Cross-domain Knowledge Transfer Schemes for 3D Human Action Recognition

Athanasios Psaltis
Centre for Research and Technology Hellas
at.psaltis@iti.gr

Georgios Th. Papadopoulos
Centre for Research and Technology Hellas
papad@iti.gr

Petros Daras
Centre for Research and Technology Hellas
daras@iti.gr

***Abstract*—Previous work in 3D human action recognition has been mainly confined to schemes in a single domain, exploiting in principle skeleton-tracking data, due to their compact representation and efficient modeling of the observed motion dynamics. However, in order to extend and adapt the learning process to multi-modal domains, inevitably the focus needs also to be put on cross-domain analysis. On the other hand, attention schemes, which have lately been applied to numerous application cases and exhibited promising results, can exploit the intra-affinity of the considered modalities and can then be used for performing intra-modality knowledge transfer, *e.g.* to transfer domain-specific knowledge of the skeleton modality to the flow one and vice verca. This study investigates novel cross-modal attention-based strategies to efficiently model global contextual information regarding the action dynamics, aiming to contribute towards increased overall recognition performance. In particular, a new methodology for transferring knowledge across domains is introduced, by taking advantage of the increased temporal modeling capabilities of Long Short Term Memory (LSTM) models. Additionally, extensive experiments and thorough comparative evaluation provide a detailed analysis of the problem at hand and demonstrate the particular characteristics of the involved attention-enhanced schemes. The overall proposed approach achieves state-of-the-art performance in the currently most challenging public dataset, namely the NTU RGB-D one, surpassing similar uni/multi-modal representation schemes.**

***Index Terms*—Action recognition, attention schemes, deep learning.**

I. Introduction

The ability of computers to recognize human actions is of paramount importance because of the wide range of possible applications (*e.g.* ranging from surveillance and robotics to health-care, entertainment and e-learning). The latter drew the attention of the research community, which has devoted over the past decades considerable resources to achieve credible solutions. However, despite the significant efforts that have been made and the abundance of published methods, the problem continues to face inherent challenges, such as the large intra-class or viewpoint variations.

The incorporation of depth sensors into conventional low-cost RGB cameras has led research advancements, beyond well-established image analysis techniques (*i.e.* appearance-based representations), to the introduction of methods that exploit either domain-specific knowledge or the complimentary of different modalities (*e.g.* skeleton, depth, 3D flow) [13], [14]. Such modalities provide robustness to illumination and appearance variations, by enriching the original RGB feature space with depth information that can be conducive for resolving ambiguous actions.

Deep Learning (DL) has contributed a huge boost to the already rapidly evolving field of computer vision. Inevitably, DL techniques have recently been applied in the field of 3D human action recognition, aiming at efficiently modeling the observed complex motion dynamics. These have been experimentally shown to significantly outperform the corresponding hand-crafted-based approaches. The majority of literature methods rely only on the use of skeleton-tracking data (*i.e.* tracked human skeleton). Several approaches have been proposed using variants of Recurrent Neural Networks (RNNs), which adapt the architecture design towards efficiently exploiting the physical structure of the human body or employ gating mechanisms for controlling the spatio-temporal pattern learning process [9], [10]. Despite the suitability of RNNs in modeling time-evolving procedures, recently CNN-based architectures have also been introduced [8], [15]. Additionally, DL techniques have also been applied to depth-based action-recognition problems, making extensive use of depth maps for estimating a representation of the human silhouette and subsequently modeling the action dynamics [17], [23], [25]. On the other hand, flow methods combine depth with RGB information for estimating more discriminative representations (namely 3D flow fields) that enable the focus of the analysis procedure on the areas where action has been observed [12], [16], [24].

Until now, very few works have concentrated on the problem of multi-modal 3D action recognition. Shahroudy *et al.* [19] propose a deep auto-encoder that performs common component analysis at each layer (*i.e.* factorizes the multi-modal input features into their shared and modality-specific components) and discovers discriminative features, taking into account the RGB and depth modalities. The latter again puts emphasis on the spatial domain analysis, relying on the initial extraction of hand-crafted features, while the method is not applicable to view-invariant recognition scenarios. In [26], a CNN architecture is presented that combines RGB and depth features, by jointly optimizing a ranking and a softmax loss. Zhao *et al.* [30] propose a two stream RNN/CNN scheme, which separately learns an RNN model, using skeleton data, along with the convolution-based model, trained using RGB

978-9-0827-9703-9/19/$31.00

information, and lately fuses the obtained features. From the above analysis, it can be deduced that the 3D action recognition-related literature has in principle concentrated on single-modality analysis, while the respective cross-modal and knowledge transfer techniques have been poorly examined, *i.e*. leaving a great potential for further performance improvement unexplored.

A recent trend in DL research are the so called attention mechanisms, which target the mimicking of the original visual attention mechanisms found in humans. Such models are designed to adaptively adjust the analysis focus (*e.g*. focus on a certain region of an image, while perceiving the surrounding image as a low-resolution background), providing also the capability to interpret and visualize what the model is learning. Inspired by the observations over human visual cognition, recent studies adopt attention mechanisms in action recognition tasks. In particular, a spatio-temporal attention based mechanism is introduced in [1] that is able to automatically focus on the hands, in order to detect the most discriminative parts of the observed action. The latter handles attention in a recurrent manner, by employing RNNs. Liu *et al*. [10] propose a class of LSTMs, termed the Global Context-Aware Attention LSTM (GCA-LSTM), which is able to selectively focus on the informative joints in the action sequence with the assistance of global contextual information. Additionally, an end-to-end spatial and temporal attention model is presented in [21], which learns to adaptively put emphasis on discriminative skeleton joints within each frame. Taking into account the above analysis it can be observed that attention-based approaches have shown considerable achievements in single domain analysis, while their potentials in cross-domain scenarios have not been explored yet.

The current study explores the problem of 3D human action recognition using DL techniques, realizing a cross-domain knowledge transfer approach. Different attention-based strategies are investigated for incarnating modeling knowledge in one domain and transferring/re-using it to/in a different one, focusing in principle on exploiting skeleton-tracking data for guiding the corresponding 3D flow modeling process. In particular, a spatio-temporal attention mechanism with informativeness gates is designed to adaptively adjust the analysis focus on different frame patches. The latter takes advantage of the increased temporal modeling capabilities of Long Short Term Memory (LSTM) models. Additionally, extensive experiments and thorough comparative evaluation provide detailed insights to the problem at hand and demonstrate the particular characteristics of the involved attention-enhanced schemes. The proposed approach achieves state-of-the-art performance in the currently broadest and most challenging public dataset, namely the NTU RGB-D [18] one, surpassing similar uni/multi-modal representation schemes.

The remainder of the paper is organized as follows: Attention-enhanced action recognition strategies are presented in Section II. Experimental results are discussed in Section III, while conclusions are drawn in Section IV.

II. Attention-enhanced action recognition

A. Single modality analysis

Prior to the application of cross-modal analysis processes, single-modality analysis is realized for each information source. More specifically, for skeleton-based analysis, the work of [9] is adopted, where spatial dependencies among joints and temporal correlations among frames are modeled at the same time, using a so called Spatio-Temporal LSTM (ST-LSTM) mechanism. For the particular case of depth- and flow-based analysis, a template matching approach, presented in the study of [16], is selected that learns spatio-temporal features from videos, by applying 3D convolutions. For performing action recognition, a composite 3D CNN-LSTM architecture is adopted, where an individual LSTM network is introduced for every considered modality. More specifically, the introduced flow and depth representations are computed by considering the spatio-temporal features from the last FC layer of the 3D CNN model [16]. The developed single-modality LSTMs are trained to predict the observed action class at every video segment, while for estimating an aggregated probability for each action for the entire video sequence, simple averaging of all corresponding probability values of all clips is performed. Multi-layer LSTMs are used in this work for efficiently encoding more long-term correlations in the input data.

B. Cross-domain knowledge transfer

Different modalities exhibit particular characteristics with respect to the motion dynamics that they encode. To this end, a truly robust action recognition system should combine multiple information sources. In this respect, several attention strategies are investigated for the integration of a complementary information stream, mainly differing at the level that attention-based modulation is applied, namely a) before (A-$LSTM_{before}$), where the attention vector is applied to the LSTM's inputs, b) after (A-$LSTM_{after}$), where the attention vector is applied to the LSTM's output, or c) using a gating mechanism inside (A-$LSTM_{gating}$) the LSTM layer. Among the three models, the A-$LSTM_{gating}$ is applied directly to the LSTM unit, controlling the spatio-temporal learning process, while achieving faster convergence and improved recognition performance. The introduced attention-enhanced scheme (A-$LSTM_{gating}$) exhibits the following advantageous characteristics: a) it retains the sequential modeling ability of the original LSTM, while reinforcing its selective attention capability, by introducing a global context memory cell derived from complementary modalities, and b) it simultaneously takes into account multi-modal information, by extending the design of the attention model in a multi-stream fusion scheme.

The proposed attention mechanism (A-$LSTM_{gating}$) utilizes informativeness gates to adaptively assign diverse levels of attention to different frames. In order to reliably identify the discriminant parts of an action, an informativeness score for the whole action sequence can be derived from complementary modalities. Inspired by the work of [10], a global context memory cell for the LSTM is defined, which maintains the

global contextual information of the action sequence that is in turn fed to subsequent LSTM processing steps. The global context memory cell models an overall representation of the whole action sequence and determines the degree of importance of each frame. The proposed attention mechanism receives as input the global context information that is derived from the first stream, denoted as Gating Signal (GS), and then appropriately modulates the processing of the second stream, denoted as Processing Signal (PS). Both GS and PS signals are modeled using LSTMs that are denoted $LSTM_{GS}$ and $LSTM_{PS}$, respectively. In particular, $LSTM_{GS}$ is used for encoding the action sequence dynamics (*i.e.* generating an attention mask) and initializing the global context memory cell, while applying different levels of attention over the inputs of $LSTM_{PS}$ (*i.e.* refining the LSTM input).

In order to better illustrate the core functionality of the proposed attention mechanism, let $\mathbf{H}'(t)$ be the LSTM state vector of $LSTM_{GS}$ at time t (*i.e.* the modality that initializes the global context memory) and $\mathbf{H}(t)$ the hidden representation of $LSTM_{PS}$ (*i.e.* the stream that is modulated by $LSTM_{GS}$). In order to initialize the global memory value $\mathbf{A}$, the average of the internal state values $\mathbf{H}'(t)$ is used, as follows:

$$\mathbf{A} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}'(t), \tag{1}$$

where T denotes the total number of frames. The informativeness degree of the input is assessed at every step in $LSTM_{PS}$. In particular, let $\mathbf{W}_1$ and $\mathbf{W}_2$ be affine transformations, comprising key model parameters (learnable weight matrices). The network's objective is to estimate an informativeness gate $\mathbf{E}(t)$ for each input value $\mathbf{H}'(t)$, by feeding the input itself and the global context memory $\mathbf{A}$ to the $LSTM_{PS}$ unit, according to the following formalization:

$$\mathbf{E}(t) = \mathbf{W}_1(\tanh(\mathbf{W}_2(\frac{\mathbf{H}'(t)}{\mathbf{A}}))) \tag{2}$$

$$\mathbf{R}(t) = \frac{\exp(\mathbf{E}(t))}{\sum_{q=1}^{T}\exp(\mathbf{E}(q))} \tag{3}$$

where $\mathbf{R}(t)$ is the normalized informativeness score of input $\mathbf{H}'(t)$, in the interval $(0, 1)$, so that it can be interpreted as probability. Using the learnt informativeness gate, the cell state $\mathbf{C}(t)$ of the $LSTM_{PS}$ unit can be updated as:

$$\mathbf{C}(t) = \mathbf{F}(t)\mathbf{C}(t-1) + \mathbf{I}(t)\mathbf{G}(t) \tag{4}$$

$$\mathbf{C}(t) = (1 - \mathbf{R}(t))\mathbf{F}(t)\mathbf{C}(t-1) + \mathbf{R}(t)\mathbf{I}(t)\mathbf{G}(t) \tag{5}$$

where $\mathbf{C}(t-1)$ is the 'internal memory' of the $LSTM_{PS}$ at previous time step and the gates $\mathbf{I}(t)$ and $\mathbf{F}(t)$ control the degree to which the memory accumulates new input $\mathbf{G}(t)$ and attenuates its memory, respectively.

By comparing the cell state equations (4) and (5), it can be easily observed that the proposed informativeness gate $\mathbf{R}(t)$ controls the degree to which the memory accumulates new input. More specifically, if the input $\mathbf{H}'(t)$ is in accordance with the global context memory (2) (*i.e.* normalized informativeness score > 0.5), then the learning algorithm updates the memory cell of $LSTM_{PS}$. On the contrary, if the input $\mathbf{H}'(t)$ is irrelevant (*i.e.* normalized informativeness score < 0.5), then its effect is suppressed.

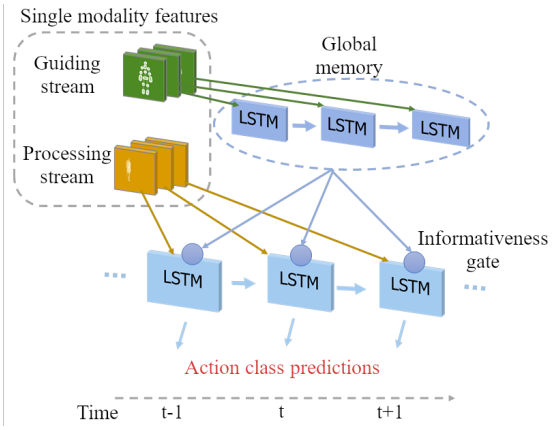

Fig. 1: Attention-enhanced scheme $A\text{-}LSTM_{gating}$: The $LSTM_{GS}$ generates an attention mask and initializes the global context memory cell, while the $LSTM_{PS}$ applies different levels of attention over the input.

Fig. 1 depicts the proposed attention scheme $A\text{-}LSTM_{gating}$ reinforced with the informativeness gate, which can be trained similarly to other LSTM unit gates, using back-propagation. The proposed multi-modal attention-based technique is generic and can be applied with any possible modality combination (*e.g.* $A\text{-}LSTM^{s\rightarrow f}_{gating}$, $A\text{-}LSTM^{f\rightarrow s}_{gating}$, $A\text{-}LSTM^{s\rightarrow d}_{gating}$, $A\text{-}LSTM^{f\rightarrow d}_{gating}$, *etc.*). Methods indicated with superscript 's', 'c', 'd' and 'f' incorporate skeleton, color, depth and flow data, respectively, while superscript $\rightarrow$ denotes the modulation direction between the two involved modalities. In this way, cross-domain knowledge transfer is realized (*i.e.* enabling the transfer of knowledge across modalities in different domains), which facilitates both more efficient action pattern modeling and faster Neural-Network (NN) learning.

III. Experimental Results

A. Dataset and implementation details

In this section, experimental results as well as comparative evaluation from the application of the proposed attention-enhanced method are presented. For the evaluation, the 'NTU RGB+D' [18] dataset was used (Table I), which supports a total of 60 action types. A set of 64 frames were uniformly selected for feature extraction, which roughly corresponds to one third of the average number of frames per action. With respect to the implemented attention-enhanced approaches, the 'Torch[1]' scientific computing framework and a Nvidia Tesla K40 GPU were used. Zero-mean Gaussian distribution with standard deviation equal to 0.01 was used to initialize all NN weight and bias matrices. All class predictions were passed

[1]http://torch.ch/

	Method	Accuracy	
		Cross-subject	Cross-view
a)	Depth	69.57%	71.29%
	Skeleton	66.87%	73.59%
	Flow	**76.43%**	**82.96%**
b)	$A\text{-}LSTM^{s\to f}_{before}$	77.34%	83.79%
	$A\text{-}LSTM^{s\to f}_{after}$	77.82%	84.11%
	$A\text{-}LSTM^{s\to f}_{gating}$	**82.16%**	**88.48%**
	$A\text{-}LSTM^{f\to s}_{gating}$	72.85%	81.26%
	$A\text{-}LSTM^{f\to f}_{gating}$	79.53%	86.10%
	$A\text{-}LSTM^{s\to s}_{gating}$	71.26%	79.73%
	$A\text{-}LSTM^{s\to d}_{gating}$	75.49%	77.74%
c)	HBRNN-L^s [2]	59.07%	63.97%
	SFAM^f [24]	57.36%	59.14%
	LieNet^s [3]	61.37%	66.95%
	P-LSTM^s [18]	62.93%	70.27%
	Atomic3DFlow^d [12]	66.20%	-
	ST-LSTM^s [9]	69.20%	77.70%
	JL_d^s [29]	70.26%	82.39%
	Two-Stream RNN^s [22]	71.30%	79.50%
	Conv3D-Flow^f [16]	73.27%	79.64%
	DSSCA-SSLM^{cd} [19]	74.86%	-
	GCA-LSTM^s [10]	74.40%	82.80%
	Res-TCN(Temporal Conv)^s [6]	74.30%	83.10%
	Adaptive Tree^s [7]	74.60%	83.20%
	MTLN^s [5]	79.60%	84.80%
	ResNet-56^s [15]	78.20%	85.60%
	View invariant CNN^s [11]	80.03%	87.21%
	VA-LSTM^s [28]	79.20%	87.70%
	ST-GCN^s [27]	81.50%	88.30%
	Two-Stream CNN^s [8]	83.20%	89.30%
	DPRL+GCNN^s [20]	83.50%	89.80%
	STA-Hands^{cs} [1]	82.50%	88.60%
	Mul-Score fusion^{cs} [30]	82.89%	90.1%
	Proposed approach^{sf}	82.16%	88.48%
d)	Slow^{sdf}	78.62%	83.86%
	Late^{sdf}	81.94%	88.36%
	A-Slow^{sdf}	85.65%	90.94%
	A-Late^{sdf}	**86.46%**	**92.47%**

TABLE I: Action recognition results in NTU: a) Single-modality analysis, b) Attention-enhanced analysis, c) Comparative evaluation, and d) Attention-enhanced multi-modal analysis. Underlined methods indicate multi-modal schemes. Methods indicated with superscript 's', 'c', 'd' and 'f' incorporate skeleton, color, depth and flow data, respectively.

through a softmax operator (layer) to estimate a probability distribution over the supported actions. Stochastic Gradient Descent (SGD) was used during training, along with a multinomial logistic loss function. The batch size was set equal to 256, while the momentum value was equal to 0.9. Weight decay with value 0.0005 was used for regularization, while the training procedure lasted 30 epochs. Additionally, an adaptive learning rate approach [4] was followed during training.

B. Attention schemes evaluation

Quantitative evaluation from the application of the proposed cross-modal attention schemes (Section II) is provided in Table I. The exhibited results [group (b) in Table I] suggest that incorporating additional information inside the LSTM unit improves the overall classification performance by approximately 5.52%, compared to the best performing single-modality results (3D flow-based ones); highlighting the importance of the gating mechanism (*i.e*. the informativeness gate). Among the three proposed attention schemes, $A\text{-}LSTM_{gating}$ is shown to be the best performing one, mainly due to its advantageous characteristic of combining the attending ability of the gating mechanism with the increased temporal modeling capabilities of LSTM, *i.e*. directly affecting the LSTM unit state during the learning process. Examining the behavior of the attention-enhanced schemes in more details, it can be observed that performance is maximized when attention information from the skeleton modality is used to guide the training of the flow modality $A\text{-}LSTM^{s\to f}_{gating}$. Intuitively, this is mainly due to the elegant combination of domain-specific features (skeletal joints) with highly rich and expressive features (3D flow vectors) that are highly complementary in nature. Additionally, it can be seen that different attention combinations (*e.g*. $A\text{-}LSTM^{s\to s}_{gating}$ and $A\text{-}LSTM^{f\to f}_{gating}$) perform reasonably well, even if the same information stream is used for both guiding and learning.

C. Comparative evaluation

The proposed action recognition schemes are also comparatively evaluated with numerous methods of the literature [third (c) group of experiments in Table I]. From the presented methods, multi-modal information processing is realized in the works of DSSCA-SSLM [19] (color, depth), STA-Hands [1] (color, skeleton) and Mul-Score fusion (LSTM+CNN) [30] (color, skeleton). The remaining single-modality methods make use of depth (Atomic3DFlow [12]), 3D flow (SFAM [24] and Conv3D-Flow [16]) or skeleton-tracking [HBRNN-L [2], LieNet [3], P-LSTM [18], ST-LSTM [9], JL_d [29], Two-Stream RNN [22], GCA-LSTM [10], Res-TCN (Temporal Conv) [6], Adaptive Tree [7], MTLN [5], ResNet-56 [15], View invariant (Synthesized CNN) [11], VA-LSTM [28], ST-GCN [27], Two-Stream CNN (Motion+Trans) [8] and DPRL+GCNN [20]] data. The recognition performance of literature methods is indicated as reported in [3], [12], [15], [16], [19], [29]. Overall, from the presented results, it can be seen that $A\text{-}LSTM^{s\to f}_{gating}$ compares favorably with most uni/multi-modal techniques (often surpassing them with a large margin), despite the fact of not using the best-performing literature representation for the skeleton modality. This justifies the fundamental claim of the current work that for achieving improved action recognition results, knowledge transfer across different domains is beneficial.

The fundamental aim of the proposed attention schemes is to model, transfer and re-use knowledge across different domains. However, in order to realize a fair comparison with the respective multi-modal literature approaches, these attention schemes need to be inevitably combined with typical multi-modal fusion strategies. In this respect, conventional slow and late fusion schemes (similar to the ones presented in

[4]) were implemented and the obtained results are indicated in Table I [fourth (d) group of results]. In particular, for the case of slow fusion (Slow^{sdf}), a composite multi-layer LSTM was developed, by introducing additional layer(s) on top of the single-modality ones; the additional LSTM layer(s) received as input a composite vector that resulted from the simple concatenation of the state signals. On the contrary, the late fusion approach (Late^{sdf}) aggregated uni-modal LSTM state vectors in a given time window, while adopting a fully-connected-based mechanism for exploiting correlations among the features of the involved modalities. The above-mentioned slow and late fusion schemes were also combined with the best performing attention mechanisms (methods A-Slow^{sdf} and A-Late^{sdf} in Table I), where instead of the original single-modality data the attention-enhanced counterparts $A\text{-}LSTM^{s\rightarrow f}_{gating}$, $A\text{-}LSTM^{s\rightarrow d}_{gating}$ and $A\text{-}LSTM^{f\rightarrow s}_{gating}$ were used. From the presented results, the following main observations can be made: a) the attention-enhanced fusion schemes perform better than the proposed attention approaches alone, with A-Late^{sdf} being the best multi-modal fusion scheme, b) the attention-enhanced fusion schemes exhibit significantly improved performance compared to the conventional fusion ones (namely Slow^{sdf} and Late^{sdf}). Moreover, it can be seen that A-Late^{sdf} exhibits state-of-the-art performance. The latter observations again highlight the significant added value of incorporating spatio-temporal attention mechanisms in the realization of 3D human action recognition.

IV. Conclusions

In this study, the problem of 3D human action recognition using DL techniques was investigated following a cross-domain knowledge transfer approach. In particular, different spatio-temporal attention mechanisms, controlled by informativeness gates, were introduced to adaptively adjust the analysis focus on different frames. Extensive experiments and thorough comparative evaluation were reported. The overall proposed approach accomplished state-of-the-art performance in the currently broadest and most challenging public dataset, namely the NTU one. Future work includes the investigation of incorporating more sophisticated attention mechanisms that will support training with multi-modal data streams, while evaluating with less or just one.

Acknowledgment

The work presented in this paper was supported by the European Commission under contract H2020-787061 ANITA.

References

[1] F. Baradel, C. Wolf, and J. Mille. Human action recognition: Pose-based attention draws focus to hands. *CoRR*, abs/1712.08002, 2017.

[2] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[3] Z. Huang, C. Wan, T. Probst, and L. Van Gool. Deep learning on lie groups for skeleton-based action recognition. *CVPR*, 2017.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. 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.

[5] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4570–4579, July 2017.

[6] T. S. Kim and A. Reiter. Interpretable 3d human action analysis with temporal convolutional networks. *CoRR*, abs/1704.04516, 2017.

[7] W. Li, L. Wen, M. Chang, S. N. Lim, and S. Lyu. Adaptive rnn tree for large-scale human action recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1453–1461, Oct 2017.

[8] H. Liu, J. Tu, and M. Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. *CoRR*, abs/1705.08106, 2017.

[9] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[10] J. Liu, G. Wang, L. Duan, P. Hu, and A. C. Kot. Skeleton based human action recognition with global context-aware attention LSTM networks. *CoRR*, abs/1707.05740, 2017.

[11] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.*, 68(C):346–362, Aug. 2017.

[12] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. *CVPR*, 2017.

[13] G. T. Papadopoulos, A. Axenopoulos, and P. Daras. Real-time skeleton-tracking-based human action recognition using kinect data. In *International Conference on Multimedia Modeling*, pages 473–483. Springer, 2014.

[14] G. T. Papadopoulos and P. Daras. Human action recognition using 3d reconstruction data. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(8):1807–1823, Aug 2018.

[15] H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Exploiting deep residual networks for human action recognition from skeletal data. *Computer Vision and Image Understanding*, 170:51–66, 2018.

[16] A. Psaltis, G. T. Papadopoulos, and P. Daras. Deep 3d flow features for human action recognition. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, Sep. 2018.

[17] H. Rahmani and A. Mian. 3D action recognition from novel viewpoints. In *CVPR, June*, 2016.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *Technical Report, arXiv preprint arXiv:1603.07120*, 2016.

[20] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Adaptive spectral graph convolutional networks for skeleton-based action recognition. *CoRR*, abs/1805.07694, 2018.

[21] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *CoRR*, abs/1611.06067, 2016.

[22] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. *CoRR*, abs/1704.02581, 2017.

[23] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3D human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 97–106. ACM, 2014.

[24] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. *CVPR*, 2017.

[25] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. *in ChaLearn Looking at People (LAP) Challenge, ICPR*, 2016.

[26] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu. Cooperative training of deep aggregation networks for RGB-D action recognition. *CoRR*, abs/1801.01080, 2018.

[27] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *CoRR*, abs/1801.07455, 2018.

[28] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *CoRR*, abs/1703.08274, 2017.

[29] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *Proc. IEEE Winter Conference on Application of Computer Vision*, 2017.

[30] R. Zhao, H. Ali, and P. van der Smagt. Two-stream RNN/CNN for action recognition in 3d videos. *CoRR*, abs/1703.09783, 2017.