Enhanced Spatio-Temporal Image Encoding for Online Human Activity Recognition

Anonymous authors
Paper under double-blind review

Abstract

Human Activity Recognition (HAR) based on sensors data can be seen as a time series classification problem where the challenge is to handle both spatial and temporal dependencies, while focusing on the most relevant data variations. It can be done using 3D skeleton data extracted from a RGB+D camera. In this work, we propose to improve the spatio-temporal image encoding of 3D skeletons captured from a Kinect sensor, by studying the concept of motion energy which focuses mainly on skeleton joints that are the most solicited for an action. This encoding allows us to achieve a better discrimination for the detection of online activities by focusing on the most significant parts of the actions. The article presents this new encoding and its application for HAR using a deep learning model trained on the encoded 3D skeleton data. For this purpose, we proposed to investigate the knowledge transferability of several pre-trained CNNs provided by Keras. The article shows a significant improvement of the accuracy of the learning according to the state of the art.

1 Introduction

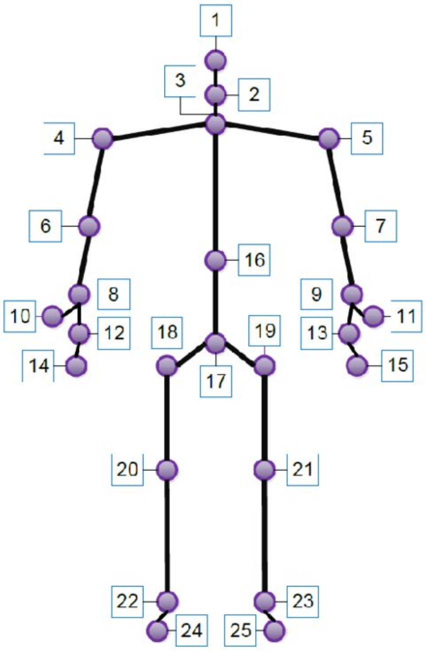
Human Activity Recognition (HAR) is a field of research that covers many application domains such as surveillance, security, assistance, health or training. HAR has become a very active field of research due to the accessibility of data provided by technological advances in the Internet of Things field and their omnipresence in everyday life (Ronao & Cho, 2016). Automatic learning techniques like deep learning, can be exploited in order to recognise activities ("cooking" for instance), which implies to recognise actions that constitute the activity (opening fridge, chopping vegetables, etc.).

Wang et al. (2019) separates HAR in two types : sensor-based and video-based. The former category relies on the use of data emitted by sensors such as accelerometers, gyroscope, bluetooth, sound sensors or Inertial Measurement Unit (IMU). The second category is based on the analysis of videos or images containing human movements, such as 3D skeleton data capture from a Kinect (Figure 1).

Choosing a good representation for the captured data is an important part of the learning process in order to achieve an accurate recognition, since the model's performances are related to this data encoding. For this purpose, several previous works, such as Liu et al. (2017c); Laraba et al. (2017); Ludl et al. (2019) or Mokhtari et al. (2022) proposed to represent the skeletal data as images by encoding a sequence of frames from RGB-D sensors. Resulting images contain spatio-temporal information about user's actions. Another important part of the learning process is the extraction of the most interesting features from these representations in order to focus on them to recognise the activities (Ronao & Cho, 2016). Several studies have investigated the characteristics of an activity, in the same way as finding the differences between the same activity performed by several people (Duong et al., 2009).

Convolutional neural networks have demonstrated their potential in feature extraction and classification in the field of image classification (Martins et al., 2020; Cao et al., 2020) and speech recognition (Zhang et al., 2021; Mustaqeem & Kwon, 2020). They seem to be the best suited technique to perform human activity recognition.

A difficulty in HAR remains in the real-time recognition of human actions. Indeed, considering the data as a continuous stream, it may be difficult to identify the beginning and end of an action. Several previous works like Liu et al. (2019); Weng et al. (2017); Delamare et al. (2021) and Mokhtari et al.



1. Head
2. Neck
3. SpineShoulder
4. ShoulderLeft
5. ShoulderRight
6. ElbowLeft
7. ElbowRight
8. WristLeft
9. WristRight
10. ThumbLeft
11. ThumbRight
12. HandLeft
13. HandRight
14. HandTipLeft
15. HandTipRight
16. SpineMid
17. SpineBase
18. HipLeft
19. HipRight
20. KneeLeft
21. KneeRight
22. AnkleLeft
23. AnkleRight
24. FootLeft
25. FootRight

Figure 1: Skeleton Joints Captured by Kinect V2 Sensor (Ahmed et al., 2015).

(2022) proposed to use a sliding window to train the model on this continuous data stream, as a solution to overcome this problem.

In this work, we propose to improve the Spatio-Temporal Images Encoding (STIE) of 3D skeletons introduced by Mokhtari et al. (2022), by studying the concept of motion energy which focuses mainly on skeleton joints that are the most solicited for an action. Our system is intended to be interactive with its users, so getting the best compromise between recognition rate and online detection is very important, to be able to offer a real-time interaction with a user.

The rest of the document is organised as follows: Section 2 introduces a synthesis of the various works carried out in the field skeleton data representation. In Section 3 we detail our proposed method for skeleton data representation. Section 4, will present the chosen data set that will be used for the experimental part of the work, addressed in Section 5. Finally, we will discuss the results of this work as well as the possible developments to improve the human activity recognition in real time in Section 6.

2 Related Works

Data encoding is an important part of the HAR based on skeletal data where handling both spatial and temporal dependencies is mandatory to achieve a good recognition. Two main categories of approaches are proposed in the state of the art: using graphs or using images.

Some works like Yan et al. (2018) and Delamare et al. (2021) proposed to represent skeletal data as graphs. Each joint at timestamp t is represented by a node, connected to the others according to the connectivity of human body. Then each joint will be connected to the same joint in the consecutive frame (Figure 2). This representation can handle both spatial and temporal dependencies, since each node is connected to its spatial neighbours (according to the skeletal) and also to its temporal one (the previous and following state of the joint) . This representation is usually used with Graph Convolutional Network (Yan et al., 2018; Delamare et al., 2021).

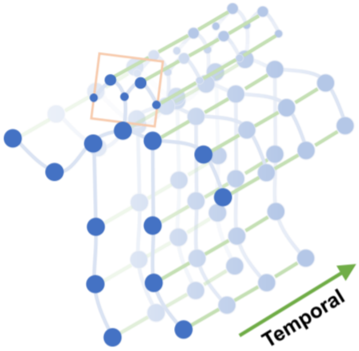


Figure 2: Spatial temporal graph of a skeleton sequence (Yan et al., 2018).

Various works used images to represent the skeleton data (Ludl et al., 2019; Laraba et al., 2017; Pham, 2019; Mokhtari et al., 2022) by encoding each joint as a pixel of an image, where the coordinates (X, Y, Z) of a joint are first normalised, then used to compute the value for the (R,G,B) color. In these studies, the Encoded Human Pose Image (EHPI) proposed by Ludl et al. (2019) was used to represent skeleton extracted from video images (RGB only), where the B value of the pixel was always fixed to 0 (Figure 3).

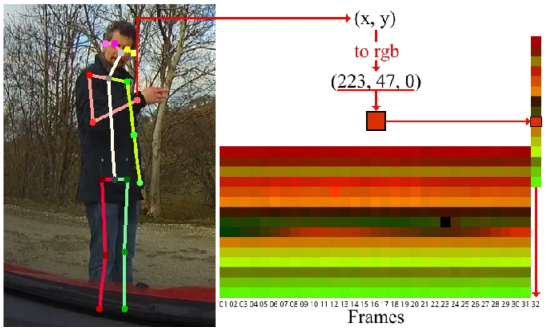

(x, y)
to rgb
(223, 47, 0)
C1 02 C3 04 05 0E 07 C8 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 3C 31 32
Frames

Figure 3: From skeletal joints to an Encoded Human Pose Image (EHPI) Ludl et al. (2019).

Pham (2019) proposed first, to reorder the skeleton joint according to Du et al. (2015) proposition, dividing the human skeleton into five parts : two arms (P1, P2), two legs (P4, P5), and one trunk (P3). In each part from P1 to P5, the joints are concatenated according to their physical connections, and rearranged in a sequential order ($P1 \rightarrow P2 \rightarrow P3 \rightarrow P4 \rightarrow P5$). Then, Pham (2019) introduced the SPMF (Skeleton Pose-Motion Feature) where he combined the 3D skeleton poses and their motions by using the distance between joints. Finally, Pham (2019) proposed to enhance the SPMF by increasing the contrast and highlighting the texture and edges of the motion maps through color enhancement (Figure 4). This method was used with ResNet model, and achieved good performance on HAR common datasets. A limitation of their work includes Online Action Recognition (OAR) task.

To overcome this limitation, Delamare et al. (2021) proposed to use a sliding window. This solution propose to segment the training set into window of equal length in order to get sequences referring

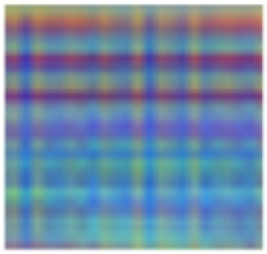


Figure 4: Enhanced-SPMF encoding of a forward kick (Pham, 2019).

to the execution of an action. Training the deep learning model on this segmented training set avoids identifying the start and the end of an action.

In their work, Mokhtari et al. (2022) combined this sliding window approach with the Spatio-Temporal Image Encoding to represent a sequence of human skeleton as a single image, relying on the encoding of the (X, Y, Z) coordinates of a joint as an (R,G,B) pixel, they proposed to reorder the skeleton according to the human body, starting from the left foot to the head through the left arm, then from the head to the right foot through the right arm. This reordering allows to create spatio-temporal neighborhoods in the image (Figure 5), that helps the CNNs to perform better by handling both spatial and temporal dependencies.

Figure 5: Writing encoded using the STIE proposed by Mokhtari et al. (2022).

Liu et al. (2017c) proposed to represent a movement sequence in a 5D space by using the 3D data as well as the joints numbers and the frames numbers. These five pieces of information are used to generate ten separated images, by using two of them as x and y coordinates and the remaining three as (R,G,B) values. This ten images were exploited by ten parallel AlexNets in order to recognise human activities from 3D skeleton data. They also introduced the concept of motion energy, used to highlight the effect of motion on generating color images by weighting skeleton joints according to their motion. Given a skeleton joint $p_n^f = (x_n^f, y_n^f, z_n^f)$, Liu et al. (2017c) estimate its motion energy by:

$$\xi_n^f = ||p_n^f - p_n^{f-1}|| \tag{1}$$

where $f > 1$ and operator $|| \cdot ||$ calculates the Euclidean metric (Liu et al., 2017c). The accumulated motion energy of p_n^f is defined as:

$$\xi_n = \sum_{J=2}^{F} \xi_n^f \tag{2}$$

To control the effect of motion on color images, Liu et al. (2017c) introduced a parameter ρ and define the weight of p_n^f as:

$$w_n = \rho.norm\xi_n + (1 - \rho) \tag{3}$$

where $0 \leq \rho \leq 1$ and function norm normalizes ξ_n to [0, 1]. Then, Liu et al. (2017c) used the following formula to weight the color values of a pixel using :

$$[r\ g\ b] = (1 - w_n)[255\ 255\ 255] + w_n[r\ g\ b] \tag{4}$$

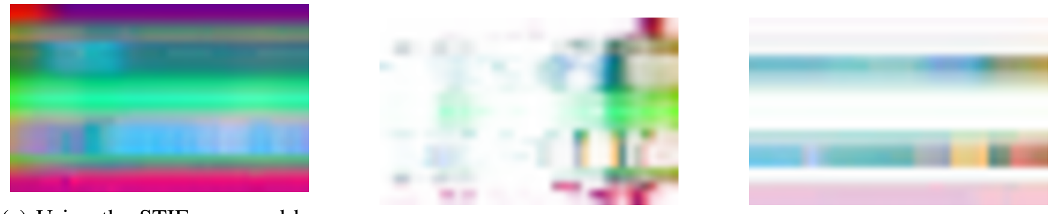
In this work, we are interested combining the works of Mokhtari et al. (2022) (reorganisation of the skeletal joints and sliding window) and Liu et al. (2017c) on the motion energy.

3 Proposed Method

We propose to enhance the Spatio-Temporal Image Encoding (STIE) proposed by Mokhtari et al. (2022) to encode sequences of skeletal data as images, using the concept of motion energy, introduced by Liu et al. (2017c) to highlight the joints that move the most during the execution of the action. This combination should help the deep learning model to perform better on recognising human actions through the spatio-temporal neighborhoods generated by the STIE, when the motion energy should help it to better distinguish between actions.

We propose to use the motion energy following two strategies:

- By computing the motion energy between each two consecutive frames, using the Eq. 1 proposed by Liu et al. (2017c). This use of energy allows to highlight micro-movements occurring during the performance of an action. In the rest of this work, we will refer to this use of the motion energy as «frame-by-frame encoding ».
- By computing the stacked motion energy, for the whole sequence, using the Eq. 2 proposed by Liu et al. (2017c). This use of stacked energy allows to highlight the joints that are the most moving during the performance of an action. In the rest of this work, we will refer to this use of stacked motion energy as «sequence encoding ».



(a) Using the STIE proposed by Mokhtari et al. (2022) of the action

(b) using the frame-by-frame encoding of the action

(c) Using the sequence encoding of the action

Figure 6: Encoding of the action "writing"

For both propositions, the color is weighted using Eq. 4 proposed by Liu et al. (2017c). The figure 6 shows our propositions to encode the action "writing", where the figure 6a refers to the use of STIE encoding, the figure 6b is the frame-by-frame encoding of the action, and the figure 6c refers its sequence encoding.

Mokhtari et al. (2022) proposed to apply transfer learning, by using a pre-trained version of the VGG16 model introduced by Simonyan & Zisserman (2015). This model was first trained on the «ImageNet Dataset », then frozen and used as a feature extractor (Mokhtari et al., 2022). Since this approach preformed well in the task of recognising the human actions, we propose to investigate the knowledge transferability of the several pre-trained models proposed by Keras in order to find a model able to reach even better performances. However, instead of freezing the feature extraction part, we propose to re-train it on the new dataset, to adapt the weights of these models. We think that this adaptation can enable networks that are trained to extract the most interesting features for object detection, to extract these features for human actions recognition. We believe that training the whole network can be useful for a better handling of the spatio-temporal dependencies.

Finally, in order to be able to perform our human action recognition in real time, we propose to use a sliding window, as done in previous work (Liu et al., 2019; Weng et al., 2017; Delamare et al., 2021; Mokhtari et al., 2022)

4 THE OAD DATASET

The proposed method is evaluated on the Online Action Detection dataset (OAD) collected from a Kinect v2, including 25 joints. Since it provides unsegmented online sequences of skeleton data (Li et al., 2016), we chose to use the OAD dataset for our experimentation. It includes 59 long sequences and 10 actions, including drinking, eating, writing, opening cup- board, washing hands, opening microwave, sweeping, gargling, throwing trash, and wiping. The training is done on 30 sequences, and tested on 20 sequences. The remaining 9 sequences are ignored in our work, since they are used for the evaluation of the running speed Li et al. (2016).

Mokhtari et al. (2022) proposed to use a sliding window length of 40 frames (the average duration of all actions in this dataset) in order to segment the training set into sequences. The labels of a sequence is the action performed at the middle of the window, the offset of this sliding window is fixed to one Kinect's frame, which allows them to have several encodings for the same action (starting and ending at different points). It provides 4152 samples for the train subset, and 4671 samples for the test subset.

In the present work, we propose to increase the number of samples, by including the frames before and after the actions. As an example, if the recording starts at frame 1, and the first action starts at frame 50, Mokhtari et al. (2022) segmented this stream starting from the frame 50 when using a sliding window of 40 frames, while we propose to start segmentation from the frame 30 (including half of the window length before the action). Indeed, it can be useful to consider some frames coming before (and even after) the action, in order to recognise its start (or its end) and distinguish between the one starting (or ending) differently, or to use shorter sequences. To test this, we propose to segment the dataset using four different window lengths : 10, 20, 30 and 40 frames, ending up having 11935 samples for the train subset and 10298 for the test subset. In all the cases, 25% of the training set is left for the validation.

5 RESULTS AND DISCUSSION

In this part of the work, we present our experimental results, obtained on the OAD dataset. First, a comparison between our skeleton encoding propositions (frame-by-frame and sequence) with the STIE is done, then we will compare between the four window length propositions. Furthermore, a comparison of the different pre-trained models proposed by Keras is done, and finally, we compare our results to existing works.

5.1 ENCODING COMPARISON

We compared our encoding propositions to the STIE proposed by Mokhtari et al. (2022), on the OAD dataset, using a pre-trained VGG16 model as a feature extractor, followed by a fully connected layer and a classification layer. Unlike Mokhtari et al. (2022), we choose to train the whole network. We used a sliding window of 40 frames, and we fixed the ρ value to 0.5 for the color weighting. The obtained results are presented in the table 1.

Encoding	**Test accuracy**
STIE	94.77%
frame-by-frame	91.17%
sequence	94.65%

Table 1: Results obtained for our encoding propositions and the STIE on the OAD dataset.

According to Table 1, we notice that our proposals are less effective than the STIE. This may mean that the color weighting causes some information loss, because the model is more focusing on joints micro-movements than on the body movement during the whole sequence.

Therefore, we proposed to combine the encoded images vertically, in two ways: combine the 3 encodes, or combine STIE with the frame-by-frame. This combination is done to keep the spatio-temporal information of both body movement and joints micro-movements. The obtained results are presented in the table 2.

Encoding	Test accuracy
STIE	94.77%
Combination (3 encodes)	94.55%
Combination (STIE + frame-by-frame)	95.22%

Table 2: Results obtained for our proposed combinations and the STIE on the OAD dataset.

The results presented in table 2 show that combining the STIE, which focuses on the body movement, with a representation that focuses on joints micro-movements using the frame-by-frame encoding, can be advantageous for human action recognition, since the model is performing a little bit better with an improvement of 0.45%

For the rest of the experimentation, we will thus use the combination of the STIE and the frame-by-frame encoding, with $\rho = 0.5$. We will refer to it as Enhanced Spatio-Temporal Image Encoding (ESTIE)

5.2 Sliding Window Length

We compared between our different sliding window length propositions: 10, 20, 30 and 40 frames (the average duration of all actions in this dataset) using the ESTIE, on the OAD dataset. The obtained results are presented in the table 3.

Window Length	Test accuracy
10	88.59%
20	91.24%
30	93.53%
40	95.22%

Table 3: Results obtained for our sliding window length propositions on the OAD dataset.

According to the results presented in table 3 we can see that the performance of the model increase when the sliding window length is increasing, with the best result obtained using 40 frames as window length, which is equal to the average duration of all actions in the OAD dataset. In consequence, considering larger window for this dataset would introduce to much noise in most of the sequences since the average actions duration is 40 frames.

As regards to these results, we will use a sliding window of 40 frames for the rest of the experimentation.

5.3 Deep Learning Model

In this part of the work, we will investigate the knowledge transferability of several pre-trained deep learning models, proposed by Keras, All of them were trained on the ImageNet dataset to perform object detection. For each model, we removed the classifier part to keep only the feature extractor, that we connected to a fully connected block followed by a classification layer, before training the whole network on the OAD dataset.

Table 4 presents the obtained results, on the OAD dataset, using the ESTIE. We can see that the tested models are well performing on the OAD dataset, giving more than 92% for most of the test accuracy. The best result is obtained by the VGG16 with 95.22%, followed by the DenseNet201 with 94.70%.

5.4 Comparison with the State-of-art methods

We compared our proposed ESTIE combined with a pre-trained VGG16 model as a feature extractor, to several existing works on the OAD dataset. The Table 5 summarises the obtained results. We notice, on the one hand, that increasing the number of samples as proposed in Section 4 and re-training the whole network allows the VGG16 to reach 94.77% on the OAD dataset when using the STIE, while Mokhtari et al. (2022) got 86.81% using the same model and the same encoding. On

Model	**Test Accuracy**
VGG16	95.22%
VGG19	94.06%
EfficientNetV2B3	92.78%
ResNet50V2	92.44%
InceptionResNetV2	93.41%
DenseNet201	94.70%
MobileNetV2	85.95%
InceptionV3	86.73%
NASNetMobile	88.89%
Xception	92.43%
DenseNet121	94.28%
EfficientNetV2L	92.86%

Table 4: Comparison between several pre-trained models proposed by Keras.

the other hand, it shows that our proposed ESTIE combined with the VGG16 model outperforms the other state-of-art methods., by obtaining 95.22%. It improves the best known accuracy on the OAD dataset by 8.41%.

Method	Authors	Accuracy
JCR-RNN	Li et al. (2016)	78.8%
ST-LSTM	Liu et al. (2017a)	77.5 %
Attention Net	Liu et al. (2017b)	78.3%
FSNet	Liu et al. (2019)	81.3 %
SSNet	Liu et al. (2019)	82.8%
VGG16 + STIE [1]	Mokhtari et al. (2022)	86.81 %
VGG16 + STIE [2]	our proposition	94.77 %
VGG16 + ESTIE	our proposition	**95.22%**

Table 5: Comparison with related works on the OAD dataset according to accuracy.

6 Conclusion

Handling both spatial and temporal dependencies is important in order to achieve a good recognition when performing human activity recognition in real time, based on skeleton data provided as a continuous stream.

In this work, we proposed to improve the Spatio-Temporal Image Encoding (STIE), which is a skeleton data representation under an image that preserves both spatial and temporal information. Our proposition enhances this representation, by adding the information of skeleton joints micro-movements, occurring when a person is performing an action.

In the online HAR, identifying the beginning and the end of an action is an important element that might be difficult when the data is coming in a stream way. We chose to combine the proposed encoding with a sliding window approach, in order to perform an online HAR.

We also proposed to investigate the knowledge transferability of several pre-trained deep neural network models provided by Keras for feature extraction, combined with classification layers, in order to achieve online action recognition. The whole network was re-trained on a dataset that is segmented using a sliding window.

The experimentation done on the Online Action Detection dataset showed that training the whole network can improve the model performances compared to the proposition of Mokhtari et al. (2022)

[1] This result is obtained by Mokhtari et al. (2022), where the VGG16 part was frozen during the training.

[2] This result is obtained by increasing the number of samples as proposed in Section 4 and re-training the whole network.

which trains only the fully connected and the classifier layers, by getting 94.77% when the previous work reached 86.81%. This experimentation also showed that our encoding proposition outperforms existing methods on the OAD dataset, by getting 95.22% accuracy improving the best known accuracy on this dataset by 8.41%.

In future work, we would like to further use the temporal information present in the encoding of our skeleton data, since each image line represents the evolution of a joint over time, which can help to improve the recognition of human actions. For that purpose the use of a recurrent neural network combined with a convolutional neural network could be relevant..

REFERENCES

Faisal Ahmed, Padma Polash Paul, and Marina L. Gavrilova. Kinect-based gait recognition using sequences of the most relevant joint relative angles. *J. WSCG*, 23:147–156, 2015.

Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4604–4616, 2020. doi: 10.1109/TGRS.2020.2964627.

Mickael Delamare, Cyril Laville, Adnane Cabani, and Houcine Chafouk. Graph convolutional networks skeleton-based action recognition for continuous data stream: A sliding window approach. 02 2021. doi: 10.5220/0010234904270435.

Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Thi Duong, Dinh Phung, Hung Bui, and Svetha Venkatesh. Efficient duration and hierarchical modeling for human activity recognition. *Artificial Intelligence*, 173:830–856, 05 2009. doi: 10.1016/j.artint.2008.12.005.

Sohaib Laraba, Med Brahimi, Joëlle Tilmanne, and Thierry Dutoit. 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. *Computer Animation and Virtual Worlds*, 28, 05 2017. doi: 10.1002/cav.1782.

Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. volume 9911, pp. 203–220, 10 2016. ISBN 978-3-319-46477-0. doi: 10.1007/978-3-319-46478-7_13.

Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates, 2017a.

Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3671–3680, 2017b. doi: 10.1109/CVPR.2017.391.

Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex Kot. Skeleton-based online action prediction using scale selection network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 02 2019. doi: 10.1109/TPAMI.2019.2898954.

Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017c.

Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Simple yet efficient real-time pose-based action recognition, 2019.

Vitor Martins, Amy Kaleita, Brian Gelder, Hilton Silveira, and Camila Abe. Exploring multiscale object-based convolutional neural network (multi-ocnn) for remote sensing image classification at high spatial resolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 168:56–73, 10 2020. doi: 10.1016/j.isprsjprs.2020.08.004.

Nassim Mokhtari, Alexis Nédélec, and Pierre Loor. Human activity recognition: A spatio-temporal image encoding of 3d skeleton data for online action detection. pp. 448–455, 01 2022. doi: 10.5220/0010835800003124.

Mustaqeem and Soonil Kwon. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167, 10 2020. doi: 10.1016/j.eswa.2020.114177.

Huy-Hieu Pham. Architectures d'apprentissage profond pour la reconnaissance d'actions humaines dans des séquences vidéo rgb-d monoculaires: application à la surveillance dans les transports publics. *HAL https://hal.inria.fr/hal-01678006*, 2019.

Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244, 2016. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.04.032. URL https://www.sciencedirect.com/science/article/pii/S0957417416302056.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, Mar 2019. ISSN 0167-8655. doi: 10.1016/j.patrec.2018.02.010. URL http://dx.doi.org/10.1016/j.patrec.2018.02.010.

Junwu Weng, Chaoqun Weng, and Junsong Yuan. Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition. 07 2017. doi: 10.1109/CVPR.2017.55.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. 01 2018.

Nan Zhang, Jianzong Wang, Wenqi Wei, Xiaoyang Qu, Ning Cheng, and Jing Xiao. Cacnet: Cube attentional cnn for automatic speech recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2021. doi: 10.1109/IJCNN52387.2021.9533666.