Artificial Intelligence and Social Computing, Vol. 28, 2022, 143–154
https://doi.org/10.54941/ahfe1001457

Automatic Labeling of Human Actions by Skeleton Clustering and Fuzzy Similarity

Chao-Lung Yang, Shang-Che Hsu, Si-Hao Wang, and Jing-Feng Nian

Department of Industrial Management National Taiwan University of Science and Technology No 43, Keelung Rd, Sec 4, Daan District, Taipei 106, Taiwan

ABSTRACT

Nowadays, human action recognition (HAR) has been applied in multiple fields with the rapid growth of artificial intelligence and machine learning. Applying HAR onto industrial production lines can help on visualizing and analyzing the correlation between human operators and machine utilization to improve overall productivity. However, to train HAR model, the manual labeling of certain actions in a large amount of the collected video data is required and very costly. How to label a large amount of video automatically is an emerging practical problem in HAR research domain. This research proposed an automatic labeling framework by integrating Dynamic Time Warping (DTW), human skeleton clustering, and Fuzzy similarity to assign the labels based on the pre-defined human actions. First, the skeleton estimation method such as OpenPose was used to jointly detect key points of the human operator's skeleton. Then, the skeleton data was converted to spatial-temporal data for calculating the DTW distance between skeletons. The groups of human skeletons can be clustered based on DTW distance among skeletons. Within a group of skeletons, the undefined skeletons will be compared with the pre-defined skeletons, considered as the references, and the labels are assigned according to the similarity against the references. The experimental dataset was created by simulating the human actions of manual drilling operations. By comparing with the manual labeled data, the results show that all of accuracy, precision, recall, and F1 of the proposed labeling model can achieve up to 95% with 40% saving time.

Keywords: Automatic labeling, Skeleton spatial temporal data, Dynamic time warping, Fuzzy similarity

INTRODUCTION

In the wave of Industry 4.0 automation, because of the mature development of artificial intelligence (AI) technology, how to introduce the related technology in the factory is the emerging research trend. Particularly, the application of applying computer vision with AI is the most common area. By utilizing cameras installed in factories, the image information of workers can be used to identify and judge the behavior of workers based on Human Action Recognition (HAR) model, one of the popular AI technology. The HAR technology can be used to investigate whether workers are completing tasks correctly, compare the differences in behavior among workers, calculate the working time, conduct work skill education training and so on.

HAR is considered as a image-based classification problem (Poppe, 2010). The human action categories need to be labeled when building the training dataset. Then, applying supervised learning, one of Machine Learning (ML) method that has gained tremendous progress in the last two decades to train the model for classifying or recognizing the human action. The accuracy of classification has been improved dramatically recently due to the fast development of deep learning network model (LeCun et al., 2015, Sarkar et al., 2022).

However, due to the difficulty and complexity of labeling the human actions in video collected in the factory, applying HAR during the manufacturing facing multiple challenges. First, labeling the human actions is very costly. In a series of multiple actions, labeling operation needs to not only clearly identify the actions in each image frame but also define the separation among the multiple actions. The human operator needs to review all of image frames in a video and label the actions in multiple image frames. In addition, the complexity of defining actions also lies in the fact that operators in the factory usually change their actions based on the order of production. Although workers need to follow standard operation procedure (SOP) which contains multiple actions for each production sequence, different workers might have very different motions to complete the same work based on their habitually practice or body size which can also lead to bias in human labeling. Imagine that different people labeling the same video for the same action may create different labels for the same action.

In order to create a large amount of labeled training data for training HAR deep learning model, this work aims to develop an automatic labeling system to label the human actions in production without the human bias. First, in each image frame, the human skeleton is created based on human action estimation method such as OpenPose (Cao et al., 2017). Then, applying Dynamic Time Warping (DTW) to calculate the pair-wise DTW distance of human skeletons among video. The calculated DTW distance is used to perform the clustering on grouping the different videos. For each video cluster, one or more video is randomly selected to be labeled by human operator as the reference. For each cluster, the temporal coordinate areas of unlabeled videos and the reference video are calculated. By applying Fuzzy similarity, the action composition ratio of the unlabeled video can be estimated and the actions in the video can be label by following the Fuzzy similarity. With this proposed technique, we can quickly label the action in the collected considerable videos by the relatively few reference videos.

LITERATURE REVIEW

Human Action Recognition

HAR is a technology that utilizes information of human body to understand human behavior in machine vision. According to the different devices that collect human information, HAR can be divided into the sensor-based HAR and the vision-based HAR (Dang et al., 2020). In the sensor-based HAR, data are captured by the embedded sensors that subjects wear. On the other hand, the vision-based HAR analyzes images or videos of subjects that optical equipment obtained. Due to difficulty of applying wearable devices on human operators, in this work, the vision-based HAR is our focus.

No matter for the sensor-based or vision-based HAR, deep learning model has been applied widely (LeCun et al., 2015, Sarkar et al., 2022). Vinyes Mora and Knottenbelt used Inception Neural Network to extract the features of each action from tennis action image data, and established a 3-layer Long Short-Term Memory (LSTM) network to classify actions in 2017 (Vinyes Mora and Knottenbelt, 2017). Convolution Neural Network (CNN) was also applied to detect the action of human falls to reduce health risks in the life of the elderly by Núñez Marcos et al. in 2017 (Núñez-Marcos et al., 2017). Jin-Miao Cai et al. proposed a method which combined skeleton information with Joint-aligned optical Flow Patches (JFP) technology to capture the detailed motion around joints as key visual information, and applied it to the open image dataset to confirm that this method improves the accuracy effectively (Cai et al., 2021). Chao-Lung Yang et al. developed an image classification model based on Spatial Temporal Graph Convolution (STGCN) and Support Vector Machine (SVM) to classify standard operating procedures and exceptional movement in factory production lines in 2021 (Yang et al., 2021).

Image Set Labeling Problem

Numerous studies have proposed publicly available image datasets which facilitate researchers to develop and train their HAR models. For example, Amir Shahroudy and others built a dataset NTU RGB+D in 2016, which includes 60 action categories, 80 camera angles and more than 56,000 videos (Shahroudy et al., 2016). Jun Liu et al. continued to build the NTU RGB+D 120 dataset in 2020, which extends the action categories from 60 to 120, camera angles from 80 to 155, and more than 110,000 videos (Liu et al., 2019). However, in the most of public datasets, a single video usually contains only a single or a small number of actions, which illustrates the insufficiency of large detection data which contains multiple actions (Liu et al., 2017). In addition, for multi-action videos, each frame of video needs to be labelled, which requires more manpower for video labeling than data sets with single or few actions. Therefore, the automatic labeling technique is urgently needed for HAR research.

Auto Labeling

For example, Parijat Dube et al. built a neural network in 2019 to solve auto-labeling problem by referring to the VGG16 architecture, and eventually automatically annotation ImageNet images with possible tags in five ways such as Nearest-N (Dube et al., 2019). In contrast, Trung-Nghia Le et al. in 2020, used recurrent self-supervised learning to continuously learn how to generate higher quality annotations (Le et al., 2020). Bowen Chen et al. similarly used an interactive framework for annotating object instances in 2020 in order to address the problem of overly complex image segmentation by splitting the steps of image segmentation into annotating objects with tracked boxes, and labeling masks inside these tracks (Chen et al., 2020). It can be found that more and more different methods have been proposed to solve the automatic labeling problem. However, most of these research methods are based on the features between images, but less on the similarity between videos. Therefore, this study tries to evaluate the similarity of two videos by DTW grouping and Fuzzy similarity. More details can be found in the following sections.

METHODOLOGY

Auto-Labeling Framework

The purpose of this research is to develop an automatic image labeling tool which can quickly label the SOP actions performed in the manufacturing/production site. The proposed framework is shown in Figure 1 and described as below:

1. The human joints data in every frame of a video is extracted by using pose estimation tool OpenPose (Cao et al., 2017). Then, the joints data is encoded into the x and y coordinate data.
2. Based on the x and y coordinate data, the pair-wise DTW distance d_{ab} between the video a and video b, each of which consists multiple frames, are calculated to form a DTW distance matrix from video v_1 to video v_q.
3. After calculating the DTW distance of all the videos, the videos can be clustered by DTW clustering method, so that the DTW distances within a cluster can be minimized and the clusters with similar cut-off points are obtained. Then, for each video cluster, a few videos are randomly selected from each cluster for manual labeling. These manually labeled videos are represented as Video Action Reference (VAR), and the remaining unlabeled videos are represented as Video without human labeling (VWHL).
4. Fuzzy similarity between VWHL and VAR can be evaluated. For example, one VWHL can be compared with multiple VARs and determine which VAR is closer in terms of Fuzzy similarity.
5. Then, for a human action, the number of frames in VWHL can be derived based on the VAR's associated action frames. This proposed framework can be used to develop an automatic labeling tool to reduce the cost of manual labeling. In the following subsections, the more detailed information will be presented.

Data Preprocessing

To convert the collected video data into skeleton information regarding the human actions, this study detects the position of the joint points of the face, arms, legs and torso of the character in each frame of the original video by utilizing OpenPose tool (Cao et al., 2017). After the conversion is completed, the time series data of the skeleton information of the video will be obtained. During the conversion, if the hands are blocked by the human torso,

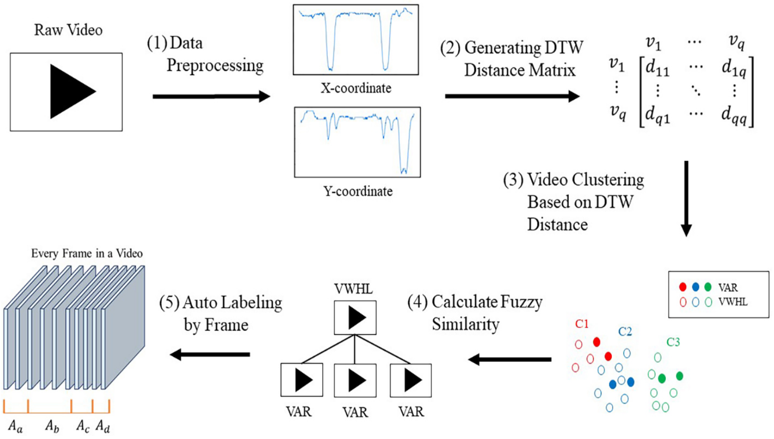
Raw Video
(1) Data Preprocessing
X-coordinate
Y-coordinate
(2) Generating DTW Distance Matrix
v_1 ⋯ v_q
v_1 d_{11} ⋯ d_{1q}
v_q d_{q1} ⋯ d_{qq}
(3) Video Clustering Based on DTW Distance
VAR
VWHL
C1
C2
C3
(4) Calculate Fuzzy Similarity
VWHL
VAR
VAR
VAR
(5) Auto Labeling by Frame
Every Frame in a Video
A_a A_b A_c A_d

Figure 1: The proposed auto-labeling framework consist of (1) Data Preprocessing (2) Generating DTW Distance Matrix (3) Video Clustering Based on DTW Distance (4) Calculate Fuzzy Similarity (5) Auto Labeling by Frame.

which will cause null values in specific parts of the skeleton information. In this study, the point value of the frame before the null value point will be supplemented in sequence.

This study utilized the standardized process to compensate the deviation of size of human body. For each collected video, the coordinates of human skeleton joins are converted from the original pixel coordinates to the coordinates between (0,0) and (1,1). The Eq. (1) and Eq. (2) shows the standardization formulas. After performing the image standardization and coordinate filtering, all video data are presented on the same datum in a standardized way, and have a series of the corresponding time-series coordinates for human skeleton.

$$X_{\mathrm{new}} = (X_I - X_{\mathrm{MIN}})/(X_{\mathrm{MAX}} - X_{\mathrm{MIN}}) \tag{1}$$

$$Y_{\mathrm{new}} = (Y_I - Y_{\mathrm{MIN}})/(Y_{\mathrm{MAX}} - Y_{\mathrm{MIN}}) \tag{2}$$

Generating DTW Distance Matrix

Dynamic Time Warping algorithm (DTW) was proposed by Hiroaki Sakoe and Seibi Chiba in 1978, which can compare the similarity of time series data of different lengths. First, the time-series data of different lengths are normalized by the time warp function, and then the distance corresponding to the two time-series data can be obtained by calculating the distance formula (Sakoe and Chiba, 1978).

Due to the deviation of human operation, the time duration of the performed SOP action will be very different by SOP operators. In this work, the DTW distance is applied to compare the human pose/joint coordinates time series based on varied length of videos. Suppose two skeleton time series matrices A and B are given, as shown in Eq. (3) and Eq. (4). For each matrix, $\left(X_i^j, Y_i^j\right)$ indicates x and y coordinates of the joint j in i-th frame of a video where $j \in (j_1, j_2, \ldots, j_h, \ldots, j_m)$, j_h is the h-th joint point, the sequence has totally m joint points collected, and a and b are the lengths of the two skeleton time series data (Senin, 2008).

$$A = \begin{bmatrix} \left[\left(X_1^{j1}, Y_1^{j1}\right), \left(X_1^{j2}, Y_1^{j2}\right), \cdots, \left(X_1^{\mathrm{jm}}, Y_1^{\mathrm{jm}}\right)\right] \\ \vdots \\ \left[\left(X_a^{j1}, Y_a^{j1}\right), \left(X_a^{j2}, Y_a^{j2}\right), \cdots, \left(X_a^{\mathrm{jm}}, Y_a^{\mathrm{jm}}\right)\right] \end{bmatrix}, \ a \in N \tag{3}$$

$$B = \begin{bmatrix} \left[\left(X_1^{j1}, Y_1^{j1}\right), \left(X_1^{j2}, Y_1^{j2}\right), \cdots, \left(X_1^{\mathrm{jm}}, Y_1^{\mathrm{jm}}\right)\right] \\ \vdots \\ \left[\left(X_b^{j1}, Y_b^{j1}\right), \left(X_b^{j2}, Y_b^{j2}\right), \cdots, \left(X_b^{\mathrm{jm}}, Y_b^{\mathrm{jm}}\right)\right] \end{bmatrix}, \ b \in N \tag{4}$$

In order to align the timing data of the two skeletons, a pair-wise distance matrix C of two skeleton time series matrices is first established, where A_i is the skeleton information of the i^{th} frame in sequence A, while B_u is the skeleton information of the u^{th} frame in the sequence B, and the element $c_{i,u}$ in matrix $C(C \in R^{a \times b})$ is the distance between the i^{th} frame in sequence A and the u^{th} frame in sequence B, as shown in Eq. (5).

$$c_{i,u} = \sum \sqrt{(A_i - B_u)^2}, \ i \in [1:a], u \in \left[1:b\right] \tag{5}$$

$$p = \left(p_1, p_2, \ldots, p_K\right), \ p_l = \left(p_i, p_u\right) \in [1:a] \times \left[1:b\right] \text{ for } l \in [1:K] \tag{6}$$

After calculating the distance matrix of the two skeleton time series data, the alignment path of the two sequences of point-to-point can be planned, and the point of the pat under DTW method is a sequence p, as shown in Eq. (6), where K is the length of converted DTW sequence p.

The generation of the sequence p must satisfy the following criteria (Senin, 2008):

1. Boundary condition: The start and end points of the warping path must be the first point $p_1 = (1,1)$ and the last point $p_K = \left(a, b\right)$ of the aligned sequence.
2. Monotonicity condition: This condition preserves the time-ordering of points, thus $n_1 \leq n_2 \leq \ldots \leq n_K$ and $g_1 \leq g_2 \leq \ldots \leq g_K$.
3. Step size condition: Step length was defined in this study as $p_{l+1} - p_l \in \{(1,1), (1,0), (0,1)\}$.

After planning the DTW path, the cost function $c_p\ (A, B)$ of A and B can be used to calculate the cost of the path, as shown in Eq. (7). The path with the minimum cost can be calculated according to Eq. (8) to obtain the optimal path with $C_{p^*}\ (A, B)$ distance between the two skeleton time series data. Based on this pair-wise DTW distance calculation shown in Eq. (8), all videos can be compared with each other to establish the DTW distance matrix, named as *DTW*.

$$c_p\left(A,B\right)=\sum_{l=1}^{L}c\left(A_{\mathrm{nl}},B_{\mathrm{gl}}\right) \tag{7}$$

$$DTW_{A,\ B}=C_{p^*}\left(A,B\right)=\min\left\{c_p\left(A,B\right),p\in P^{a\times b}\right\} \tag{8}$$

Video Clustering Based on DTW Distance

For clustering videos, K-means method was applied based on the generated *DTW* distance matrix (Niennattrakul and Ratanamahatana, 2007). Essentially, *k* videos are randomly selected in DTW distance matrix as the centroid, and the remaining videos are clustered by comparing their DTW distances with the *k* centroids. The videos will be clustered into the same cluster with the centroid with the smallest distance. The determination of *k* can be followed by the methods found in the literature (Kodinariya and Makwana, 2013). Different from the traditional K-means method, the iteration process used in this work uses the total distance of all clusters as the decision factor to determine if the new clusters should be generated. Basically, for each cluster, a new centroid video in the same cluster is randomly selected to re-cluster until the total distance is smaller than it of the pervious. The iteration is stopped when the total distance is no longer changed after 1000 times of calculation which is much larger than the number of videos in this work. When the iteration stops, it also means that the videos in each cluster are the closest to each other.

Calculate Fuzzy Similarity

Based on the DTW clustering result, three videos would be randomly selected from each cluster for manual labeling as VAR. Then, the areas of all joints of VWHL and VAR in each cluster are compared as shown in Figure 2. For example, in Figure 2, the trajectory of human elbow in x-axis of VAR and VWHL are plotted. Please note the area constituted by the VAR moving trajectory and time in space is called Trajectory area of labeling video (TAL) while the area constituted by the VWHL moving trajectory and time in space is called Trajectory area of un-labeling video (TAU).

The Fuzzy similarity is calculated as shown in Eq. (9), and the similarity of the VWHL to the VAR is calculated for a node j_h (j_h is the h^{th} joint). Basically, in Eq. (9), Fuzzy similarity can be calculated as one minus the ratio of the difference between TAL and TAU, against the TAL. The larger the Fuzzy similarity is, TAU is more similar to TAL.

$$S_{v,j_h}^{f,s}=1-\frac{\left|\mathrm{TAU}_{v,j_h}^{f}-\mathrm{TAL}_{j_h}^{f,s}\right|}{\mathrm{TAL}_{j_h}^{f,s}} \tag{9}$$

Then, the Fuzzy set can be obtained by calculating the similarity of all skeleton joins of the VWHL against the VAR as shown in Eq. (10). In fact,

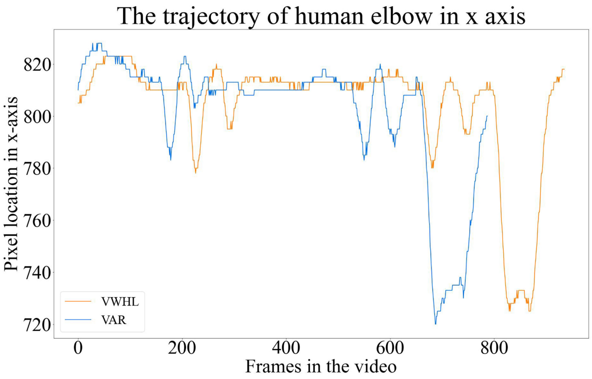

Figure 2: Comparison of trajectory of human elbow in VAR and VWHL.

Eq. (10) defines the membership degree $SF_v^{f,s}$ of v-th VWHL based on all of the Fuzzy similarities against s-th VAR of cluster f. The more details about Fuzzy membership degree can be found in (Zadeh, 1965).

$$SF_v^{f,s} = \frac{S_{v,j_1}^{f,s}}{j_1} + \frac{S_{v,j_2}^{f,s}}{j_2} + \cdots + \frac{S_{v,j_m}^{f,s}}{j_m} \tag{10}$$

Finally, we can calculate the weighted Fuzzy similarity $WSF_v^{f,s}$ of v-th VWHL based on all of the Fuzzy similarities against s-th VAR of cluster f, as shown in Eq. (11). In addition to Eq. (10), Eq. (11) utilizes $d_{j_1}^v$ as the weights of v-th VWHL on j_1 joint which is calculated by considering the total displacement of all skeleton joints as the contributor on the Fuzzy similar of a particular joint. Eq. (11) further normalizes the $WSF_v^{f,s}$ across all VARs. Please note that only three VARs were used in each group.

$$WSF_v^{f,s} = \frac{\sum_{h=1}^{m} d_{j_h}^v * S_{v,j_h}^{f,s}}{\sum_{h=1}^{m} d_{j_h}^v} \tag{11}$$

$$WSF_v^{f,s*} = \frac{WSF_v^{f,s}}{\sum_{s=1}^{3} WSF_v^{f,s}} \tag{12}$$

Auto Labeling by Frame

After comparing each VWHL against three VARs in each cluster, there will be three corresponding Fuzzy similarities, and finally these three Fuzzy similarities will be used to determine the labeling of action in VWHL. For action n of v-th VWHL in cluster f, Eq. (13) shows the number of image frames $Fr_{v,n}^f$ of action n is equal to the summation of the weighted $WSF_v^{f,s*} \times F_n{}^{f,s}$ for all of VARs, where $F_n{}^{f,s}$ is the number of frames for action n of s-th VAR.

$$Fr_{v,n}^f = \sum_{all\ s} (WSF_v^{f,s*} \times F_n{}^{f,s}) \tag{13}$$

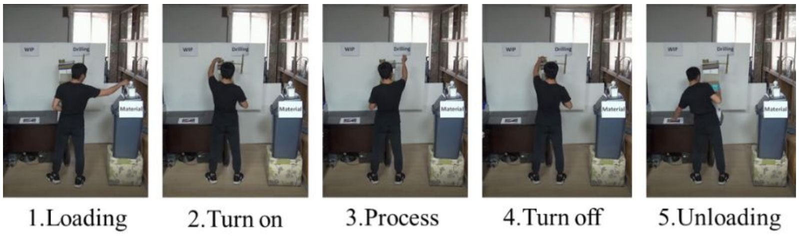

Figure 3: Five different actions performed in the simulated SOP operations: 1) loading, 2) turning on the drilling, 3) processing, 4) turn off the drilling, and 5) unloading.

PRELIMINARY RESULT

In order to verify the effectiveness of the proposed automatic labeling framework, this study simulated a complete operation of a vertical drilling machine in a standing position by an operator in a factory environment. Figure 3 shows five different actions performed in this simulated SOP operations. Basically, these five actions are SOP process when workers operate the drilling machine by following the sequences of 1) loading, 2) turning on the drilling, 3) processing, 4) turn off the drilling, and 5) unloading. Totally, 67 videos with a total length of 45,566 frames were used in the study. In order to evaluate the performance of the proposed auto-labeling framework, all of the collected videos were manually labeled by local experts.

In order to judge the final automatic labeling result, the manual labeling information of five experimenters will be used as the basis for labeling judgment. Due to the discrepancy among the labeling personnel, there might be deviation of action frames labeled by different labeling personnel. Therefore, the upper and lower boundaries of the manual labeling segments will be used as the tolerance interval of labeling, considering as human labeling error range. Figure 4 illustrates the error range of human labeling. Suppose two experts: expert #1 and #2 manually label the one video considering five different action from A_1 to A_5. Their labels might be different and cause the discrepancy shown as grey area in the bottom of Figure 4. It also means that once the action segmentation result of auto-labeling falls within the range of human error range, the labeling can be regarded as accurate. The final auto-labeling accuracy rate can be calculated as the ratio of the total number of correct frames of auto-labeling divided by the total number of frames of the video.

By conducting the proposed framework mentioned in the previous sections, 67 videos are clustered as two groups: 14 and 53 videos, respectively. For each group, 3 videos were randomly selected for manually labeling. It means that the labeling accuracy of 61 videos were compared with the manual labeling results conducted by 5 experts. If the auto-labeling cut-off point between action segments is located within the human error range, the labeling was considered as accurate. In this preliminary study, totally, 41,372 frames were evaluated with the human labeling results. Figure 5 shows the confusion matrix of the evaluation results. Based on this result, Accuracy, Precision, Recall, and F1 were 0.949, 0.948, 0.949, and 0.948. Obviously, the

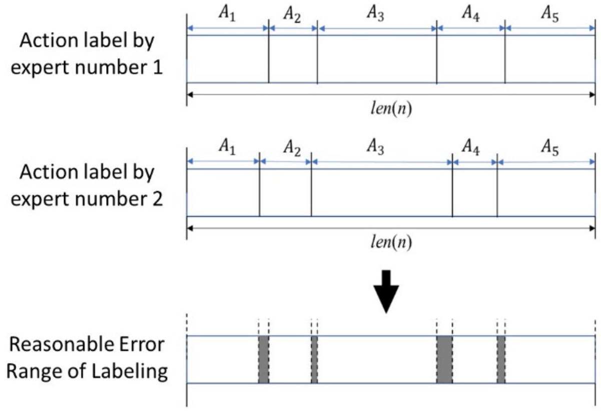
A_1
A_2
A_3
A_4
A_5
Action label by expert number 1
$len(n)$
Action label by expert number 2
$len(n)$
Reasonable Error Range of Labeling

Figure 4: Reasonable error range of labeling by human experts.

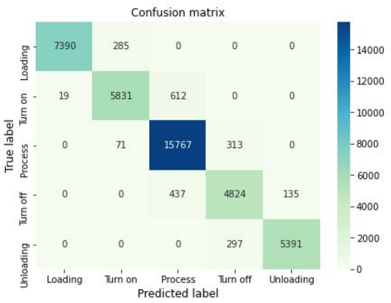

Figure 5: Confusion matrix of 41,372 frames when comparing the auto-labeling result against manual labeling by human experts.

miss-labeling can be found to be located near previous or next actions. Particularly, actions “Turn on”, “Process”, and “Turn off” have relatively more mis-labeling, although Accuracy, Precision, Recall, and F1 of those actions are ~95%. The possible reason might be the hand movements of “Turn on”, “Process”, “Turn off” are very similar while human body still maintains the similar poses when conducting “Turn on”, “Process”, and “Turn off”.

The total time of manual labeling 67 videos was around 300 minutes. If running the proposed auto-labeling framework under Intel i7 8700, 4.6GHz, 32 GB memory with graphic card GTX 1070Ti which will be used for OpenPose estimation, the skeleton generation will take around 172 minutes and auto-labeling process including DTW clustering and Fuzzy similarity will take only around 11 minutes. Obviously, the most of calculation time is on skeleton generation. The DTW clustering and Fuzzy similarity are relatively fast and efficient.

CONCLUSIONS AND DISCUSSION

With the wave of Industry 4.0 automation, the introduction of artificial intelligence technology, particularly HAR, create the potentials of auto-recognizing the human actions. major challenge for most companies. In addition to the need for professional manpower and equipment, the large number of training sets required to train AI models is a major burden. In order to solve the problem of insufficient training data sets, this study proposed a auto-labeling framework which utilizes DTW clustering technique and Fuzzy confidence to automatically label the videos with a small amount of manual labeling data. Essentially, DTW distance for video clustering were computed, and then the composition of unlabeled video labels is calculated by fuzzy similarity against the VAR. Based on the Fuzzy similarities against multiple VARs, VWHL can be labeled by considering the similarity distributions of different actions in VAR.

In this work, the human actions of drilling operations were simulated for validating the proposed framework. 67 videos of drilling SOP consisting of five actions were collected. The automatic labeling results of 61 VWHL were obtained with around 95% for Accuracy, Precision, Recall, and F1. Comparing all of manual labeling time, the proposed framework can save around 40% of labeling time. If the skeleton data can be retrieved fast, the time efficiency of the proposed framework can be further improved. In this study, we decided to select three videos for manual labeling based on the execution results to obtain good automatic labeling results. In the future work, it is interesting to investigate the optimal number of clusters and automatically decide the number of manually labeled videos for each group.

ACKNOWLEDGMENT

The authors appreciate the financial support from the Ministry of Science and Technology of Taiwan, R.O.C. (Contract No. 110-2221-E-011 -140) and the "Center for Cyber-Physical System Innovation" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. We also thank Wang Jhan Yang Charitable Trust Fund (Contract No. WJY 2020-HR-01) for their financial support.

REFERENCES

CAI, J., JIANG, N., HAN, X., JIA, K. & LU, J. JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, January 5-9 2021 Hawaii, USA. IEEE, 2735-2744.

CAO, Z., SIMON, T., WEI, S.-E. & SHEIKH, Y. Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE conference on computer vision and pattern recognition, July 21-26 2017 Hawaii, USA. Computer Vision Foundation, 7291–7299.

CHEN, B., LING, H., ZENG, X., GAO, J., XU, Z. & FIDLER, S. Scribblebox: Interactive annotation framework for video object segmentation. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, August 23-28 2020. Springer, 293–310.

DANG, L. M., MIN, K., WANG, H., PIRAN, M. J., LEE, C. H. & MOON, H. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition,* 108, 107561.

DUBE, P., BHATTACHARJEE, B., HUO, S., WATSON, P., BELGODERE, B. & KENDER, J. 2019. Automatic labeling of data for transfer learning. *nature,* 192255, 122-129.

KODINARIYA, T. M. & MAKWANA, P. R. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal,* 1, 90-95.

LE, T.-N., SUGIMOTO, A., ONO, S. & KAWASAKI, H. Toward interactive self-annotation for video object bounding box: Recurrent self-learning and hierarchical annotation based framework. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, March 1-5 2020 Colorado, USA. Computer Vision Foundation, 3231–3240.

LECUN, Y., BENGIO, Y. & HINTON, G. 2015. Deep learning. *nature,* 521, 436–444.

LIU, C., HU, Y., LI, Y., SONG, S. & LIU, J. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, October 23-27 2017 California, USA. Association for Computing Machinery, 1-8.

LIU, J., SHAHROUDY, A., PEREZ, M., WANG, G., DUAN, L.-Y. & KOT, A. C. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence,* 42, 2684-2701.

N*Wireless Communications and Mobile Computing,* 2017, 9474806.

NIENNATTRAKUL, V. & RATANAMAHATANA, C. A. On clustering multimedia time series data using k-means and dynamic time warping. 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07), April 26-28 2007. IEEE, 733–738.

POPPE, R. 2010. A survey on vision-based human action recognition. *Image and Vision Computing,* 28, 976-990.

SAKOE, H. & CHIBA, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, signal processing,* 26, 43-49.

SARKAR, A., BANERJEE, A., SINGH, P. K. & SARKAR, R. 2022. 3D Human Action Recognition: Through the eyes of researchers. *ELSEVIER,* 193, 116424.

SENIN, P. 2008. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa,* 855, 40.

SHAHROUDY, A., LIU, J., NG, T.-T. & WANG, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE conference on computer vision and pattern recognition, June 27-30 2016 Nevada, USA. IEEE, 1010–1019.

VINYES MORA, S. & KNOTTENBELT, W. J. Deep learning for domain-specific action recognition in tennis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, July 21-26 2017 Hawaii, USA. IEEE, 114–122.

YANG, C.-L., HSU, S.-C., HSU, Y.-W. & KANG, Y.-C. Human Action Recognition on Exceptional Movement of Worker Operation. International Conference on Applied Human Factors and Ergonomics, July 25-29 2021 New York, USA. Springer, 376–383.

ZADEH, L. 1965. Fuzzy sets. Information and Control, 8, 338–353.