCA heterogeneous features fusion-based low-resolution human detection method for outdoor video surveillance. *International Journal of Automation and Computing*, 14(2):136–146, 2017.

[3] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*, 2018.

[4] Michael S Ryoo, Kiyoon Kim, and Hyun Jong Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2018.

[5] Saimunur Rahman, John See, and Chiung Ching Ho. Exploiting textures for better action recognition in low-quality videos. *EURASIP Journal on Image and Video Processing*, 2017(1):page 74, 2017.

[6] Michael S Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pages 4255–4262, 2017.

[7] Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 139–147, 2017.

[8] Mingze Xu, Aidean Sharghi, Xin Chen, and David J Crandall. Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1607–1615, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[11] Saimunur Rahman, John See, and Chiung Ching Ho. Deep CNN object features for improved action recognition in low quality videos. *Advanced Science Letters*, 23(11):11360–11364, 2017.

[12] Ji Dai, Jonathan Wu, Behrouz Saghafi, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–76, 2015.

[13] Christian Herrmann, Dieter Willersinn, and Jürgen Beyerer. Low-resolution convolutional neural networks for video face recognition. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 221–227, 2016.

[14] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[15] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, page 132, 2018.

[16] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the European Conference on Computer Vision*, pages 620–636, 2018.

[17] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6026–6035, 2018.

[18] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):pages 295–307, 2016.

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[20] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017.

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[22] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018.

[23] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1012–1020, 2017.

[24] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[25] Amor Ben Tanfous, Hassen Drira, and Boulbaba Ben Amor. Coding kendall's shape trajectories for 3D action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2018.

[26] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun. Attention in convolutional LSTM for gesture recognition. In *Proceedings of the Neural Information Processing Systems*, pages 1953–1962, 2018.

[27] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[28] Ting Zhang, Guo-Jun Qi, Bin Xiao, and Jingdong Wang. Interleaved group convolutions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4373–4382, 2017.

[29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proceedings of the Neural Information Processing Systems Workshop*, 2015.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[32] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[34] Shang Gao, Arvind Ramanathan, and Georgia Tourassi. Hierarchical convolutional attention networks for text classification. In *Proceedings of The Workshop on Representation Learning for Natural Language Processing*, pages 11–23, 2018.

[35] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[36] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2556–2563, 2011.

[37] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *Proceeding of the European Conference on Computer Vision*, pages 635–648, 2010.