ANNALS of Faculty Engineering Hunedoara – International Journal of Engineering
Tome XX [2022] | Fascicule 1 [February]

1. C. CHANG, 2.P. HUANG

COMPUTER VISION-BASED ACTION MONITORING SYSTEM IN ASSEMBLY LINES

[1.]Fu Jen Catholic University, Graduate Institute of Business Administration, Taipei City, TAIWAN
[2.]Fu Jen Catholic University, Taipei City, TAIWAN

Abstract: The emergence of computer vision technology with consumer red, green, blue (RGB) cameras provides more information than motion sensors and has created more advantages in developing new solutions. This current study develops an action monitor system (AMS) by applying computer vision techniques for detection and recognition of assembly line operator actions using depth information gained from RGB cameras and the application of a dynamic time warping (DTW) algorithm. The AMS extracts the features from the depth information of the operators' hand action characteristics, while the DTW algorithm is used to recognize and detect the specific actions. The AMS monitors operators executing every action defined by the working instructions. Production engineers can improve action efficiency based on the data collected, while supervisors can allocate operators to suitable workstations based on their action performance as monitored by the AMS. Manufacturing quality is thus guaranteed and efficiency improved.

Keywords: computer vision, dynamic time warping, action detection, behavior recognition, skeleton

1. INTRODUCTION

In the electronics industry, manual assembly is an important process. One missing screw or user manual results in customer complaints. It is challenging to ensure that operators execute each activity according to the standard operating procedures (SOPs). Computer vision is one of the promising intelligent solutions to overcome this shop-floor management issue. With the deployment of intelligent solutions surrounding the working environment, operators can gain sufficient assistance and their quality of work can be improved (Chua, Chang, Jaward, Parkkinen, and Wong, 2014).

Over the past decades, most computer vision-based applications and systems have been developed using 2D video cameras. Advances in technology is resulting in a strong interest in and motivation for conducting research related to the development of improved systems with computer vision techniques that can determine depth as well, in order to be able to detect and recognize manufacturing activity through the use of video cameras. The advantage of video cameras with depth sensor information is that they can provide more visual-based information such as precisely detecting the position of the body in space. Moreover, the depth information can accurately detect the body parts even when some of the parts are overlapping each other.

This current study aims to develop an action monitoring system (AMS) to intelligently detect and recognize operator actions and assist operators in their daily work. An AMS can remind an operator that skips an action to ensure every action is executed as per the work instructions. The objectives of this research are:

- To design a computer vision-based system to automatically identify operator actions through depth information from RGB cameras.
- To monitor the execution of every action as defined by the work instructions and measure time spent on each action.

2. CLASSICAL VERSUS MODERN COMPUTER VISION

There are few previous works carried out that combine both sensor-based and visual-based techniques to collect data in order to assist operators in performing manufacturing activities (Spasova and Iliev, 2012). However, computer vision and image processing techniques have become a major research interest due to the advances in the field of digital signal processing. In the study by Cardinaux, Bhowmik, Abhayaratne, and Hawley (2011), most of the visual-based approaches mainly focus on body detection, such as walking, opening a door, standing, etc. (Cardinaux et al., 2011; Chaaraoui, Climent-Pérez, and Flórez-Revuelta, 2012; Sun, Yao, Jia, and Sun, 2013; Ren, Yuan, Meng, and Zhang, 2013). In Zhou, Chen, Chung, He, Han, and Keller (2009), a single fisheye camera without calibration is used to capture the physical location and the moving speed of a person and a hierarchical action decision tree is used to classify the human actions.

The proposed technique in Zhou et al. (2009) uses the locally linear embedding method to map the primitive visual features based on the motion vector search from the high-dimensional into low-dimensional space for discovering a small set of composite features. An automatic detection of chewing events by using video surveillance is proposed by Cadavid and Abdel-Mottelab (2012). This system first applies an active appearance model for tracking the face in the video sequence and observing the changes of parameters, and then uses a support vector machine classifier to detect the desired actions.

Iosifidis, Marami, Tefas, and Pitas (2012) indicate that human activities in videos can be described by a sequence of video frames. This forms a 3D volume representation for each frame where the third dimension refers to time. A fuzzy vector quantization is performed on the 3D volumetric human body representation and linear discriminant analysis is used to map the activity representations into a low dimensional discriminant feature space. Nearest-centroid classifications are then used to classify human activities.

Multiple objects in a scene or other noise may affect aspect ratio measurements, leading to some error of detection. As such, Nasution and Emmanuel (2007) propose to compare the histogram of the posture by applying the K-nearest neighbors (KNN) algorithm to classify similar postures between the frames. After the postures are identified, the movement or action can be recognized. Motion history image-based solutions are commonly used to estimate the movement of an object within a given time frame or image sequence (Foroughi, Naseri, Saberi, and Yazdi, 2008; Zhou et al., 2009).

The technological issues of computer vision that need to be addressed are: (1) Choosing the right image and video processing techniques and algorithms; (2) Handling multiple subjects and target objects; (3) Solving occlusion issues; (4) Fusing data from multiple cameras or sensors; and (5) Handling the presence of other non-human moving objects (Spasova and Iliev, 2012). Based on this, Ben, Mohamed, Val, Andrieux, and Kachouri (2013) develop a monitoring system using the Microsoft Kinect V1 sensor to control and monitor activities. With the use of an RGB camera, the location and skeleton information of a human can be obtained. In addition, the segmentation of the foreground, background, and motion can easily be performed as well. Moreover, the depth information obtained can accurately provide distance information that is not sensitive to the change of lighting conditions and illumination. This current study is inspired by the above monitoring system and develops an AMS with further improvements in the algorithms.

3. ACTIVITY IDENTIFICATION THROUGH THE USE OF DEPTH INFORMATION

This current study applies a visual-based system for automatic identification of manufacturing activities using depth information from a RGB camera. Compared with conventional RGB cameras, the depth sensor provides a greater advantage related to the detection of human skeletons because it is not affected by the change of lighting conditions. In order to recognize and differentiate between manufacturing and non-manufacturing activities, a dynamic time warping (DTW) algorithm is used to analyze and recognize the depth data obtained.

The hand postures and actions of a person are the main key points during the assembly activity of manufacturing. When a person is performing an assembly activity, the hand posture changes and the sequence will repeat in a cyclical manner until the SOP is completed. By using depth information, the hand distance data can be modelled in a repeating signal, which forms the principle of this proposed technique. Through the application of a Microsoft Kinect V1 RGB-D camera, the skeleton and joint position of the operator can be detected. The depth information of the operator's manufacturing sequence, which is the distance from the camera to the interest point in the frame, is collected. The depth information sequence of the operator's hand that is used for grabbing objects (e.g., screws) is recorded. The DTW algorithm, which is commonly used in 1-dimensional (e.g., speech) signal recognition (Giorgino, 2009; Vullings, Verhaegen, and Verbruggen, 1998; Sakoe and Chiba, 1978), is then used to analyze the depth information of the manufacturing sequence in order to determine whether the activities are assembly related. The overall diagram of the proposed system contains four different stages: (1) Data acquisition: The application of Microsoft Kinect Windows SDK 1.8 to initialize the necessary sensor to collect the depth frame data and information; (2) Normalization: The normalization of the variant in the range of the depth distance captured by the camera; (3) Segmentation: A manual segmentation is carried out to segment out every manufacturing activity for recognition in order to enhance the accuracy of the recognition; and (4) Recognition: The DTW algorithm is applied to calculate the distance between two different signals and to identify the dissimilarity between the signals, allowing the activities to be recognized.

4. TARGET OBJECT DETECTION AND RECOGNITION

This current study applies multi-access edge computing (MEC) to execute the AMS. The MEC specifications are Intel i7core, 16 GB DDR RAM, 256GB SSD, and 4 VPUs. The specifications for the video storage hard drives are 6TB, 2.5 inches, and SATA 3.0Gb/s. A wide-range 3D camera is mounted on top of the operator work station along with a monitor, speaker, and alarm. A 21 inch electronic board with touch functions displays the live status through an HDMI interface, designed so that supervisors can review any issue triggered by the AMS. The topology of the AMS is shown in Figure 1.

The AMS executes the methodology in two phases. Phase one is the depth image object extraction, which consists of (1) Depth image acquisition using the Kinect sensor; (2) Background subtraction to extract the object's depth image; (3) Histogram feature representation to facilitate the object recognition analysis; (4) Redundant information removal from the histogram pattern obtained; and (5) Noise reduction to enhance the

image quality. Phase two consists of feature extraction and object recognition. This is achieved through the application of KNN classification to the histogram dissimilarity measurements of the features, and recognition of moment invariant features through the application of a statistical approach.

In Microsoft Kinect V1, the depth information from the sensor is represented as 11-bit long data. The conventional techniques to display video frames on screen are normally represented as 8-bit or 256 levels of color depth. The use of an 8-bit color to display the depth data from the depth sensor results in some of the information being clipped or eliminated. Hence, in the system developed in this study a 16-bit greyscale color depth is used for display purposes in order to maintain the accuracy of the depth information. Microsoft Kinect V1 human skeleton detection technique is used for acquiring the necessary depth data. From the existing skeleton detection technique in Holmquest (2012), up to 20 skeleton points can be detected. In the AMS, a seated mode skeleton detection is used, as assembly actions usually only consist of upper body movements. To ensure a low computational complexity to the proposed system, only the depth information from the hand is recorded. Figure 2 presents the live monitoring of the AMS.

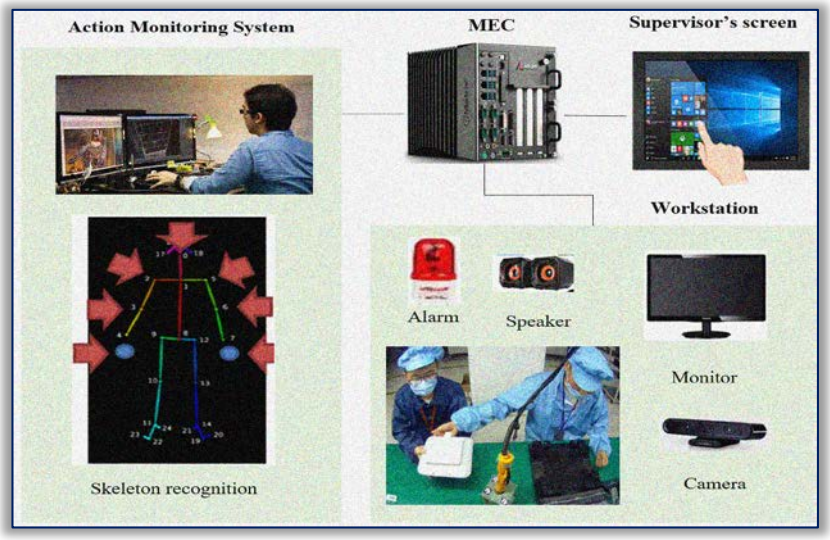

Figure 1. AMS topology

Figure 2. Live monitoring of the AMS

The AMS first records the operator's real actions and compares these with the work instructions to determine the norm for each action. The AMS monitors the time of each action in order to notify when there are missing actions. The AMS reminds operators of the missing actions immediately through a speaker when the operator delivers finished parts to the next workstation. Meanwhile, the AMS calculates the longest and shortest time spent executing the same action. Supervisors can check and determine the areas that can be improved for each operator. This leads to the operators being better able to meet the standard times defined by the work instructions. An important factor to take note of, is that the speaker only reminds operators of missing actions and not that time spent exceeds the standard. This is to reduce the psychological pressure of the operators.

The AMS records each operator's learning curve over time for each action. That is, it determines how long it takes until the operator can complete the action within the set target time. Production engineers review the operators' actions and then design tools to assist the operators in completing the actions more efficiently. Supervisors can thus now understand the physical limitations of each operator based on his or her learning curve and can therefore arrange the operators and workstations based on individual strengths. For example, some operators can complete a 12-second action in 8 seconds, while other operators will still require 12 seconds or more to complete the same action even after he or she has been working at the same workstation for two weeks. This means that this operator may not be suitable for the action due to physical limitations. The supervisors can thus assign the operator to another workstation with a better individual fit.

The output of an assembly line is limited by the bottleneck workstation. If the operators of the bottleneck workstation can complete every action within the standard time, the output from the assembly line can meet the production targets. The AMS should therefore first be used on the bottleneck workstation to ensure that the specific operators can complete every action within the standard time. In addition, production engineers can review the workstation's actions monitored by the AMS and then improve the actions further. These

improvements will allow the operators of the bottleneck station to complete the actions more effectively and thus improves the efficiency of the entire production line.

5. CONCLUSION

This current study develops a vision-based AMS for manufacturing activity recognition. A DTW algorithm is applied to analyze and recognize the depth data obtained. The technical advantage of using depth information as the data for manufacturing activity recognition is that the detection will not be affected by the lighting and illumination conditions.

The manufacturing activities are first normalized and segmented into manufacturing actions, which then form the reference and test sequence database. The practical benefits of the AMS developed in this study is that it ensures operators execute every action as defined by the SOP. Assembly quality is thus guaranteed. Supervisors can also allocate operators to suitable workstations based on their action performance in terms of time. The visual-based AMS thus provides technical advantages as well as practical benefits.

References

[1] Ben, A.; Mohamed, H.; Val, T.; Andrieux, L. and Kachouri, A.: Assisting people with disabilities through kinect sensors into a smart house, in Computer Medical Applications (ICCMA), 2013 International Conference, 1–5, 2013.

[2] Cadavid, S. and Abdel-Mottaleb, M.: Exploiting visual quasi-periodicity for automated chewing event detection using active appearance models and support vector machines, in Pattern Recognition (ICPR), 20th International Conference, 1714–1717, 2010.

[3] Cardinaux, F.; Bhowmik, D.; Abhayaratne, C. and Hawley, M. S.: Video based technology for ambient assisted living: A review of the literature, Journal of Ambient Intelligence and Smart Environments, 3 (3), 253–269, 2011.

[4] Chaaraoui, A. A.; Climent-Pérez, P. and Flórez-Revuelta, F.: A review on vision techniques applied to human behaviour analysis for ambient-assisted living, Expert Systems with Applications, 39 (12), 10873–10888, 2012.

[5] Chua, J.-L.; Chang, Y. C.; Jaward, M. H.; Parkkinen, J. and Wong, K.-S.: Vision-based hand grasping posture recognition in drinking activity, in Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium, 185–190, 2014.

[6] Foroughi, H.; Naseri, A.; Saberi, A. and Yazdi, H. S.: An eigenspace-based approach for human fall detection using integrated time motion image and neural network, in Signal Processing, 9th International Conference, 1499–1503, 2008.

[7] Giorgino, T.: Computing and visualizing dynamic time warping alignments in r: the dtw package, Journal of statistical Software, 31 (7), 1–24, 2009.

[8] Holmquest, L.: Starting to develop with Kinect, MSDN Magazine, 27 (6), 2012.

[9] Iosifidis, A.; Marami, E.; Tefas, A. and Pitas, I.: Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes, in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, 2201–2204, 2012.

[10] Nasution A. H. and Emmanuel, S.: Intelligent video surveillance for monitoring elderly in home environments, in Multimedia Signal Processing, IEEE 9th Workshop, 203–206, 2007.

[11] Ren, Z.; Yuan, J.; Meng, J. and Zhang, Z.: Robust part-based hand gesture recognition using kinect sensor, Multimedia, IEEE Transactions, 15 (5), 1110–1120, 2013.

[12] Sakoe, H. and Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition, Acoustics, Speech and Signal Processing, IEEE Transactions, 26 (1), 43–49, 1978.

[13] Spasova, V. and Iliev, I.: Computer vision and wireless sensor networks in ambient assisted living: State of the art and challenges, Journal of Emerging Trends in Computing and Information Sciences, 3 (4), 585–595, 2012.

[14] Sun, X.; Yao, H.; Jia, W. and Sun, M.: Eating activity detection from images acquired by a wearable camera, in Proceedings of the 4th International Senser Cam Pervasive Imaging Conference, 80–81, 2013.

[15] Vullings, H.; Verhaegen, M. and Verbruggen, H.: Automated ECG segmentation with dynamic time warping, in Engineering in Medicine and Biology Society, Proceedings of the 20th Annual International Conference of the IEEE, 163–166, 1998.

[16] Zhou, Z.; Chen, X.; Chung, Y.-C.; He, Z.; Han, T. X. and Keller, J. M.: Video-based activity monitoring for indoor environments, in Circuits and Systems, IEEE International Symposium, 1449–1452, 2009.

ISSN 1584 – 2665 (printed version); ISSN 2601 – 2332 (online); ISSN-L 1584 – 2665