Human Action Recognition Dataset for ergonomic risk assessment applications

Coruzzolo A.M. *, Forgione C., Lolli F., Zhao Q., Rimini B.

** Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Via Amendola 2, 42122– Reggio Emilia (antoniomaria.coruzzolo@unimore.it, francesco.lolli@unimore.it, elia.balugani@unimore.it, bianca.rimini@unimore.it)*

Abstract: Marker-less sensors are widely used for body-tracking in different applications, as Human Action Recognition (HAR) and ergonomic risk assessment. For HAR different datasets were developed to train Neural Networks (NN) able to recognize actions based on labeled skeleton videos. However, state-of-the-art datasets include many tasks but rarely there are work activities like lifting of loads or manual handling. This paper presents an expansion of Northwestern-UCLA dataset with manual handling and lifting of loads. Our contribution includes the addition of 540 acquisitions to the existing dataset where 3 different tasks are performed by 5 different subjects. The first task is not new but useful to generalize task recognition and regards picking up a box from different height and surfaces; these activities were classified as an existing label of Northwestern-UCLA dataset. While two completely new labels were added to the existing dataset, and they are: leaving a box on a table and leaving a box inside a container. In this way, the final dataset contains 12 labels and 2034 acquisitions, and it was used to train and to test a graph convolutional neural network. Results from the testing of the NN reveal a high accuracy in the recognition of the newly added actions. The work here presented aims at supporting automatic ergonomic evaluation in picking activities.

Keywords: Human Action Recognition, Depth camera, skeletal data, deep learning

I. INTRODUCTION

Industry 5.0 represents the new industrial paradigm. By putting human factor at the centre of any process, the fifth industrial revolution restructures human tasks in manufacturing, monitoring and reducing risks [1]. There are many technologies to satisfy workers' needs and to ensure a safe environment: smart and connected sensors can provide a real-time overview of climate, temperature and consumption of energy [2]. Along with environment control, Industry 5.0 technologies allow an automatic ergonomic risk assessment to reduce work related musculoskeletal disorders (WMSDs)In this context there was a substantial development of sensors. An important category is represented by marker-less sensors, which are more affordable and less intrusive than marker-based ones [3]. Specifically, the most used technology is Kinect: a depth-RGB camera able to provide real-time body segmentation [4]. Three different generations of depth-RGB camera were developed by Microsoft: the first one was Kinect v1, followed by Kinect v2 and the latest version is Azure Kinect [5]. These sensors are used to get skeleton data through body tracking. Recent applications of Kinect are in the field of Human Action Recognition (HAR). HAR is the process able to decipher human action and automatically analysing, understanding, and classifying tasks taking as input depth-RGB camera data [6]. HAR could be useful in health monitoring, industrial applications for human-robot interaction [7], but also monitoring and identify hazardous postures during working activities. To improve the automatization of ergonomic risk assessment, our work presents an extension of Northwestern-UCLA dataset [8] with 2 new classification labels that are: leaving a box on a table and leaving a box inside a container. We also generalized task recognition performing activities in different conditions: in the original dataset a box is picked up from the ground in all the samples, so we added videos where same task is performed but the picking was performed from different heights and surfaces. Then, we used a graph

convolution neural network with the dataset to classify tasks and to predict when the activity starts and when it stops. Our purpose is to use the output of a NN to identify human posture and to automatically calculate National Institute for Occupational Safety and Health (NIOSH) Lifting Equation [9].

This paper describes the steps we made:

1. Data gathering using Azure Kinect.
2. Data pre-processing to standardize our data to Northwestern-UCLA dataset.
3. CTR-GCN [10] training and testing on our dataset.
4. Prediction of the start and the stop of each classified activity.

The paper is structured as follows: the first part is a brief review of the *state-of-the-art*, then we describe our methods. We started with Northwetern-UCLA Multiview Action3D Dataset [8] description and an explanation of how we expanded it. In this paper there is also the description of data pre-processing and the following implementation of the final dataset on a neural network. In the end there are results and conclusions.

II. STATE-OF-THE-ART

Many datasets for human action recognition have been built using different sensors. The type of the data can be classified in Uni-modal and Multi-modal. Uni-modal data gathering involves the use of a single type of sensor, while in Multi-modal method multiple type of sensors are combined [6]. Microsoft Kinect combines RGB-Depth sensors so it can be classified as Multi-modal [11]. The first dataset built for HAR is named HDM05 [12]. It is based on an optical marker-based technology: Vicon MX system comprising 12 cameras, 6 of which operated in the visible red and 6 of which operated in the infrared spectral range. [12] Body segmentation in HDM05 Dataset includes 24 joints. This dataset contains 2337 samples performed by 5 subjects and 130 classes [12]. One of the first dataset populated using Microsoft Kinect v1 is HOJ3D in 2012 [13]. HOJ3D contains 200 samples classified in 10 labels. The dataset is cross-subject because tasks were performed by 10 different people and human body was tracked by identifying 20 body joints [13]. Northwestern-UCLA Multiview Action3d Dataset [8] was populated by three Kinect v1 acquisitions. This dataset contains 1475 samples classified in 10 different labels of common action in daily life. Also, in Northwester-UCLA dataset, tasks were performed by 10 different persons, and it is also cross-view and cross-environment.

The second generation of Kinect was used in NTU-RGB+D dataset. In the first version (RGB+D 60) [14] it was made by 56880 samples in 60 different classes, but it was updated in RGB+D 120 [15] where 114480 samples are classified in 120 different labels. An additional innovative contribution that NTU-RGB+D has brought to the field of HAR is data gathering by Kinect v2, which tracks the human body through 25 joints. However, the 120 actions performed in the dataset are not usable for industrial purpose; there are daily tasks like: put on a jacket, kick backward, arm circles, high five, cheers and drink and brush teeth [15]. For this reason, we did not use it in our project. After a thorough analysis of the *state-of-the-art* in the field of human action recognition datasets we decided to use for our scope Northwestrn-UCLA Multiview Action3D Dataset [8] because it has enough samples to train and validate a neural network and it has a limited number of different labels. In this way, each action is characterized by specific movements and the classification is almost unequivocal.

Kinect v1 and v2 were widely used for data gathering but at the best of our knowledge and as confirm in [6] there are not dataset with Azure Kinect data. An additional gap that we found is the absence of a skeleton-based dataset containing labels usable in industrial field. We decided to add some working tasks to an existing dataset because we would use HAR for automatic ergonomic risk assessment during manual handling of loads, by using NIOSH Lifting Equation [9]. Our work proposes a data collection through Azure Kinect with a special focus on picking up and leaving loads in different conditions.

III. METHODS

We extended the Northwestern-UCLA dataset [16] with 540 acquisitions collected in our laboratory under three settings: cross-subject, cross-view and cross-environment. The aim of this work is to build a useful dataset for automatic risk assessment starting from an existing one. We had to use an existing dataset since we did not have enough data to train and test a neural network, with the aim to build our dataset in future. We chose the dataset mentioned above because it is easy to understand, there are already some labels useful for our scope (e.g. *pick up with two hands*, *carry, pick up with one hand*), there are enough data to implement it on a NN, dataset is cross-subject, cross-view and cross-environment [16].

A. Northwestern-UCLA Multiview Action3D Dataset

Northwestern-UCLA dataset [16] contains RGB, depth and human skeleton data captured by three Microsoft Kinect v1. In the dataset there are 1494 samples classified in 10 labels:

1. *Pick up with one hand.*
2. *Pick up with two hands.*
3. *Drop trash.*
4. *Walk around.*
5. *Sit down.*
6. *Stand up.*
7. *Donning.*
8. *Doffing.*
9. *Throw.*
10. *Carry.*

These activities were performed by 10 different subjects and data were taken from different viewpoints. Some images of dataset are shown in Fig. 1. Using Kinect v1 for data gathering, skeleton segmentation contains 20 joints.

Fig. 1. Example of 10 labels in the Northwestrn-UCLA dataset [8] where: (1) is pick up with one hand, (2) is pick up with two hands, (3) is drop trash, (4) is walk around, (5) is sit down, (6) is stand up, (7) is donning, (8) is doffing, (9) is throw, (10) is carry

B. Experiment setting and data gathering

To collect and process data, we used Alienware PC M17 R4, with a 10th generation Intel® Core™ i7-10870H processor (8 cores, 16 MB cache memory, maximum turbo frequency 5.0 GHz), Windows 10 Home (64-bit), NVDIA® GeForce RTX™ 3060 graphics card with 6GB of GDDR6, and 32GB of DDR4 RAM at 2,933 MHz. We installed the Visual Studio Code editor on this machine to work with Python 3.4. Experiment setting includes one Azure Kinect, a 90 cm high table, a 10 cm high shelf, a 50 cm high shelf, three box with dimensions: 33 cm x 18 cm x 17 cm; 50 cm x 30 cm x 25 cm; 41 cm x 13 cm x 41 cm, the setup is shown in Fig. 3.

Azure Kinect settings during acquisitions were the following:

- Color mode: On 720p.
- Depth mode: On NFOV_2X2BINNED.
- No depth delays.
- Frames per second: 15.
- IMU: ON.
- External sync: Standalone.

- Sync delay: 0.
- Exposure: Auto.
- Gain: Auto.

540 videos were collected where 5 different subjects performed 5 different tasks:

1. *Pick up a box with two hands from a height of 10 cm.*
2. *Pick up a box with two hands from a height of 50 cm.*
3. *Pick up a box with two hands from the ground.*
4. *Leaving a box on a table.*
5. *Leaving a box inside a container.*

In Fig.2 new tasks are shown.

Fig. 2 Example of 5 labels of our dataset: (1) pick up a box with two hands from a height of 10 cm, (2) pick up with two hands from a height of 50 cm, (3) pick up with two hands from the ground, (4) leaving a box on a table; (5) leaving a box inside a container

The first three tasks were classified with label 2 of the existing dataset: *pick up with two hands*. By analysing videos of Northwestern-UCLA dataset [8], we found a gap: each subject performed in same way and the picking was always in the same condition. Specifically, they picked up the same box in each video and they picked up the box from the ground in all the acquisitions. So, the dataset was not very generalized and it was not useful for our purpose. We wanted to have a good accuracy in prediction of picking up activities of different loads in different conditions, to use the dataset for automatic ergonomic risk assessment. Task 4 and task 5 were classified with two new labels. In this way, the final dataset contains 12 labels and 2034 samples.

C. Data pre-processing

Our data were collected using Azure Kinect. The new technology of Microsoft is able to recognize 32 body joints, while Kinect v1 recognizes 20 body joints. We had to build a process for the conversion of our data, unavoidably losing some information. In Fig.3 the two body hierarchies are compared. To obtain Azure v1 hierarchy we had to ignore 12 body joints and to re-build joint connections.

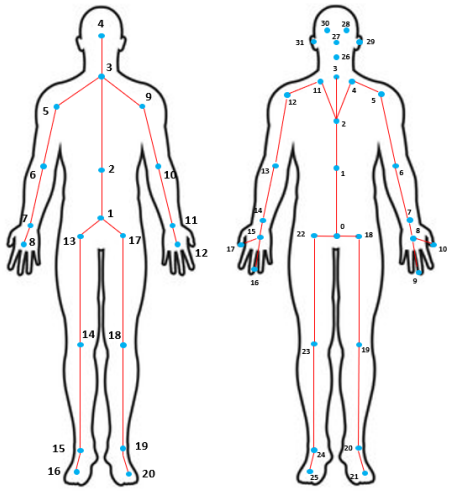

Fig. 3. Comparison between Microsoft Kinect v1 body hierarchy (on the left) and Azure Kinect body hierarchy (on the right)

Pre-processing follows these steps:

1. Conversion of the Azure Kinect output: files from Kinect are *.mkv* type, by using Powershell we get *.json* files, which are the type used in the dataset.
2. Checking for completeness and correctness of the data for each frame.
3. Elimination of any compromised frames.
4. Editing of body joints list to keep only 20 joints and them coordinates.
5. Re-order of the list to create the new joints hierarchy.
6. Coordinate conversion in the reference system used.

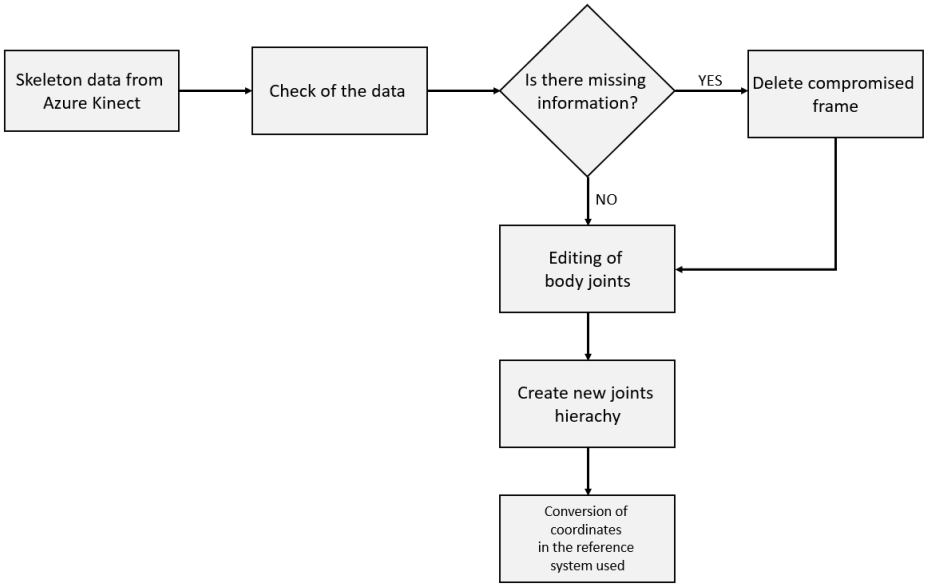

Fig. 4. Flowchart representing pre-processing steps

The development of this procedure has been an extremely iterative process. We tested it on 24 acquisitions before getting congruent data with Northwestern-UCLA dataset.

D. Data implementation on a graph convolutional neural network

Our contribution is aimed to train a neural network to classify useful actions for automatic risk assessment. To test this, we implemented the final dataset on a Channel-wise Topology Refinement Graph Convolution Network (CTR-GCN) [10]. CTR-GCN was already trained and tested on Northwestern-UCLA Multiview Action3D Dataset [8]. On our work, training set contains 1372 samples where 350 are our acquisitions and 1022 are from existing dataset. Validation set contains 623 samples, where 151 are from new samples. In the best epoch we got following results: mean error of 0,309; Top1 accuracy of 92,33%; Top5 accuracy of 99,02%. In Fig.5 and Fig.6 accuracy and loss function trend are shown.

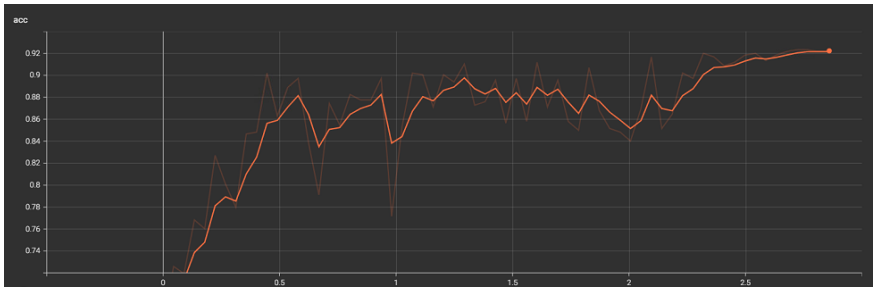

Fig. 5. Accuracy trend during training phase. There is accuracy in ordinates and steps in abscissa

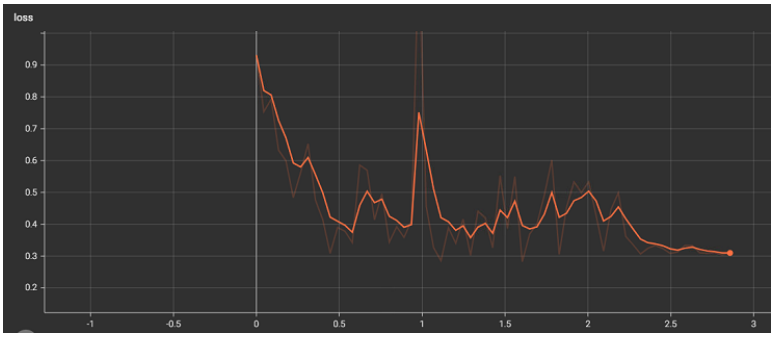

Fig. 6. Loss function trend during training phase. There is error value in ordinates and steps in abscissa

Specifically, we focused on the accuracy for label 2 (*pick up with two hands*), label 11 (*leaving a box on a surface*) and label 12 (*leaving a box inside a container*) because we extended the dataset adding these tasks. Results are shown in TABLE I.

TABLE I
ACCURACY OF INTEREST LABELS

Label	Accuracy
Label 2	0,992
Label 11	0,979
Label 12	0,929

Testing phase was done with 39 samples classified with label 2, label 11 and label 12. In this way we could test the accuracy of CTR-GCN [10] for our scope. Samples in test set were performed by two different subjects to avoid overfitting through a cross-subject analysis. Results we obtained are: Top1 accuracy of 90,32% and Top5 accuracy of 96,77%.

IV. CONCLUSIONS

We propose an extended version of an existing dataset for human action recognition. We added some working activities like manual handling of loads with the scope to use HAR for automatic risk assessment. We implemented the final dataset on a neural network, and we obtained high accuracy for labels of our interest (*pinking up a box* and *leaving a box on different surfaces*). The proposed dataset is one of the steps to automatically calculate NIOSH Lifting Equation through Azure Kinect. Future developments involve task recognition to classify working actions and to automatically identify start time, stop time, duration, frequency of various activity. Then, by using this output the ergonomic risk assessment will be fully automated.

REFERENCES

[1] F. Longo, A. Padovano, and S. Umbrello, "Value-oriented and ethical technology engineering in industry 5.0: A human-centric perspective for the design of the factory of the future," *Applied Sciences (Switzerland)*, vol. 10, no. 12, pp. 1–25, Jun. 2020, doi: 10.3390/APP10124182.

[2] V. Martos, A. Ahmad, P. Cartujo, and J. Ordoñez, "Ensuring agricultural sustainability through remote sensing in the era of agriculture 5.0," *Applied Sciences (Switzerland)*, vol. 11, no. 13. MDPI AG, Jul. 01, 2021. doi: 10.3390/app11135911.

[3] A. Altieri, S. Ceccacci, A. Talipu, and M. Mengoni, "A low cost motion analysis system based on RGB cameras to support ergonomic risk assessment in real workplaces," in *Proceedings of the ASME Design Engineering Technical Conference*, American Society of Mechanical Engineers (ASME), 2020. doi: 10.1115/DETC2020-22308.

[4] Microsoft, "About Azure Kinect DK," 2021. https://learn.microsoft.com/en-gb/azure/kinect-dk/about-azure-kinect-dk

[5] M. Tölgyessy, M. Dekan, Ľ. Chovanec, and P. Hubinský, "Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2," *Sensors (Switzerland)*, vol. 21, no. 2, pp. 1–25, Jan. 2021, doi: 10.3390/s21020413.

[6] A. Sarkar, A. Banerjee, P. K. Singh, and R. Sarkar, "3D Human Action Recognition: Through the eyes of researchers," *Expert Systems with Applications*, vol. 193. Elsevier Ltd, May 01, 2022. doi: 10.1016/j.eswa.2021.116424.

[7] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," *Journal of Imaging*, vol. 6, no. 6. MDPI, Jun. 10, 2020. doi: 10.3390/jimaging6060046.

[8] J. wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view Action Modeling, Learning and Recognition," May 2014, [Online]. Available: http://arxiv.org/abs/1405.2941

[9] T. R. Waters, V. Putz-Anderson, A. Garg, and L. J. Fine, "Revised NIOSH equation for the design and evaluation of manual lifting tasks," *Ergonomics*, vol. 36, no. 7, pp. 749–776, 1993, doi: 10.1080/00140139308967940.

[10] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.12213

[11] Zhengyou Zhang, "Microsoft Kinect Sensor and Its Effect," 2012. [Online]. Available: www.microsoft.com/

[12] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation Mocap Database HDM05".

[13] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 20–27. doi: 10.1109/CVPRW.2012.6239233.

[14] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," Apr. 2016, [Online]. Available: http://arxiv.org/abs/1604.02808

[15] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," May 2019, doi: 10.1109/TPAMI.2019.2916873.

[16] J. wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view Action Modeling, Learning and Recognition," May 2014, [Online]. Available: http://arxiv.org/abs/1405.2941