Upper Body Action Classification for Multiview Images using K-means

Siew Cheng Lai and Phooi Yee Lau
Faculty of Computing and Communication Technology (FICT); Universiti Tunku Abdul Rahman, Kampar, Malaysia
{laisc, laupy}@utar.edu.my

***Abstract*—This paper proposes to extend the framework presented in [1]. In this paper, we proposed to label the upper body pose in the acquired multiview images. These human pose can be used to represent different body action, such as punching, which involve upper arm to be further from the body. Our extended framework here focuses on the upper body parts especially actions involving the upper and lower hand(s). Two steps were added to the previous framework, namely, (1) pose estimation, and (2) hand action classification. Here, we employ a simple estimation method, where three points will be identified, namely, shoulder *S*, elbow *E* and wrist *W,* forming two lines. In the hand action classification, this paper focuses on three types of hand actions, namely (1) *HSS* (hand stretching), *HNS* (hand not stretching). For this, the k-means clustering is being employed, whereby, images from different view (front camera, left camera and right camera) will be classified into one of the two possible action. Preliminary results show that our proposed framework is quite promising in classifying the upper body actions.**

Keywords—action recognition, pose estimation, hand action classification, k-means clustering

I. INTRODUCTION

Human action recognition is still an active area in computer vision due to its wide applications in everyday life. Due to this reason, there have been many algorithms developed for human recognition since in the 1980s. Furthermore, with the Kinect devices in the market, there are more action recognition develop using depth and skeleton data. Paper [1] – [3] show various methods and algorithms reviewed for action recognition. In 2014, Guo et al. categorized human action recognition using high-level cues such as human body, body parts, objects, human object interaction and scene [1]. These high level cues are characterized using low level features such as Dense sampling of Scale Invariant Feature Transform (DSIFT), Histogram of Oriented Gradient (HOG), shape context, and some other features. In 2014, Aggarwal et al., and in the 2016, Presti et al., review works focus on using depth data in image to classify an action [2-3]. Aggarwal et al. categorized human action recognition based on depth data based on the following five categories of features: 3D silhouettes, skeletal joints/body parts, local spatio-temporal features, local occupancy pattern and 3D optical flow. Meanwhile, Presti et al. classified human action based on the following three categories of 3D data in skeleton-based human: joint-based representation, mined joints based representation and dynamics-based descriptors. Even though, most of the methods reviewed have its advantages and disadvantages, there are no reviews on works with multiview datasets. Multiview datasets can help us to overcome occlusion in action recognition because most of the time using image from one view might not be sufficient to determine the action. Thus, our work will focus on multiview datasets.

In this paper, we will extend our framework in [4] with pose estimation and action classification process. The focus of our action classification is to recognize upper body part, namely, based-on the hand. The pose estimation algorithm proposed focuses on the 3-point joints location detection of the upper arm and lower arm, namely shoulder, elbow and wrist. The idea used here is based on the work [5], i.e. using joints location, to determine the action of the person. Then taxonomy is built according to the relative position of the 3 joints. There are 4 categories defined as follow: elbow on bottom left of shoulder, elbow on bottom right of shoulder, elbow on top left of shoulder and elbow on top right of shoulder. Then each category is divided into smaller categories based on whether the wrist is in left, between or right relative to shoulder and elbow joints. Finally, this group is further categorized into three sub-partitions based on whether the wrist is below, between or above the shoulder and the elbow joints for pose estimation.

This paper will be divided to few sections as described next. Section II describe and discusses the proposed framework. Section III shows the experimental results. Section IV concludes with future work.

II. ACTION RECOGNITION FRAMEWORK

Figure 1 shows the extended framework for our work. The framework consists of image acquisition, pre-processing of image acquired, skeletonization, pose estimation and hand action classification. This paper extended our work in [4].

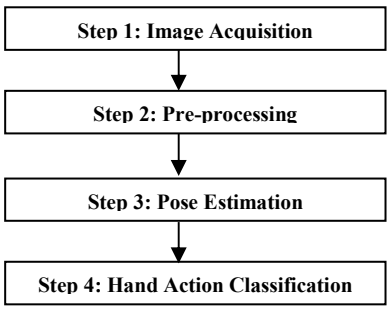


Fig. 1. Extended Framework for [4]

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

A. *Step 1: Image Acquisition*

Figure 2 shows the background setup area for image acquisition. In Figure 3 the three cameras position are shown with their position in the setup area. The cameras are placed in these positions so we can take the image from different position which can shows same action in different views.

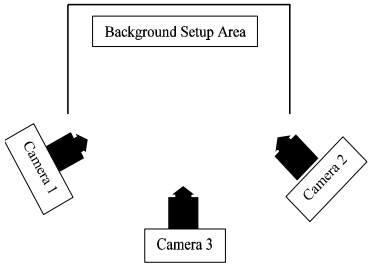
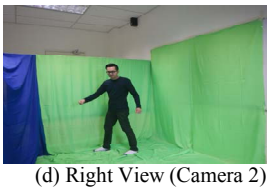


Fig. 2. Camera Position in Acquisition

(a) Background Setup Area

(b) Front View (Camera 3)

(c) Left View (Camera 1)



(d) Right View (Camera 2)

Fig. 3. Acquisition Image: (a) Background of Setup Area and (b) Camera 1, (c) Camera 2, and (3) Camera 3

B. *Step 2: Preprocessing*

The acquired image in *Step 1* will be sent for preprocessing before pose estimation. Firstly, the mask for background color will be applied to the image to remove the background color. Then the image will be converted to binary image. Thereafter, the small hole in the image is filled. This is followed by clearing the border and small areas from the image before it can be processed further. The clean image from *Step 1* will be processed in *Step 3*.

C. *Step 3: Pose Estimation*

Figure 4 shows the pose estimation process. The output image from *Step 2* will be used to find the human shape from head to leg in a bounding box. This is due to foreground pixels are assumed to resemble the human pose. Here, we obtain the starting point, S, of the (1) upper hand pose, and (2) lower hand pose, based on the average ratio of the head to neck calculated manually from the image. The average ratio from head to hand is also calculated manually first. Later, the following sub-steps will determine three points, namely S (shoulder), E (elbow), and W (wrist).

P1: From the starting row until end, check each column until the first column with value 1 is encountered. Store this column position in a vector. Initialize End as the maximum row value in the Bounding Box. End is used to determine the wrist point.

P2: Starting from the row position specify in S in vector from P1, we will check the difference between the current value and the next value. If the difference can be accepted, then go to next value in the vector and repeat the process.

P3: If the difference cannot be accepted due to a big change between the current value and next value in vector, then set End to the current value position in the vector. This can occur when the next value belongs to other body part like body side which is far from the stretch hand.

P4: Check if the End is more than the ratio of head to hand. If more than the ratio, then set End to the ratio value. This can occur when the hand is at the side. Else, the End value remained the same in P3. When the process ended, the starting position S in the vector represents the shoulder point and the End value represents the position of wrist point.

When the shoulder point and the wrist point are detected, the elbow point can be calculated using the middle row position between shoulder and wrist. Thus, the hand pose can be obtained by using the stored vector. A sample of the hand pose for different pose is shown in Figure 5 where left column shows the ground truth image and the right column shows the hand pose drawn in line label with shoulder, elbow and wrist points.

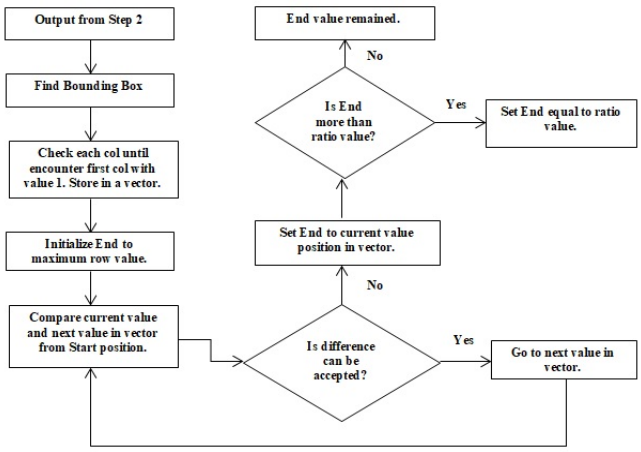
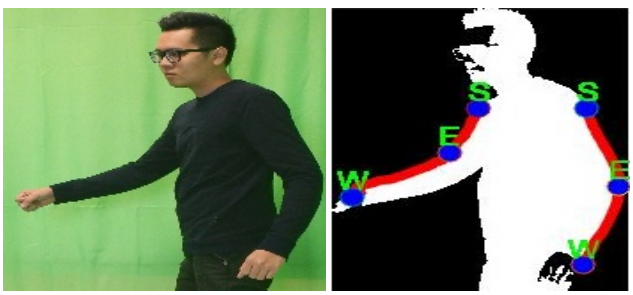


Fig. 4. Pose Estimation Process



(a) (b)

Fig. 5. Output: Pose Estimation

D. *Step 4: Hand Action Classification*

The left and right hand pose obtained from *Step 3* will be used in this step. The hand pose from *Step 3* will be further divided into 2 segments. The first segment is from shoulder point to elbow point and the second segment is from elbow point to wrist point. Then the gradient from these 2 segments are used to classify the hand action using k-means algorithms. The classification for left hand and right hand will be clasify Figure 6 and Figure 7 shows the k-means classification for the left hand and right hand respectively.

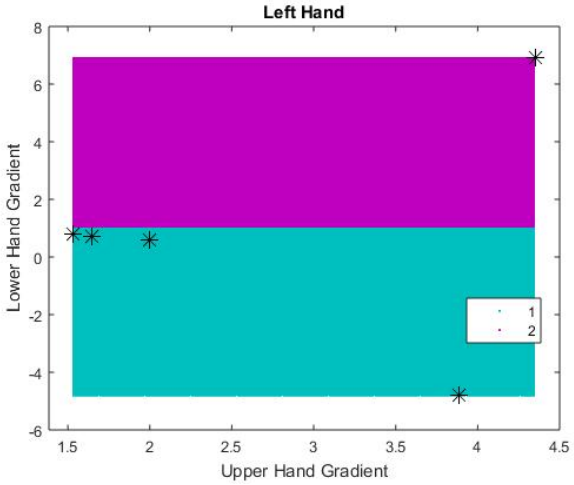


Fig. 6. K-means Classification for Left Hand

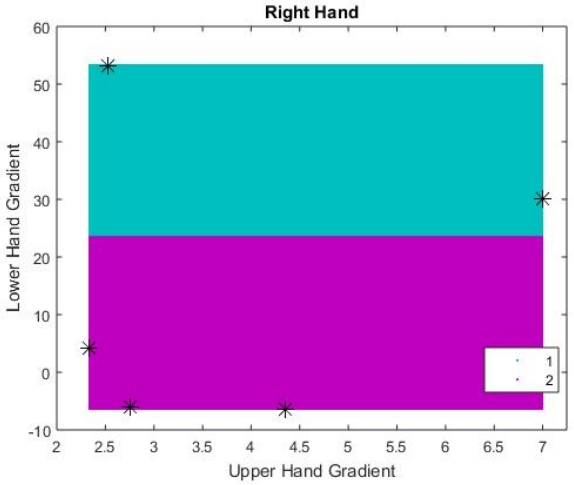


Fig. 7. K-means Classification for Right Hand

III. EXPERIMENTAL RESULTS

The experimental dataset used in this work consist of two action acquired from one person. For each action, three images were acquired based on the 3 views (front, left and right), i.e. using 3 cameras set-up as shown in Figure 2. The action (whether a person is stretching hand or not) from each view might be different. For example, Figure 8(a) (Left Camera), 8(b) (Right Camera) and 8(c) (Front Camera) are from the same dataset. Table I discussed the dataset used in the experiments in detail, elaborating the action from each view for both datasets.

TABLE I. DATASETS WITH ACTION FROM THREE CAMERAS

Dataset	Action		
	Left Camera	Center Camera	Right Camera
1	Left Hand Stretch. Right Hand Not Stretch.	Left Hand Stretch. Right Hand Not Stretch.	Left Hand Stretch. Right Hand Stretch.
2	Left Hand Not Stretch. Right Hand Stretch.	Left Hand Not Stretch. Right Hand Stretch.	Left Hand Not Stretch. Right Hand Stretch.

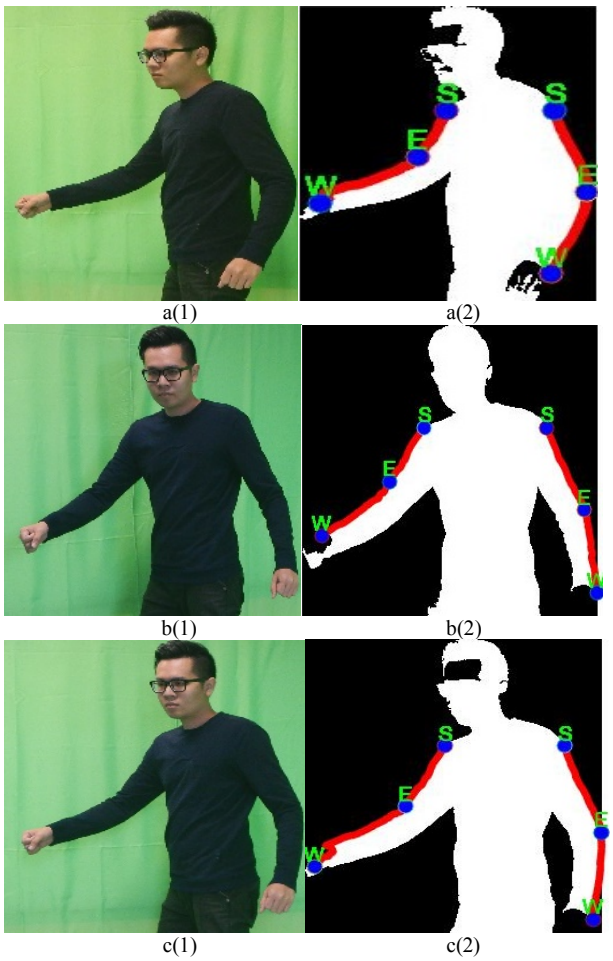


Fig. 8. Output: Pose Estimation

Figure 8 and Figure 9 show the result of the pose estimation step. The left hand column shows the ground truth image and the right hand column shows the result from the pose estimation of the hand annotated with the 3 joints (S, E and W). Figure 8(a), (b) and (c) shows the result from dataset 1 with the hand pose annotated. Meanwhile, we also run additional experiment for 2 more images shown in Figure 9. Figure 9(b) and 9(c) are images with non-stretch hand.

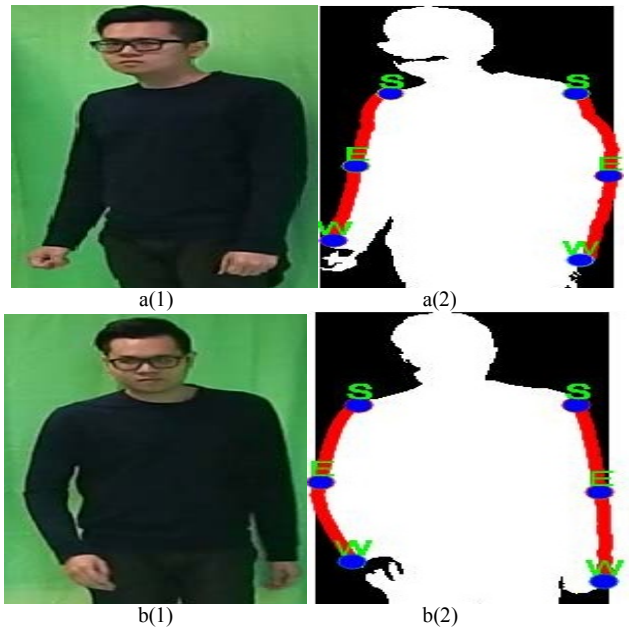


Fig. 9. Output: Pose Estimation

Table II below shows the classification results for the left and right hand based on the images in Figure 8 and Figure 9.

TABLE II. CLASSIFICATION RESULTS FOR LEFT AND RIGHT HAND

Image	Ground Truth		K-means Classification Class		Accuracy
	Left	Right	Left	Right	
8(a)	HSS	HNS	1	2	Accurate
8(b)	HSS	HSS	1	2	Partially
8(c)	HSS	HNS	1	1	Partially
9(a)	HNS	HNS	2	2	Accurate
9(b)	HNS	HNS	1	1	Not Accurate

***HSS – Hand Stretch**
***HNS – Hand Not Stretch**

The classification result is compared with the ground truth images. The accuracy for each image is shown as (1) accurate, (2) partially accurate, or (3) not accurate. The classification is accurate when both the left and right hand is classified correctly. Partially accurate is for when either left or right hand is correcly classify to the right class. Finally, if both left and right hand are not classified correctly to the right class, then it is consider not accurate. Based on our result in Table 2, there are 2 images accurately classify for both hand and 2 are partially accurate and only 1 not accurate. If we consider each side separately, it is found that 80% of the left hand action are correctly classified while for right hand there are only 40% correctly classified. In overall, 60% of both left hand and right hand are classified to the correct class and 40% is not correctly classify.

IV. CONCLUSION AND FUTURE WORKS

This paper has proposed a framework for upper body action classification based on the hand action. Here we have proposed a simple pose estimation algorithm to find the 3 joints in hand to segment the hand to upper segment and lower segment. The gradient of the 2 segments are used in the classification for the hand using k-means. The results shown most of the images are correctly classified. Even though the result shown here is quite promising, but we still need to do the experiment with more images that have different hand actions to get a more accurate result. Besides that, the pose estimation can be more accurate if we can use depth data from the image which can help to overcome occlusion in the image

ACKNOWLEDGMENT

This work is supported by the UTAR Research Fund Project No. IPSR/RMC/UTARRF/2017-C1/L01 “A Study on Recognizing Human Action in a Multi-View Controlled Environment Using Depth Map” from the Universiti Tunku Abdul Rahman, Malaysia.

REFERENCES

[1] G. Guo and A. Lai, “A survey on still image based human action recognition,” Pattern Recognition, vol. 47, no. 10, pp. 3343-3361, 2014.

[2] J.K. Aggarwal and L. Xia, “Human activity recognition from 3D data: a review,” Pattern Recognition Letters, vol. 48, pp. 70-80, 2014.

[3] L. Lo Presti and M. La Cascia, “3D skeleton-based human action classification: a survey,” Pattern Recognition, vol. 53, pp. 130-147, 2016.

[4] SC Lai, PY Lau and MC Lim, “A framework for multiview synthesis for action labeling,” 2017 Proc. of International Workshop on Advanced Image Technology (IWAIT 2017), Penang, Malaysia, pp. , 8-10 January 2017.

[5] Y. Wu, H. Chen, W. Tsai, S. Lee and J. Yu, “Human action recognition based on layered-HMM,” 2008 IEEE International Conference on Multimedia and Expo, Hanover, pp. 1453-1456, 2008.