



Human Activity Recognition: A Spatio-temporal Image Encoding of 3D Skeleton Data for Online Action Detection

Nassim Mokhtari, Alexis Nédélec, Pierre de Loor

► To cite this version:

Nassim Mokhtari, Alexis Nédélec, Pierre de Loor. Human Activity Recognition: A Spatio-temporal Image Encoding of 3D Skeleton Data for Online Action Detection. 17th International Conference on Computer Vision Theory and Applications, Feb 2022, Online Streaming, France. pp.448-455, 10.5220/0010835800003124 . hal-03586862

HAL Id: hal-03586862

<https://hal.science/hal-03586862>

Submitted on 27 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Human Activity Recognition : A Spatio-Temporal Image Encoding of 3D Skeleton Data for Online Action Detection

Nassim Mokhtari¹^a, Alexis Nédélec¹^b and Pierre De Loor¹^c

*Lab-STICC (UMR CNRS), ENIB,
Centre Européen de Réalité Virtuelle, Brest
mokhtari@enib.fr; nedelec@enib.fr; deloor@enib.fr*

Keywords: 3D Skeleton Data, Spatio-temporal Image Encoding, Sliding Window, Online Action Recognition, Human Activity Recognition, Deep learning

Abstract: Human activity recognition (HAR) based on skeleton data that can be extracted from videos (Kinect for example) , or provided by a depth camera is a time series classification problem, where handling both spatial and temporal dependencies is a crucial task, in order to achieve a good recognition. In the online human activity recognition, identifying the beginning and end of an action is an important element, that might be difficult in a continuous data flow. In this work, we present a 3D skeleton data encoding method to generate an image that preserves the spatial and temporal dependencies existing between the skeletal joints. To allow an online detection of the actions we combine this encoding system with a sliding windows on the stream of the data. By this way, no start or stop timestamp is needed and the recognition can be done at any moment. A deep learning CNN algorithm is used to achieve actions online detection.

1 INTRODUCTION

The extraction of knowledge from sensor data has become a very active field of research in part due to the accessibility of data generated by technological advances in the field of the Internet of Things and their pervasiveness in everyday life. Research on human activity recognition (HAR) has become more widespread in recent years due to its use in several areas such as surveillance-based security and life support (Ronao and Cho, 2016). This research area exploits different automatic learning techniques for HAR like recognising "do sport" activity, which implies to recognise actions that constitute the activity like running, jumping, etc...


According to (Wang et al., 2019), there are two types of human activity recognition, sensor-based and video-based. The first category is concerned with data emitted by sensors such as accelerometers, gyroscope, bluetooth, sound sensors, Inertial Measurement Unit (IMU), etc... The second category is concerned with the analysis of videos or images containing human movements, including depth cameras that can provide skeletal data such as Kinect (Figure 1) .


In this work, we propose a spatio-temporal image encoding for online action recognition (OAR) based on 3D skeletal data, focusing on the use of depth cameras (Kinect sensor), since they are less invasive than wearable sensors (IMU). Our OAR system will be exploited to allow interaction between a user and a computer system.


Studies have been made in this research field to determine the characteristics of an activity as well as the differences between form and style of the same activity performed by several people (LeCun et al., 2015; Duong et al., 2009). These features as well as the time series constitute the basic information, if they are well used, using a classifier and an extraction of the most important features, they allow activity recognition (Ronao and Cho, 2016).

Recent advances in the field of image classification (Martins et al., 2020; Cao et al., 2020) and speech recognition (Zhang et al., 2021; Mustaqeem and Kwon, 2020) related to deep learning research, particularly convolutional neural networks, have demonstrated their interest in feature extraction and classification and seem to be best suited to our problem on the recognition of human activities.

Data encoding is an important part of the learning process, since the model's performances are related to the data encoding, choosing a good data represen-

^a  <https://orcid.org/0000-0002-9402-3638>

^b  <https://orcid.org/0000-0003-3970-004X>

^c  <https://orcid.org/0000-0002-5415-5505>

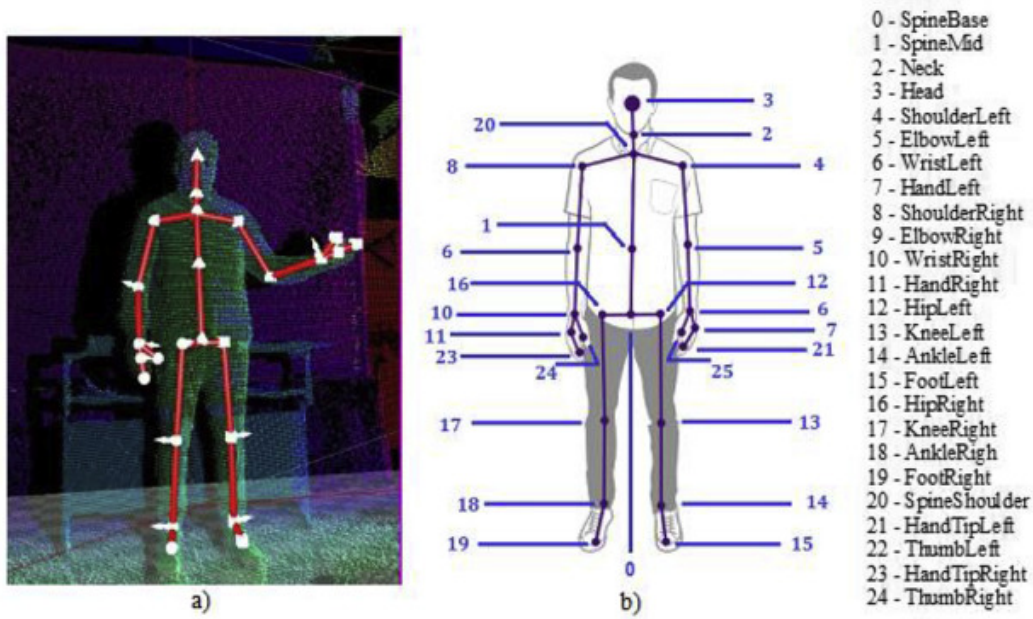


Figure 1: Skeleton Joints Captured by Kinect V2 Sensor (Manghisi et al., 2017).

tation is crucial to achieve an accurate recognition. There are several skeletal data representation propositions, such as using images (Liu et al., 2017c; Laraba et al., 2017; Ludl et al., 2019).

Performing Online Action Recognition (OAR) implies to consider sensor's data as a continuous stream, where identifying the beginning and the end of an action is an important element, that might be a difficult task. One solution to overcome this problem is the use of a sliding window on continuous data in order to train the model (Delamare et al., 2021; Liu et al., 2019; Weng et al., 2017; Kulkarni et al., 2014).

Activity recognition has a huge potential in many areas. Our system can be used in different fields such as sport (gesture training) or healthcare (functional rehabilitation, Assistance Daily Living). Our system is intended to be interactive with its users, so the important thing is to get the best compromise between recognition rate and online detection, 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 present our proposed method for skeleton data representation. Section 4, will presents the chosen data set that will be used for the experimental part of the work, presented in Section 5. Finally, we will present 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

In the case of human activity recognition based on skeletal data, and in order to achieve a good recognition, the evolution of the different joints must be considered on both spatial and temporal domain, therefore, choosing an encoding that fulfils these criteria is mandatory.

There are many proposals for encoding the skeletal data, such as using images, like the Encoded Human Pose Image (EHPI) proposed by (Ludl et al., 2019), which encodes each joint of the skeleton, extracted from a video, as a pixel, where the x and y coordination where first normalised, then used as values for the (R,G,B) color, where the third value was fixed to 0 (Figure 2). This encoding was used with a CNN to recognise actions from video data.

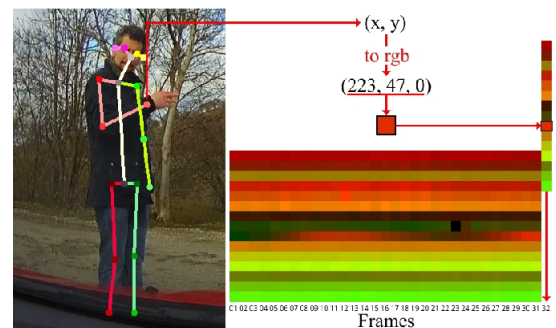


Figure 2: From skeletal joints to an Encoded Human Pose Image (EHPI) (Ludl et al., 2019).

(Laraba et al., 2017) and (Pham, 2019) propositions, also make a transformation of the coordinates (X, Y, Z) of a joint into components (R, G, B) of a color. Each joint is represented by a pixel and each frame is represented by a column. The result is an image that corresponds to a specific action or activity.

(Laraba et al., 2017) proposed to duplicate rows of the images, to enhance their information (Figure 3), and used their encoding with a CNN on 3D Skeleton data.

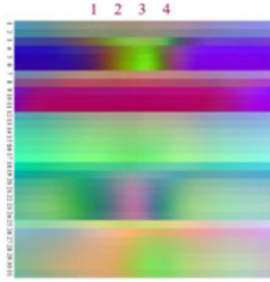


Figure 3: Illustration of the proposed RGB representation of a 3D skeleton sequence from (Laraba et al., 2017) representing a front kick.

(Pham, 2019) proposed first, to reorder the skeleton joint according to (Du et al., 2015) proposition, which divide human skeleton into five parts, including 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, then rearrange these parts in a sequential order (P1 → P2 → P3 → P4 → P5). Then Pham proposed SPMF (Skeleton Pose-Motion Feature) combining 3D skeleton poses and their motions, using the distance between joints. Finally, an Enhanced-SPMF was proposed, using a color enhancement, for increasing contrast and highlighting the texture and edges of the motion maps (Figure 4). This method was used with ResNet model, and achieved good performance on human activity recognition common datasets, but, one of the limitations listed by Pham is how to scope with Online Action Recognition (OAR) task.

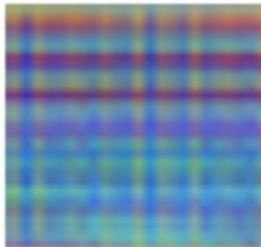


Figure 4: Enhanced-SPMF encoding of a forward kick (Pham, 2019).

(Liu et al., 2017c) also proposed to encode the data as images. Where the 3D data as well as the joint numbers and the frame numbers were used to represent a movement sequence in a 5D space. These five pieces of information are encoded in different ways in ten separate frames. This ten images were exploited by ten parallel AlexNets in order to recognise human activities from 3D skeleton data.

(Yan et al., 2018) proposed a different way to represent skeletal data, using graphs, where each node represents a joint in a timestamp t . The first step consist of connecting joint with edges according to the connectivity of human body structure, Then each joint will be connected to the same joint in the consecutive frame (Figure 5). A such representation can handle both spatial and temporal dependencies, since each node is connected to his spatial neighbours (according to the skeletal) and also his temporal one (the previous and following state of the joint) .

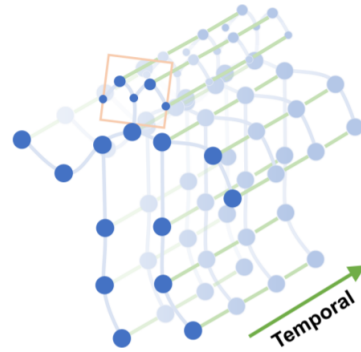
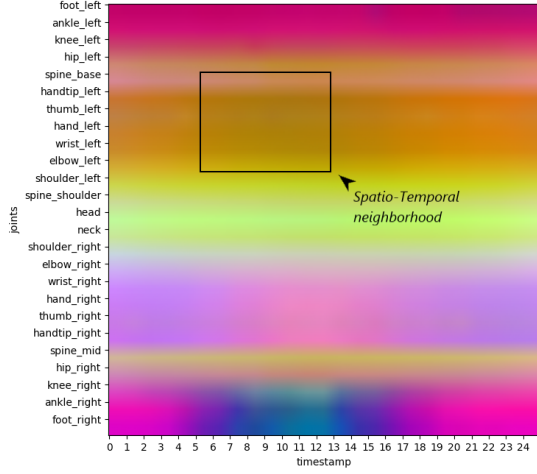


Figure 5: Spatial temporal graph of a skeleton sequence (Yan et al., 2018).

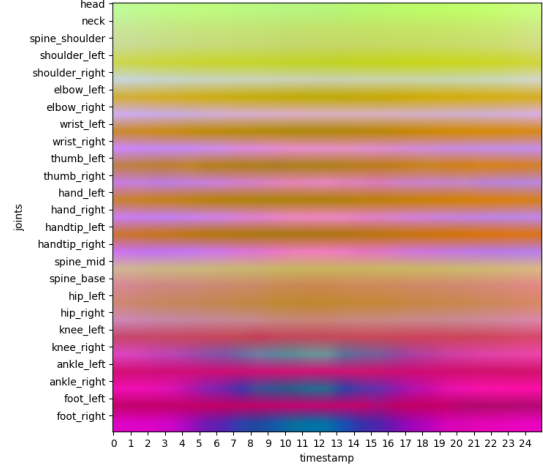
3 PROPOSED METHOD

We propose to encode the sequences of skeletal data as images, using the Encoded Human Pose Image introduced by (Ludl et al., 2019), but, since the Kinect can provide 3D data for each joint, we choose to use each of (X,Y,Z) components of the joint in order to calculate the (R,G,B) value, like (Laraba et al., 2017; Pham, 2019) did. The (X,Y,Z) components are first normalized, in order to make the encoding insensitive to the size of the people, then encoded into (R,G,B).

In order to preserve the spatial dependencies existing between the skeletal joints, we propose to reorder the joint provided by the Kinect (initially order as show in Figure 1) according to the human skeletal. We follow two different strategies :



(a) Using foot to foot order.



(b) Using head to feet order.

Figure 6: Forward-Kick from UOW Online Action 3D data set using Skeletal encoding based on (EHPI).

- foot to foot : Starting from the left foot to the head through the left arm, then from the head to the right foot through the right arm.
- head to feet : Starting from the head to the feet through the arms, in the other hand.

These representations ensure that both of spatial and temporal dependencies are preserved, since each pixel (joint at a timestamp t) is near to his temporal neighbours (following the image width) and his spatial one (following the image height).

The Figure 6 shows the representation of a forward kick from the UOW Online Action 3D data set, using the foot to foot order (Figure 6a) and the head to feet order (Figure 6b), where we can notice that the most important color changes are done in the right foot, ankle and knee region.

The black square in Figure 6a represents a spatio-temporal neighbourhood of a joint, applying a convolution filter on a such image region, ensures to handle the spatial dependency existing between joints, and temporal one existing between frames.

Since we proposed to encode skeletal data as images, we choose to use a CNN model (presented in Section 5.1) to perform this online action recognition, due to their well-known performance in the field of image processing, specifically in the task of features extraction. We propose to use a pre-trained version of VGG16, as we believe that using transfer learning for feature extraction can be helpful and time saving,

As our main goal is to perform an Online Action Recognition, this implies that the skeletal data have to be treated in a stream. In this case, finding the start of each action is a crucial point, that's why, we used the sliding window approach proposed by (Delamare

et al., 2021).

The training set is segmented into window of equal length (which is the average of actions' duration in term of Kinect's frames) in order to get sequences, the labels of this sequences 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 us to have several encodings for the same action (starting and ending at different points). Training our model using this approach avoids identifying the start and the end of an action, and ensure the action recognition even when the data is provided as a stream.

4 DATASET

The proposed method is evaluated on two datasets, the UOW Online Action 3D dataset, and the OAD dataset. Both datasets have unsegmented online sequences of skeleton data, collected from a Kinect.

4.1 UOW Online Action 3D

The UOW Online Action 3D Contains 20 different actions (21 including the "No-Action"), performed by 20 different subjects with up to 3-5 different executions. From each of the 48 sequences, the 25 joint positions per frame were used as inputs (Tang et al., 2017). The train is done on the repeated sequences, and test on the continuous ones.

As proposed by (Delamare et al., 2021) The data have been reorganised into windows of 50 frames as it is the average duration of all actions in this dataset, with a frame-by-frame offset, ending up having 91525

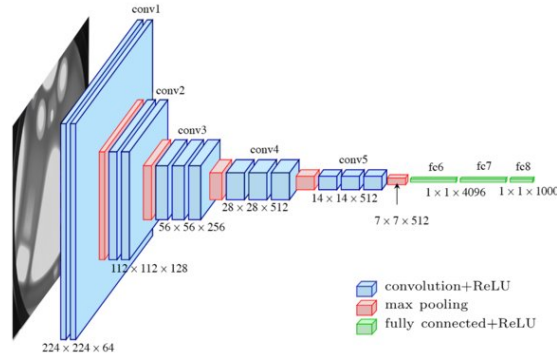


Figure 7: The standard VGG-16 network architecture (Fergusson et al., 2017).

samples for the train subset (Where 25% is left for the validation), and 22555 samples for the test subset.

4.2 The OAD Dataset

The Online Action Detection dataset (OAD) was captured using the Kinect V2, 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 train is done on 30 sequences, and test 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).

The data have been reorganised into windows of 40 frames as it is the average duration of all actions in this dataset, with a frame-by-frame offset, ending up having 4152 (Where 25% is left for the validation), and 4671 samples in the test subset.

5 RESULTS AND DISCUSSION

In this part of the work, we will present our experimental results, obtained on the UOW Online Action 3D dataset and the OAD dataset. First, a comparison between our skeleton encoding propositions (foot to foot, head to feet) with "no reorder" is done on the UOW dataset, and then, we will compare our results to existing works on both datasets.

5.1 Deep Learning Model

The Keras pre-trained version of the VGG16 model (Figure 7), proposed by (Simonyan and Zisserman, 2015), is used as a feature extractor. We use the convolution/pooling part of the VGG16 that will be frozen during the training phase (no weight adjustment will be done on it), followed by two Dense layers of 4096

units, with a Dropout probability of 0.5 and a batch normalisation for each layer's output. Finally a classifier layer using *softmax* function is used for the class inference.

The model is trained for 100 epochs, with batches of 64 samples, using ADAM as our stochastic gradient descent method, with the following parameters :

- learning rate=0.001
- decay rate for the 1st moment = 0.9
- decay rate for the 2nd moment = 0.999
- epsilon = 1e-07

The weights giving the best the validation accuracy during the training process are stored.

5.2 Encoding Comparison

Our proposed method for skeleton data encoding will produce images of dimension $h \times w$, where h refers to the number of joints, and w to the sliding window length. Since we are working on data provided by Kinect giving 25 joints, and the the sliding window length is fixed to 50 for the UOW Online Action 3D dataset, the result of our encoding will be images of 25×50 .

We choose to resize our images to 224×224 , which is the default VGG16 input shape, since this model does not support images bellow 32×32 . The resizing is done using the bi-linear interpolation, or the area interpolation.

The comparison between the proposed encoding methods is done according to the accuracy obtained on the test subset of the UOW dataset. The results are summarised in Table 1, where the "no reorder" refers to the use of Encoded Human Pose Image (EHPI) proposed by (Ludl et al., 2019).

From the results presented in Table 1, we can notice that the resizing using the bi-linear interpolation offers better performances then the area interpolation

Table 1: Results of encoding method comparison on the UOW dataset.

| Encoding method | Accuracy |
|--|----------------|
| foot to foot + bi-linear interpolation | 71.38 % |
| foot to foot + area interpolation | 70.54 % |
| head to feet + bi-linear interpolation | 69.01 % |
| head to feet + area interpolation | 66.82 % |
| no reorder + bi-linear interpolation | 69.35 % |
| no reorder + area interpolation | 68.88 % |

in all cases. Reordering the skeleton data from the left foot to the head through the left arm, then from the head to the right foot through the right arm gives the best accuracy, this result demonstrates the contribution of our proposal to the handle the spatio-temporal dependencies.

We decided to introduce some sparsity in our model, by adding L1 regularisation to the last layer, in order to improve the classification by using only the relevant features. This regularisation rises our overall accuracy to **73.54 %**.

For the rest of tests, we will use a foot to foot re-order combined with a bi-linear interpolation.

5.3 UOW Online Action 3D dataset Comparison

The UOW Online Action 3D dataset is recent and does not propose a method similar to ours, except (Delamare et al., 2021) proposition, which was not tested according to the UOW protocol, since they choose to train on 46 sequences, validate on 1, and test on the remaining one.

(Delamare et al., 2021) proposed the Sliding Window Graph Convolutional Network (SW-GCN), that represents a sequence of skeletal data obtained using a sliding window, as a graph, then use the GCN proposed by (Yan et al., 2018), they also implemented a Sliding Window Convolutional Neural Network (SW-CNN) to compare with their proposition.

In order to compare our proposition to the SW-CNN and SW-GCN, we tested our proposition on the last sequence of the UOW dataset. The Table 2 summarises the comparison between our proposition and the related works on the UOW dataset.

Table 2: Comparison with related works on the UOW dataset according to accuracy.

| Method | Accuracy |
|--------------------------------|---------------|
| SW-CNN (Delamare et al., 2021) | 68 % |
| SW-GCN (Delamare et al., 2021) | 75.5% |
| VGG16 + our encoding | 81.9 % |

From the results presented in Table 2, we can no-

tice that our proposition outperform the SW-CNN, and SW-GCN, by improving the best accuracy by 6.4 %. This result show that an image can handle both spatial and temporal dependencies, and can even offer better performances compared to a graph representation.

5.4 OAD Dataset Comparison

The Table 3 summarises the comparison between our proposition and the related works on the OAD dataset.

Table 3: Comparison with related works on the OAD dataset according to accuracy.

| Method | 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 + our encoding | 84.18 % |
| VGG16 + our encoding + L1 | 86.81 % |

(Delamare et al., 2021) tested their proposition on the OAD dataset and obtained 90% of overall accuracy, but since it does not follow the dataset testing protocol proposed by (Li et al., 2016), we did not compared our proposition to their on this data set.

The best known overall accuracy for the OAD dataset, following the testing protocol, is 82.8%, obtained by (Liu et al., 2019) with their Scale Selection Network (SSNet), our method could improve this overall accuracy, by getting 84.18 %. The use of L1 regularisation on the last layer rises our overall accuracy to 86.81 %.

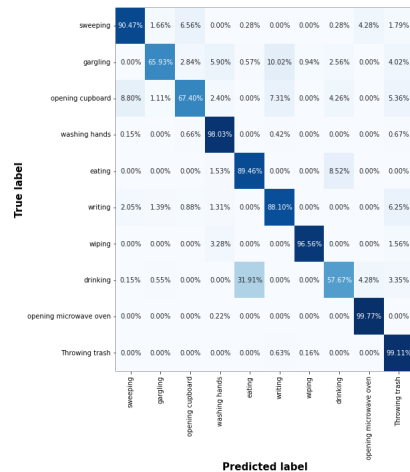


Figure 8: Confusion Matrix obtained by our method on OAD dataset.

Table 4: Comparison with related works on the OAD dataset according to F1-Score.

| Actions | SVM-SW (Li et al., 2016) | RNN-SW (Zhu et al., 2016) | CA RNN (Li et al., 2016) | JCR RNN (Li et al., 2016) | our proposition |
|------------------------|-----------------------------|------------------------------|-----------------------------|------------------------------|--------------------|
| Drinking | 0.15 | 0.44 | 0.58 | 0.57 | 0.67 |
| Eating | 0.47 | 0.55 | 0.56 | 0.52 | 0.81 |
| Writing | 0.65 | 0.86 | 0.75 | 0.82 | 0.85 |
| Opening cup-board | 0.30 | 0.32 | 0.49 | 0.50 | 0.76 |
| Washing hands | 0.56 | 0.67 | 0.67 | 0.71 | 0.92 |
| Opening Microwave oven | 0.60 | 0.67 | 0.47 | 0.70 | 0.96 |
| Sweeping | 0.46 | 0.59 | 0.60 | 0.64 | 0.90 |
| Gargling | 0.44 | 0.55 | 0.58 | 0.62 | 0.77 |
| Throwing trash | 0.55 | 0.674 | 0.43 | 0.46 | 0.89 |
| Wiping | 0.86 | 0.75 | 0.76 | 0.78 | 0.98 |

The Figure 8 shows our confusion matrix for the OAD dataset, where we can see that actions like Washing hands, Wiping, Opening Microwave oven, Throwing trash are well recognised, with more than 96% accuracy, while drinking action has the worst accuracy with 57.67 %, confused with eating action (31.91 %) of the time. This can be explained by the fact that both actions are kind of similar, since both implies the movement of the hand to the mouth.

The Table 4 shows a comparison between our proposed method, and related works on the OAD dataset, according the the F1-Score. We can see that our proposition outperforms all methods, on all actions, except the result of X and their RNN-SW or Writing action, which is 0.01% better.

A short video showing our model performing online action detection on a sequence from the OAD dataset can be found on Youtube

6 CONCLUSION

The human activity recognition based on skeleton data is a time series classification problem, where handling both spatial and temporal dependencies is a crucial task, in order to achieve a good recognition.

In this work, we presented a skeleton data representation under image, based on an existing skeleton encoding, the Encoded Human Pose Image (EHPI). Our proposition preserves the spatial and temporal dependencies existing between the skeletal joints, by reordering the joints according to the human skeleton.

In the online human activity recognition, 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 proposed encoding with a sliding window approach, in order to perform an online human activity recognition.

We proposed a transfer learning approach, by using a pre-trained deep neural network model (VGG16 provided by Keras) for feature extraction, combined with classification layers, in order to achieve online action recognition,

The experimentation done on the UOW Online Action 3D Dataset showed that our encoding proposition can improve the baseline technique proposed by (Ludl et al., 2019), by reordering the joints according the human skeleton, starting from the left foot to the right one, through the head.

Our proposition outperforms existing methods on two challenging datasets : The UOW Online Action 3D and the OAD dataset, by getting 81.9% accuracy on the first one, and 86.81% on the second, improving the best known accuracy by respectively 6.5% and 4.01%.

On the OAD dataset most of actions where well recognised (with at least 90%) where some actions like "Opening Microwave oven" are recognised with 99%, while "drinking" action has the worst accuracy with 57.67 %, confused with "eating", which can be caused by the similarity of both actions. A video was uploaded to Youtube showing our online action detection on this dataset

As a future work, we aim to improve our skeleton data encoding method by enhancing motion information focusing on the most important joints. We also aim to use Recurrent Neural Networks combined with CNNs to a better use of the temporal information present in our skeleton data encoding, since each

image row represents the evolution of a joint in time, using RNNs might be relevant.

ACKNOWLEDGEMENTS

This work has been carried out within the French-Canadian project DOMAID which is funded by the National Agency for Research (ANR-20-CE26-0014-01) and the FRQSC

REFERENCES

- Cao, X., Yao, J., Xu, Z., and Meng, D. (2020). Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4604–4616.
- Delamare, M., Laville, C., Cabani, A., and Chafouk, H. (2021). Graph convolutional networks skeleton-based action recognition for continuous data stream: A sliding window approach.
- Du, Y., Wang, W., and Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Duong, T., Phung, D., Bui, H., and Venkatesh, S. (2009). Efficient duration and hierarchical modeling for human activity recognition. *Artificial Intelligence*, 173:830–856.
- Ferguson, M., ak, R., Lee, Y.-T., and Law, K. (2017). Automatic localization of casting defects with convolutional neural networks. pages 1726–1735.
- Kulkarni, K., Evangelidis, G., Cech, J., and Horaud, R. (2014). Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 112(1):90–114.
- Laraba, S., Brahimi, M., Tilmanne, J., and Dutoit, T. (2017). 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. *Computer Animation and Virtual Worlds*, 28.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.
- Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C., and Liu, J. (2016). Online human action detection using joint classification-regression recurrent neural networks. volume 9911, pages 203–220.
- Liu, J., Shahroudy, A., Wang, G., Duan, L.-Y., and Kot, A. (2019). Skeleton-based online action prediction using scale selection network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP.
- Liu, J., Shahroudy, A., Xu, D., Kot, A. C., and Wang, G. (2017a). Skeleton-based action recognition using spatio-temporal lstm network with trust gates.
- Liu, J., Wang, G., Hu, P., Duan, L.-Y., and Kot, A. C. (2017b). Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3671–3680.
- Liu, M., Liu, H., and Chen, C. (2017c). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362.
- Ludl, D., Gulde, T., and Curio, C. (2019). Simple yet efficient real-time pose-based action recognition.
- Manghisi, V. M., Uva, A. E., Fiorentino, M., Bevilacqua, V., Trotta, G. F., and Monno, G. (2017). Real time rula assessment using kinect v2 sensor. *Applied Ergonomics*, 65:481–491.
- Martins, V., Kaleita, A., Gelder, B., Silveira, H., and Abe, C. (2020). 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.
- Mustaqeem and Kwon, S. (2020). Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167.
- Pham, H.-H. (2019). 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>.
- Ronao, C. A. and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.
- Tang, C., Wang, P., and Li, W. (2017). Online action recognition based on incremental learning of weighted covariance descriptors.
- Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11.
- Weng, J., Weng, C., and Yuan, J. (2017). Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition.
- Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition.
- Zhang, N., Wang, J., Wei, W., Qu, X., Cheng, N., and Xiao, J. (2021). Cacnet: Cube attentional cnn for automatic speech recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.
- Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., and Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks.