Human Action Recognition through the First-Person Point of view, Case Study Two Basic Task

Mohammad Almasi
Dept. of Advanced Mathematics and Mathematical Engineering
Polytechnic University of Catalonia
Barcelona, Spain

Hamed Fathi
Dept. of Civil Engineering
Islamic Azad University Branch of Karaj
Karaj, Iran

Sayed Adel Ghaeinian
Dept. of Computer engineering and Information technology
Amir Kabir University of Technology
Tehran, Iran

Samaneh Samiee
Dept. of Computer engineering and Information technology
Amir Kabir University of Technology
Tehran, Iran

ABSTRACT
In this study, a human motion dataset is built and developed based on indoors and outdoors actions through a bounded-on-head camera and Xsens for tracking the motions. The key point here to structuring the dataset is utilized to set the sequence of a Deep Neural Network and order an arrangement of frames in the performed task (washing, eating, etc.). As a final point, a 3D modeling of the person suggested at every frame centered with the comparable structure of the first network. More than 120,000 frames constructed the dataset, taken from 7 different people, each one acting out different tasks in diverse indoor and outdoor scenarios. The sequences of every video frame were 3D synchronized and segmented 23 parts.

Keywords
Machine learning; deep learning; Computer vision; LSTM; Recurrent neural network; ResNet; motion recognition.

1. INTRODUCTION
Wearable cameras are increasingly popular for both entertaining and scientific point of view. Many fields, such as biomedical and conventional engineering utilized in two main features, i.e., footage approving and body pose analyzing of the camera wearer. In Sports and running for estimating the path and velocity [2]. This complimentary issue can make benefit in different desired applications, for instance for at-home-observing of rehabilitating patients by doctors from the hospital, for defining the items by reinforcement application in the robotic science, reduce to bare bones of simulated learning established from the audiovisual inputs to an operative organization by assessing the different operator placement. Three methods are used to envisage human pose with an egocentric video: calculation of third-person's pose, investigation of egocentric video, and appraisal of the first-person body pose from the video. It is thought-provoking to supposing the first-person pose for the body from an egocentric video; meanwhile, the physique of the person is not often on screen. Information is mutually extracted from the camera motion and utilized toward building a regression model for assessing the 3D-Pose [11] and merging that specific pose by using recurrent neural networks. The camera motions are taken from the video and the scene itself, using CNN's- based feature extractors, taking into account the preceding frames and poses. Recently, a new approach [6, 18] is reported to prophesy the 3D-pose through the egocentric video. In this approach, the imitation method of learning is used for understanding a policy controller for the extrapolation of a pose. Despite the clean results, it put the results in interactions of the person and other persons in the sight of the camera. It should be noted that low content of information would be achieved in the case of utilizing videos of no interaction with more people. Therefore, the joints of the person were watched in the frame, in the present study. Then, there is no need for person to person interactions, and expedient information is obtained from the total pose. In recent years, first-person videos have gained widespread admiration due to intelligent devices with different insight in people undertakings than the mutual third-person videos. It should be considered that the number of egocentric video datasets in first-person videos is noticeably smaller than third-person ones, but it's growing steadily. In the set of accessible datasets of egocentric videos, two types are considered based on object interaction tasks and person interaction tasks [1, 5, 8, and 16]. In recent years, assistive robots and humanoid robots are getting lost of scientist's attention that the reason is the numerous services that they provide. Since most of these robots are reaction-based, they have to realize the action of [the person whom they serve/ their partner] correctly and precisely. Various vision-based human action recognition algorithms (HAR) have been proposed so far. These approaches can be categorized into holistic and local. Holistic approaches are those who attend to extract features from human joint angles after background subtraction. The latter approach utilizes interest points (IP) to do its job. Despite holistic approaches, local approaches neither need background subtraction nor body tracking; thus they are more reliable and efficient [7].

Computer scientists proposed some deep neural network methods to reduce deficiencies (E.g., high dataset dependency) while using those approaches aforesaid in the previous paragraph [3, 4]. The correlational Convolutional LSTM unit has been proposed in reference [13]. This unit operates (with a combination of convolution and cross-correlation) for regarding spatial and motion features, while it constructs temporal dependencies too. For some actions such as washing, eating, it is favorable to predict the starting and ending point of the activity for a more significant reaction (E.g., passing a spoon before eating). Also, in [12], a method applied for joint classification and some regression optimization with the aid of the LSTM network to deal with the action forecasting dilemma mentioned earlier.

Kinect cameras are vastly used for recording the human pose due to their affordability [11]. In Panoptic Studio, the dome is also used for recording the utilized sequences. Passive or active

markers of visual approaches are also meant to capture human motions. The common feature of various mentioned datasets is using the chest-mounted camera as an alternative of a head-mounted one. The disadvantage of these cameras is that despite the right images, they are not well meet the reality, as the massive content of egocentric videos is recorded using head-mounted wearable cameras.

The recent report [19] is used the Convolutional Neural Networks (CNN) to gain the 3D-poses from images via a direct regression approach. CNN is reported in other studies to subtract buoyancy maps in the image for each joint. It is indicated that the possibility of a particular joint discovery in the image areas, can be presented. In [3], this approach is combined with part affinity fields (PAF's). The ways of connections of joints with each other are considered to improve outcomes. However, PAF's are not desirable for this study, since only a small set of joints will be hardly realized on the screen. Besides, a certain joint that is revealed on screen can help confidence maps to be detectable. It would offer valuable data about the imperceptible pose of the camera wearer. Video analysis is approached using many Deep Neural Networks (DNN) infrequent ways, like joining CNN with using Recurrent Neural Networks or dealing out stacks of video frames with 3D CNN.

In this study, it is aimed to prepare a human motion dataset using a Motion capture system and a GoPro for human motion recognition. The practicability of the prepared database has been evaluated and used to sequence a neural network and organize the sequences of egocentric video frames. In the plausibility of preparing such a dataset.

2. METHODOLOGY

Here, the 3D coordinates of all the joints of the subject were developed throughout a sequence of activities and images of the same arrangements using video-camera. In the condition of the presence of a third person and static camera, the PnP algorithm could be functional to a particular frame. A visible frame is a set of the subject joints, and the arrangements of these joints could be gotten by physical selection in the image or using a Deep Learning approach as it is in [3]. Then and there, the camera alignment and location would not modify through the sequence in the 3D world organize the system. The results obtained from adopting the EPnP for a single frame could be practical to all the video.

The egocentric camera used here is not constant in terms of its position and orientation expressed in the world coordinate system because it transports with the head of the person. Fortunately, it can be considered that the Motion capture system affords the alignment and the location of the head part. The head-mounted camera also has constant related to the pose of the camera to the head. Consequently, a local coordinate system can be well-defined up on the alignment and location of the head joint. Then, the 3D points provided by the Motion capture system would be articulated in the head local coordinate system and the applied EPnP algorithm. The obtained matrices could be used all over the video.

The EPnP algorithm has been used to obtain the data, and the Motion capture system has been recorded using a GoPro camera to get hold of the camera pose during the sequence. The camera is adjusted for achieving the intrinsic matrix and the coefficients of distortion. Total frames were then disseminated from the video by the distortion coefficients. Furthermore, the camera pose is gotten through a PnP algorithm on a single or a minor set of frames. The selected joints were visible on the frame, and the projection action is then performed.

Before prognostic of obtained data onto egocentric videos, they were sequenced with the camera in an immobile pose while during this projection, a third-person point of view was utilized. In these sequences, there is no need for points to be expressed in the local direct system of the head because the pose of the camera is fixed in the 3D world coordinate system. The third person sequences provide a complete 3D view of the frame projected onto the subject, as long as more information is recorded about the accuracy of the joints.

The measurements were used altogether from the subject that can be presented in the Motion capture system Analyze software, to gain the projections. The lengths of all the segments were designed using measurements from the legs and the length of the arm, in the range of the shoulder to the wrist, however the leaves indeterminate the precise length of the upper arm and the forearm. Likewise, the most precarious fragment of the calibration of the Motion capture system is in the hands, so it's tranquil to have some ignored calibration error.

The two former elements illuminate most of the error perceived during the projections of the data gotten from the Motion capture system. More error in the projections has occurred on the wrists, which often objected to some displacement. Owing to this movement, the 2D- 3D communications were done on the wrists while the EPnP observed some error and led to a prominent, slightly improper outcome. Then, the Motion capture system has propagated the error to the EPnP algorithm. During the projection of sequences recorded in the third person, this error is typically understood, is where the minimal errors of the camera position were heightened.

Using seven persons with approximately similar age, who were asked to perform various tasks, the data is collected. It should be considered that for preparing the dataset, working with a healthy adult's applicant is the necessity to record accurate data.

The 3D-pose of the participants and the egocentric video images were apprehended independently. The motion capture system obtains the 3D-pose during all the sequences. It is required to use a portable wireless camera due to the first person point of view as the recorded video. The methodology used is moderately manual. A neural system is prepared with the dataset recently made to check its helpfulness. The system needed to group the movement performed in a chronicle dependent on the arrangement of frames of an egocentric video. For this issue, highlights were separated from every frame and its past ones and were utilized as a contribution for an LSTM (Long Short Term Memory). The general design depended on the one utilized in, however, without evaluating the posture of a subsequent individual associating with the camera wearer. Alternatively, the situation of the hands and feet of the person is looked at, which gave critical data about the human posture.

The system engineering utilized in this task which is depended on the one used in the approach. Previously, a lot of highlights were extricated from the casings that work as "pieces of information" for the LSTM. The active highlights were extricated, figuring a lot of homographies from the 15 past frames and the present one that gave data about the latest human movement. As observed in [11], various dislocations of the person produce diverse movement designs in the video. These movement examples are acquired by discovering point correspondences between casings. A while later, these movement examples are communicated as a grouping of homographies between progressive casings.

The Lucas-Kanade optical flow method is utilized for discovering the point correspondence between frames [15]. The focuses in frame f_t were at first chosen, and at that point, their

correspondences were found at frame f_{t+1}. These correspondences were utilized for figuring out the following correspondences at casing f_{t+2}, as beginning focuses intending to spare processing time. Particular focuses may move out of the edge as the grouping is handled; if the quantity of origin focuses is not precisely a discretionary limit higher than 4 (the insignificant number of focuses to process a homography), at that point, the initials focuses were figured utilizing. For this study, the frame is set to 6. In this exploration, 15 homographies were processed between 16 back to back frames (the present one and the past 15). Each homography is standardized with the upper left corner. At that point, all the homographies were vectorized and joined into a singular vector $m_t \in R^{135}$.

The scene appearance can give applicable data of the assignment that is being executed. To remove these static highlights, a ResNet-18 pre-prepared on ImageNet is utilized [9, 14]. The center thought behind ResNet is presenting a personality alternate way association that skirts at least one layer. At that point, while back-propagating, the inclinations can course through the additional way associations with the underlying stages. The last completely associated layer of the pre-prepared ResNet-18 is dropped, utilizing the normal pool one as a yield to get a vector $s_t \in R^{512}$.

3. RESULTS AND DISCUSSION

A pre-designed variant of the model made in [3] is utilized to search for joints of the camera wearer. This CNN model uses two branches to foresee the joint places of the pictures: one predicts a lot of 2D certainty maps (one for each joint), and one predicts a lot of 2D-vector fields of part affinities. The Part Affinity Fields (PAF's) encrypt the place and direction of the appendages in the picture, which enables the model to improve the expectations for the body joints' position. The aftereffects of these two branches are then connected and sent for the

Accompanying stage. For this study, the PAF's were not helpful, as the appendages of the camera were out of view more often than not, and no affiliation could be made. Consequently, the branch that forecasts 2D certainty maps for the essential step is utilized. At that point, the situation of each joint is resolved from the certainty maps by finding a top in every certainty map, just as its likelihood. There is no compelling reason for utilizing a lot of scales, and a solitary worth (0.2) is used, as the joints that ordinarily can be found in an egocentric perspective (hands and feet) are in a similar range as far as the size on the screen. As just hands and feet are seen some of the time on the casing, every other joint was sifted through to wipe out any false-positive. In this way, four points in the frame were searched for, the two wrists and ankles. If a joint is missed, u and v directions were equivalent to zero. Something else, by the width and its tallness, individually, u and v directions of the joint in the picture were standardized. At that point, these four 2D vectors were linked in a singular vector $b_t \in R^8$ in the accompanying request: right wrist, left wrist, right lower leg, and left lower leg. The highlights removed from the present frame were at long last linked in an element vector $f_t \in R^{512+135+8}$.

This vector filled in as a contribution for an LSTM network, a sort of repetitive neural system (RNN). Conventional neural systems can't process serial data [17], and RNNs deal with this issue. Essential RNNs work fine when ongoing data from past sources of info is needed; however, in situations where more setting is required, RNN miss the mark in memory [10]. At the point when more memory is required, Long Short Term Memory systems (LSTM's) can be utilized, as they don't present this issue. The cell state is the central idea behind LSTM's, which fills in as a transport line going through the LSTM. The phone state conveys the data from the past information; it is deliberately changed by structures called entryways and go to the accompanying state. The cell state does not substitute the concealed express that could be found in RNN; the shrouded state is additionally present in LSTM's. The members were used for the dataset isolation that is every one of the groupings of five members for preparing and two of them for testing. At any blend utilized, it is guaranteed that all errands were available in the preparation and the testing set. During preparing, 85% of the preparation set is being used for training and 15% for validation. For training the model, a clump size of 64 is utilized. The LSTM is characterized by a fixed installing measurement of two and a concealed state measurement of 16. The model is prepared for 11 epoch's all together, with a learning rate equivalent to 0.0001 for the initial eight epochs, and after that, it is diminished to 0.00001.

The loss for the system is the cross-entropy loss over the prepared succession for foreseeing the best possible errand. The dataset could not be contained an equivalent number of frames per task. At the point when this occurs, it's reasonable to weight the loss for every class. As found in Equation 6, the weight for every classification is determined, where *total* represents all information in the dataset, and n_i represents the number of events in the dataset of the classification i.

The last model prepared had a 92.58% of accuracy in the training set, 93.65% in the validation set, and 61.10% in the testing set. The accuracy of the model during the test for every classification is exceeding the expectations in characterizing in washing arrangements (68% of them) and expected results with the eating (7%), as can be observed in figure one.

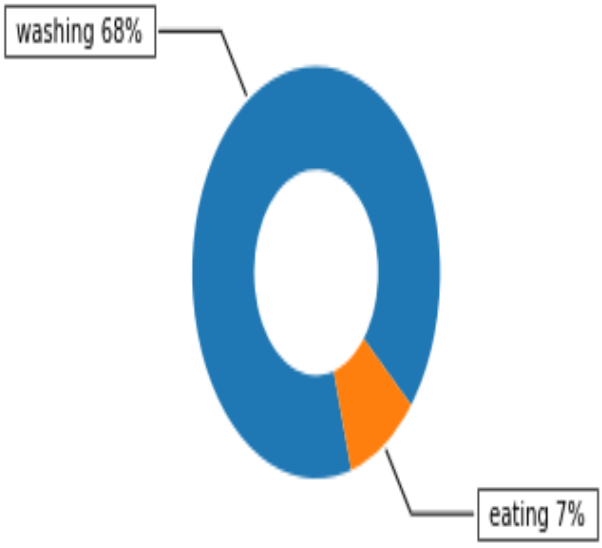

Fig1. Demonstration of two tasks with their correctness percent.

The confusion matrix of the forecasts during the test has appeared in Table 1. Each row is assigned to the anticipated yield of the model and every section to the ground truth. In a perfect world, it is required that every count is put in the askew of the lattice, and the remainder of the components outside the slanting is equivalent to zero. The model is by all accounts distinguishing the extensive majority of the running groupings as walking ones. While it is misclassifying many running models, it's misconception them with assignments of comparable dynamism, which relates to taking after camera movement designs.

Table I. Confusion matrix

Output Class	**19**	**41**	31.7%
	1.3%	2.9%	68.3%
	194	**175**	47.4%
	13.7%	12.4%	52.6%
	7.0%	68.1%	**61.1%**
	93.0%	31.9%	**38.9%**
	Target Class		

The graphs of loss and accuracy during train and validation can be found in Figure 2. The last outcomes are moderately comparable. This model had a 93.97% of exactness in the preparation set, 97.20% in the approval set, and 68.36% in the testing set. The model predicts preferable outcomes while testing over the principal model (61.10%).

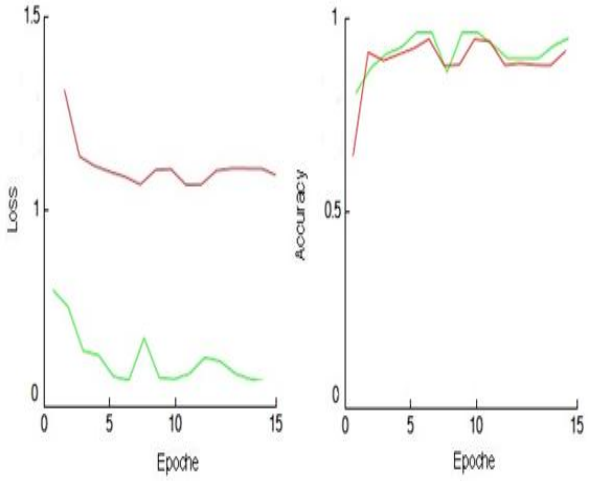

Fig2. Loss and accuracy representation during the test (green) and train (red)

More than 120,000 frames make a total of the dataset with its parallel human poses. Like what is done here, a lot of poses in a dataset can be measured to a narrower set of postures. For doing this, all the poses from the dataset were grouped using K-means. For determining the number of groups K, various values in the range K [1, 500] were examined. For choosing the optimal, the mean distance from all the poses to the centroid of their parallel group is estimated. Since all postures were adjusted in scale and orientation, the distance considered as a percentage of the width of shoulders.

In the Current study, K = 500 is used. The average error of the shoulders width is equal to 3.6%. Five examples out of the 500 clusters computed are indicated in figure 3. In each plot, not only both the centroid (in red), but also some of the random postures are overlapped, which belong to the cluster of that centroid (in gray). Hence, it can be regarded that the poses in the same group were partly similar to each other.

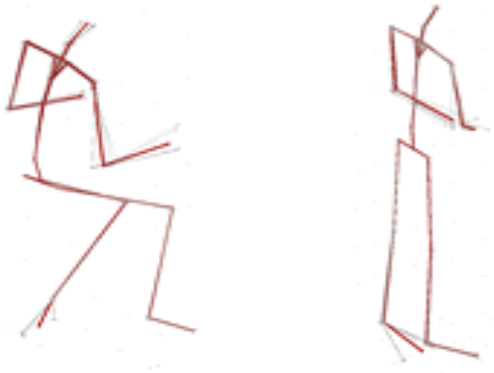

Fig3. The results of two tasks made by the motion capture system

4. CONCLUSION

The current study has indicated that the Motion capture system, which is based on inertial and wireless sensors, is adequate good enough to make a dataset of human motion in diverse environments. It makes the creation of the datasets with motion data recorded outdoors and in demanding environments possible, where there is not any effective performance for visual-based movement capture systems. As a result, more original and relevant information would be achievable. It comes right at the cost of unhanding a minor part of segment length accuracy. Notably, the network developed and trained for classification of the task has indicated that a dataset made by Motion capture system and a GoPro has the potential of utilizing for Deep Learning purposes adequately. A larger and more balanced dataset built using the same equipment is expected in later works to train a similar model with better outcomes. Moreover, there is the possibility of preparing a network to measure the person 3D-pose implementing the same model to the one suggested in the current project successfully.

5. REFERENCES

[1] Alletto, S., Serra, G., Calderara, S., & Cucchiara, R. (2015). Understanding social relationships in egocentric vision. *Pattern Recognition*, *48*(12), 4082-4096. DOI:10.1016/j.patcog.2015.06.006

[2] Almasi, M. (2018). Investigating the Effect of Head Movement during Running and Its Results in Record Time Using Computer Vision. *International Journal of Applied Engineering Research*, *13*(11), 9433-9436

[3] Cao, Z., Simon, T., Wei, S., & Sheikh, Y. (2017). Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI:10.1109/cvpr.2017.143

[4] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, & Wanqing Li. (2017). Skeleton-based action recognition using LSTM and CNN. *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. DOI:10.1109/icmew.2017.8026287

[5] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E, Wray, M. (2018). Scaling Egocentric Vision: The Dataset *Computer Vision – ECCV 2018*, 753-771.doi:10.1007/978-3-030-01225-0_44

[6] Ekvall, S., & Kragic, D. (2006). Learning Task Models from Multiple Human Demonstrations. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*. DOI:10.1109/roman.2006.314460

[7] El-Yacoubi, M. A., He, H., Roualdes, F., Selmi, M., Hariz, M., & Gillet, F. (2015). Vision-based Recognition of Activities by a Humanoid Robot. *International Journal of Advanced Robotic Systems*, 1. DOI: 10.5772/61819

[8] Fathi, A., Li, Y., & Rehg, J. M. (2012). Learning to Recognize Daily Actions Using Gaze. *Computer Vision – ECCV 2012*, 314-327. DOI: 10.1007/978-3-642-33718-5_23

[9] Hara, K., Kataoka, H., & Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI:10.1109/cvpr.2018.00685

[10] Hochreiter, S. & Schmidhuber, Jü. (1997). Long short-term memory. *Neural computation*, 9, 1735--1780

[11] Jiang, H., & Grauman, K. (2017). Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI:10.1109/cvpr.2017.373

[12] Li Y., Lan, C., Xing, J., Zeng, W., Yuan, C., & Liu, J. (2016). Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks. *Computer Vision – ECCV 2016*, 203-220. DOI: 10.1007/978-3-319-46478-7_13

[13] Majd, M., & Safabakhsh, R. (2019). Correlational Convolutional LSTM for human action recognition. *Neurocomputing*. DOI:10.1016/j.neucom.2018.10.095

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[15] Patel, D., & Upadhyay, S. (2013). Optical Flow Measurement using Lucas Kanade Method. *International Journal of Computer Applications*, *61*(10), 6-10. DOI: 10.5120/9962-4611

[16] Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. DOI:10.1109/cvpr.2012.6248010

[17] Squartini, S., Hussain, A., & Piazza, F. (n.d.). Preprocessing based solution for the vanishing gradient problem in recurrent neural networks. *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03*. DOI:10.1109/iscas.2003.1206412

[18] Yuan, Y., & Kitani, K. (2018). 3D Ego-Pose Estimation via Imitation Learning. *Computer Vision – ECCV 2018*, 763-778. DOI: 10.1007/978-3-030-01270-0_45

[19] Zhu, L., & Wan, W. (2018). Human Pose Estimation Based on Deep Neural Network. *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. DOI:10.1109/icalip.2018.8455245