Multi-View Skeleton Analysis for Human Action Segmentation Tasks

Laura Romeo*, Cosimo Patruno, Grazia Cicirelli and Tiziana D'Orazio
Institute of Intelligent Industrial Systems and Technologies for Advanced Manufacturing (STIIMA), National Research Council of Italy (CNR), Bari, Italy

Keywords: Action Segmentation, Action Recognition, Skeleton Data, Manufacturing, Human Monitoring.

Abstract: Human Action Recognition and Segmentation have been attracting considerable attention from the scientific community in the last decades. In literature, various types of data are used for human monitoring, each with its advantages and challenges, such as RGB, IR, RGBD, and Skeleton data. Skeleton data abstracts away detailed appearance information, focusing instead on the spatial configuration of body joints and their temporal dynamics. Moreover, Skeleton representation can be robust to changes in appearance and viewpoint, making it useful for action segmentation. In this paper, we focus on the use of Skeleton data for human action segmentation in a manufacturing context by using a multi-camera system composed of two Azure Kinect cameras. This work aims to investigate action segmentation performance by using projected skeletons or "synthetic" ones. When one of the cameras fails to provide skeleton data due to occlusion or being out of range, the information coming from the other view is used to fill the missing skeletons. Furthermore, synthetic skeletons are generated from the combination of the two skeletons by considering also the reliability of each joint. Experiments on the HARMA dataset demonstrate the effects of the skeleton combinations on human action segmentation.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Human Action Recognition (HAR) and Human Action Segmentation (HAS) are gaining more interest in the literature, as they are crucial topics in several real-world applications, such as visual surveillance, human-robot interaction, healthcare, and entertainment (Cicirelli et al., 2015; Ma et al., 2022; Sun et al., 2023; Benmessabih et al., 2024).

Both HAR and HAS can be performed by using several typologies of data. Most of the works in the literature focus on using RGB videos due to the large and easy availability of this type of data (Jegham et al., 2020). Infrared imaging systems have recently been considered as they have a lower sensitivity to lighting conditions and appearance variability (Manssor et al., 2021). The availability of low-cost RGB-D sensors has also made possible approaches based on multimodal data, which take advantage of the synergistic use of different typologies of data (Shaikh and Chai, 2021). In addition, according to the complexity of the performed actions it could be necessary to use multiple systems to observe people from different points of view, to solve or reduce occlusion problems, and to highlight the variability of human movements.

Other important aspects have affected the development of HAR and HAS approaches. First, the high dimension of data to be processed could prevent the real-time application of these techniques. Methods based on RGB, depth, and IR data must store and process vast quantities of information. In addition, privacy constraints could impose strict rules on videos that observe people performing work or daily actions. For these reasons, researchers have started to explore the use of synthetic information extracted from videos. Instead of directly using raw image or video data, representations such as skeletal data can encode relevant information about human actions while abstracting away details that could compromise privacy. Additionally, these representations require less storage and computational resources compared to raw image or video data.

Many approaches are available in the literature for skeleton extraction from RGB video (Beddiar et al., 2020). In this case, the skeletons contain 2D information on several joints that represent the positions, among others, of hands, shoulders, elbows, knees, and feet and characterize human movements. Over the past few years, the rapid progress of low-cost RGB-D

* Corresponding author

Romeo, L., Patruno, C., Cicirelli, G. and D'Orazio, T.
Multi-View Skeleton Analysis for Human Action Segmentation Tasks.
DOI: 10.5220/0013129500003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 579-586
ISBN: 978-989-758-730-6; ISSN: 2184-4313
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

cameras has made it possible to have easy access to 3D data at a higher frame rate resolution and to conduct research on 3D human action recognition with the advantages of illumination invariance and high usability (Khaire and Kumar, 2022; Filtjens et al., 2022). RGB-D cameras such as the Microsoft Kinect also provide a set of SDK routines that, during real-time acquisitions, extract the 3D skeletal information allowing the storage of only these kinds of features, preventing privacy concerns.

One of the most crucial issues about the Skeleton data is the tendency to present inaccuracies in skeleton pose estimation that could alter the performances of HAR and HAS approaches. The application of multiple Kinect sensors in a workspace can help mitigate inaccuracies in skeleton pose estimation by combining the measurements from the different sensors. However, it introduces new challenges concerning the integration of the information extracted from the different sensors that could have variable reliability.

In (Moon et al., 2018) the authors developed a human skeleton tracking system based on Kalman filtering to face the problem of poor skeleton pose estimation due to self-occlusion. They propose a method to determine the reliability of each tracked 3D position of joints and combine multiple observations, acquired from multiple Kinect sensors, according to measurement confidence.

This paper studies the effects of using multi-view data on a temporal action segmentation approach. A visual system consisting of two Azure Kinect cameras observes people performing an assembly task. The aim is to segment the acquired untrimmed videos into seven predefined actions. The cameras are placed in a Frontal and Lateral position to the operator's workplace. The proposed method extracts features by considering the skeleton provided by the Azure Kinect SDK (v1.1.2). Due to self-occlusions or out-of-range problems, the SDK could give skeleton joints with low confidence values in one or the other view. So, the proposed method, after an initial camera calibration, projects the skeleton from one view to the other to have both in the same reference system. Then, the method computes a "synthetic" skeleton by combining, for each frame, the projected skeleton and the extracted one by considering the reliability of each corresponding joint from the two views. The experiments conducted over the HARMA dataset (Romeo et al., 2024), prove that the proposed approach reaches high performance in action segmentation compared to using the features of skeletons directly extracted from the Kinect SDK.

The remainder of the paper is structured as follows. Section 2 describes the camera setup and the acquired data. Section 3 presents the proposed method for synthetic skeleton computation. Then, experimental results are provided in Section 4. Section 5 concludes the paper.

2 DATA ACQUISITION

This section describes the setup of the camera and the data used in this work, which are included in the HARMA dataset (Romeo et al., 2024). The data are a collection of videos relative to actions performed by different subjects in collaboration with a cobot for assembling an Epicyclic Gear Train (EGT). The videos have been recorded in a laboratory scenario by using two Microsoft® Azure Kinect cameras positioned in frontal and lateral positions. The Microsoft® Azure Kinect camera provides several types of data, including RGB frames, Depth maps, IR frames and Skeleton data. In this work, we are primarily interested in skeletal data as they include the 3D coordinates of joints over time, thus providing an adequate representation of human body motion.

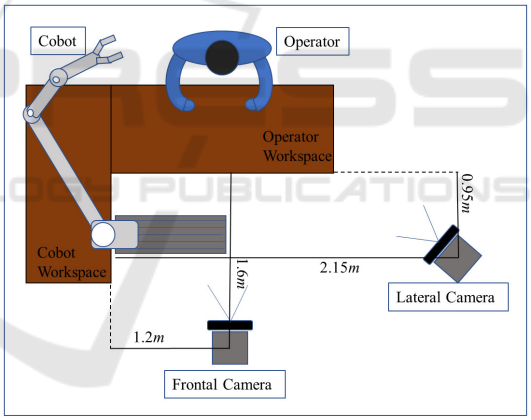

Figure 1: Acquisition setup. Two Azure Kinect cameras are placed in a Frontal and Lateral position to the operator's workplace.

2.1 Camera Setup

The acquisition setup is pictured in Fig. 1. The two Microsoft® Azure Kinect cameras have been placed on a tripod in Frontal and Lateral positions to the Operator Workplace. The Frontal Camera is at a height of $1.72\,m$ above the floor and down tilted by an angle of $6\,degrees$, while the Lateral Camera is at a height of $2.07\,m$ and $19\,degrees$ down tilted. Two typical RGB frames captured by both cameras are shown in Fig. 2. As shown in Fig. 2, the EGT components are spread over the Operator Workplace, so the operator can pick

(a) (b)

Figure 2: Sample frames captured by the (a) Frontal and (b) Lateral camera, respectively, during the assembly task.

(a) (b)

Figure 3: Frame with the calibration pattern acquired by the (a) frontal and (b) lateral camera, respectively.

up one component at a time to perform the assembly task in seven pick-and-place actions (Romeo et al., 2024).

2.2 Data Description

Skeleton joint data are returned by the Azure Kinect Body Tracking SDK (version 1.1.2) (Microsoft, 2021) exploiting the depth map (Brambilla et al., 2023). Data include, among others, joint 3D coordinates and joint confidence levels. The joint coordinates (X,Y,Z) are estimates relative to the reference frame of the depth sensor of each Azure Kinect and are in millimetres. The current SDK provides three confidence levels C for each joint: None, when the joint is out of range or is not detected by the camera; Low, when the joint is not observed, likely due to occlusion, but is predicted; and Medium when the joint is observed.

3 METHOD

This section describes the developed method for generating a robust skeleton representation. It principally exploits the two views of the Frontal and Lateral cameras and the calibration data.

3.1 Camera Calibration

To calibrate the two-camera system we apply the methodology proposed in (Romeo et al., 2022). First, a chessboard pattern was placed in the workspace so that it was visible to both cameras as shown in Fig. 3. The two Azure Kinect cameras provide both RGB and IR images of the pattern, so it is possible to detect and process the chessboard corners. Moreover, taking advantage of the ToF principles, these points can be directly computed from the depth map.

Considering the Frontal and Lateral camera, the method estimates the transformation matrix T that relates the two coordinate systems of the cameras:

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \quad (1)$$

where $R \in \mathbb{R}^{3\times3}$ is the rotation matrix and t is the translation vector. Different transformation matrices can be computed by using the RGB or the IR sensors of the Azute Kinect cameras, both in 2D and 3D. The matrix $T_{3Dinfrared}$, which transforms points detected in the IR image and projected in the 3D space, is the one that produces the best performance as proved in (Romeo et al., 2022). So, in this work, matrix $T_{3Dinfrared}$ will be used for projecting the skeleton joints from one camera reference frame to the other.

3.2 Synthetic Skeleton Generation

The Azure Kinect Body Tracking SDK (version 1.1.2) returns 32 skeleton joints (Microsoft, 2021) as shown in Fig. 4. Let $J_i^F(X_i,Y_i,Z_i)$ and $J_i^L(X_i,Y_i,Z_i)$, with $i = 1,..,32$, be the joints of the detected skeletons S^{Front} and S^{Lat} extracted by the Frontal and Lateral camera, respectively. Let $T_{3Dinfrared}^{Front}$ be the transformation matrix that projects skeleton joints from the Frontal camera reference frame to the Lateral camera reference frame (see Fig. 5). Analogously, let $T_{3Dinfrared}^{Lat}$ be the transformation matrix that projects skeleton joints from the Lateral camera reference frame to the Frontal camera reference frame. So, it is possible to generate two projected skeletons named S_{Proj}^{Front} and S_{Proj}^{Lat} having the following projected joints:

$$J_{Proj_i}^F(X_i,Y_i,Z_i) = T_{3Dinfrared}^{Lat} * J_i^L(X_i,Y_i,Z_i) \quad (2)$$

and

$$J_{Proj_i}^L(X_i,Y_i,Z_i) = T_{3Dinfrared}^{Front} * J_i^F(X_i,Y_i,Z_i) \quad (3)$$

respectively.

As introduced in section 2.2, the Azure Kinect Body Tracking SDK also provides the confidence levels for each joint. Let C_i^F and C_i^L be the confidence levels of joints J_i^F and J_i^L, respectively. Considering these confidence levels, we can combine skeleton

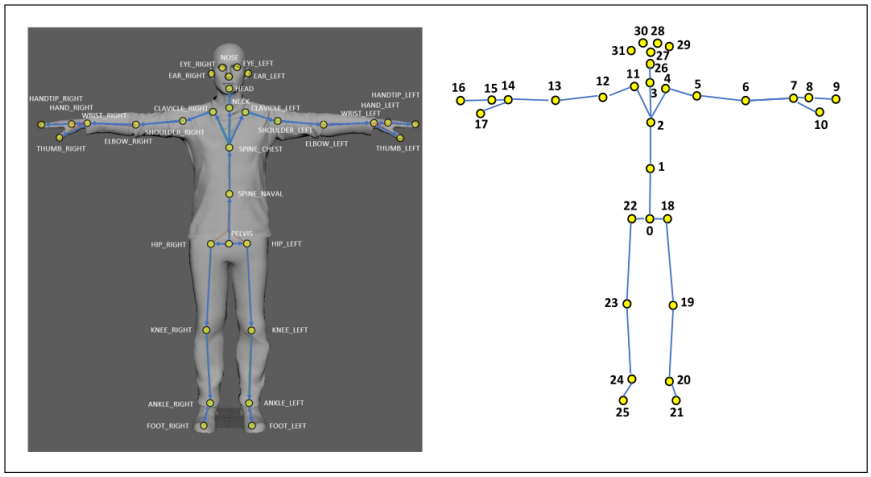

Figure 4: Skeleton joints extracted by the Azure Kinect Body Tracking SDK (version 1.1.2).

S^{Front} with S^{Front}_{Proj} and S^{Lat} with S^{Lat}_{Proj} estimating two new skeletons that we will name "synthetic" and indicate with S^{Front}_{Synt} and S^{Lat}_{Synt}, respectively. The joint coordinates of the synthetic skeletons are computed as the weighted mean of the joint coordinates of the projected and detected skeletons as follows:

$$J^F_{Synt_i}(X_i,Y_i,Z_i) = \frac{C^F_i J^F_i(X_i,Y_i,Z_i) + C^L_i J^F_{Proj_i}(X_i,Y_i,Z_i)}{C^F_i + C^L_i} \quad (4)$$

$$J^L_{Synt_i}(X_i,Y_i,Z_i) = \frac{C^L_i J^L_i(X_i,Y_i,Z_i) + C^F_i J^L_{Proj_i}(X_i,Y_i,Z_i)}{C^F_i + C^L_i} \quad (5)$$

Notice that the projected skeleton joints $J^F_{Proj_i}$ and $J^L_{Proj_i}$, defined in (2) and (3), inherit respectively the confidence level C^L_i and C^F_i of the detected skeleton joints used in the projection.

The confidence level returned by the Azure Kinect SDK (Microsoft, 2021) can have values equal to 0 if the joint is out of the depth range or field of view. Therefore, in the case where both confidence levels are 0 in (4) and (5), the joint coordinates of the "synthetic" skeletons are computed as the arithmetic mean of the joint coordinates of the projected and detected skeletons and inherit a 0 value confidence level.

4 EXPERIMENTS

This section presents the experiments carried out to segment the untrimmed videos of HARMA dataset. First, we describe the features that have been used to train and test two different deep architectures: the AS-Former (Yi et al., 2021) and MS-TCN++ (Li et al., 2023). Then a quantitative and qualitative analysis of results is provided.

4.1 Feature Definition

As the assembly task involves principally movements of the top section of the body, the (X_i,Y_i,Z_i) joint coordinates of 23 joints composing the top body section (see Fig. 6) have been selected from the whole skeletons extracted by the Kinect SDK. Then, we analyzed the confidence levels of this subset of joints in both cases of the Frontal and Lateral cameras. Considering all the videos, the average confidence level of the considered skeletal joints is 0.65 in the case of the Frontal camera and 0.49 in the case of the Lateral camera. Furthermore, by analyzing all the frames of the videos it has emerged that in some cases the Kinect SDK fails to provide skeleton data due to occlusion or being out of range. So, in some video frames, the skeleton is missing. These observations (i.e. different joint confidence levels depending on the viewpoint and skeleton missing) have driven our experiments considering the following combinations of features:

S^{Front*} (risp. S^{Lat*}):
: use of top-body skeleton joints extracted by the Kinect SDK of the Frontal (risp. Lateral) camera. In case of missing skeletons, those of the neighboring frame are copied.

S^{Front*}_{Proj} (risp. S^{Lat*}_{Proj}):
: use of top-body skeleton joints extracted by the Kinect SDK of the Frontal (risp. Lateral) camera. In case of missing skeletons those projected from the Lateral (risp. Frontal) camera reference frame to the Frontal (risp. Lateral) camera reference frame are considered.

S^{Front}_{Synt} (risp. S^{Lat}_{Synt}):
: use of the top-body joints of the "synthetic"

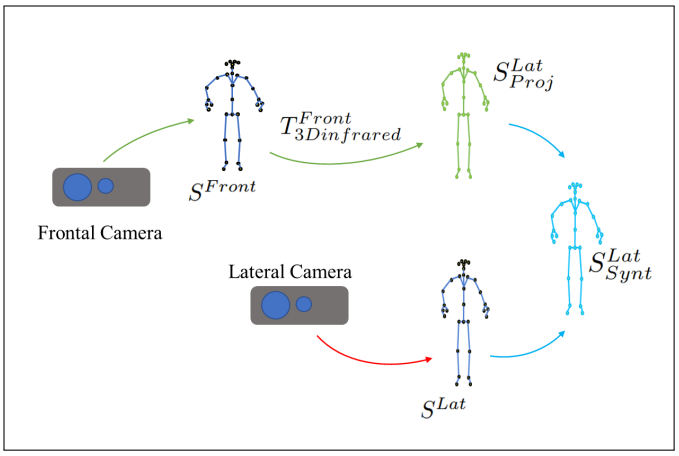

Figure 5: Skeleton projection from the Frontal camera to the Lateral camera reference frame to obtain the projected skeleton S^{Lat}_{Proj} and then the "synthetic" one S^{Lat}_{Synt}. Analogously, the same procedure can be applied to the Lateral camera obtaining the "synthetic" skeleton S^{Front}_{Synt}.

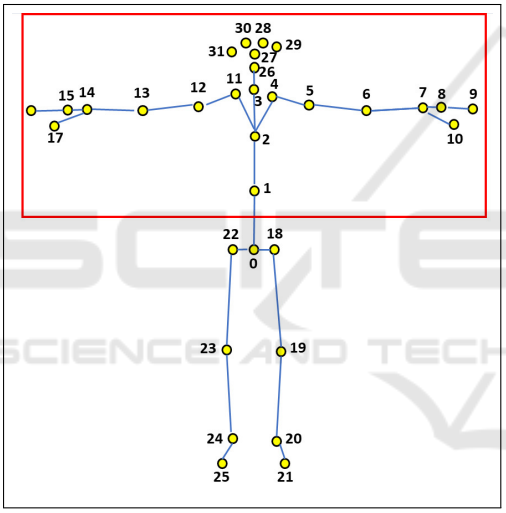

Figure 6: The joints of the top body skeleton (red box) used in the experiments.

skeleton of the Frontal (risp. Lateral) camera as defined in (4) (risp. in (5)). Notice that, in this case for each frame the skeletal joints are computed as the weighted mean of detected and projected skeletons.

Fig. 7 and 8 show a graphical representation of the top body skeleton joints considering one frame of the Frontal view and one frame of the Lateral view. In Fig.7 and 8, the top body skeletons are displayed as these are used to extract the features for the action segmentation phase.

In particular, Fig.7 displays the skeleton as detected by the Kinect SDK in the Frontal View (green line); the one obtained by projecting the skeleton acquired by the Lateral camera (red line) and the "synthetic" skeleton (blue line) obtained as the weighted mean of the previous ones. As can be seen, some joints of the projected skeleton (red line), such as for instance those of the hands, have low confidence values, so the obtained "synthetic" skeleton aligns with the S^{Front} one which has higher confidence values. Analogously, in Fig. 8 the "synthetic" skeleton aligns with the one projected from the Frontal view which has the higher confidence values. This is mainly evident by considering the head joints. Notice that, in this case, as the joints of the Lateral view have, generally, lower confidence values, both S^{Lat}_{Proj} and S^{Lat}_{Synt} are the better representations than those of S^{Lat} as they take advantage of the skeleton joints of the Frontal view.

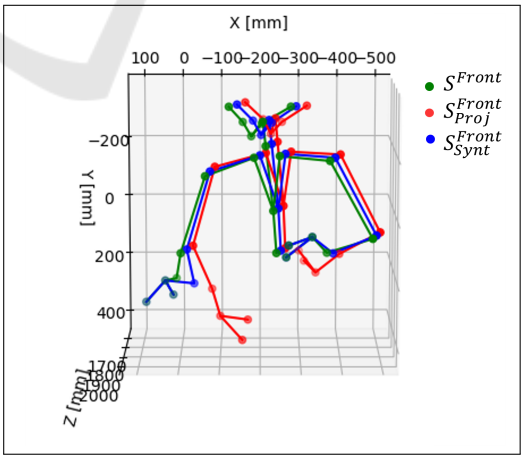

Figure 7: Graphical representation of top body skeleton projection to compute the "synthetic" skeleton joints in one frame acquired by the Frontal view.

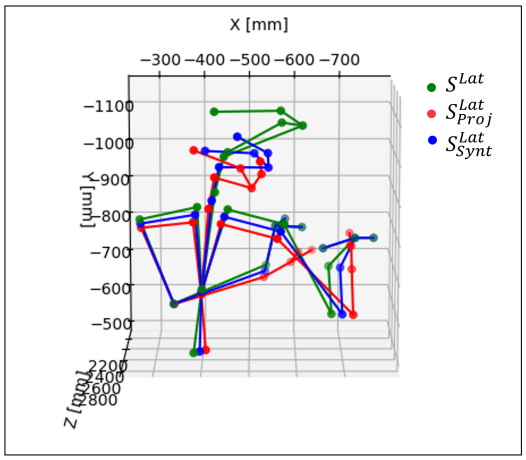

Figure 8: Graphical representation of top body skeleton projection to compute the "synthetic" skeleton joints in one frame acquired by the Lateral view.

4.2 Performance Analysis

To test the influence of the previously defined combinations of features on action segmentation tasks, two deep learning architectures, the ASFormer (Yi et al., 2021) and MS-TCN++ (Li et al., 2023), have been trained and tested.

First, the HARMA dataset has been split into non-overlapping training and testing sets by considering the 70% of videos for training and the remaining 30% for testing ensuring that videos of the same operator do not appear in both training and testing sets. The ASFormer (risp. the MSTCN++) models were trained over 120 (risp. 100) epochs, collecting losses for each iteration. The best model is chosen as the one with the lower loss within the total number of iterations and is used in the test phase.

Tab. 1 lists the performance rates of temporal action segmentation in terms of Accuracy, Edit Distance, and F1-score (Grandini et al., 2020). The Accuracy is a frame-wise metric that measures the proportion of correctly classified frames in the entire video sequence without capturing the temporal dependencies between action segments. The Edit Score, instead, measures how well the model predicts the ordering of action segmentation without requiring exact frame-level alignment.

Finally, F1-score with a threshold τ, often denoted as F1@τ, accounts for the degree of overlap between the Intersection over Union (IoU) of each predicted segment and ground truth segments (Ding et al., 2023). Segments with IoU greater than or equal to τ threshold are considered correctly predicted, while segments with IoU below τ are considered false positives. In this work, the τ threshold was set to 60%, 70%, and 80%.

Focusing on the results listed in Tab. 1, it can be noticed that all the considered models succeeded in correctly segmenting the actions for the assembly task. In general, the features extracted by the skeletons provided by the Frontal camera outperform those of the Lateral camera. This is principally due to the orientation of the Lateral camera to the operator, which affects the extraction of the skeleton. As previously stated, this is further proven by the average confidence levels of the skeletal joints, which are better in the case of the Frontal camera (0.65) compared to the Lateral camera (0.49).

Furthermore, regardless of applying one or the other network architecture, a very interesting result emerges from Tab. 1. As can be noticed, the percentage rates vary depending on the combination of features considered. When the original skeleton joints (S^{Front*}) directly extracted by the Frontal Kinect SDK are used, the Accuracy rates are 91.74% for the ASFormer and 93.17% for the MS-TCN++. With the application of the proposed approach that resolves the problem of missing skeleton joints (S^{Front*}_{Proj}), the Accuracy rate improves in both cases: 94.51% for the ASFormer and 94.45% for the MS-TCN++. The Accuracy rates of the Lateral camera, either in the case of S^{Lat*} and in the case of S^{Lat*}_{Proj}, follow in principle the same trend as those of the Frontal camera. Therefore, the Accuracy rates are as follows: 87.38% and 91.18% for the ASFomer; 90.64% and 91.59% for the MS-TCN++, respectively.

The use of the features of the "synthetic" skeletons S^{Front}_{Synt} and S^{Lat}_{Synt} requires a deeper analysis. When using the features of S^{Front}_{Synt}, the Accuracy rates are worse than those of S^{Front*} and S^{Front*}_{Proj}. This can be fully explained by considering how the "synthetic" skeletons were generated as described in section 3.2. The skeleton joints of S^{Front}_{Synt} are computed as the weighted mean of the joint coordinates of the skeleton projected from the Lateral view (S^{Lat}_{Proj}) and those of the detected skeleton (S^{Front}). So, S^{Front}_{Synt} is negatively affected by the influence of the projected Lateral skeleton, which has lower confidence levels. Indeed, the Accuracy is 89.05% for the ASFormer and 92.75% for the MS-TCN++. Edit and F1-score metrics follow the same trend, too.

Conversely, the Accuracy rates of S^{Lat}_{Synt} improve as, in this case, the "synthetic" skeleton S^{Lat}_{Synt} is positively affected by the influence of the projected Frontal skeleton, which has higher confidence levels. Indeed, the accuracy reaches 93.26% for the ASFormer and 92.22% for the MS-TCN++. Analogously, Edit and F1-score values reach maximum values in this case.

Table 1: Performance rates on action segmentation obtained by applying ASFormer and MS-TCN++ architectures and using different combinations of features.

	Accuracy	*Edit Score*	*F1 @ {60, 70, 80}*		
ASFormer					
S^{Front*}	91.74%	86.43%	80.10%	73.72%	61.13%
S^{Front*}_{Proj}	94.51%	95.08%	91.03%	87.97%	78.24%
S^{Front}_{Synt}	89.05%	84.01%	73.78%	65.70%	52.06%
S^{Lat*}	87.38%	78.75%	66.14%	58.96%	45.31%
S^{Lat*}_{Proj}	91.18%	90.34%	79.62%	72.89%	59.45%
S^{Lat}_{Synt}	93.26%	91.41%	87.69%	81.69%	73.05%
MS-TCN++					
S^{Front*}	93.17%	94.69%	89.58%	84.04%	77.55%
S^{Front*}_{Proj}	94.45%	93.89%	90.24%	87.80%	81.80%
S^{Front}_{Synt}	92.75%	92.12%	88.20%	85.39%	76.59%
S^{Lat*}	90.64%	94.02%	85.93%	79.69%	68.74%
S^{Lat*}_{Proj}	91.59%	92.89%	87.54%	83.20%	75.47%
S^{Lat}_{Synt}	92.22%	93.18%	87.14%	83.17%	74.85%

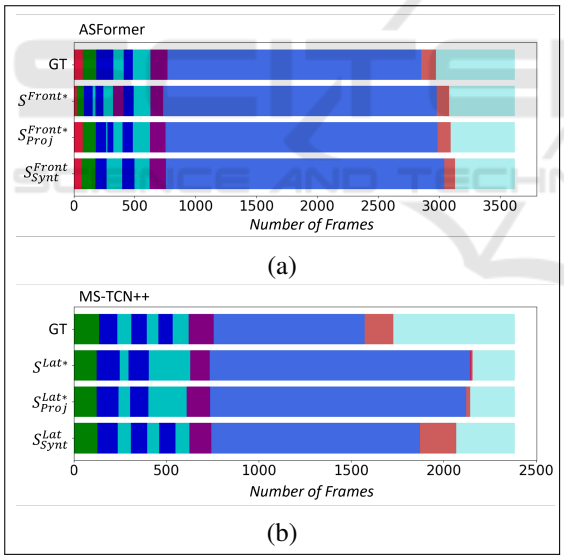

Figure 9: Qualitative representation of action segmentation for 2 videos within the HARMA dataset. (a) result obtained when applying the ASFormer model to skeleton features obtained from the Frontal view; (b) result obtained when applying the MS-TCN++ model to skeleton features obtained from the Lateral view. GT stands for Ground Truth.

To further support the obtained segmentation results, Fig. 9 shows a qualitative representation of action segmentation obtained by applying ASFormer and MS-TCN++, respectively, on two videos within the HARMA dataset. These videos have been chosen to display challenging situations when using skeleton features obtained from the Frontal (Fig. 9(a)) and Lateral (Fig. 9(b)) camera, respectively. As can be seen, both models, ASFormer and MS-TCN++, perform better when using the "synthetic" skeleton features (S^{Front}_{Synt} and S^{Lat}_{Synt}). This result is mainly evident when considering Action2 (dark blue bars) and Action3 (light blue bars), that are optimally segmented in the case of "synthetic" skeleton features rather than the other skeleton features (S^{Front*} and S^{Front*}_{Proj}, S^{Lat*} and S^{Lat*}_{Proj}).

In detail, the action segmentation Accuracy, Edit, and F1@{60, 70, 80} reach 89.28%, 100%, 66.67%, 66.67% and 44.44% for the ASFormer and 77.08%, 100%, 81.81% 81.81% and 54.54% for the MS-TCN++, respectively, compared to the Ground Truth (GT) of the videos considered in Fig. 9.

5 CONCLUSIONS

Human action recognition and segmentation are active topics of research in many fields of application such as healthcare, agriculture, surveillance, and manufacturing where the monitoring of human actions is fundamental. In this paper, we focus on the use of skeleton data for human action segmentation in the manufacturing context by using a multi-camera system composed of two Azure Kinect cam-

eras. Skeleton data represents the human pose and movement, focusing on body joints' spatial configuration and temporal dynamics. In particular, this work aims to investigate action segmentation performance by estimating a projected skeleton and a "synthetic" skeleton, which can be either a combination of the skeleton information provided by both cameras or an estimate when one of the cameras fails to provide skeleton data due to occlusion or being out of range. When using a multi-camera system, one view can provide more reliable data than others due to different factors, such as the camera's particular orientation or the software's ability to extract particular features, such as skeleton information. As proved by the experiments, the proposed approach addresses these issues by estimating new skeletons, taking advantage of the most reliable view.

ACKNOWLEDGMENT

The authors are deeply thankful to Michele Attolico and Giuseppe Bono for their technical and administrative support. Research funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Commission under the NextGeneration EU programme.

REFERENCES

Beddiar, D. R., Nini, B., Sabokrou, M., and Hadid, A. (2020). Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79:30509–30555.

Benmessabih, T., Slama, R., Havard, V., and Baudry, D. (2024). Online human motion analysis in industrial context: A review. *Engineering Applications of Artificial Intelligence*, 131:107850.

Brambilla, C., Marani, R., Romeo, L., Nicora, M. L., Storm, F. A., Reni, G., Malosio, M., D'Orazio, T., and Scano, A. (2023). Azure kinect performance evaluation for human motion and upper limb biomechanical analysis. *Heliyon*, 9(11).

Cicirelli, G., Attolico, C., Guaragnella, C., and D'Orazio, T. (2015). A Kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3).

Ding, G., Sener, F., and Yao, A. (2023). Temporal Action Segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Filtjens, B., Vanrumste, B., and Slaets, P. (2022). Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, pages 1—11.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. *arXiv preprint arXiv:2008.05756*.

Jegham, I., Ben Khalifa, A., Alouani, I., and Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901.

Khaire, P. and Kumar, P. (2022). Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey. *Journal of Visual Communication and Image Representation*, 86:103531.

Li, S.-J., AbuFarha, Y., Liu, Y., Cheng, M.-M., and Gall, J. (2023). MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6647–6658.

Ma, N., Wu, Z., Cheung, Y.-m., Guo, Y., Gao, Y., Li, J., and Jiang, B. (2022). A Survey of Human Action Recognition and Posture Prediction. *Tsinghua Science and Technology*, 27(6):973–1001.

Manssor, S. A., Ren, Z., Huang, R., and Sun, S. (2021). Human Activity Recognition in Thermal Infrared Imaging Based on Deep Recurrent Neural Networks. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–7.

Microsoft (2021). Azure Kinect DK documentation. https://docs.microsoft.com/en-us/azure/kinect-dk/.

Moon, S., Park, Y., Ko, D., and Suh, I. (2018). Multiple Kinect Sensor Fusion for Human Skeleton Tracking Using Kalman Filtering. *International Journal of Advanced Robotic Systems*, 13(2).

Romeo, L., Marani, R., Perri, A., and D'Orazio, T. (2022). Microsoft Azure Kinect Calibration for Three-Dimensional Dense Point Clouds and Reliable Skeletons. *Sensors*, 22(13):4986.

Romeo, L., Maselli, M., Domínguez, M. G., Marani, R., Nicora, M. L., Cicirelli, G., Malosio, M., and D'Orazio, T. (2024). A dataset on human-cobot collaboration for action recognition in manufacturing assembly. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 866–871. IEEE.

Shaikh, M. B. and Chai, D. (2021). RGB-D Data-Based Action Recognition: A Review. *Sensors (Basel)*, 21(12):4246.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2023). Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225.

Yi, F., Wen, H., and Jiang, T. (2021). ASFormer: Transformer for Action Segmentation. In *The British Machine Vision Conference (BMVC)*.