Motion Capturing

Koushik Chowdhury
2572865
Motion Synthesis for Virtual Characters Seminar
German Research Centre for Artificial Intelligence
Saarbrcken, Saarland.

ARTICLE HISTORY
Compiled July 31, 2019

ABSTRACT
Motion capture is the technique of recording human movements with the help of a camera and finally mapping these movements into a character model. Sensing, digitalizing and recording are the main steps for the motion capture technique. Motion capture is used in so many places such as in sports, film, medical application, etc. This report describes the high-level idea, summarization, connection and personal opinion of 3 states of art motion capturing research papers. The first paper illustrates the survey finding of depth and inertial sensor for human action recognition (*1*). A large number of review survey research papers were published that are based on either vision sensors or inertial sensors. The researcher of this paper published a survey paper which investigates some recent works where both vision and inertial sensors have been used simultaneously to perform the human action recognition more effectively. The other 2 paper demonstrates the machine learning technique on vision or sensor data to predict human action recognition where one of the paper is PhD paper.

KEYWORDS
Vision sensor; inertial sensor; survey; dataset; machine learning;

1. Introduction

Motion capture was first introduced in the late 70s and at that time, it was used for entertainment purposes. Star Wars: Episode I- The Phantom Menace was the first film to introduce a character developed using a motion capture technique (*2*). The name of the character is Jar Jar Blinks. The film was released in 1999. There are so many animated films that are nominated by the OSCAR committee used the motion-capture technique. Films such as Monster House (2006), Happy Feet (2006), etc (*3*). There are three types of Motion capture technology such as optical, electromagnetic and mechanical. Optical motion capture requires a reflective sensor and a motion capture suit. Electromagnetic motion capture requires several magnetic receivers in which transmit and calculate the movement of the subject and in mechanical motion capture, the subject wears an exoskeleton suit to track the subject's movement (*4*). The main objective of this report is to summarize the survey paper. The second objective of this report is to illustrate 2 recent works based on the vision or inertial datasets. The final objective of this report is to give a personal opinion and show pros-pons. At

the end, this report discusses the future works based on relevant research.

2. Main Approaches: Survey Paper

Human action recognition is a process of detecting and analyzing human actions from the data that are gathered from sensors. There are mainly two approaches in human action recognition. They are vision and inertial based action recognition. The vision action recognition is based on the action performed in the videos. The challenges of vision-based action recognition are moving cameras, viewpoint variations, human shape, etc. RGB cameras and depth cameras are examples of a vision sensor. RGB camera creates orthomosaic maps that show the entire scenario at once (*1*). Depth camera provides real-time 3D data which are cost-effective and also, it gives the highest performance in the human action recognition. The main challenges of depth sensor are camera position, subject variation and occlusion. Microsofts Kinect is one of the examples for the depth sensor which consists of the color camera, infrared emitter, tilt monitor, microphone, LED light and infrared depth sensor (*1*). The Survey paper

Figure 1.: Microsofts Kinect (*1*).

mentions some previous works on depth sensors such as projected depth map, space-time occupancy, etc (*1*). In this work, the depth sensor has been used for human action recognition.

In the case of inertial based action recognition, the subject wears the inertial sensor in via shoes, wristwatches, clothes, mobile, etc. It provides 3D Data. Xsens motion tracking is one of the examples for the inertial sensor. The following picture (a) tells the skeleton joints and skeleton frame in 3D. The second picture (b) is an example of an inertial sensor that is developed in ESP Laboratory, USA. The survey paper

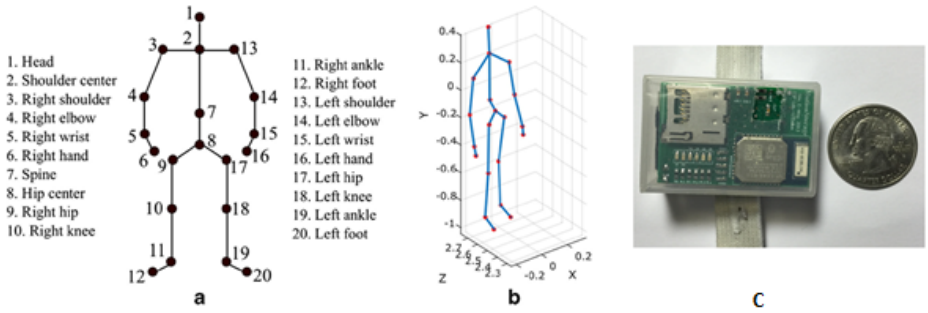

Figure 2.: (a) Skeleton joints, (b) skeleton frame in 3D, (c) inertial sensor (*1*).

mentions some previous work based on inertial sensors such as fall detection system, hand-twist and open hand action (*1*). The main advantage of the inertial sensor is that it can work in a dark environment and also, it provides a high sampling rate as

well as it can capture the 3D pose estimation in the wild. To capture the overall body movement, multiple sensors are required. On the other hand, Depth is not sensitive to color change and also, it does not provide color information. So, both inertial and depth sensor has some individual advantages and disadvantages. It is possible to overcome this limitation with the combination of both sensors because the combined sensor can improve each others limitation to get better recognition performance. This can be done by data synchronization and preprocessing, action segmentation, feature extraction and classification and fusion. The survey paper describes the datasets that are collected with both vision and inertial sensor (*1*). The following table describes the human action datasets involving both vision and inertial sensors.

Dataset	Modality					# Sub	# Act	# Seq
	M	V	D	A	I			
Berkeley MHAD [47]	1	12	2	4	6	12	11	660
URFD [36]	–	2	2	–	1	5	>5	70
UTD-MHAD [12]	–	1	1	–	1	8	27	861
50 salads [58]	–	1	1	–	7	25	17	966
ChAirGest [52]	–	1	1	–	4	10	10	1200
TST Fall detection database [27]	–	–	1	–	2	11	8	264
Huawei/3DLife dataset [59]	–	5	5	5	8	17	22	3740

Figure 3.: List of data-sets based on different modality (*1*).

3. Other Datasets

There are other two interesting motion capturing datasets which are not mentioned in the survey paper. They are SFU (*5*) and MOSH (*6*) dataset. SFU stands for Simon Fraser University and MOSH stands for motion and shape capture database. There are various categories in the SFU database such as locomotion which has subcategories like jumping, rolling, running, etc. Another category is interaction and obstacles which have subcategories like jump and roll, and jumping over an obstacle, etc. It has also other categories like dancing, martial arts. In MOSH, the motion and shape collect from sparse markers.

4. Pros-Cons and Comparision with Recent Works

It is true that data coming from combined sensors provide better action recognition performance. So, when we do some motion capture or human action recognition, it is good to use different sensors for good performance. The main shortcoming of this survey paper (*1*) is that the researcher presents various types of motion capture data but the researcher did not mention the importance of these datasets such 'why they were collected?' and 'where these datasets were used?' There are no recent works on these data sets mentioned in the survey paper. Also, the authors did not go through any experiment that can tell that multimodal data can provide better action recognition performance. In this section, I go through some recent relevant works and try to compare the results of these relevant works to find out the ultimate goal of the survey

paper. The ultimate goal of the survey paper was that 'the combination of vision and inertial senor data improve the accuracy of human action recognition'. They represent 7 multimodal human action datasets (see figure 3). From figure 3, we can notice that only Berkeley multimodal human action database has all features. The Berkeley MHAD data was made as a part of the National Science Foundation project (*7*). Recently, a skeleton-based action recognition approach was published (*8*) using deep learning and the Berkeley MHAD dataset as a benchmark. They have applied the convolutional neural network technique on the Berkeley MHAD to build an end to end hierarchical framework for skeleton-based action recognition (*8*). They compare their proposed method with 5 previous methods which are based on Berkeley MHAD dataset (see figure 4). There is also another research paper on machine learning which is based on vision-based data. The paper was submitted for PhD thesis (*9*). The paper predicts human action recognition from video data. The authors of this paper have applied both deep learning and non-deep learning approaches. A deep belief network trained on the vision-based dataset was compared with 7 previous approaches. Both papers used different datasets and different methods. There is a research on hand gesture recognition in which the authors applied the same method to depth, inertial and fusion-based datasets (*10*). The authors worked with two models called the Hidden Markov Model (HMM) and Dynamic Time Wrapping (DTW). The authors applied these two methods to depth, inertial, and fusion datasets. The following table shows the results.

Table 1.: Results of two methods (*10*).

Method	Kinect	Inertial	Fusion
DTW	69	71	80
HMM	84	88	93

*Average recognition Rates (%)

Kinect is one of the examples of a depth sensor. From the table, we can see that data coming from fusion provide the highest average recognition rates for both HMM and DTW technique. For the HMM method, fusion-based dataset provides 93% average recognition rate which is higher than other average rates. DTW and HMM measure the lowest average recognition rates for Kinect. After comparing this result, it is clearly shown that the fusion of vision and inertial data improves human action recognition.

5. Conclusion and Future Work

Deep learning is a very common approach for human action recognition. There are several human action recognition works based on deep learning. There are deep learning based human action recognition works using either vision sensors or inertial sensors or a combination of both sensors are used. In this report, I have tried to show some comparisons between depth, vision and fusion of vision and inertial sensors data. To do that I go through a previous work on hand gesture recognition. In that work, HMM and DTW measure the highest average action recognition rates for fusion datasets but the main limitation of work is that HMM and DTW are the very basic model in statistical learning. Therefore, the future work is that the same deep learning algorithm should be applied in vision or inertial and fusion data. Then, the comparison result would be more justified and there would be no obligation to say that the fusion data

improves the accuracy of action recognition than individual sensor data.

References

(1) Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. *A survey of depth and inertial sensor fusion for human action recognition*; Multimedia Tools and Applications 76.3 (2017): 4405-4425.

(2) Wikipedia. *https://en.wikipedia.org/wiki/Motion_capture#Applications*

(3) Wikipedia. *https://en.wikipedia.org/wiki/Academy_Award_for_Best_Animated_Feature*

(4) Slideshare. *https://www.slideshare.net/murlidharsarda/motion-capturing-technology*

(5) SFU Motion Capture Database.*http://mocap.cs.sfu.ca/*

(6) MOSH Database. *http://mosh.is.tue.mpg.de/*

(7) Ofli, Ferda, et al. *Berkeley mhad: A comprehensive multimodal human action database*; 2013 IEEE Workshop on Applications of Computer Vision (WACV). IEEE, 2013

(8) Du, Yong, Yun Fu, and Liang Wang. *Skeleton based action recognition with convolutional neural network*; 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015.

(9) Bux, Allah. *Vision-based human action recognition using machine learning techniques*; Diss. Lancaster University, 2017.

(10) Liu, Kui, et al. *Fusion of inertial and depth sensor data for robust hand gesture recognition.*; IEEE Sensors Journal 14.6 (2014): 1898-1903.