Skeleton Based Action Analysis on Manufacturing Assembly Site

Ching Yuan Kung
National Taiwan University of Science and Technology, chingyuan1215@gmail.com

TING-YANG CHOU
National Taiwan University of Science and Technology, tychou100@gmail.com

Follow this and additional works at: https://aisel.aisnet.org/pacis2022

Recommended Citation
https://aisel.aisnet.org/pacis2022/87

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2022 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Skeleton Based Action Analysis on Manufacturing Assembly Site

Short Paper

Ching-Yuan Kung
Industrial Management, NTUST
chingyuan1215@gmail.com

Ting-Yang Chou
Industrial Management, NTUST
tychou100@gmail.com

Abstract

This study intends to establish a human action comparison analysis framework that can be practically applied to factory production lines. This research is based on a human skeleton motion recognition technology (OpenPose), which captures the hand motion skeletons of the operators on the production line when assembling products, and then uses the dynamic time warping algorithm (DTW) in the time series data analysis, and K-Means clustering to distinguish the assembly proficiency of the operator for the assembly operation, as a reference for planning education and training. It is expected that the tools provided can help the factory quickly and accurately grasp the movement differences of each operator and obtain effective improvement suggestions. It is also hoped that the research results will help reduce the cost of education and training, improve the speed, accuracy, and quality of products produced by operators, thereby greatly increasing the overall performance of the manufacturing plant.

Keywords: Spatial–temporal data, Dynamic Time Warping algorithm, Cluster analysis, Skeleton, Manufacturing

1. Introduction

Under the influence of the wave of Industry 4.0 and customer demand, manufacturing factories are moving towards automation and small-scale and diverse production. In the complex manufacturing process of customized products, manual assembly and inspection are still indispensable. In principle, new assembly workers need to undergo pre-employment training before they can enter the production line to operate. At present, teaching assembly skills to new operators by means of human guidance has its limitations.

Therefore, this study develops an action comparison model driven by a dynamic time warping algorithm based on the spatial-temporal data of the human skeleton. The model can be trained by importing images of standard assembly actions for the links that require manual assembly on the production line. Operators can also be used to analyze the most suitable movements of operators when assembling production lines, improve the efficiency of pre-employment training, and assist supervisors in making decisions on division of labor.

By calculating the difference value of the spatial-temporal data of the 44 joint points of the operator’s hands, combined with the statistical hypothesis test, the operator who should correct the action is identified, and then the priority improvement items for each operator are sorted out according to the spatial coordinate data. The results obtained can not only be used as effective suggestions, but also new operators are expected to rapidly improve to the level of senior operators in a short period of time, or even surpass the work
efficiency of old senior operators. In addition, grouping technology can also be used to group employees with similar problems into a group to give managers a basis for arranging training directions.

2. Related Work

2.1 Skeleton-based action recognition

In the common human motion recognition algorithm, there are two methods, one is "Top-down" detection, the time is proportional to the number of people detected, and the accuracy is high, but the calculation performance is not good. The other is "bottom-up" detection. The number of people has little impact on the computing time. It is a method with extremely high running speed and maintaining accuracy. This study focuses on the efficiency of motion recognition, so the OpenPose package, which is the second solution with better computing efficiency, is used as a tool for capturing spatial-temporal information of human joints.

![Figure 1: Inference time comparison between OpenPose, Mask R-CNN, and Alpha-Pose (fast Pytorch version).][1]

OpenPose is a human limb recognition model developed based on Convolution neural network (CNN) and supervised learning. It can run on different platforms, including Ubuntu, Windows, Mac OSX and embedded systems, and also supports a variety of hardware. In the 2018 version, the prediction of the partial affinity field (Part affinity field) was carried out in the previous stage, and then the prediction of the pixel coordinate confidence map (Confidence map) of a specific part of the human body in the image was carried out. As shown in Figure 2, the calculation amount of each stage is reduced by half.

![Figure 2: Architecture of the multi-stage CNN.][1]

2.2 Human action recognition

Human action recognition is an extremely important field of artificial intelligence applications, and due to the successive establishment of related open-source databases, such as UCF Data Set, HMDB, and NTU RGB+D human motion database, recent related research has made great progress. For example: in the field of sports, Masato Naka et al. in 2018, used the OpenPose model to apply the logistic regression model to the skeleton position to analyze the performance of the shot or not [2]; in the industrial field, Khawla Mallat et al. in 2019 In 2008, the OpenPose model was also applied to the Internet of Things (IoT). When an emergency occurs in the factory, the camera will detect whether someone passes by, and if so, it will be uploaded to the cloud alarm system in time to minimize occupational disasters [3].
Table 1: Related research on human action recognition based on skeleton.

<table>
<thead>
<tr>
<th>Title</th>
<th>Model</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition[4]</td>
<td>MST-GCN (GCN-based)</td>
<td>2021</td>
</tr>
<tr>
<td>Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition[6]</td>
<td>MS-G3D (Spatial-based GNNs)</td>
<td>2020</td>
</tr>
<tr>
<td>Skeleton-Based Action Recognition with Shift Graph Convolutional Network[7]</td>
<td>shift-GCN (GCN-based)</td>
<td>2020</td>
</tr>
</tbody>
</table>

2.3 Dynamic Time Warping (DTW)

Dynamic time normalization is a kind of distance metric, which can scale and correct different time series, find the most matching trajectory between actions, and calculate the distance. Dynamic time rounding first uses Euclidean distance to calculate the distance matrix corresponding to each time series. Then use dynamic programming (DP) to find the shortest path and accumulate the distance to form the final cost matrix as shown in Figure 3.

Figure 3: Cost matrix after cumulative distance. [8]

Fast-DTW consists of three operators: Coarsening, Projection, Refinement. (1) Use fewer data points to represent the same time series curve as accurately as possible to achieve lower The length of the time series; (2) Use dynamic programming (DP) to find the shortest distance path in a low-resolution grid, and generalize this path to a higher resolution, and this generalized path is called a projected path; (3) In order to optimize and improve the accuracy, a new grid is added near the projected path, and the operation optimization is performed together. As shown in Figure 4, a linear algorithm with a time complexity of $O(n)$ is completed.

Figure 4: The Fast-DTW algorithm with time complexity of $O(n)$. [8]
3. Materials & Methods

In this study, through the erection of photographic equipment, the motion video of the assembly operator's hands in the production line is recorded, and the video is cut into several assembly steps. First, we use OpenPose to extract the joint point coordinates of the hand skeleton and use it as the input data of the model. Then, after standardizing the data, the non-linear DTW distance between the time series data of the character skeletons in the film is calculated through DTW, and the working condition of the operator is then quantified, and then the statistical hypothesis test is used to determine whether there is a gap between the operator's assembly action and the selected standard action. Finally, according to the statistical results, K-Means clustering is carried out for unqualified operators, and the grouping results are sent to the quality control personnel for feedback analysis.

Figure 5: Flow chart of the implementation

3.1 Image data collection and OpenPose human skeleton identification

The working environment of this study is shot by a V8 camera set at a 45° depression angle in front of the operator, which can clearly record the operator's nose tip, shoulders, arms, palms and their complete working area, and the simulated operation consists of five designated (Standard) of the motherboard assembly steps, and record the video at 60 frames per second to ensure the integrity of the action data record, as shown in Figure 6. A total of 79 videos with about 60 seconds of assembly action were filmed in this topic as research data.

Fig. 6: Workstation setup diagram.
Since the study intends to use the limb recognition kit for intensive hand assembly, the output data of the whole-body recognition (including the hand) has been revised, only the joint points and finger joint points of the upper body are considered, and the coordinates of all the joint points of the lower body are deleted, as shown in Figure 7.

Fig. 7: Human skeleton diagram for this study.

### 3.2 Preprocessing of human skeleton data

First, the skeletal coordinates of all frames of each operator are integrated into a data set. Due to the fine hand movements, the camera may not be able to detect some joint points when shooting a few movements, resulting in null values. To reduce the calculation error caused by null values in the follow-up data, in this experiment, all joint coordinate positions with null values are replaced with the coordinate positions of the previous frame of the coordinate point. In addition, the experiment also needs to consider the differences in the operator's body shape, body movements and postures, as well as the different assembly speeds, so the data needs to be standardized before the OpenPose model is put into the DTW image distance calculation. The normalization method is to normalize the X-axis and Y-axis respectively through the number of each skeleton point in the image, take the ratio of the maximum value to the minimum value, and transform its coordinates to be between (0,0) and (1,1).

### 3.3 Import DTW model and image distance calculation

In this study, a standard video will be selected from most of the videos as a comparison benchmark, and the assembly actions of the rest of the operators will perform DTW distance calculation with the standard video to produce the distance between the two data. In this study, a new calculation method is designed: use the matrix of k joints to calculate the respective Warping Path, then divide the respective total distance (Sum) by the length of the Warping Path, and then add up the k values and divide by k to get out the final distance. This method does not need to calculate the Global Path, so it can reduce its time complexity.

Fig. 8: Motion trajectory generated by DTW.

### 3.4 Similarity between operators and standard operator.

After calculating the DTW distance matrix between the operator's assembly action and the standard video, the next step is to calculate the similarity between the assembly action of each operator and the standard operator, as shown in formula (2). Among them, there are 21 joints (Nodes) for OpenPose unilateral hand detection, and the assembly action of this experiment is two-hand operation plus two joints of the elbow, a total of 44 joints are used, and the corresponding joints of each joint are calculated. The distance difference
of the same joint point of the standard operator, the distance difference of 44 joint points is summed up to form the cost matrix (Cost Matrix) of formula (1), which is shown as $C$ in formula (2).

In addition, the Optimal Warping Path and the final distance difference are obtained when calculating the DTW distance. However, the DTW distance difference is calculated cumulatively, so the length of the time series directly affects the final distance. As the length of time increases, the distance difference will be larger, and then the error will be derived. Therefore, the solution is to add the accumulation times of the optimal corresponding path (Optimal Warping Path) when designing the similarity formula, which is represented by $S$ in formula (2), and then divide the final distance by DTW by the accumulation times $S$ to eliminate. The time length is represented by $T$ in the formula (2) due to the error caused by accumulation. This study uses the above method to calculate the final distance proportion of each DTW and the similarity percentage (0~100%) required for subsequent calculations.

$$\text{Dist} = \sum_{m=1}^{M} d(X(i), Y(j))$$  \hspace{1cm} (1)

$$\text{Performance} = \frac{\sum_{k=1}^{K} \left[ 1 - \frac{C(s_{ik}, t_{jk})}{44} \right]}{S \times \sqrt{2}} = 1 - \frac{\text{DTW}(S, T)}{44 \times S \times \sqrt{2}}$$  \hspace{1cm} (2)

### 3.5 Hypothesis test of actions qualification

To compare and analyze the assembly actions of other operators and the standard operator, first calculate the average and standard deviation of the movement difference between each operator and the standard video, and then use the hypothesis test to determine whether the operator's assembly action is consistent with the standard action. differences, and whether operators need to improve their assembly movements.

The statistical hypothesis testing process of this study is to perform the variance hypothesis test (F-test) between the operator and the standard operator, and then decide whether to perform the mean value hypothesis test (Z-test) according to the result of whether the variances are equal. Assuming that the variance of an operator is not equal to that of the standard operator, it is then determined that the result is not similar to the assembly action of the standard operator and needs to be improved. Conversely, if the variances are equal, the mean hypothesis test is accepted. If the null hypothesis of the mean value is still not rejected, that is, the two mean values are equal, the operator's assembly action is judged to be qualified; if it is rejected, it is also judged to be unqualified when the variance is not equal.

### 3.6 K-means clustering analysis

After identifying all unqualified operators through statistical hypothesis testing, managers can refer to the relevant data to arrange education and training for employees who should strengthen their assembly activities. Choosing reasonable teaching content and proper grouping will help the factory to achieve the improvement goal in a more efficient and cost-effective training method. After obtaining the optimal number of clusters through the inflection points of the line chart, randomly set $K$ cluster centers, and assign all data points to the center points of the closest distance, and then reset the cluster centers until the cluster centers do not change after each update, achieving the purpose of convergence is to complete the clustering. This study applies the fuzzy concept to the results of the unqualified operator grouping, indicating how high the probability that the operator belongs to each cluster, which is different from the traditional grouping, all operators can be divided into different probability belong to multiple classes at the same time. According to formula (3), calculate the probability that the joint points that each unqualified operator needs to improve is the same as the joint points that each group worker needs to improve. From the results of the grouping and the analysis feedback of the quality control personnel, it is possible to know the need for improvement.
Suppose that among the 79 films, there is a standard operator. In this experiment, the DTW distance is used to calculate the difference between 78 videos and the standard video assembly action and calculate the average and standard deviation of each operator’s assembly action difference, as shown in TABLE 2.

\[
S_k = 1 - \frac{Distance_k}{\sum_{i=1}^{n}Distance_i}
\]  

(3)

4. Experiments

Then, the difference F test of the variance of the two population was carried out, and the value was taken to the seventh decimal place. It was concluded that only the third and fourth workers had the same variance as the standard worker, and the variance of the rest of the workers was equal. It can be seen from the data in TABLE 3 of this experiment that only the third and fourth operators need to do the mean value hypothesis test, and the rest of the work has been judged to be unqualified in the variance hypothesis test stage. After calculation, it is shown that only the third operator's assembly action is similar to that of the standard operator, and the fourth operator was rejected in the mean value test, and the null hypothesis was also regarded as unqualified, as shown in TABLE 4.

<table>
<thead>
<tr>
<th>Operator</th>
<th>Average of difference</th>
<th>SD of difference</th>
<th>Number of videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard worker</td>
<td>0.0463855</td>
<td>0.0062361</td>
<td>13</td>
</tr>
<tr>
<td>1st worker</td>
<td>0.0729345</td>
<td>0.0058841</td>
<td>8</td>
</tr>
<tr>
<td>2nd worker</td>
<td>0.0639717</td>
<td>0.0038757</td>
<td>10</td>
</tr>
<tr>
<td>3rd worker</td>
<td>0.0477967</td>
<td>0.0025528</td>
<td>10</td>
</tr>
<tr>
<td>4th worker</td>
<td>0.0714534</td>
<td>0.0022276</td>
<td>11</td>
</tr>
<tr>
<td>5th worker</td>
<td>0.0740704</td>
<td>0.0044448</td>
<td>10</td>
</tr>
<tr>
<td>6th worker</td>
<td>0.0635332</td>
<td>0.0053556</td>
<td>9</td>
</tr>
<tr>
<td>7th worker</td>
<td>0.0636780</td>
<td>0.0045403</td>
<td>2</td>
</tr>
<tr>
<td>8th worker</td>
<td>0.0647844</td>
<td>0.0050540</td>
<td>2</td>
</tr>
<tr>
<td>9th worker</td>
<td>0.0794864</td>
<td>0.0047178</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: The difference between the assembly action of the operator and the standard film

<table>
<thead>
<tr>
<th>Operator</th>
<th>p-value</th>
<th>Result</th>
<th>t-value</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st worker</td>
<td>0.8583535</td>
<td>Reject</td>
<td>-15.3696</td>
<td>Reject</td>
</tr>
<tr>
<td>2nd worker</td>
<td>0.1563081</td>
<td>Reject</td>
<td>-10.1952</td>
<td>Reject</td>
</tr>
<tr>
<td>3rd worker</td>
<td>0.0114552</td>
<td>Accept</td>
<td>-0.7697</td>
<td>Accept</td>
</tr>
<tr>
<td>4th worker</td>
<td>0.0028210</td>
<td>Accept</td>
<td>-14.4728</td>
<td>Reject</td>
</tr>
<tr>
<td>5th worker</td>
<td>0.3058330</td>
<td>Reject</td>
<td>-16.0458</td>
<td>Reject</td>
</tr>
<tr>
<td>6th worker</td>
<td>0.6481568</td>
<td>Reject</td>
<td>-9.9325</td>
<td>Reject</td>
</tr>
<tr>
<td>7th worker</td>
<td>0.7970288</td>
<td>Reject</td>
<td>-10.0029</td>
<td>Reject</td>
</tr>
<tr>
<td>8th worker</td>
<td>0.8768083</td>
<td>Reject</td>
<td>-10.6353</td>
<td>Reject</td>
</tr>
<tr>
<td>9th worker</td>
<td>0.5792748</td>
<td>Reject</td>
<td>-19.1583</td>
<td>Reject</td>
</tr>
</tbody>
</table>

Table 3: The result of F test  

TABLE 4 The result of t test
Operator | Unqualified joints
--- | ---
1st worker | BD6, LH16, LH20, LH18, LH19
2nd worker | LH16, LH12, LH11, LH15, LH19
3rd worker | None, None, None, None, None
4th worker | BD6, LH9, LH0, LH1, LH6
5th worker | LH0, LH2, LH6, LH1, LH5
6th worker | BD6, LH20, LH16, LH19, LH12
7th worker | BD6, LH9, LH0, LH1, LH20
8th worker | LH1, BD6, LH0, LH9, LH20
9th worker | LH0, LH7, BD6, LH2, LH17

TABLE 5: Unqualified joints of each operator (LH: left hand nodes, RH: right hand nodes, BD: body nodes)

5. Conclusion & Suggestion

5.1 Conclusion

This study establishes a system. First, the dynamic time warping algorithm is used to analyze the time series data of the production line operator's hand skeleton. Then use statistical tools to check whether the assembly actions of the operators are standard, and then use the K-means clustering technology to classify the unqualified operators. Finally, provide the information to managers, to improve the efficiency of production line training and employee selection. In the part of the hand skeleton time series data, the standard action is first formulated, then the workstation is set up and the shooting operator assembles the video data, and the standard operator and the standard video are selected. Then, the two-branch multi-stage CNN included in the OpenPose package is used to output these image data as spatiotemporal information of human joints. In the dynamic time warping algorithm part, first divide the coordinate axis of the space-time information into two dimensions (X and Y), and perform numerical normalization processing, and then input the data into the dynamic time warping algorithm for data length alignment and distance from the standard video. Compare, output the final difference distance and the joint node that contributed the most distance. In the statistical verification part, the difference F test of the variance of the two mothers is performed on the distance between the standard operator and the general operator to determine whether the variance of the two is the same, and then the average test is performed on those with the same variance, to achieve the goal of determining whether the operator is similar to the standard operator. Operators whose variance and average value are the same as those of standard operators are judged as qualified operators, while unqualified operators can be trained according to the joints with the most differences.

In addition, this study conducts two verification experiments on its own algorithm. First, compare the same video of the standard operator with the original video after speeding up and slowing down. The difference between the results is very small, and it can be judged that it is a calculation error, and the similarity is close to 100%. Second, after reducing the scale of the same film by the standard operator, and comparing it with the original film, the difference between the results is very small, and it can also be judged that it is a calculation error, and the similarity is close to 100%.

5.2 Suggestion

In terms of follow-up research, this topic proposes three major aspects, focusing on "data", "algorithmic operation", "architecture" and "result analysis". First, in terms of data, the image data used in this topic contains a small number of operators, and the number of videos shot by each operator is also inconsistent, so the calculation is prone to insufficient samples. It is
suggested that researchers can increase the number of operators in the future and pursue the consistency of the number of video researchers’ videos. Second, in terms of algorithm operation, the DTW algorithm used in this topic is an algorithm with a relatively complete calculation range in dynamic programming, and the Fast DTW algorithm has appeared on the same basis. Fast DTW is a DTW algorithm with a simplified calculation range. It is better than the original DTW in terms of operational efficiency, but it also loses accuracy while reducing the calculation time. The degree of misalignment has no fixed ratio and needs to be tested in practice. It is suggested that future researchers can test the effect of Fast DTW and compare it with this topic. At the same time, optimization methods such as parallel computing should also be considered to reduce the calculation time of the algorithm. Third, in terms of architecture, the framework of this topic is based on the DTW algorithm as the main axis for data comparison. The results under this framework can only provide decision-making aids with joint points and frame numbers. Therefore, it is suggested that researchers in the future can add ST-GCN human skeleton action recognition to provide more humanized auxiliary information.

References