Deep learning architectures applied for recognizing human motion primitives from the Ergonomic Assessment Worksheet

A. Aubert, B. E. Olivas-Padilla, V. Syrris, S. Manitsaris

Abstract— **On-site ergonomic prevention aims at decreasing work-related musculoskeletal disorders in the industry. Recognition of hazardous posture during the execution of manual tasks by humans moves towards this direction by helping to classify those tasks as potential assignments to a robot. Sensor-driven motion capture using body-mounted inertial sensors associated with pattern recognition enables to track joints all along with a hazardous task. In this work, different deep neural networks were compared for time series classification of motion primitives based on the Ergonomic Assessment Worksheet. Moreover, to investigate the possibility of using a reduced number of sensors for the recognition, the performance achieved using only an optimal set of sensors is contrasted as well. This optimal set of sensors was formed using a stochastic-biomechanic approach, the Gesture Operational Model. Compared to all-sensor configuration, the best performance was achieved using the optimal set of sensors and the Rocket transformation.**

I. INTRODUCTION

The growing implementation of human-robot collaboration in industry has led to the development of human motion recognition for better interaction and increased safety. More than a tool, the robot becomes an extension of the human body which will ultimately free the worker from any hazardous task. Preventing the worker from repeated exposure to high- or low-intensity loads over a long period of time, is key to avoiding most work-related musculoskeletal disorders (MSDs). MSD prevalence is highest among "blue-collar workers", such as plant and machine operators and assemblers (66%) [1]. From lower and upper limbs to the back, shoulders and neck, the spectrum of muscular pains reported in work-related MSDs is broad and frequently associated with mental well-being low-levels. Preventive measures are effective not only on the workers' health but also on economic and societal dimensions by reducing healthcare costs, work absenteeism or productivity loss.

Any prevention strategy begins with an on-site ergonomic risk assessment, which can be conducted by worker self-assessment, observation by ergonomists, direct measurements on the worker's body, or computer-based assessment from camera recordings [2, 3]. Real-time continuous ergonomic assessment based on inertial measurement units (IMUs) in an industrial environment enables both identifying the most exposed joints during task execution and detecting the most MSD-dangerous tasks [4]. A restrained number of IMUs embedded in an arm-mounted smartphone provides pattern-distinguishable data that reflect complex activity pertained to occupational tasks and can be exploited by machine learning classification models [5]. The breakthrough of human activity recognition (HAR) fuelled by the progress in recent years of both sensor-based motion capture and machine learning approaches has paved the way to a real-time evaluation of ergonomic risks on industrial sites which would use robots to human monitoring and assistance in the riskiest tasks [6].

In this framework, the present study implements deep learning recognition of motion primitives based on the EAWS Ergonomic Assessment Worksheet (also called European Assembly Worksheet [7]) with the aim to reliably detect motion primitives with known ergonomic risk levels. The motion primitives' detection allows estimating accurately the ergonomic risks workers are exposed to throughout the execution of any professional tasks. This information may enable the design of ergonomically effective human-robot collaboration systems in which the robots execute the most hazardous motions while preventing workers from exposing themselves to ergonomic risks.

Design a Human Robot Collaboration (HRC) which is both human-centred and compliant to physical and cognitive ergonomic principles is a challenging research area [8]. General HRC principles for system design (ISO 9241-110:2019) stresses "suitability for individualization", namely "the possibility of adapting the robot to the workers' needs and abilities" [9]. Personalised levels of assistance entail robots' flexibility and mobility in the collaborative workspace either by straightforwardly leading the worker towards an optimized ergonomic body posture to fulfil a task [10, 11] or by providing assistance as soon as a fatigue threshold is exceeded [12]. Any human-robot interaction requires a human kinodynamic monitoring coupling whole-body motion recording with forward dynamics musculoskeletal modelling based for instance on electromyography (EMG) [9]. EAWS-related activities recognition algorithms based on whole-body kinematic tracking enable bypassing musculoskeletal models by providing an automatic ergonomic scoring to workers [13]. Implemented in task scheduling strategy, this feedback could enable fine-tuned adjustment of human–robot task sequence assignment and planning based on physical ergonomics aspects [14].

Deep learning has proven successful in recognition of human movements and professional tasks [15]. Thus, this study explores diverse deep learning approaches to identify the most efficient method for recognizing the motion primitives. Moreover, it evaluates the possibility of training deep learning models with data from a reduced set of sensors

A. A. is with the Centre for Robotics, MINES ParisTech, PSL Université Paris (corresponding author; e-mail: agnes.aubert@mines-paristech.fr).

B.E. O.P. is with the Centre for Robotics, MINES ParisTech, PSL Université Paris (e-mail: brenda.olivas@mines-paristech.fr).

V. S. is a partner of the AIMOVE Post Master's professional program, MINES ParisTech, PSL Université Paris (e-mail: vasileios.syrris@mines-paristech.fr).

S. M. is with the Centre for Robotics, MINES ParisTech, PSL Université Paris (e-mail: sotiris.manitsaris@mines-paristech.fr).

solely (optimal set), defined by the Gesture Operation Model as proposed by Olivas-Padilla et al. [16]. The next section presents the HAR pipeline on which this study is based, notably the generated data set, the statistical analysis for feature selection, and the baseline Hidden Markov Models (HMM) classification. The different deep learning time series classifiers are described in the methodology section. Accuracy performance and discussion of the main results are presented in Section IV. Section V summarizes the present work and sets the way forward.

II. RELATED WORK

Sensor-based motion capture using body-mounted inertial sensors allows rebuilding the kinematic skeleton through the orientation and acceleration of the measured body segments. The motion trajectories are thus directly connected to the body joint movements involved in the worker's motions, providing critical information regarding body posture during a hazardous task. Based on this hypothesis, Olivas-Padilla and coll. developed an occupational vocabulary based on the EAWS evaluation including symmetric and asymmetric working postures for standing, sitting, and kneeling with rotation, lateral bending and far reach [7, 16, 17] (Table I). This EAWS data set is composed of 28 motions recorded in ten healthy human agents, with three repetitions, six seconds each, using IMU full-body suit, and described by computed whole-body joint angles (Euler angles) on the three axes x, y, and z. Based on state-space modelling, the Gesture Operational Model (GOM) proposed by Manitsaris and coll. [18] has been applied to describe the cooperation of body parts in the execution of the EAWS motions. A statistical analysis over the models that compose GOM enabled the selection of an optimal set of five sensors contributing the most to the 28 motion primitives and maximizing the F-score of HMM classification by 91.77% [16].

III. METHODOLOGY

To compare deep learning recognition of hazardous motions to the HMM approach, a range of deep neural networks (DNNs) have been applied to the EAWS data set collected by Olivas-Padilla and coll. both in its full and its optimal-GOM sensor configurations, namely Spine 2, Left Arm, Right shoulder, Right UpLeg, Left Fore Arm [16]. In addition to those 3D joint angles coordinates ("Full-angles" and "Optimal-angles"), 3D local joint positions coordinates have been computed through the full-sensor configuration, i.e. "Full-positions". To leverage computer vision dedicated models as well, 3D joint full-positions have been converted into RGB images according to the process defined by [19]. All in all, four variants of the EAWS (Table II) have been used to train the selected DNNs.

TABLE I. TWENTY-EIGHT MOTION PRIMITIVES (M_{01} TO M_{28}) DERIVED FROM EAWS DATA SET [7] ACCORDING TO [16, 17].

	Standing postures	**Seated**	**Kneeling**
Upright	$^{*}M_{01}$, M_{02}^{a}, M_{03}^{b}	$^{**}M_{15}$	$^{***}M_{21}$, M_{22}^{a}, M_{23}^{b}
Bent forward (20-60°)	M_{04}, $M_{05}^{a,b}$, $M_{06}^{a,b,c}$		
Strongly bent forward (>60°)	M_{07}, $M_{08}^{a,b}$, $M_{09}^{a,b,c}$	M_{16}, $M_{17}^{a,b}$, $M_{18}^{a,b,c}$	M_{24}, $M_{25}^{a,b}$, $M_{26}^{a,b,c}$
With elbow at / above shoulder level	M_{10}^{d}, $M_{11}^{a,b,d}$, $M_{12}^{a,b,c}$		M_{27}^{c}, $M_{28}^{a,b,c}$
With hands above head level	M_{13}, $M_{14}^{a,b,c}$	M_{19}^{c}, $M_{20}^{a,b,c}$	

Initial postures: *standing, no body support; **upright with back support slightly bent forward or backwards; ***upright. Variations in the postures: a. with torso rotation; b. with laterally torso bending; c. arms stretching; d. with forearms 90° bending.

Different DNNs for multivariate time series classification have been implemented according to the overview displayed by Fawaz and coll. [20] (Table III). DNNs apply non-linear transformations on the input time series to finally output probability distribution over classes. In generative DNNs, the time series are first represented in a latent space before classification [21, 22], while in discriminative ones, hidden features are directly learnt from time series. Outside of DNNs framework, one of the most effective machine learning algorithms, the Gradient boosting (namely gradient boosted decision trees), has been also tested. Overall, 14 models have been trained with four variants of the EAWS data set and transfer learning has been applied by employing four pretrained models on ImageNet with the "RGB images". Transfer learning is consisted of shallow retraining of only the last layers of the model. Lightweight model architectures have been employed to mitigate the small size of the EAWS data set of 840 samples, with only slight adaptations in the number of convolutional or fully connected layers, filters and kernel size, according to the specificity of the data set (e.g., low or high dimensionality, sequences or RGB images).

The inputs of the models (I) vary according to both the models' requirements and the variants of the EAWS data set, namely (1) Optimal-angles (I1), (2) Full-angles and Full-positions (I2), and (3) RGB images (I3) (Table III). Among the end-to-end models dedicated to times series models, the family of Gated Recurrent Units (GRU) and Bidirectional-GRU (Bi-GRU) is represented with its hybrids formed with the Convolutional Neural Networks (CNNs) (CNN-GRU and CNN-BiGRU). Well-known in computer vision, the CNNs are also effective for time series classification. CNN 1D concatenates the 3D joint coordinates in one vector for every time step while CNN 2D is trained on three time-series corresponding to x, y, z (e.g., for "Optimal-angles", I_1 = (540,15) and (540,5,3) respectively). Rocket model developed by Dempster and coll. uses 10,000 random convolutional kernels to transform the time series into features to be trained by a linear classifier [23]. The MiniRocket model is a more deterministic and optimized variant of the former one [24]. Finally, the Temporal Convolutional Network (TCN) exploits the features of CNNs and Recurrent Neural Networks (RNNs) by using causal and dilated convolutions to capture long-range temporal patterns with an output of the same length as the input [25].

TABLE II. FOUR VARIANTS OF THE EAWS DATA SET COMPUTED FROM JOINT ANGLES [16].

Joints computation on x, y, z	**Full-angles**	**Optimal-angles**	**Full-positions**	**RGB images[a]**
Joints number	65	5	65	65
Time series length	540 (6 seconds x 90 frames/second)			
Data design	840 (10 human agents x 28 motions x 3 repetitions)			

a. Conversion of joint full-positions on x, y, z according to [19].

TABLE III. DEEP NEURAL NETWORKS SELECTED FOR MULTIVARIATE TIMES SERIES CLASSIFICATION [20].

Models	(1) Optimal-angles I_1 = (540,15)	(2) Full-angles and Full-positions I_2 = (540,195)	(3) RGB images I_3 = (224,672)
I. Discriminative models – End-to-End			
Convolutional Neural Networks			
CNN 1D K7[a]	$I_{1,2,3}$→C(16,7)→C(64,7)→MP(5)→F→FC(128)→FC(28)		
CNN 1D K540[a]	$I_{1,2}$→C(16,540)→F→FC(195)→FC(28)		
	I_3→C(16,3)→C(64,3)→F→FC(128)→FC(28)		
CNN 2D 1c[b]	I_1→C(16,5)→F→FC(128)→FC(28)		
	$I_{2,3}$→C(16,7)→MP(5)→F→FC(128)→FC(28)		
CNN 2D 2c[b]	I_1→C(16,5)→C(64,1)→F→FC(128)→FC(28)		
	$I_{2,3}$→C(16,7)→C(64,7)→MP(5)→F→FC(128)→FC(28)		
Rocket and MiniRocket	kernels=10,000, ridge classifier		
Temporal Convolutional Network	$I_{1,2,3}$→TCN(11)→FC(28)		
Gated Recurrent Units and Bidirectional-GRU			
GRU	$I_{1,2,3}$→GRU(256)→GRU(128)→FC(28)		
Bi-GRU	$I_{1,2,3}$→Bi-GRU(256)→Bi-GRU(128)→FC(28)		
CNN-GRU	$I_{1,2,3}$→C(16,7)→C(64,7)→MP(5) →GRU(256)→GRU(128) →FC(28)		
CNN-BiGRU	$I_{1,2,3}$→C(16,7)→C(64,7)→MP(5) →Bi-GRU(256)→Bi-GRU(128)→FC(28)		
II. Discriminative models – Time series transformation into RGB images			
Pretrained models on ImageNet with input I_3=(224,224,3), shallow retraining			
I_3→VGG16/MobileNet/DenseNet201/InceptionV3→F→FC(128) →FC(28)			
III. Generative models			
GRU-autoencoder			
Encoder	$I_{1,2,3}$→GRU(256)→GRU(128)→RepeatVector(time steps)…		
Decoder	→GRU(128)→GRU(256)→TimeDistributed(FC(features))		
Full-model	Encoder→F→FC(128)→FC(28)		
Echo state network	units in reservoir: 500		

I=(,): input size(time steps, features); C(,): convolutional layer(filters,kernel); MP(): max pooling(pool size); F: flatten; FC(): fully connected layer(units); TCN(): temporal convolutional layer (kernel). a. CNN 1D with Kernel 7/540. b. CNN 2D with 1/2 convolutional layer(s). Optimizer= Adam(learning rate= $1e^{-3}$); loss = sparse categorical crossentropy (except GRU-autoencoder: loss(autoencoder) = mean squared error; loss(full-model)=categorical crossentropy); earlyStopping; activation function=softmax; train/test split (90-10%).

IV. RESULTS AND DISCUSSION

Sixty trainings have been performed using the 14 models applied to the four EAWS data set variants and the 4 pretrained models on "RGB images". Table IV displays the 19 results exceeding the result (F-score of 91.77%) obtained with the GOM-HMM pipeline on 3D joint angles coordinates with the optimal sensor-set [16]. Eight out of 19 results are attributed to Rocket and MiniRocket models which confirm their high efficiency in time series classification. Eleven out of 19 results have been achieved by the selected joint angles by GOM ("Optimal-angles"), while for the other variants of the EAWS data set, almost exclusively Rocket and MiniRocket outperformed the 91.77% F-score. GRU, Bi-GRU, CNN-BiGRU, CNN-GRU, ESN, and TCN provided good performance as well as the CNN models with the lighter architectures in this small dataset framework.

Pretrained models on ImageNet do not outperform the 91.77% F-score due to the artificial RGB images from EAWS data set that differ significantly from the natural color images of ImageNet. Furthermore, the change of domain from 3D coordinate sequences to RGB images does not provide additional advantages usually pertained to image data sets. Finally, an analysis of the misclassified motions according to the different models was conducted which led to the characterization of 9 movements from the EAWS data set as outliers.

TABLE IV. MAIN RESULTS OBTAINED WITH DIFFERENT DEEP LEARNING APPROACHES AND VARIANTS OF EAWS DATA SET. SHADED BOXES MARK F-SCORES ABOVE 91.77%, THAT IS THE F-SCORE OBTAINED WITH GOM-HMM PIPELINE ON 'OPTIMAL-ANGLES' DATA SET [16].

Dataset	Full-angles	Optimal-angles	Full-positions	RGB images
Rocket	0.929	0.988	0.928	0.940
MiniRocket	0.976	0.928	0.953	0.937
ESN[a]	0.951	0.976	0.916	0.889
TCN[b]	0.873	0.952	0.904	0.863
GRU[c]	0.803	0.964	0.828	0.887
CNN[d]-GRU	0.855	0.951	0.866	0.782
Bi[e]-GRU	0.807	0.900	0.833	0.808
CNN-BiGRU	0.845	0.952	0.864	0.819
CNN 1D K7[f]	0.917	0.948	0.906	0.632
CNN 1D K540[g]	0.893	0.940	0.889	0.684
CNN 2D 1c[h]	0.861	0.940	0.928	0.863
CNN 2D 2c[i]	0.795	0.940	0.897	0.786
GRU-autoE[j]	0.731	0.797	0.658	0.717
Gradboost[k]	0.901	0.911	0.709	0.687
VGG16[l]	–	–	–	0.913
DenseNet201[l]	–	–	–	0.892
MobileNet[l]	–	–	–	0.900
InceptionV3[l]	–	–	–	0.754

a. Echo state network. b. Temporal Convolutional Network. c. Gated Recurrent Units. d. Convolutional Neural Networks. e. Bidirectional. f. CNN 1D with Kernel 7. g. CNN 1D with Kernel 540. h. CNN 2D with 1 convolutional layer. i. CNN 2D with 2 convolutional layers. j. GRU-autoencoder. k. GradBoost: Gradient boosting (not a DNN), XGBClassifier (objective='multi:softmax', num_classes=28, max_depth=7, n_estimators=100). l. Transfer learning with shallow retraining of only the last layers and without data augmentation.

V. CONCLUSION

This paper explores the application of diverse DNNs to recognize motion primitives with different ergonomic risk levels and evaluates the possibility of using a reduced set of selected sensors for tracking the joints of the human body. Choosing the most meaningful sensors with the GOM approach and training DNNs with their corresponding data results in a high classification performance of the 28 motion primitives. This opens the very promising possibility of detecting accurately dangerous motions on industrial sites using only a small number of sensors. An additional study on the selection of the joint positions which contribute the most to the 28 motions using GOM will be carried out in comparison with the results of the joints angles selection. Establishing an occupational vocabulary dedicated to human motion recognition unlocks a wide range of possibilities for on-site ergonomic risk assessment, and with respect to hazardous tasks, it leads to the future undertaking of those tasks by robots.

ACKNOWLEDGMENT

We thank Dr. Sohaib Laraba for converting the time series into RGB images. This work resulted from research supported by the EU Horizon 2020 Research and Innovation Programme under grant agreement No. 820767, project CoLLaboratE.

REFERENCES

[1] J. de Kok, P. Vroonhof, J. Snijders, G. Roullis, M. Clarke, K. Peereboom, P. van Dorst, and I. Isusi, *Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU*. European Agency for Safety and Health at Work, Publications Office of the European Union: Luxembourg, 2019.

[2] G. C. David, "Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders", *Occup. Med. (Lond.)*, vol. 55, pp. 190–199, 2005.

[3] M. Massiris Fernández, J. A. Fernández, J. M. Bajo, and C. A. Delrieux, "Ergonomic risk assessment based on computer vision and machine learning", *Comput. Ind. Eng.*, vol. 149, p. 106816, 2020.

[4] N. Vignais, F. Bernard, G. Touvenot, and J.-C. Sagot, "Physical risk factors identification based on body sensor network combined to videotaping", *Appl. Ergon.*, vol. 65, pp. 410–417, 2017.

[5] N. D. Natha, T. Chasparib, and A. H. Behzadana, "Automated ergonomic risk monitoring using body-mounted sensors and machine learning", *Adv. Eng. Inform.*, vol. 38, pp. 514–526, 2018.

[6] B. Parsa, E. U. Samani, R. Hendrix, C. Devine, S. M. Singh, S. Devasia, and A. G. Banerjee, "Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks", *IEEE Robot. Autom. Lett.*, vol. 4, p. 3153, 2019.

[7] K. Schaub, G. Caragnano, B. Britzke, and R. Bruder, "The European Assembly Worksheet". *Theor. Issues Ergon. Sci.*, vol. 14, pp. 616–639, 2013.

[8] L. Gualtieri, I. Palomba, E. J. Wehrle, and R. Vidoni, "The opportunities and challenges of SME manufacturing automation: safety and ergonomics in human–robot collaboration," in *Industry 4.0 for SMEs*, D. Matt, V. Modrák, and H. Zsifkovits, Eds. Cham: Palgrave Macmillan, 2020, pp. 105-144.

[9] A. Ajoudani, P. Albrecht, M. Bianchi, A. Cherubini, S. Del Ferraro, and P. Fraisse, et al., "Smart collaborative systems for enabling flexible and ergonomic work practices", *IEEE Robot. Autom Mag.*, vol. 27, pp. 169-176, 2020.

[10] B. Busch, G. Maeda, Y. Mollard, M. Demangeat, and M. Lopes, "Postural optimization for an ergonomic human-robot interaction", *IEEE Int. C. Int. Robot.*, pp. 2778-2785, 2017.

[11] W. Kim, M. Lorenzini, P. Balatti, Y. Wu, and A. Ajoudani, "Towards ergonomic control of collaborative effort in multi-human mobile-robot teams". *IEEE Int. C. Int. Robot.*, pp. 3005-3011, 2019.

[12] M. Lorenzini, W. Kim, E. De Momi, and A. Ajoudani, "A new overloading fatigue model for ergonomic risk assessment with application to human-robot collaboration", *IEEE Int. Conf. Robot*, pp. 1962-1968, 2019.

[13] A. Malaisé, P. Maurice, F. Colas, and S. Ivaldi, "Activity recognition for ergonomics assessment of industrial tasks with automatic feature selection," *IEEE Robot. Autom. Lett.*, vol. 4, pp. 1132–1139, 2019.

[14] L. Gualtieri, E. Rauch, and R. Vidoni, "Emerging research fields in safety and ergonomics in industrial collaborative robotics: A systematic literature review". *Robot. Com.-Int. Manuf.*, vol. 67, p. 101998, 2021.

[15] Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.

[16] B. E. Olivas-Padilla, S. Manitsaris, D. Menychtas, and A. Glushkova, "Stochastic-biomechanic modeling and recognition of human movement primitives, in industry, using wearables", *Sensors*, vol. 21, p. 2497, 2021.

[17] B. E. Olivas-Padilla, D. Menychtas, A. Glushkova, and S. Manitsaris, "Hidden Markov modelling and recognition of Euler-based motion patterns for automatically detecting risks factors from the European assembly worksheet", in *2020 IEEE Int. Conf. on Image Processing*, pp. 3259–3263.

[18] S. Manitsaris, G. Senteri, D. Makrygiannis, and A. Glushkova, "Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories", *Front. Robot. AI*, vol. 7, pp. 1–20, 2020.

[19] S. Laraba, M. Brahimi, J. Tilmanne, T. Dutoit, "3D skeleton-based action recognition by representing motion capture sequences as 2D-RGB images", *Comput. Anim. Virtual Worlds*, vol. 28, p. e1782, 2017.

[20] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Deep learning for time series classification: a review", *Data Min. Knowl. Disc.*, vol. 33, pp. 917–963, 2019.

[21] F. M. Bianchi, S. Scardapane, S. Løkse, and R. Jenssen, "Reservoir computing approaches for representation and classification of multivariate time series", arXiv:1803.07870v3 [cs.NE], 2020.

[22] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, "Spatial-Temporal Data Augmentation Based on LSTM Autoencoder Network for Skeleton-Based Human Action Recognition", in *2018 IEEE Int. Conf. on Image Processing*, pp. 3478–3482.

[23] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: Exceptionally fast and accurate time classification using random convolutional kernels." *Data Min. Knowl. Disc.*, vol. 34, pp. 1454–1495, 2020.

[24] A. Dempster, D. F. Schmidt, and G. I. Webb, "MiniRocket: A very fast (almost) deterministic transform for time series classification", arXiv:2012.08791v1 [cs.LG], 2020.

[25] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling", arXiv:1803.01271v2 [cs.LG], 2018.