Real-Time Human Motion Detection, Tracking and Activity Recognition with Skeletal Model

Sandar Win
Faculty of computing(UCSY)
University of Computer Studies,Yangon
Yangon, Myanmar
sandarwin@ucsy.edu.mm

Thin Lai Lai Thein
Faculty of Computer Science (UCSY)
University of Computer Studies,Yangon
Yangon, Myanmar
tllthein@ucsy.edu.mm

***Abstract*—Human activity recognition with 3Dskeletal model has been attracted in a lot of application area. Representations of human based on 3D perception have been occurred prevalent problems in activity recognition. In recent work with RGB-Depth cameras, expensive wearable sensors and illuminator array have been used to construct the 3D human skeleton model in recognition system. But these systems have been defined specific lightening condition, limited range. and great constraint in outdoor applications. To overcome this restriction, the proposed system is considered on the real-time video sequences of the human movement to understand human behavior in indoor and outdoor environment. The proposed method is constructed human detection and motion tracking by using framewise displacement and recognition is based on skeletal model with deep learning framework. The result is to become an efficient detection, tracking and recognition system for real-time human motion. The performance and accuracy of the system is analyzed with the various videos to show the results.**

Keywords—human recognition, skeletal model, deep learning

I. INTRODUCTION

Real-Time human motion detection, tracking and activity recognition is one of the active research areas in computer vision and has applied in many application areas. The component of human pose can interpret visual information from the surrounding environment to real-world problem. Human activity recognition system also applies in many fields such as security, safety and human activity monitoring in many environments. Human motion analysis in computer vision system consists of face recognition, gesture recognition and body recognition. 3D skeleton-based human representations generally analyzed into four categories based on raw position, joint displacement, orientation, and combined information. Local features and skeleton based representation are useful for human representation systems [6].

Although 3D modelling and skeletonization system have been proposed in past decades, there are unsatisfied for outdoor application [11]. To get reliable manner, the proposed system is structured 3D skeletal model according to the component of human body parts contain 18 fundamental points defined as head, neck, shoulders, elbows, wrists, hips, knees and ankles, etc. to represent human body skeletal structure and to recognize human activity. 3D skeletal model performs us to focus hidden human body parts in 2D images and that can show people interact with each other, overlapping groups and human activity. Currently human activity recognition with 3D skeletal model has limitations in various aspects. Most of the system require the generation of data for better performance with development method. Our system is intended to perform robust 3D human skeleton system based on deep neural network that is without being altered by different situations and environmental changes. The system is constructed a skeletal model from the perception of data through joint estimation and pose recognition. Our goal is a robust and efficient approach in human recognition system from training and testing on different data. Fig. 1 shows input video with detection, tracking and recognition result of the proposed system. The arrangement of this paper as follow: Section II concerned with about the related papers. Section III focus on detection and tracking process of human movement. Section IV shows human recognition with skeletal model. In session V, expresses the experiment and accuracy results in detail. Finally, Session VI describes the conclusion and future work.

Fig. 1. Example result of human detection, tracking and recognition with 3D skeletal model

II. RELATED WORK

Along with several numbers of human representation approaches, most of the existing of human recognition system have been proposed with skeletal model. Hussein et al. [5] proposed human skeleton by using Covariance of 3D joints (Cov3DJ) matrix over the sub-sequences frame in a temporal hierarchy manner. That has fixed length and not depend on sequence length. They computed random joint probability distribution variable correspond to different feature maps in the region. The system deployed joint location over time and action classification with Support Vector Machine (SVM). Du et al. [8] proposed Recurrent Pose Attention Network (RPAN) that predicts the related features in human pose. The system used end to end recurrent network layers for temporal action modeling to construct a skeleton-based human representation. As the number of layers increase, the representations extracted

by the subnets are hierarchically fused to figure a high-level feature to represent human in 3D space.

Plagemann et al. [1] analyzed human shape from interest points of salient human body. The system directly estimated 3D orientation vector from body part in space and learned the estimated body part locations with Bayesian network. Luvizon et al. [3] proposed a multitask framework for 2D joint and 3D pose estimation from still images and human action recognition from video sequences. They trained multi type of dataset to generate 3D predictions from 2D annotated data and proved an efficient way for action recognition based on skeleton information. Hou et al. [4] proposed shadow silhouette-based skeleton extraction (SSSE) method that analyzed silhouette information. Skeleton synthesis and 3D joint estimation is working on extracted 2D joint positions from shadow area on the ground. Major joint position is defined and result is compared with RGB-D skeletonization. Various systems with different improved methods have been occurred on skeleton based human representation.

Our system is intended to build more robust system for human motion with fast and accurate detection, tracking and activity recognition by using skeleton based system. The system flow of proposed system is shown in Fig. 2.

The proposed system proceeds as the following steps:

1) Detect human silhouette information from background for each frame.
2) Extract 2D joint projected positions.
3) Extracted 2D joint positions are categorized as sequence of body parts.
4) Generate 3D human skeleton using spatial-temporal integration of 2D joint positions.
5) Define activity recognition according to skeleton joint position.

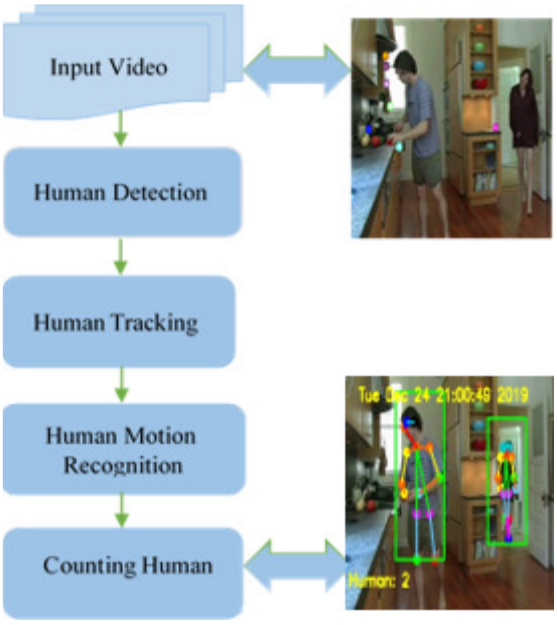


Fig. 2. System flow of the proposed system

III. Human Detection

Human detection is to know human shape, pose, and motion from perception data in a frame. An efficient and accurate detection method rely on the application of human attributes. The system detects silhouette image from the background by using threshold value and the skeleton feature points is extracted from spatial-temporal evolution of human body parts. The system finds the location of each human joint on all the feature maps [7]. And the feature map with the highest-activated value at the joint location is selected for the corresponding joint. The position of the joints is defined by using local minimum and maximum value of distance transformation and that value should be greater than their neighbor's points. The system uses iterative method on the human body until a skeleton leftover with deletion or retaining the value. The arrangement of points is based on nearest neighbor distance order according to body parts. Then, define a bounding box and detect around the human body.

A. Graph Modeling

For spatial-temporal graph modeling, the system initiates a novel adaptive dependency matrix and learns it through node embedding on connected human body parts. Our model can precisely capture the hidden spatial dependency in the data. As the number of layers increases, whose receptive field grows exponentially and handles the very long sequences for pose estimation.

B. 3D Pose Estimation

Pose estimation is an important class in action recognition. Partially, previous pose is carried as the known pose to fit and serves as a preprocessing step for next processing [2]. The estimation of 3D human pose from 2D joint locations is central to many vision problems on the analysis of people in images and video. Pose estimation can deal with any composition of rigidly moving parts connected to each other according to joints. The system identifies human body parts to estimate the pose, as in:

$$L_{pose}=\sum_I \sum_{t=1}^{T} \sum_{k=1}^{K_1\times K_2} \left(N_t^I(k)-\alpha_t^I(k)\right)^2 \quad (1)$$

where T is total time steps, N_t^I is high visual data for all joints, α_t^I is joint attention score which can be computed, as in:

$$\alpha_t^I(k)=\frac{\exp\{\tilde{\alpha}_t^I(k)\}}{\sum_k \exp\{\tilde{\alpha}_t^I(k)\}} \quad (2)$$

The related feature of human body parts is described, as in:

$$F_t^P=\sum_{I\in P} \sum_k \alpha_t^I(k) V_t(k) \quad (3)$$

$V_t(k)$ is the feature vector of V_t at the k^{th} spatial location $(k=1,\ldots, K_1\times K_2)$, $\tilde{\alpha}_t^I(k)$ is the unnormalized attention score of $V_t(k)$ for each joint $I\in P$, P is the body part. That describes the human figure which has relative positions of the limbs and generate 3D pose estimation that contains frontal, lateral, backwards and forwards displacement. And then tracking algorithm is applied to track the detected object across different frame.

C. Frame-wise Displacement

In Frame-wise displacement, Frame-wise Motion Fields (FMF) and Frame-wise Appearance Features (FAF) are motion representations that estimate appearance and contextual information of the video between two frames. Frame-wise displacement relative to previous frame and reference frame is computed, as in:

$$FD_i=|\Delta d_{ix}|+|\Delta d_{iy}|+|\Delta d_{iz}|+|\Delta\alpha_i|+|\Delta\beta_i|+|\Delta\gamma_i| \quad (4)$$

Where $\Delta d_{ix}=d_{(i-1)x}- d_{ix}$, these variable measures the movement of any given frame. We have to consider the cross-entropy method to fix uncertain matches in information. This leads to very fast and reliable matching among a large number of objects bounding boxes with the significant achievements made in object tracking.

D. Human Motion Tracking

The main work of motion tracking is to find detected object in each frame and search motion path of the object to track in time sequences. That is changing location with respect to its background. By using inference algorithm. To increase reliability and robustness, analyzing temporal information is good track for trajectories. The motion sequences move each target from frame to frame, which is the key of moving object tracking to locate the search regions of a target in the next frame. When the next frame, this distribution has changed due to the new mode and tracks the object correctly. This system gives a high precision of detection and tracking for moving object concerned with variation of appearance such as presence of noise in foreground image, different poses, changes of size, shape and scene in indoor or outdoor environment. Input video and result of outdoor environment as shown in Fig. 4.

IV. Human Motion Recognition

In human recognition systems, the input features have structured in various ways in motion sequences. Since human actions are highly dynamic, partially visible, occluded, or cropped targets and very closely resemble with ambiguities. For the understanding and analysis of human behavior, recognition with skeleton system are robust to variations of viewpoints, body scales and motion speeds. It is essential for general shape representation and that information can be access in real-time. The system first extracts simple but effective frame-level features from the skeletal data and build a recognition system based on deep learning neural network to obtain the recognition results on defined action sequence level. Human body parts representation from spatial-temporal parts as expressed in Fig. 3.

A. Deep Learning Approach

The system applies Deep Neural Network (DNN) to represent human in 3D space and to recognize skeleton joint locations. DNN is capable of capturing the full context of each body joint in the full image as a signal. Since, skeleton feature points are scattered in human body and joint coordinates system has different coordinates due to translation and rotation and then depend on different body sizes. The system is learned in neural network for connecting the skeleton feature points and

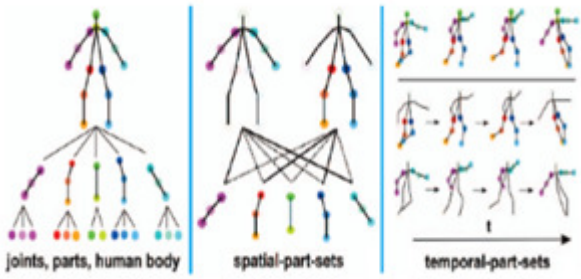


Fig. 3. Human body parts representation from spatial-temporal parts

estimating joints to increase precision result of joint localization.

B. The Gradient of Distance Transform

Human body is connected by joints and human action present continuous evolution of spatial configurations of joints [9]. We normalize the positions of the joints of each subject using the neck point and then concatenate (x, y) coordinates into a feature vector. The system is trained in a deep network using similar approach to the above shape context method. To obtain the connected lines, we may need to explore the local features that is using with the gradient of distance transform image, as in:

$$\nabla DT=\left(\frac{\partial DT}{\partial x}, \frac{\partial DT}{\partial y}\right) \quad (5)$$

The norm of gradient vector is obtained in:

$$|\nabla DT|=\sqrt{\left(\frac{\partial DT}{\partial x}\right)^2+\left(\frac{\partial DT}{\partial y}\right)^2} \quad (6)$$

To overcome unmatched point in skeletonization, the system is defined on constraint condition when searching the successive feature points that means $DT(p)>0$, where p is a rigid part that is joined by joint. It is reliable for any lines connecting two points must locate on the foreground. Roughly, the proportions of human body parts are the same.

C. Skeletonization

Evolving the skeleton feature points in the body parts is a specific routine of skeletonization. By capturing spatial-temporal representation of body parts in one frame as well as movements across a sequence of frames through graph model [10]. The graph model captures the geometric and appearance variations of human at each frame and that represent the motion of human with 3D skeleton joints. The structural data is represented by pairwise features, relative to the position of each pair of joints relate to each other. The orientation between two joints *i* and *j* are computed, as in:

$$\theta(i, j)= \arcsin\left(\frac{i_x-j_x}{dist(i, j)}\right)/2\pi \quad (7)$$

where dist(i,j) is define for the geometry distance between two joints i and j in 3D space.

Fig. 4. Input walking video and recognition result in outdoor environment

V. Experimental Results

The results obtained in the implementation are shown in this section by using HMDB51 dataset consists of 6,766 videos. That is collected from various sources from YouTube, Google videos, mostly movies and small public database. From that we experiment on 500 videos with 4 actions. The frame rate of videos is 30 fps. Testing on dataset with different conditions are shown in Table I. The confusion matrix on different activities are described in Fig. 5. The performance of the system is critical point in any dataset. The accuracy for human recognition results on different iterations are expressed in Fig. 6.

VI. Conclusion and Future Work

In many application area, real-time information is very important and require for efficient human motion detection, tracking and activity recognition system. That can record useful information and analyze the environment in many scopes. There are many challenges that are concerned with different variation in human pose, shape, illumination changes and background appearance. In this paper, the system is implemented by using deep neural network framework to get high accuracy recognition of human movement in indoor and outdoor areas. The experimental results have been concluded that all method have a big dependency on different backgrounds, camera calibration and illumination changes. We trained and tested video data on different changes that are significantly increased the detection, tracking and recognition rate of our results.

Future research directions will continue 3D skeletal model for moving object with various dataset containing different activities to describe the accurate result of human motion detection, tracking and activity recognition system.

TABLE I. Testing The Sequences on HMDB51 Dataset

Activity	Frame Per Second	Length of Time	Number of Track Box	Conditions
Walking	30	00:00:05	12	sunny
Standing	30	00:00:03	13	indoor
Sitting	30	00:00:05	10	cloudy
Running	30	00:00:10	5	outdoor

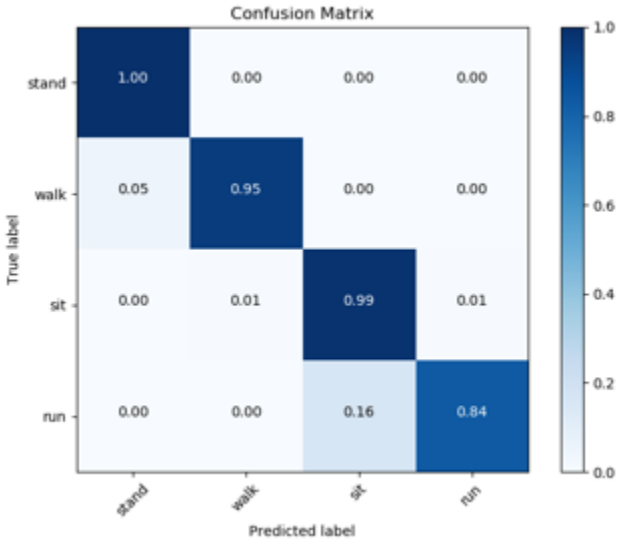
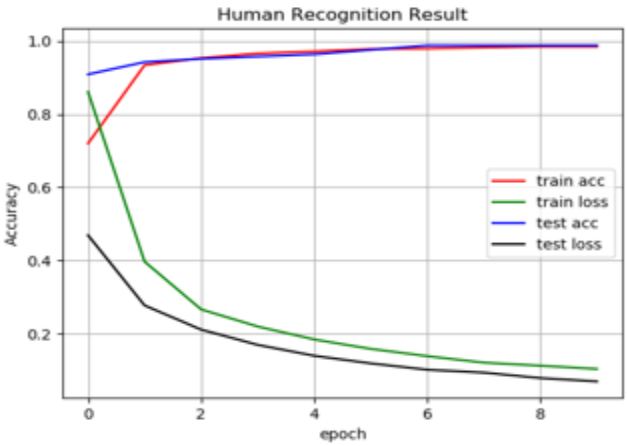


Fig. 5. The result of detection, tracking and recognition with confusion matrix on different activities



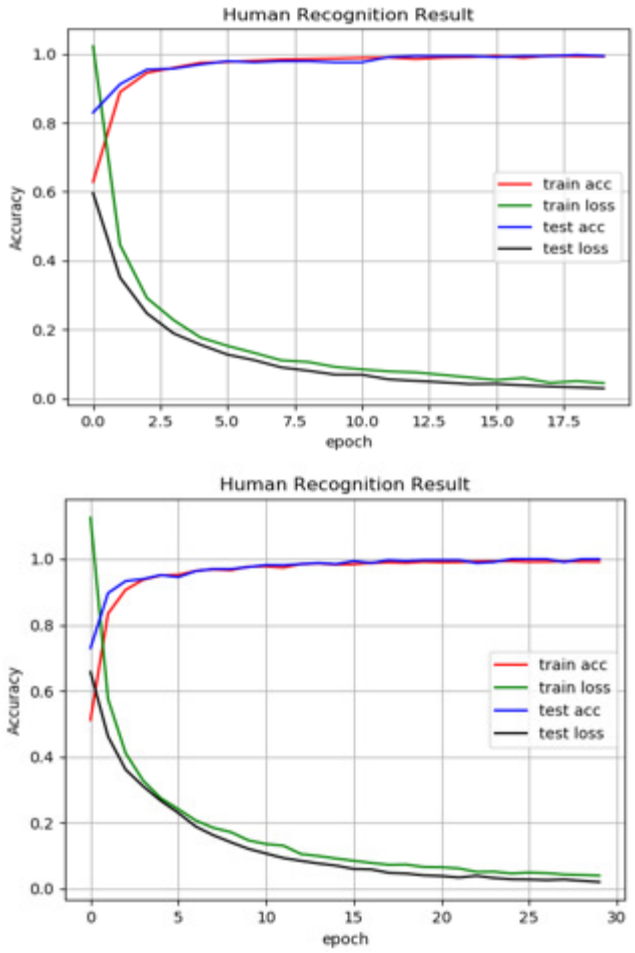


Fig. 6. Human recognition result on training and testing with different iteration

REFERENCES

[1] C. Plagemann et al., "Real-time identification and localization of body parts from depth images", in: IEEE International Conference on Robotics and Automation, 2010.

[2] C. Wang, Y. Wang, A. L. Yuille, "An approach to pose-based action Recognition", in: IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[3] D. C. Luvizon et al., "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning", IEEE Conference on Computer Vision Foundation, 2014.

[4] J. Hou et al., "3D Human Skeletonization Algorithm for a Single Monocular Camera Based on Spatial-Temporal Discrete Shadow Integration", Appl. Sci. 2017.

[5] M. E. Hussein et al., "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations", in: International Joint Conference on Artificial Intelligence, 2013.

[6] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection," in IEEE International Conference on Computer Vision (ICCV), 2017.

[7] S.H. Rezatofighi, A. Milan, Z. Zhang, Qi. Shi, An. Dick, and I. Reid, "Joint probabilistic data association revisited", in ICCV, 2015, pp. 3047–3055.

[8] W. Du, Y. Wang, and Y. Qiao, "RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos," in IEEE Int. Conf. on Computer Vision (ICCV), Oct. 2017, pp. 3745–3754.

[9] Xia, L. Aggarwal, J. "Spatial–temporal depth cuboid similarity feature for activity recognition using depth camera", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.

[10] Y. Du, W. Wang, L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition", in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] Zhang, Z. "Microsoft Kinect sensor and its effect", IEEE Multimedia 2012.