Human Fall-down Event Detection Based on 2D Skeletons and Deep Learning Approach

Wen-Nung Lie
National Chung Cheng University
Department of Electrical Engineering
Chia-Yi, Taiwan, ROC.
ieewnl@ccu.edu.tw

Anh Tu Le
Ho Chi Minh City University of Technology
Faculty of Electrical & Electronics Engineering
Vietnam
eng.leanhtu@gmail.com

Guan-Han Lin
National Chung Cheng University
Department of Electrical Engineering
Chia-Yi, Taiwan, ROC.
linshing00@gmail.com

***Abstract* - The goal of this research is to apply the state-of-the-art deep learning approach to human fall-down event detection based on 2D skeletons extracted from RGB sequence. In this paper, we adopt convolutional neural network (CNN) to extract humans' 2D skeletons for each input frame and then employ recurrent neural network (RNN) with Long Short Term Memory (LSTM) state cells to process temporal skeleton series to make best use of not only spatial features but also temporal information to classify each short-term action to five categories, i.e., standing, walking, falling, lying, and rising. After simple rule processing, the consecutive RNN outputs can be used to detect human's long-term actions (falling down event) and determine whether to issue an alarm or not. The accuracy of classification into 5 sub-actions is capable of achieving 90%. Our contributions lie on two aspects: (1) improving the performance on short-term human action recognition based on the combination of CNN and RNN/LSTM, (2) excluding the fall-down events that actually need no help and achieving a lower false alarm rate.**

Keywords – Fall-down event detection, human action recognition, deep learning, human skeleton.

I. INTRODUCTION

Nowadays, action recognition plays an important role in computer vision and has a wide range of applications such as human-computer interaction, video surveillance, robotics, game control, and so forth. For the most part, human body can be regarded as an articulated system with rigid bones and hinged joints so that human actions can be represented as the motions of the skeleton.

Human fall-down event detection is important in homecare technology, especially for aged persons. Though some cost-effective depth sensors, like Microsoft Kinect, combining with real-time "*3D skeleton*" estimation algorithms, are available to provide relatively reliable joint coordinates for posture/action recognition, extraction of "*2D or 3D skeleton*" from a single RGB sensor still has its advantage of lower cost, but is more challenging on the processing speed and recognition performance. Most of the existing skeleton-based action recognition approaches model actions based on well-designed hand-crafted local features. Recently, an end-to-end approach based on deep learning was proposed for action recognition. This releases us from designing sophisticated image features.

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

II. RELATED WORKS

In recent years, we have witnessed the achievements of deep learning in image and video processing, classification, image captioning, semantic segmentation, object detection, etc., especially, recognizing person's actions from the video [1, 2]. Some researches focus on identifying human actions using the depth values derived from the Kinect camera [3]. There are several methods determining human actions after joint skeleton mapping from depth maps, such as the Hidden Markov Model algorithm in [4]. Thanks to the Kinect camera, a 3D human skeleton map can be obtained, from which human actions can be identified easier.

Nevertheless, the cost of the Kinect camera is more expensive than traditional RGB color camera. Therefore, some researchers try to derive human's 2D or 3D skeletons from simple RGB images. Deriving the skeleton can be done in several ways. To avoid designing hand-crafted feature extraction sophistically, some recent works applied the neural network to extract the maps of human skeleton as in [5,6,7]. By using convolutional neural networks, the accuracy of the identification can be improved in case of the presence of image noises or distortions.

Du et. al. [8] did a research with convolutional neural network (CNN) to extract the skeleton joint map from images. However, the result can be improved if recurrent neural network (RNN) is applied properly as in [9,10].

III. OUR SYSTEM DESIGN

In this research, we divide the fall-down event detection process into two stages. The first stage is known as the 2D skeleton extraction. As an input, we used 2D RGB images of 1920 x 1080 pixels.

Each frame will go through a pre-trained CNN model built on DeeperCut [9] (an extremely deep, 152 layers, Residual Network (ResNet) [10]) for estimating a 2D skeleton for any person in the image (see Fig.1). The 2D skeleton is composed of 14 2D joints which are actually key points in human body, called "forehead", "chin", and left and right "shoulder", "elbow", "wrist", "hip", "knee", "ankle".

The second stage is to recognize actions from the skeleton motions. We employed an RNN with LSTM (Long Short Term Memory, [11]) as the state cell to capture features from temporal observations (here, 8 skeletons at 8 prior time instants) and provide longer-range context for current action prediction. The output of prediction by our model is a classification of an action into five classes: standing, walking, falling, lying, and rising (see Fig. 2). Based on some simple rules, the short-term action label output at each time instant can be integrated to determine whether to issue an alert for nursing requirement.

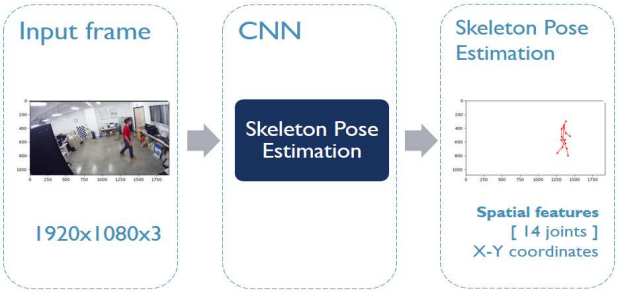

Fig 1. The 1st stage: extracting 2D skeleton from each input RGB frame.

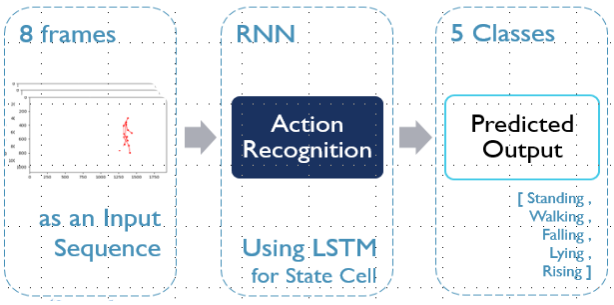

Fig 2. The 2nd stage: classifying short-term action (8 time instants) into 5 classes by an RNN with LSTM.

1. Extracting 2D skeleton by using very deep Convolution Neural Network

In this research, we applied a pre-trained model built on the recent advances in object classification and adopted the extremely deep Residual Network (ResNet) for human body part detection. This pre-trained model, called DeeperCut (152 layers), is capable of achieving excellent results to extract joint skeleton map for humans in the captured image.

It might be that 14 joints calculated from DeeperCut were sometimes affected by background. So we limit the detected human body to a constrained rectangular size to eliminate irrelevant/wrong estimations.

Every frame is input to this model to get 14 joints which are represented as a skeleton. To make LSTM in the next stage more robust, we calculate the mean of the left-hip and right-hip, called central-hip, and refer the mid-point between the chin and central-hip as the centroid of the human skeleton. Coordinates of all the 14 joints are converted from the image coordinates into skeleton-centered coordinates with respect to the centroid. In this way, a translation-invariant skeleton model which will be more suitable to robust action recognition can be achieved.

2. Predicting actions from 2D skeleton motions by using RNN with LSTM

Denote the 28 coordinates (corresponding to 14 joints) at each time instant t as $\mathbf{x}_t$. The input $\mathbf{x}_t$'s are multiplied by a weight **W.** Then an output state $\mathbf{h}_t$ will be memorized by RNN and forwarded to the input of next time for estimating a new output state $\mathbf{h}_{t+1}$. Each output $\mathbf{h}_t$ has actually considered the past input skeletons.

LSTM cell units are built to tackle the vanishing and exploding gradient problem. The input sub-sequence $\mathbf{x}_{t-K} \sim \mathbf{x}_t$ for determining each $\mathbf{h}_t$ is designed to have a fixed length of 8, i.e., K=7. The coefficients **W** are obtained through training to optimize/minimize errors between the outputs and the ground truths. That is, the training datasets will be composed of samples, each contains 8 skeletons (28 coordinates), which are input in parallelism to RNN/LSTM model for classification. For test, the system will collect 8 consecutive skeletons and feed them in parallelism for short-term action prediction.

IV. EXPERIMENTAL RESULTS

For experiments, sub-sequences, each is composed of 8 consecutive input frames, are collected to form samples in the dataset. Each frame has a size of 1920x1080 pixels. The dataset is partitioned into training (800), validation (255), and test (250) sets for 2nd stage RNN/LSTM training. The RNN/CNN model was implemented by using Tensorflow library based on Python programing language. Note that,

we don't need to train 1st stage for skeleton extraction since it is a pre-built model.

The results were obtained at a training of 500 epochs and a learning rate of 0.0001. The average test accuracy achieves 89.9% (for several rounds of experiments, ranging from 87.6% to 92.8%), see Table 1.

Table.1 Evaluation on testing data after training.

Epochs	Learning rate	Test Accuracy
500	0.0001	92.4 %
		88.4 %
		89.2 %
		92.8 %
		88.8 %
		87.6 %
Average Test Accuracy		**88.9 %**

We implement our fall-down event detection system into an on-line or live application version. For on-line systems, two problems will be faced.

First, every frame is processed to extract its 2D skeleton, which, in accompany with the latest 7 ones, is sent to RNN/LSTM model for short-term event prediction/ classification. Wrong skeletons from the pre-trained CNN/ DeeperCut model should be removed to prevent performance degradation of the RNN/LSTM stage. For example, when no or incomplete human body exists in the frame, the system might still output a most probable skeleton as the result. To solve this problem, a rule judging the validity of the skeleton is used before it is sent to the 2nd stage for RNN/LSTM. For example, sum of the distances between each joint and the centroid is calculated, to which an empirical threshold will be applied. With this screening process, incorrect or unreliable skeletons can be removed to increase system performance.

Second, an alarm decision rule is designed to determine whether an alarm should be issued or not. It may be the cases that the LSTM gets wrong classification temporarily, a person bends down to pick up something or fasten his shoelace, or even someone falls and lies down, but rises after a reasonable period. In this case, our rule will determine not to issue an alarm so that a lower false-alarm rate can be achieved. In this rule, we consider a long-term history (the last 20 outputs) for decision. The "standing" and "walking" are considered as safe status, and "falling" and "lying" are considered as danger one. When the current-time output is "lying" and the previous 19 outputs contain more than 14 "danger" statuses, an alarm will be triggered. When the current output belongs to "safe" status and the previous 8 outputs contain over 5 "safe" statuses, then the alarm signal will be reset.

In our on-line implementation, a processing speed of 8 frames/second can be achieved with GTX 1060 6G GPU. Each processed frame can get one output (both for 5-class classification and alarming decision).

V. CONCLUSION

In this paper, we successfully apply deep learning algorithms to resolve fall-down event detection problem. The problem is decomposed into: 1) 2D human skeleton estimation and action recognition based on temporal skeleton motions. We adopt CNN to extract 2D skeleton pose from a single RGB for each person in the scene, and simultaneously discard possible background clutter noise. In addition, based on the RNN with LSTN state cells, our system is able to process the dynamic sequence and predict short-term actions effectively with 5 classes. Based on these 5-class classification, we are thus able to design an on-line alarm system that achieves a lower false-alarm rate.

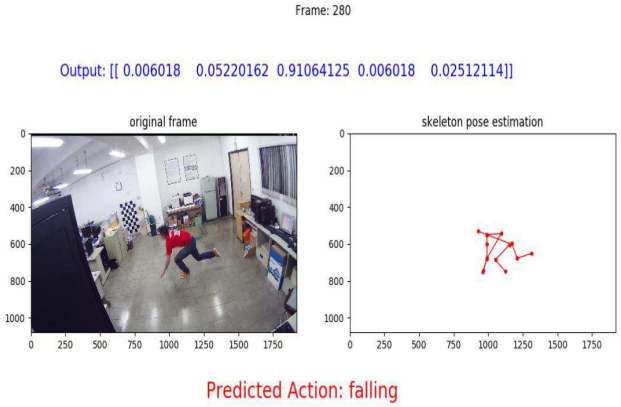

Fig 3. A real on-line application of our system for fall-down event detection.

REFERENCES

[1] Alex Krizhevshy, Ilya Sutskever, and Geoffrey Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. of the 25th Int'l Conf. on Neural Information Processing Systems, NIPS'12*, Vol.1, pp.1097-1105, 2012.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, and Marcus Rohrbach, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," CVPR, 2015.

[3] Enea Cippitelli, Samuele Gasparrini, Ennio Gambi, Susanna Spinsante, "A Human Activity Recognition System Using Skeleton Data from RGBD Sensors," *Computational Intelligence and Neuroscience*, 2016.

[4] P. T. Hai and H. H. Kha, "An efficient star skeleton extraction for human action recognition using hidden markov models," *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE).*

[5] Varun Ramakrishna, Daniel Munoz, Martial Hebert,

J.A. Sheikh, “Pose Machines: Articulated Pose Estimation via Inference Machines,” ECCV 2014.

[6] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, “Convolutional Pose Machines,” CVPR 2016.

[7] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele: “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation,” *Proc. of IEEE Int’l Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Yong Du, Yun Fu, and Liang Wang, “Skeleton Based Action Recognition with Convolutional Neural Network,” 3rd IAPR Asian Conference on Pattern Recognition, 2015.

[9] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele: “DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model,” *Proc. of European Conference on Computer Vision (ECCV)*, 2016.

[10] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. of CVPR*, 2016.

[11] Sepp Hochreiter, J. Schmidhber (1997), “Long short-term memory,” *Neural Computation* 9(8), pp.1735-1780, 1997.