Volume: 12 Issue: I Month of publication: January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58130

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Unified Spatio-Temporal Attention Models for Advanced Human Action Recognition & Detection

Midhilesh Momidi[1], Priyanka Jadhav[2], Shrey[3], Astha Patel[4], Aaditya Modi[5]
[1]*Senior Machine Learning Engineer, Dell Technologies, Bangalore, India*
[2]*Deputy Manager - Data Analytics and PMO, Cipla Ltd, Mumbai, India*
[3]*Systems Engineer, Tata Consultancy Services, New Delhi, India*
[4]*Undergraduate Student, (Information and Communication Technology), Pandit Deendayal Energy University, Gujarat, India*
[5]*Undergraduate Student, (Computer Engineering), Institute of Advanced Research, Gujarat, India*

Abstract: This research paper introduces innovative approaches to enhance human action analytics in computer vision through the development of spatial and temporal attention models. The first model is based on Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units, aiming to extract discriminative spatial and temporal features from skeleton data for improved human action recognition and detection. The model selectively focuses on discriminative joints within each frame, utilizing a regularized cross-entropy loss for effective training. Additionally, a joint training strategy is proposed to optimize the learning process. The second model, a spatio-temporal attention (STA) network, is designed to address the limitations of existing 3D Convolutional Neural Networks (3D CNNs) in treating all video frames equally. The STA network characterizes beneficial information at both the frame and channel levels, leveraging differences in spatial and temporal dimensions to enhance the learning capability of 3D convolutions. The proposed STA method can be seamlessly integrated into state-of-the-art 3D CNN architectures for video action detection and recognition. Evaluation of diverse datasets, including SBU Kinect Interaction, NTU RGB+D, PKU-MMD, UCF-101, HMDB-51, and THUMOS 2014, demonstrates the effectiveness of both proposed models in achieving state-of-the-art performance for action recognition and detection tasks. The spatial and temporal attention models contribute significantly to capturing discriminative features, showcasing their potential in advancing the field of human action analytics in computer vision.

Index Terms: RNN, LSTM, 3DCNN, Channel Level Analysis, STA network

I. INTRODUCTION

The utilization of attention mechanisms has become pervasive across diverse domains [34]. Notably, Bahdanau et al. incorporated attention into neural machine translation, while Rush et al. introduced a neural attention model for abstractive sentence summarization [35]. Chorowski et al. applied attention-based models to enhance speech recognition [36], and Xu et al. devised an attention-based framework for image caption generation [37]. In the context of image classification tasks, the attention mechanism has garnered significance in the literature. Wang et al. recently developed a residual attention network by stacking attention modules, effectively capturing image attention through softly weighted output features [38]. Hu et al. introduced a SE-block, a lightweight gating mechanism, to model channel-wise relationships [39]. Furthermore, the Convolutional Block Attention Module (CBAM), proposed by others [40], represents a simple yet effective attention module for feed-forward convolutional neural networks. CBAM incorporates both channel and spatial attention, demonstrating efficacy in image classification and object detection tasks. Extending attention mechanisms to video data, Sharma et al. presented a soft attention-based recurrent model for action recognition. Their findings illustrate the model's capability to recognize pivotal elements within video frames based on detected activities [41]. This extensive application of attention mechanisms underscores their versatility and effectiveness across various domains, providing valuable insights for advancing research in attention-based models.

The distinction in input characteristics leads to the classification of human action analytics into two primary categories: RGB video-based methods and skeleton-based methods. In the case of RGB videos [1]–[5], each frame captures a highly articulated human within a 2D space. While this modality may lose some depth information inherent in the 3D space, it offers the advantage of flexibility in achieving human location and scale invariance. Conversely, an alternative approach involves harnessing high-level information derived from skeleton data, wherein a person is represented by the 3D coordinates of key joints such as the head and neck. Notably, this method lacks the incorporation of RGB information, resulting in a deficiency of appearance-related details.

Fortunately, insights from biological observations [6] support the notion that the positions of a limited set of joints can effectively convey human behavior even in the absence of appearance information. This recognition of the significance of joint positions provides a rationale for the utilization of skeleton-based methods in human action analytics, demonstrating their efficacy in capturing essential behavioral aspects without reliance on appearance cues.

The human body can be effectively characterized by the spatial coordinates of several key joints within the 3D space. The articulated arrangements of these joints give rise to diverse postures, and the trajectories of skeletal joints serve as informative cues for identifying various human actions. In leveraging skeletons as explicit high-level representations of human posture, numerous studies have formulated algorithms that utilize the positions of these joints as input features. Within this context, several works have been dedicated to designing algorithms that extract and leverage discriminative features from skeletal data. Examples include the histograms of 3D joint locations (HOJ3D) [7], pairwise relative position features [8], relative 3D geometry features [9], and co-occurrence feature learning [10]. These approaches aim to discern and exploit key patterns and relationships within the skeletal data, contributing to the robust identification and characterization of human actions based on joint trajectories.

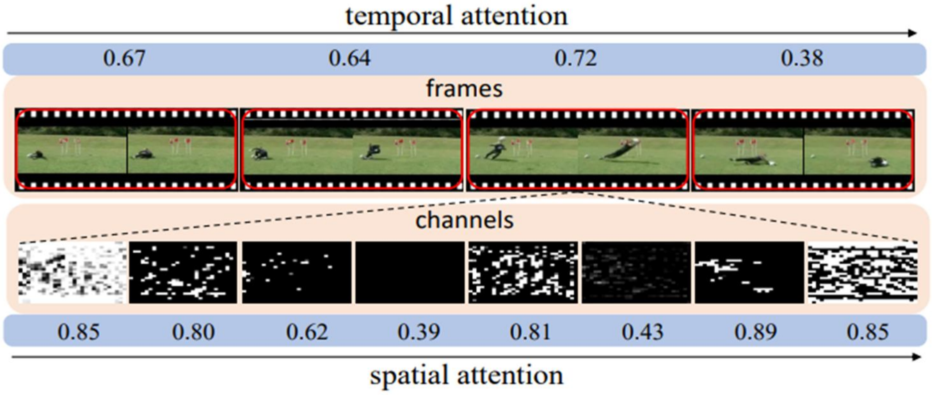


Fig. 1. Spatio-temporal attention of STA-ResNeXt-101(64f) on “Jump” action from HMDB-51 dataset.

The top row shows the successive frames in the video and the corresponding temporal attention weights learned by our STA module, where the frames with the most discriminative features for the “Jump” action can be identified with the larger weights. The bottom row shows the feature maps (channels) for the specified frame and their spatial attention weights, where the more informative feature maps play more important roles in action recognition and detection.

In practical applications, the adoption of 3D Convolutional Neural Networks (3D CNNs) has demonstrated significant enhancements in video processing speed, achieving performance levels at least an order of magnitude faster than real-time processing [42]. Numerous studies have sought to augment and refine 3D CNNs, aiming to amplify their capacity to represent finer temporal relations within local spatio-temporal windows [43]. Despite these advancements, the current 3D CNN solutions still face limitations in effectively capturing discriminative spatio-temporal feature representations for actions without explicit extraction of the most informative information in both spatial and temporal dimensions of videos. Traditional 3D CNNs typically treat consecutive frames equally when learning temporal features, overlooking the potential variability in contributions to action recognition among different frames. For instance, frames with motion blur may offer fewer cues for activity recognition. Similarly, along the spatial dimension, existing 3D CNN solutions often fail to adequately distinguish between the visual information derived from different channels. Consequently, these challenges underscore the need for more refined approaches to 3D CNNs that can discern and leverage the most informative details in both spatial and temporal dimensions for improved action recognition.

The primary contributions of this study can be delineated as follows:

1) Development of an LSTM network featuring two distinct attention modules for action recognition and detection. The spatial attention module incorporates joint-selection gates to dynamically allocate varying levels of attention to different skeleton joints within each frame. Simultaneously, the temporal attention module, facilitated by a frame-selection gate, allocates distinct attention levels to different frames.

2) Introduction of spatio-temporal regularizations to enhance network learning. The spatial regularization is devised to promote comprehensive exploration of all joints, mitigating the tendency to overly emphasize specific joints. Temporal regularization is implemented to prevent unbounded increases in temporal attention and the potential vanishing of gradients.

3) Formulation of a joint training strategy for efficient network training. An iterative training scheme is devised to mitigate the mutual influence between the main network and subnetworks, facilitating the approach toward optimized solutions.
4) Proposal of a method for generating action temporal proposals based on temporal attention. This method accurately localizes the start and end points of actions in untrimmed sequences, demonstrating state-of-the-art performance in action detection.

Moreover, the central contribution of this paper lies in the introduction of an innovative and efficient spatio-temporal attention mechanism within 3D CNNs for action recognition and detection tasks. Diverging from the conventional approach of employing separate temporal and spatial attention mechanisms, our proposed mechanism concurrently captures both temporal and spatial attention within a unified module. Operating as a form of soft attention, the mechanism adaptively learns attention weights, demonstrating its efficacy in enhancing the network's performance.

II. RELATED WORK

Action recognition [44] and action detection [45] constitute pivotal research domains within the broader field of visual understanding [46]. Notably, researchers have increasingly focused on enhancing the accuracy of action recognition on widely adopted datasets such as HMDB-51, UCF-101, and Sports-1M. Despite notable progress in action recognition, the outcomes achieved by action detection methods, employing conventional hand-crafted features [47], CNN features [48], and the fusion of two feature types [49], have not been deemed satisfactory thus far. In recent years, the exponential growth of video data has introduced significant challenges, primarily characterized by computationally intensive requirements and insufficient annotations in video action recognition and detection tasks. Consequently, a growing number of researchers have redirected their efforts toward addressing challenges related to rapid recognition or detection with limited supervised information [38]. Additionally, researchers have explored unsupervised methods [50], which have proven effective in addressing the challenges posed by large-scale recognition and detection tasks in the context of the expanding volume of video data.

The concept of 3D filter field descriptors was initially introduced by Kim et al. [51], and subsequently, Ji et al. extended this notion by proposing the 3D CNNs descriptor [52]. Building upon these foundational works, extensive research has been conducted, with a predominant focus on refining 3D CNNs [53], leveraging practical implementations, and releasing source code. Recent investigations have particularly emphasized strategies to enhance feature learning performance for 3D CNNs, with specific consideration for the temporal dimension. In a comprehensive analysis presented in [54], Hara et al. underscored the efficacy of deep 3D CNNs in conjunction with Kinetics, asserting their ability to replicate the successful outcomes observed in 2D CNNs and ImageNet.

In the context of action identification, not all frames within a sequence hold equal importance. Certain frames may capture less meaningful or even misleading information associated with other types of actions. Conversely, some frames convey more discriminative information [11]. Various approaches have been proposed to address this issue, with one strategy involving the use of keyframes as a representation for action recognition. For instance, one method employs conditional entropy of visual words to measure the discriminative power of a given frame, selecting the top 25% most discriminative frames for a majority vote in recognition [12]. Another approach utilizes AdaBoost to identify the most discriminative keyframes for action recognition. The learning of keyframes can also be formulated within a max-margin discriminative framework, treating them as latent variables [13]. These methodologies collectively aim to enhance the discernment of informative frames for improved action recognition accuracy.

In the pursuit of localizing and recognizing actions within untrimmed video sequences, numerous detection methods employ a sliding window approach. For instance, in [14]–[16], the observation window is systematically slid along temporal frames, and classification is conducted within each window utilizing multiple features, such as dense trajectories, CNN features, in conjunction with action classifiers. Drawing inspiration from recent advancements in object detection from static images [17], [18], the concept of generating object proposals has been adapted for action detection in video sequences [19]–[27]. Certain methods [21]–[23] generate spatio-temporal object volumes to facilitate the detection of simple or cyclic actions in the spatio-temporal domain. Notably, Peng et al. [26] exploit region proposal networks within a two-stream model to produce high-quality spatial proposals, surpassing the performance of other action detection methods. A more recent contribution [27] achieves real-time spatio-temporal action localization and early prediction, establishing new state-of-the-art results by constructing action tubes through the utilization of the Single Shot MultiBox Detector. Additionally, several works [24], and [25] concentrate on temporal action proposals, emphasizing segments likely to contain human actions within the temporal dimension.

The availability of human-labeled ground truths for explicit attention is generally limited and may not consistently align with real attention patterns relevant to specific tasks. Recently, there has been a growing interest in the utilization of attention models that implicitly learn attention across various fields.

Pioneering this approach, Bahdanau et al. [28] introduced the attention mechanism in machine translation, ushering in a new era of state-of-the-art implementations. Building upon this foundation, Xu et al. [29] incorporated soft and hard attention mechanisms for image caption generation. Notably, in [30], a model trained with reinforcement learning demonstrated the ability to selectively focus on the most relevant regions of an input image for multiple object recognition. In the context of action recognition and detection, the idea of selectively focusing on different spatial regions is proposed for RGB videos [31]. Ramanathan et al. [32] introduced an attention model for event detection in RGB videos, wherein attention is directed toward individuals responsible for the event. Another proposal involves the fusion of neighboring frames within a sliding window, employing learned attention weights to enhance the performance of dense labeling of actions in RGB videos [33]. However, it is noteworthy that all the aforementioned attention models for action analytics are grounded in RGB videos. Surprisingly, there is a dearth of investigation into attention models specifically tailored for skeleton sequences, which exhibit distinct characteristics from RGB videos.

III. RESEARCH METHODOLOGY

Proposing a multi-layered LSTM network equipped with spatial and temporal attention mechanisms for action analytics, encompassing both action recognition and detection. The network is structured to autonomously identify dominant joints within each frame via the spatial attention module, while concurrently assigning varying degrees of importance to different frames through the temporal attention module. Introducing the spatio-temporal attention that learns different focusing weights for different frames in the temporal dimension and different focusing weights for different channels in the spatial dimension.

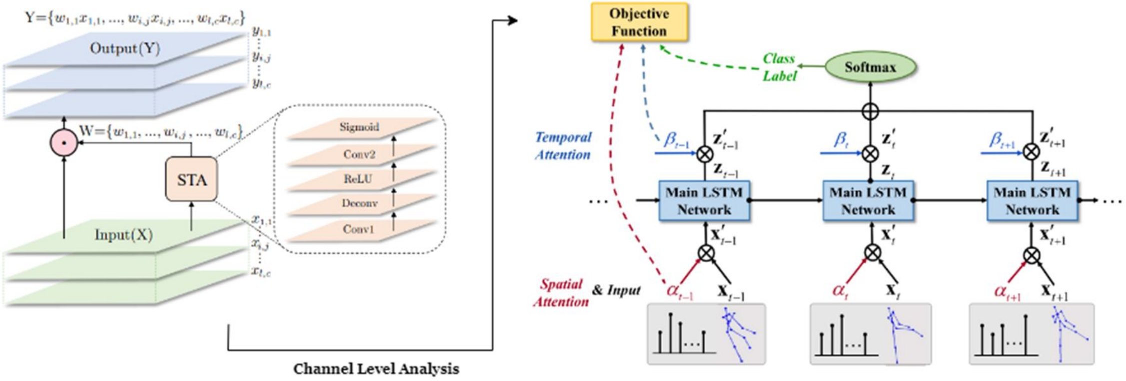


Fig. 2 illustrates the architecture of our Spatio-Temporal Attention (STA) network module integrated with the LSTM network into the 3D CNN model.

The module comprises two convolutional layers (Conv1 and Conv2), one deconvolutional layer (Deconv), and two activation functions (ReLU and Sigmoid). The output, denoted as W, represents the weights assigned to the input feature maps, and the symbol ⊙ denotes the element-wise multiplication between the input feature map and the weights. It provides a visual representation of how the spatial attention output α and temporal attention output β collectively influence the LSTM network. Combining these two attention mechanisms enhances the network's ability to selectively focus on discriminative spatial and temporal features, contributing to improved performance in action recognition and detection tasks. For a sequence, the amount of valuable information provided by different frames is generally not equal. Only some of the frames (key frames) contain the most discriminative information while the other frames provide contextual information. For example, for the action of shaking hands, the sub-stage of approaching should have lower importance than the sub-stage of hands together. Based on this observation, we design a temporal attention module to automatically pay different levels of attention β to different frames.

For sequence level classification, based on the output z_t of the main LSTM network and the temporal attention value β_t at each time step t, the scores for C classes are the weighted summation of the scores at all time steps

$$g = \sum_{t=1}^{T} \beta t.\, zt, \qquad (1)$$

where $g = (g1, g2, \cdots, gC)\ T, T$ denotes the length of the sequence. Fig. 2 illustrates how the temporal attention output β is incorporated into the main LSTM network.

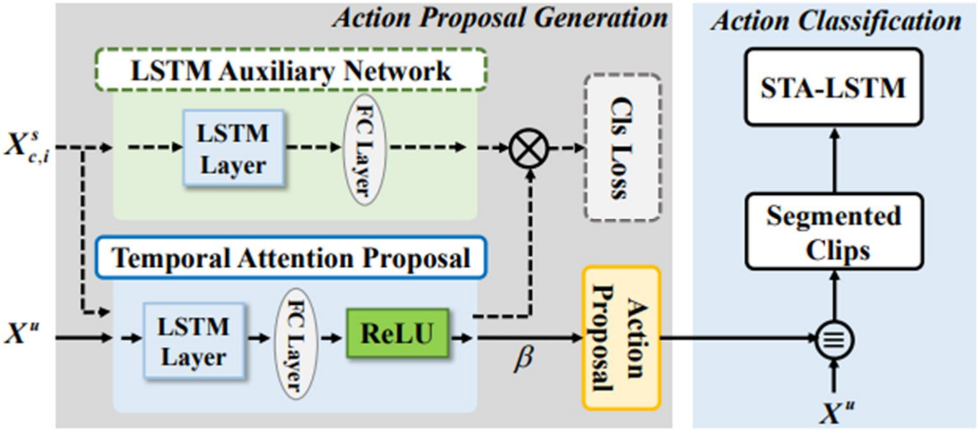


Fig. 3: Proposed action detection framework.

A temporal attention proposal subnetwork (TAP-LSTM) is introduced to produce a temporal attention curve and then generate temporal action proposals based on the attention curve. As shown in Fig. 3, we train the temporal attention proposal using the training data composed of trimmed valid action clips (with non-action clips excluded) collected from the detection training set. The training procedure is supervised by classification loss, which can be found in (1). The training set is denoted for the temporal attention network as D, where $D = \{Dc\}\ C\ c = 1$ with C valid action classes, and $Dc = \{Xs\ c, i\}\ nc\ i = 1$ is the training data corresponding to class c, i denotes the i th trimmed clip of the action class c. The training procedure in action proposal generation is shown in Fig. 3 by the Action Proposal Generation module.

Algorithm 1 STA-LSTM Training for Action Detection

Input: untrimmed training sequences, action interval labels.

//Generate training data.

1: With the ground truth of action detection, generate training sets $D, D*$ with trimmed clips, $where\ D = \{Dc\}\ C\ c = 1\ and\ D* = \{Dc\}\ C + 1\ c = 1$.

//Train Proposal Generation Network.

2: Train the temporal attention proposal subnetwork as Fig. 3 using the Action Proposal Generation module with D. Note that only classification loss is back-propagated. The LSTM auxiliary network will be removed during the test.

//Train Action Classification Network.

3: Train the proposed spatio-temporal attention LSTM network STA-LSTM (with the detailed structure shown in Fig. 2) with $D*$

Output: Well-trained model for proposal generation and action classification.

Like the temporal attention mechanism, the preference is to initially reduce the dimensions involved in the 3D convolutions. This involves squeezing both the temporal dimension and the feature map in each channel, retaining only the channel dimension. This operation enables the network module to primarily concentrate on the spatial content among different channels, mitigating potential side effects associated with other dimensions.

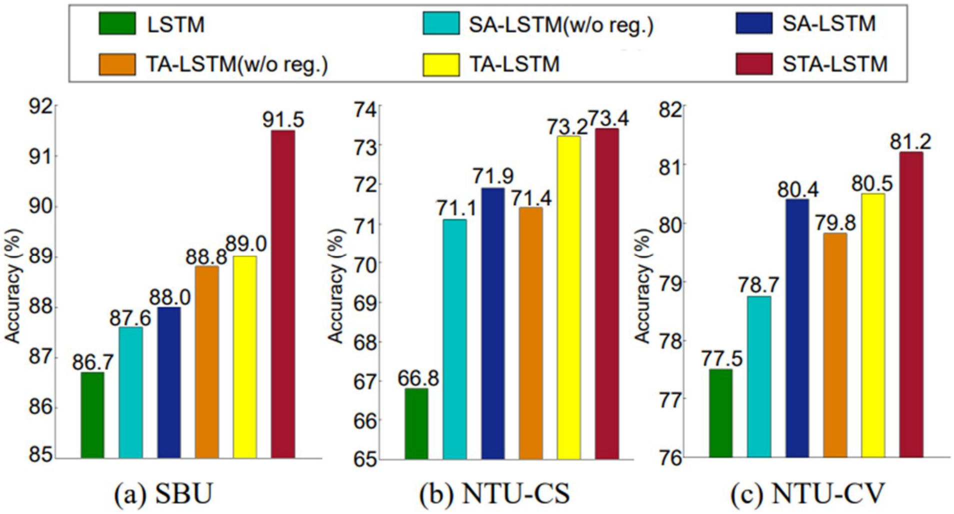
Similarly, the dimension squeezing function Ps for spatial attention, transforming the input Xs into one dimensional sequence $Zs \in R\ c \times 1$ for Ts to discriminate the most informative channels. This can be implemented by a convolution with $c \times l \times h \times w \times c$ learnable parameters $As = \{ak \in R\ l \times c \times h \times w, k = 1, \dots, c\}$:

$$Z_s(k) = \sum_{i=1}^{l}\sum_{j=1}^{c}\sum_{m=1}^{h}\sum_{n=1}^{w} a_k\,(i,j,m,n).\,x_{i,j}(m,n)$$

Combining the temporal and spatial attention modules forms the spatio-temporal attention (STA) module for 3D CNN models. The STA module can concurrently compute the corresponding frame-wise and channel-wise attention weights from the input temporal sliding window at a specified layer of the 3D CNNs.

IV. RESULTS & DISCUSSION

Experiments are performed on the following datasets: the SBU Kinect interaction dataset [54], the largest RGB+D dataset of the NTU [55], and the newly collected action detection dataset, PKU-MMD [21]. Note that we utilize PKU-MMD for both action recognition and detection.

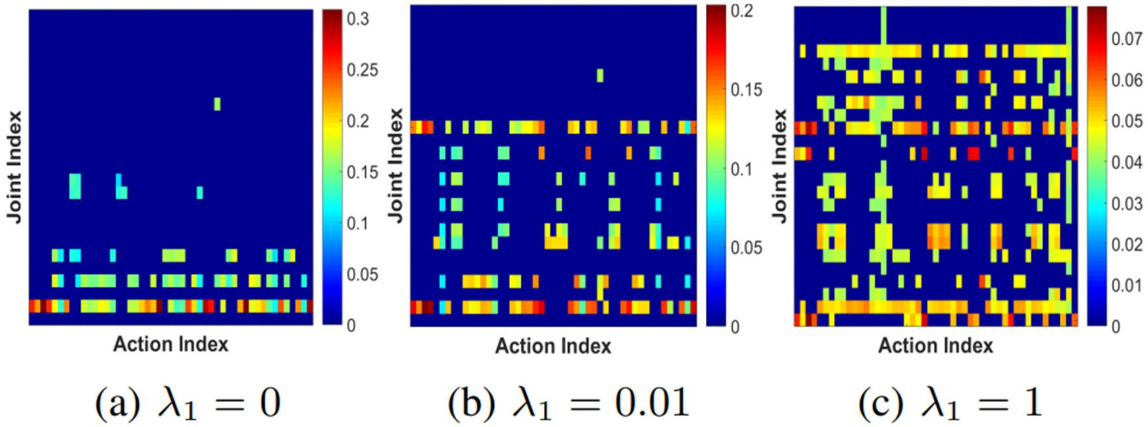


(a) SBU (b) NTU-CS (c) NTU-CV

Fig. 4: Performance evaluation of our attention models and the regularization items on two datasets in terms of accuracy (%).

- LSTM: main LSTM network without attention designs.
- SA-LSTM(w/o reg.): LSTM + spatial attention without regularization (only includes 1 st and 4 th items in (7)).
- SA-LSTM: LSTM + spatial attention network.
- TA-LSTM(w/o reg.): LSTM + temporal attention without regularization (only includes 1 st and 4 th items in (7)).
- TA-LSTM: LSTM + temporal attention network.
- STA-LSTM: LSTM + spatio-temporal attention network.

Fig. 4 presents performance comparisons on the SBU, NTU (Cross-Subject), and NTU (Cross-View) datasets. Relative to the baseline scheme LSTM, the incorporation of the spatial attention module (SA-LSTM) and the temporal attention module (TA-LSTM) results in accuracy improvements of up to 5.1% and 6.4%, respectively. The best performance is observed when both modules are combined (STA-LSTM). In the objective function defined in (1), the second and third items for regularizations are specifically tailored for the spatial and temporal attention models. Notably, these regularization terms contribute to enhanced performance for both the spatial attention model and the temporal attention model.



(a) $\lambda_1 = 0$ (b) $\lambda_1 = 0.01$ (c) $\lambda_1 = 1$

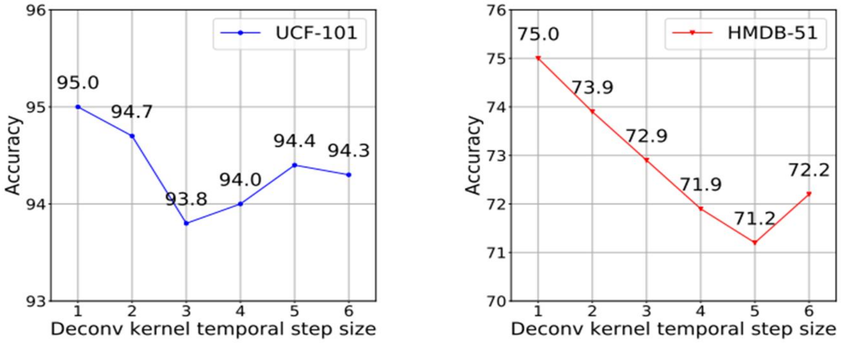
Fig. 5 illustrates the distribution of the most engaged joints corresponding to different actions under various parameter settings.

It is important to note that joints whose average attention responses are less than 80% of the maximum response value are suppressed in this representation. To analyze how the parameters affect the response of spatial attention, we count the most engaged joints for various actions under different values for λ1. As shown in Fig. 5, a larger λ1 (Fig. 5(c)) leads to spatial attention distribution on many more joints, making it hard for the network to extract discriminative joints. Whereas a smaller λ1 (Fig. 5(a)) makes the network focus on too few joints, resulting in information loss.

TABLE 1: THE EFFECT OF DECONV LAYER USING STA-RESNEXT-101(64F) ON UCF-101 AND HMDB-51.

Settings	**UCF-101**	**HMDB-51**
w/o Deconv	94.5	74.4
w/ Deconv	95.0	75.0

Table 1 displays the impact of incorporating a deconvolutional layer in the STA-ResNext-101(64F) model on two benchmark datasets, UCF-101 and HMDB-51. The table presents results under two settings: "w/o Deconv" (without deconvolutional layer) and "w/ Deconv" (with deconvolutional layer). The reported values represent the accuracy percentages achieved by the STA-ResNext-101(64F) model under each setting on the respective datasets. w/o Deconv: The model without the deconvolutional layer achieves an accuracy of 94.5% on UCF-101 and 74.4% on HMDB-51. w/ Deconv: Introducing the deconvolutional layer results in improved performance, with accuracy increasing to 95.0% on UCF-101 and 75.0% on HMDB-51. In summary, the inclusion of the deconvolutional layer enhances the accuracy of the STA-ResNext-101(64F) model on both UCF-101 and HMDB-51 datasets. These results suggest that the deconvolutional layer contributes positively to the model's ability to recognize and classify actions in video data.



(a) UCF-101 (b) HMDB-51

Fig. 6. The effect of different temporal step sizes in Deconv layer. (a) Top-1 accuracy performance for different step sizes of STA-ResNeXt-101 on UCF-101 split1. (b) Top-1 accuracy performance for different step sizes of STA-ResNeXt-101 on HMDB-51 split1.

V. CONCLUSION

Introducing an Unified spatio-temporal attention-based LSTM network for action analysis from skeleton data, featuring a spatial attention module with joint-selection gates to automatically assign varying levels of importance to different joints. Simultaneously, the proposed temporal attention module allocates distinct levels of attention to each frame within a sequence. A joint training procedure efficiently combines spatial and temporal attention using a regularized cross-entropy loss. The temporal attention facilitates the localization of action intervals, leveraging the attention response to generate multiple-scale action proposals for subsequent detection. Experimental results showcase the effectiveness of the scheme, outperforming other state-of-the-art methods in both action recognition and detection. In this paper, a novel spatio-temporal Attention (STA) module is presented to enhance 3D CNNs for action recognition and detection. The module aims to exploit discriminative information at both frame and channel levels, employing convolution and deconvolution operations to distinguish characteristics in temporal and spatial dimensions. The STA module significantly improves the feature learning capability of 3D CNNs without a substantial increase in computational cost. Experiments on various 3D CNN architectures and three datasets demonstrate that the STA module achieves state-of-the-art performance in action recognition (98.4%, 81.4%, 91.5%, 73.4%, and 81.2%) on UCF-101, HMDB-51, SBU, NTU-CS, and NTU-CV datasets, respectively) and improved performance in action detection tasks.

REFERENCES

[1] R. Poppe, “A survey on vision-based human action recognition,” Image and Vision Computing, vol. 28, no. 6, pp. 976–990, 2010.

[2] D. Weinland, R. Ronfard, and E. Boyerc, “A survey of vision-based methods for action representation, segmentation and recognition,” Computer Vision and Image Understanding, vol. 115, no. 2, pp. 224–241, 2011.

[3] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, “Human action recognition in unconstrained videos by explicit motion modeling,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3781–3795, 2015.

[4] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, “Semantic pyramids for gender and action recognition,” IEEE Transactions on Image Processing, vol. 23, no. 8, pp. 3633–3645, 2014.

[5] J. Miao, X. Xu, S. Qiu, C. Qing, and D. Tao, “Temporal variance analysis for action recognition,” IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5904–5915, 2015.

[6] G. Johansson, “Visual perception of biological motion and a model for it is analysis,” Perception and Psychophysics, vol. 14, no. 2, pp. 201–211, 1973.

[7] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012.

[8] J. Wang, Z. Liu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[9] R. Vemulapalli, F. Arrate, and R. Chellappa, “R3DG features: Relative 3D geometry-based skeletal representations for human action recognition,” Computer Vision and Image Understanding, vol. 152, pp. 155– 166, 2016.

[10] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, “Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks,” in AAAI Conference on Artificial Intelligence, 2016.

[11] L. Liu, L. Shao, and P. Rockett, “Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition,” Pattern Recognition, vol. 46, no. 7, pp. 1810–1818, 2013.

[12] Z. Zhao and A. Elgammal, “Information theoretic key frame selection for action recognition.” in British Machine Vision Conference, 2008.

[13] M. Raptis and L. Sigal, “Poselet key-framing: A model for human activity recognition,” in IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[14] S. Karaman, L. Seidenari, and A. Del Bimbo, “Fast saliency-based pooling of fisher encoded dense trajectories,” in European Conference on Computer Vision THUMOS Workshop, vol. 1, no. 2, 2014, p. 5.

[15] D. Oneata, J. Verbeek, and C. Schmid, “Action and event recognition with fisher vectors on a compact feature set,” in IEEE International Conference on Computer Vision, 2013.

[16] L. Wang, Y. Qiao, and X. Tang, “Action recognition and detection by combining motion and appearance features,” THUMOS14 Action Recognition Challenge, vol. 1, no. 2, p. 2, 2014.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards realtime object detection with region proposal networks,” in Advances in Neural Information Processing Systems, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[19] P. Siva and T. Xiang, “Weakly supervised action detection.” in British Machine Vision Conference, 2011.

[20] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, “DAPs: Deep action proposals for action understanding.” in European Conference on Computer Vision, 2016.

[21] M. Jain, J. Van Gemert, H. Jegou, P. Bouthemy, and C. G. Snoek, ´ “Action localization with tubelets from motion,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[22] G. Gkioxari and J. Malik, “Finding action tubes,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[23] W. Chen, C. Xiong, R. Xu, and J. J. Corso, “Actionness ranking with lattice conditional ordinal random fields,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[24] F. Caba Heilbron, J. Carlos Niebles, and B. Ghanem, “Fast temporal activity proposals for efficient detection of human actions in untrimmed videos,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[25] Z. Shou, D. Wang, and S. Chang, “Action temporal localization in untrimmed videos via multi-stage cnns,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[26] X. Peng and C. Schmid, “Multi-region two-stream r-cnn for action detection,” in European Conference on Computer Vision, 2016.

[27] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, “Online realtime multiple spatiotemporal action localisation and prediction,” in IEEE International Conference on Computer Vision, 2017.

[28] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in International Conference on Learning Representations, 2015.

[29] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in International Conference for Machine Learning, 2015.

[30] J. Ba, V. Mnih, and K. Kavukcuoglu, “Multiple object recognition with visual attention,” in International Conference on Learning Representations, 2015.

[31] S. Sharma, R. Kiros, and R. Salakhutdinov, “Action recognition using visual attention,” in International Conference on Learning Representations Workshop, 2015.

[32] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and F.-F. Li, “Detecting events and key actors in multi-person videos,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[33] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, “Every moment counts: Dense detailed labeling of actions in complex videos,” International Journal of Computer Vision, pp. 1–15, 2017.

[34] R. Pramono, Y. Chen, and W. Fang, “Hierarchical self-attention network for action localization in videos,” in IEEE International Conference on Computer Vision, 2019.

[35] A. Rush, S. Chopra, and J. Weston, “A neural attention model for abstractive sentence summarization,” arXiv preprint arXiv:1509.00685, 2015.

[36] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems, 2015.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in International Conference on Machine Learning, 2015.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” arXiv preprint arXiv:1704.06904, 2017.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” arXiv preprint arXiv:1709.01507, 2017.

[40] S. Woo, J. Park, J. Lee, and I. Kweon, “CBAM: Convolutional block attention module,” in European Conference on Computer Vision, 2018.

[41] S. Sharma, R. Kiros, and R. Salakhutdinov, “Action recognition using visual attention,” arXiv preprint arXiv:1511.04119, 2015.

[42] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, “Real-time action recognition with enhanced motion vector CNNs,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[43] L. Wang, W. Li, W. Li, and L. Van Gool, “Appearance-and-relation networks for video classification,” arXiv preprint arXiv:1711.09125, 2017.

[44] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in IEEE International Conference on Computer Vision, 2017.

[45] A. Piergiovanni and M. Ryoo, “Temporal gaussian mixture layer for videos,” arXiv preprint arXiv:1803.06316, 2018.

[46] H. Xu, A. Das, and K. Saenko, “R-C3D: Region convolutional 3D network for temporal activity detection,” in IEEE International Conference on Computer Vision, 2017.

[47] D. Oneata, J. Verbeek, and C. Schmid, “The lear submission at thumos 2014,” 2013.

[48] S. Yeung, O. Russakovsky, G. Mori, and F. Li, “End-to-end learning of action detection from frame glimpses in videos,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[49] L. Wang, Y. Qiao, and X. Tang, “Action recognition and detection by combining motion and appearance features,” THUMOS14 Action Recognition Challenge, vol. 1, no. 2, pp. 2, 2014.

[50] K. Soomro and M. Shah, “Unsupervised action discovery and localization in videos,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[51] H. Kim, J. Lee, and H. Yang, “Human action recognition using a modified convolutional neural network,” in International Symposium on Neural Networks, 2007.

[52] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221– 231, 2013.

[53] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S. Chang, “CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[54] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet,” in IEEE Conference on Computer Vision and Pattern Recognition, 2018.