Motion Features for Human Action Recognition Using 3D Skeleton Model

Salah R. Althloothi[*1], *Almokhtar Alazhari*[2]

***e-mail: salah668@gmail.com**

[1,2]*Computer Engineering Department, Faculty of Engineering, University of Al Zawiya*

Abstract

This paper presents the development of motion features for accurately extracting the distal segments of human limbs in visual data for human action recognition. Using the depth map provided by the Kinect sensor, motion features are extracted to classify human actions in videos. The motion features are the motion of the 3D joint positions of the human body. These 3D joint positions are used to provide precise endpoints of the distal segments of each limb which are reduced to centroids for efficient recognition. Each limb centroid is described by its angle with respect to the vertical body axis to create an action descriptor vector. The action descriptor which represents the position of the torso and four limb segments is detected and tracked without any manual initialization. It is also invariant to image resolution and video frame rates, making it suitable for a wide range of human tracking applications in real time surveillance. To evaluate our approach, a public dataset was used for human action recognition. The results of our experiments show a good direction in incorporating motion features using SVM technique for automated recognition of human actions.

Keywords: Features, Recognition, Tracking, Action description, Skeleton.

1. Introduction

Human action recognition has remained an interesting and challenging topic in the area of computer vision and pattern recognition. The topic has been motivated by many applications such as automated visual surveillance, human-robot interaction, video retrieval, and motion-based human identification. The surveys by Gavrila [23], Aggarwal et al. [24], Poppe [21], and by Moeslund et al.[22] provide a broad review of over two hundred articles published on recognizing and analyzing human actions in videos, including human motion tracking, capture, segmentation, classification and recognition.

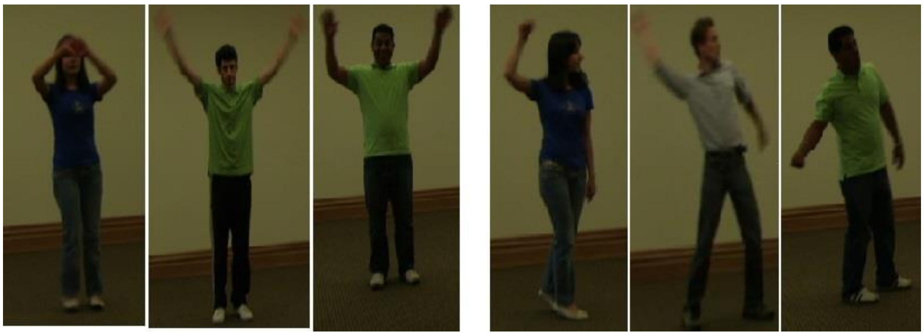
According to Moeslund et al. [22, 34], human movement can be divided into three categories: limb action, whole-body action, and activity. The limb action is a movement that can be described at the limb level. The whole-body action consists of a set of limb actions that can be described at the whole-body movement for a short period. Finally, activity contains a number of subsequent whole-body actions in a cycle which gives an interpretation of the movement that is being performed in this cycle. For example, “right leg forward” is a limb action, whereas walking is the whole-body action because it consists of a set of limb actions. Fighting is an activity that contains a number of subsequent actions such as standing, jumping, running, etc. In this paper this classification were adopted in recognizing limb actions and the whole-body actions such as walking, running, etc.

Challenges in human action recognition

Recently, significant amount of research has been done in the field of automated visual surveillance systems [13].The main purpose of these researches is to describe the human action in order to detect unusual behavior. In human action recognition, the common approach is to extract motion and shape features from the video and to classify performed actions by subjects. The classification algorithm that can deal with a variety of actions is usually learned from training data. However, human action recognition is still one of the challenges in the field of computer vision due to variations in several factors such as performance, environment and view angles variations. In addition, segmentation problems in different environments cause action recognition algorithms fail to recognize the correct actions because it relies on segmenting body parts. In the next paragraph, the challenges that influence the action representations are discussed. For further detail we refer our readers to [22, 34].

Performance variations

There are variations in performing most actions amongst most people. Each person performs an action in his way and in different rate compared with other individuals. For example, running movements can differ in stride length and rate between individuals. However, a robust human action recognition approach should be able to generalize performance variations among one action and capable of distinguishing between actions among different individuals. Additionally, recognizing a large number of actions is more challenging because the similarity between the actions is been increased.



Waving Action Throwing Action

Figure 1. Each person performs an action in his own style.

Environment variations

The most important source of variation is the environment variation where the action performance takes place. It is very challenging to detect the location of the person in cluttered or dynamic environments. Moreover, lighting conditions or illumination of the environment can further influence the appearance of the moving person. Therefore, with a static environment, it is less challenging to represent the type of motion of the foreground object.

View-angle variations

Observing the same action from different view angles can lead to different image observations and thus to a different perceived action. Multiple cameras can alleviate this issue and the occlusion issue by combining all the observations from different views into a consistent scene at the same time. Because of the synchronization the challenge here becomes harder by using multiple cameras in the same time to capture the action. Moreover, moving the cameras increase the complexity of localizing the moving object in the scene, especially when the backgrounds are dynamic. In human action recognition, all these challenges should be addressed explicitly or fixed before the action take place.

Rate variations

An important effect on temporal variations is the rate of performance, especially when motion features are utilized to extract the action. Because of the rate variations for an action it will be more challenging to know the beginning and the ending of the action. Therefore, a robust human action recognition approach should be invariant to different rates of execution and be able to generalize rate variations among different individuals.

Considering the above challenges and issues for human action recognition, an image representation should generalize the variations in localizing the person in the scene, background, view angle, rate, appearance of person, and the action performance [21]. At the same time, the representations of the image should be sufficiently rich with the information as features to increase the accuracy of the action classification algorithm. In addition, the image representation is used to represent the temporal dimension for each frame in order to be utilized in the classification algorithm [22].

A3D skeleton model that aims to address certain parts of these challenges is presented. The approach presented in this paper addresses these issues by extracting features that are less sensitive to the aforementioned variations from skeleton model and employing a Support Vector Machine (SVM) to classify the action based on the training data. Also, the 3D skeleton model presented in this paper is capable of representing the simultaneous changes in the human limbs and motion in a unified manner. The core of the proposed approach is to use features computed from spatiotemporal skeleton model as stable features for deciding what actions is performing by a person. The proposed action representation is generated in two steps: the creation of a 3D skeleton model from the silhouette of depth map, and the extraction of the action descriptor from this model in order to classify the action.

In summary, the contributions of this paper are as follow: accurate limb extraction based on a 3D skeleton model, reduction of the 3D skeleton model to the distal limbs only for action recognition, evaluation of the approach with respect to human action classification.

The remainder of this paper is organized as follows. Related work is reviewed in the next Section. Human skeleton model creation, feature vector extraction and the action classification is illustrated in Section 3. An experimental result for a public dataset is presented in Section 4. Finally, Section 5 concludes the paper.

2. Related Works

According to the surveys by Poppe [21] and Moeslund et al. [22],the recent work on action recognition can be broadly classified into two types of approaches: global approaches and local approaches. Global approaches proceeds in a top-down fashion: First, the whole body of person is localized in the region of interest (ROI) which is usually obtained through tracking or background subtraction. Then, the ROI is encoded in order to create the image descriptor. In other words, global approaches use global features such as optical flow, silhouettes or edge maps to represent the motion in the whole frame. All these features are sensitive to partial occlusions, noise and view angle variations because they focus in the motion in the whole frame. To overcome these issues, multiple images over time can be utilized to form a three-dimensional space–time volume or grid-based approaches can be used to divide the observation into cells spatially [19]. In other hand, local approaches are used to describe the observation as a collection of independent patches. Local approaches are obtained in a bottom-up fashion: First, spatial-temporal interest points are detected. Then, local patches are calculated around these interest points. Finally, the local patches are combined into a final image representation. Local approaches are less sensitive to noise and partial occlusion compared with global approaches because they focus only on relevant interest points and correlation between them [22,33,34].

In global approaches, there are two general categories for human action recognition. The first category is based upon the motion and the shape of the whole body. An example of this approach is the work of Wang et al. [18]. They build a model composed of fourteen segments each of them presented by a truncated cone. The purpose of this model is to detect local variations in shape and motion in order to recognize the action. Another example can be found in the work of Feng and Abdel-Mottaleb [7, 8]. In their work they combined both shape and motion features of the silhouette and used them in creating a set of Hidden Markov Models to identify the action. By matching these features using voting algorithm their system was capable to discriminate the action among different view angles. Tran et al. in [27] presented an optical flow as motion features and shape based approach that uses separate histograms of the optical flow as well as the silhouette of the person as a motion descriptor. An advantage of using this approach which is based on the motion and the shape of the whole body is to detect local variations in shape and motion in order to recognize the activities. However, an accurate segmentation which is not a trivial task and labeling of human body parts are necessary to recognize the correct activities especially for surveillance cameras.

The second category looks at the changes in the shape of the whole body and tries to recognize actions based on its dominant motion in the limbs. For instance, the approach pursued by Chen et al. [14] falls in this category; they proposed a star skeleton representation to recognize human motion. The center of mass of a human silhouette is extracted to represent the body as a single star. Then, extreme points corresponding to extreme contour points are detected as local peaks. In order to accurately detect extreme points for human body, Yu and Aggarwal [28] proposed a two-star skeleton representation by adding the highest contour point as the second star. Two sets of local peaks are estimated to find more precise extreme points. This work was modified later to variable star skeleton by Yu and Aggarwal [28]. The variable star skeleton was proposed to improve the accuracy of finding extreme points. They first constructed a medial axis from the human body contour. Then, they treated all junction points of the medial axis as stars. For each star, they generated a set of extreme points; each extreme point was processed according to its robustness, visibility, and proximity to its neighbors. Another example is the work of Chun et al. [29]. They proposed 3D star skeleton based on 3D information of the human posture through using eight projection maps. Although star skeleton is simple and fast for computation, its accuracy for detecting limbs needs further improvement. It cannot discriminate between the limbs and the other parts of the body compared with our approach.

In contrast to the global approaches, the local approaches describe the observation as a collection of local descriptors or patches. These patches which correspond to interesting motions are sampled at space–time interest points where the local neighborhood has a significant variation in both spatial and temporal domain. The idea of this approach is to collect the interest points as spatiotemporal features that are distinctive and descriptive. Therefore, the interest points detection algorithm play an important role in local approaches to classify the action.

One example of this category is the work done by Wu et al. [32] developed a new image representation called SIFT Motion Estimation (SIFT-ME). SIFT-ME is derived from SIFT correspondences in a sequence of video frames and adds tracking information to describe human body motion in both spatial and temporal domain. They utilized SIFT parameters for translation and rotation to describe 2D human body motion with SIFT-ME as spatial-temporal interest points. Another related work SIFT-based approach to recognize human action is known as MoSIFT by Chen et al. [29]. In this work, they proposed MoSIFT algorithm to detect and describe spatial-temporal interest points. The MoSIFT algorithm detects spatially distinctive interest points through local appearance to describe human body motion in both spatial and temporal domain.

In this paper, the focus is the global approaches in identifying actions performed by persons in a short period of time by describing the motion of the limbs and tries to recognize actions based on its dominant motion in the limbs. We believe a good approach to representing body limbs is the human 3D skeleton model, which consists of line segments linked by joints. In this case, the motion of joints provides the key to motion estimation and action recognition of the whole body. In our work, this concept is achieved in a novel way by detecting distal limb segments from silhouette of depth map to accurately fit a human 3D skeleton model. With this model, we generalize variations among different people performing the same action.

Proposed Approach

Using the silhouette or the human contour to represent a human posture is inefficient since all the pixels in the silhouette and contours are similar to each other. On the other hand, simple information from the silhouette like center, height and width of blob cannot represent a human posture and it is difficult to discriminate human posture from each other. Consequently, 3D human skeleton model seems to be a good way to represent the whole structure of a human posture. Also, it can be utilized to extract small variations in human postures to describe human action. In the following we describe our approach for extracting the 3D skeleton model from depth map sequences followed by human action classification based on six-element action descriptor as shown in Figure 2.

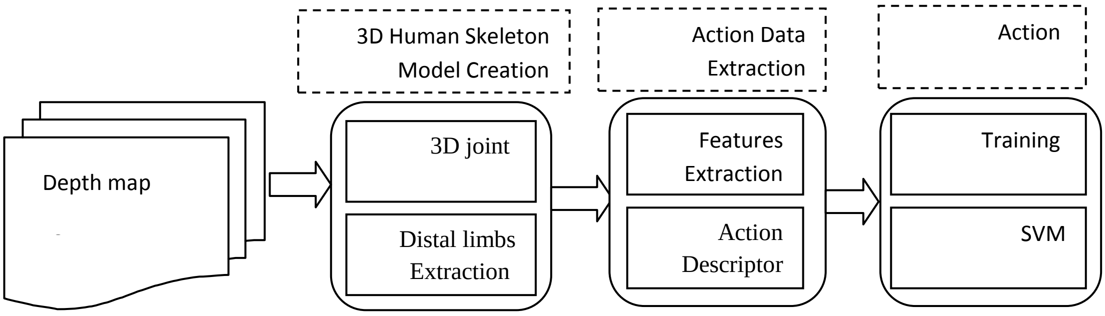


Figure 2. The process of recognizing human actions developed in this paper.

3D Human Skeleton Model Creation

Recently, the developed commodity depth sensors such as Kinect [6] have opened up new possibilities of dealing with 3D data. This type of sensor has given the computer vision community the opportunity to acquire RGB-D images at a good frame rate (30 fps) with a good resolution. As we can see in Figure 3, the depth map provides additional information as 3D data which is expected to be helpful in distinguishing poses of silhouettes. Furthermore, compared with RGB images, the depth map increases the amount of information that can help to detect the 3D joint positions.

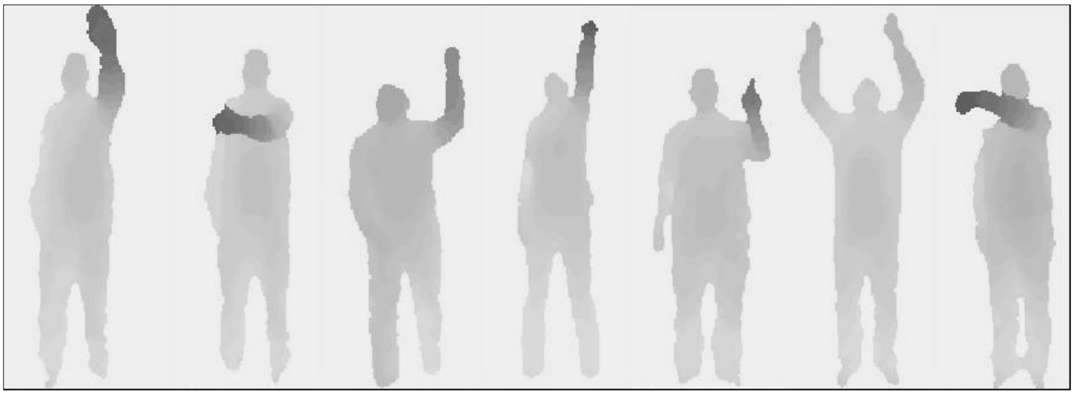


Figure 3. The depth map provides additional information as 3D data

In fact, the 3D joint positions of the distal limb segments (four limbs) provided by Kinect sensor, as illustrated in Figure 3, are utilized to create the 3D human skeleton model. Once the 3D joint positions are successfully extracted from the depth map, these 3D joint positions were used to produce the 3D skeleton model. Therefore, a simple way to represent 3D human skeleton model in the frame is to use its 3D joint positions by utilizing the line fitting algorithm. The 3D skeleton model used here to represent the human consists of Seventeen segments and eighteen joints as shown in Figure 4.

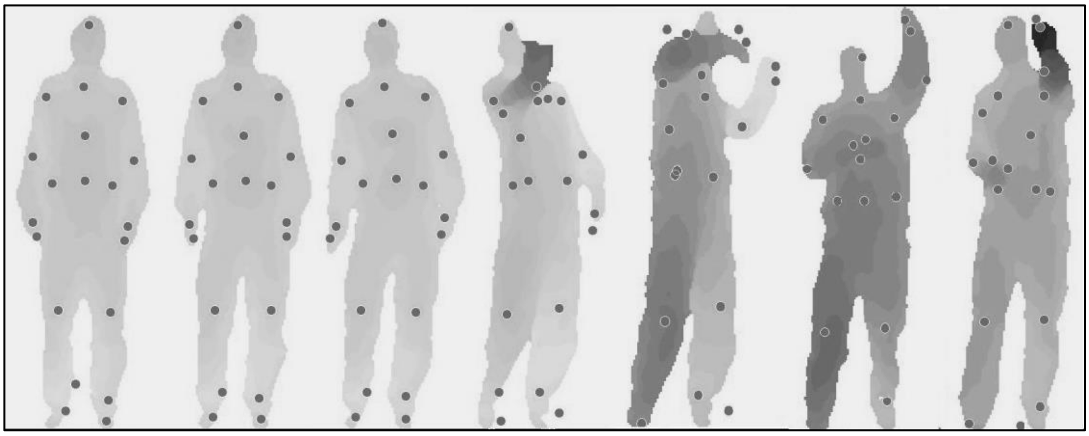


Figure 4. The 3D joint positions of the distal limb segments (four limbs).

The 3D skeleton model is employed to extract the motion features of the human limbs. In fact, the 3D joint positions of the distal limb segments (four limbs) provided by Kinect sensor, as illustrated in Figure 4, are utilized to extract the motion features which represent the orientation and the translation distance of the distal limb segments. Our key observation is that the change in positions of the distal limb segments provides sufficient information to represent the human body movement as discriminative features to classify the human actions.

Action Data Extraction

Once the 3D human skeleton model is created, the centroid angle of the distal segment on the limb can be detected as a feature to identify the orientation of the limb segment. In order to determine the centroid angle with high accuracy, the angles of the endpoints of the distal segment with respect to the vertical body axis are averaged for each segment.

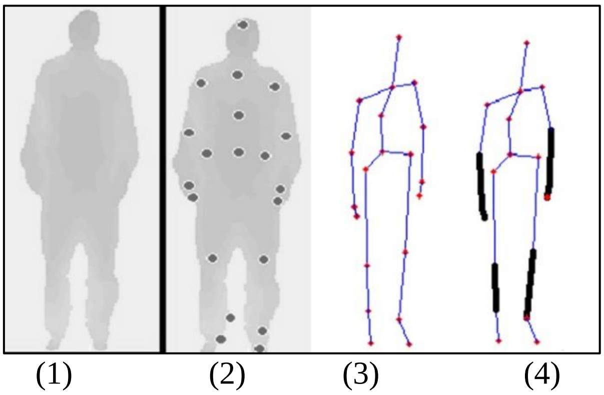


Figure 5. The 3D joint positions and it is skeleton mode with the distal limb segments (four limbs).

Since all angle measurements are relative to the vertical body axis in one frame it is necessary to add other information about the motion of the human as function of time to the classifier to discriminate between two actions such as picking up and jumping. Therefore, translation distance, which gives the difference in position between two consecutive frames, is added to the descriptor vector to increase the classification accuracy. Consequently, the resulting features extracted from each frame can be the angles of displacement from the vertical body axis and the translation distance of the position of the hip [15, 33, and 34]. This yields a descriptor vector with precise orientation data and translation data of the human body for each frame.

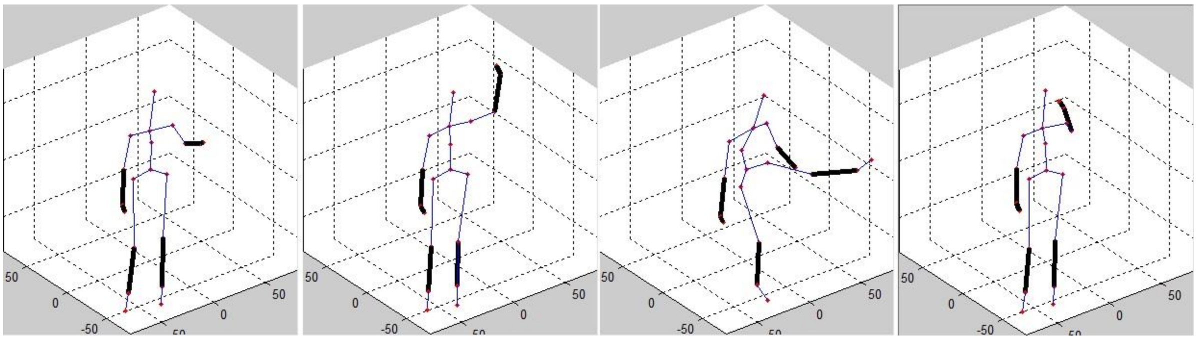


Figure 6. The 3D human skeleton model with the distal limb segments (four limbs) for some actions.

In our approach, the 3D joint positions of the distal limb segments are used to characterize the motion features. Thus, each distal limb segment {L_k} is described by its end points as 3D vector with respect to the body center (hip) in order to create the3D unit vector which represents the orientation of the distal limb segment. The orientation of the 3D vector is defined by a unit vector $\{\overrightarrow{U_{ab}}\}$ for frame {F_t}. Therefore, the length of distal segments L_k in frame F_t is$|V_{ab}| = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2}$

Since all the measurements are relative to the center of the body in one frame, it is necessary to add another information about the motion of the human limbs between the current frame F_t and the initial frame F_{t_0} , which represents to the neutral human pose. Thus, the translation, which gives the difference in position for the distal limb segments between two frames, is computed in order to create a 3D Spatio-temporal feature, and to make the classifier discriminates between the actions that have similar orientation but different positions. Practically, each human subject has four distal limb segments which are tracked by the skeleton tracker; each distal limb L_k is represented by 3D unit vector and translation vector in each frame.

In order to represent the motion features that include {t} number of frames, our motion features for all the frames are concatenated together to build a spatio-temporal features vector for motion features. This yields to a 3-D spatio-temporal feature with precis orientation and translation data for each human limb in all the frames. Thu, these motion features define the movement of distal limbs in the video. In addition, all the vectors which represent the 3D distal limb segments were normalized to reduce intra-class variations among subjects and to be invariant to the body size.

Action classification

In machine learning, support vector machines (SVM), are supervised learning models with associated learning algorithms that analyze data used for classification. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

To perform action classification using the proposed action descriptor, Support Vector Machine (SVM) [16, 17, 34] were used. First, we trained several SVMs for different actions using Leave-One-Subject-Out (LOSO) method. For all the videos, one subject was removed from the training set and the other subjects were utilized to train separate SVMs for separate actions. Then, the excluded videos were used to test the accuracy of our approach in classifying the performed actions in videos. The test subject was different from the training ones. This process was repeated for all the subjects and all the actions in the dataset separately. We tested different kernel functions such as linear function and Radial Basis Function (RBF) where RBF kernel showed better best results, As result SVM classifiers with RBF kernel were used in our experiments.

3. Experiments and Results

A challenging public dataset which is known as MSR-Action3D dataset [1] for human action recognition was chosen to evaluate our proposed method. Also, two kernel functions were used; Gaussian function and polynomial function, each with different parameters are linearly combined to classify the action using multiclass-SVM classifier. Within this frame, several experiments were conducted using different number of training samples in order to evaluate the performance of our proposed method..

MSR-Action 3D Dataset

MSR-Action 3D dataset [1] is an action dataset of depth map sequences captured by a depth camera (Kinect sensor) as illustrated in Figure 5. This dataset contains different human actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw.* It includes twenty actions performed by ten subjects. Each action was performed 2 or 3 times by each subject. The size of the depth map is 320×240. All the subjects were facing the camera during the performance, and were given a freedom to perform the actions at their own place in front of thecamera. Furthermore, most of the actions involve the movement of limbs (i.e. arm, leg) in one place, which makes most of the actions highly similar to each other.

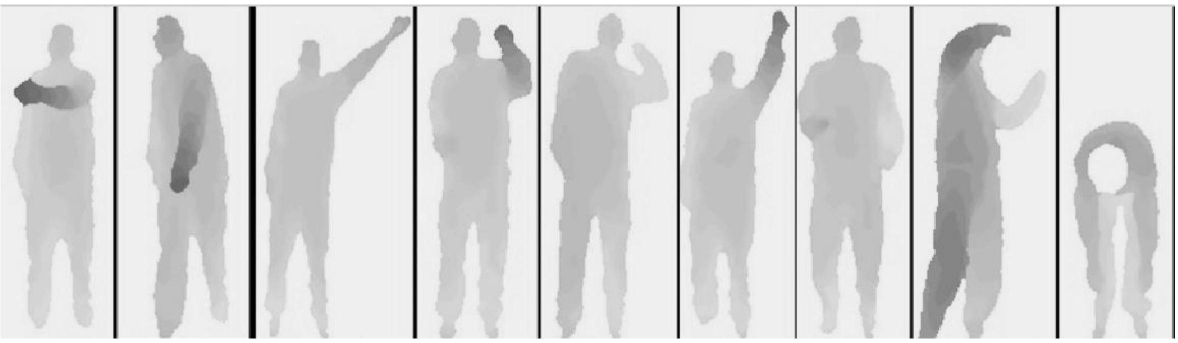
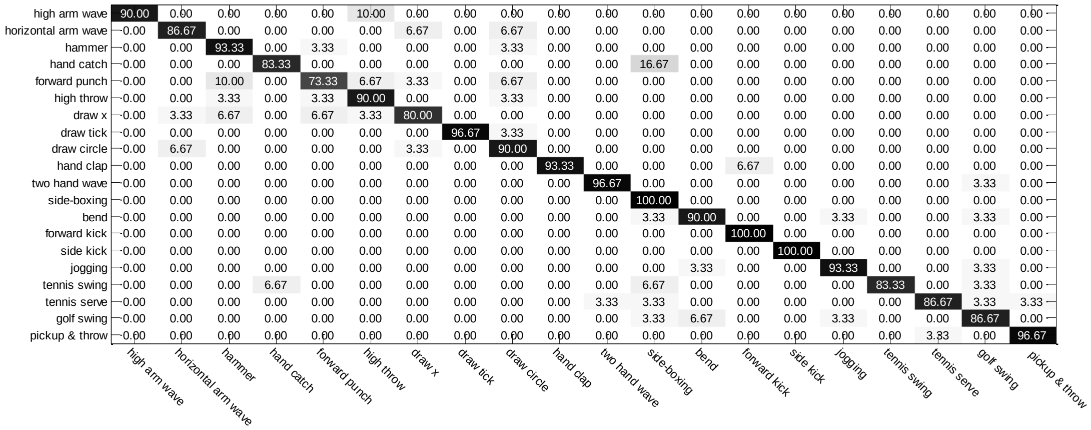


Figure 7. A few sample frames of MSR Action 3-D dataset; the activities from left to right: horizontal arm wave, hammer, high throw, draw circle, throw, hand catch, jogging, tennis

swing and bend actions.

In our experiment, the end points of distal limb segment were used to calculate the orientation as a unit vector representing the distal limb segment and the translation distance with respect to the initial frame as explained in Section 3. Since this dataset has a number of samples for training, Leave-One-Subject-Out (LOSO) test and cross subject test were applied in our experiments to verify the performance of our method. In the LOSO test, one subject is removed from the training set and the other subjects were utilized to train the multiclass-SVM classifier. The excluded subject is used to test the accuracy of our method in classifying the performed activities. The test subject was different from the training ones. This process was repeated for all the subjects (10 subjects), and the results of classifying actions are averaged among ten subjects.

The average recognition rate for LOSO test is 88.1%, for all activities together. From the confusion matrix in Figure 8, we observe that the classification errors occur if there are similarity among the actions, such as ”forward punch” and ”hand catch” or if the occlusion occur among human limbs which makes the skeleton tracker fails as in tennis swing action. Therefore, since the skeleton tracker sometimes fails and because of the high rate of similarity among the actions, we considered the recognition rate 88.1%, for twenty actions together is a good performance and comparable with other methods [1, 3]. Also, by considering the difficulties of the dataset and the noisy of 3D joint positions, we considered the result is reasonably and comparable with the state of art methods.



	high arm wave	horizontal arm wave	hammer	hand catch	forward punch	high throw	draw x	draw tick	draw circle	hand clap	two hand wave	side-boxing	bend	forward kick	side kick	jogging	tennis swing	tennis serve	golf swing	pickup & throw
high arm wave	90.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
horizontal arm wave	0.00	86.67	0.00	0.00	0.00	0.00	6.67	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hammer	0.00	0.00	93.33	0.00	3.33	0.00	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hand catch	0.00	0.00	0.00	83.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forward punch	0.00	0.00	10.00	0.00	73.33	6.67	3.33	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
high throw	0.00	0.00	3.33	0.00	3.33	90.00	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw x	0.00	3.33	6.67	0.00	6.67	3.33	80.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw tick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.67	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw circle	0.00	6.67	0.00	0.00	0.00	0.00	3.33	0.00	90.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hand clap	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	93.33	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00
two hand wave	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00
side-boxing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bend	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	90.00	0.00	0.00	3.33	0.00	0.00	3.33	0.00
forward kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
side kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	0.00	93.33	0.00	0.00	3.33	0.00
tennis swing	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00	83.33	0.00	3.33	0.00
tennis serve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	3.33	0.00	0.00	0.00	0.00	0.00	86.67	3.33	3.33
golf swing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	6.67	0.00	0.00	3.33	0.00	0.00	86.67	0.00
pickup & throw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	96.67

Figure 8.The results in a form of Confusion matrix using the orientation of distal limb segments as motion features with single kernel function (average recognition rate: 88.1%).

4. Conclusions and Future Work

This paper presents the development of a novel skeleton model for accurately extracting the distal segments of limbs of human in visual data for robust human action recognition. After the 3D human skeleton model creation, a six-element feature descriptor was extracted from the model in every video frame. The features describe a 3D spatio-temporal representation of 3D human joint positions (i.e. motion features using the end points of the distal limb segments extracted from a

3D human skeleton model. These features were trained and classified SVM classifier. Our system is able to recognize twenty different actions with an average recognition rate of 88.1%.

Clearly, there are some restrictions of the proposed approach for human action recognition. One limitation is the extraction of human silhouette. To build a robust system, a strong approach for extracting accurate human silhouette independent of environment needs to be developed. Another limitation is the viewing direction of the camera which is fixed in this work. In reality the camera view angle is not fixed and varies depending on the location of the person with respect to the camera. In the future, we plan to work on these limitations and develop a system that can track human body from different camera angles independent of environments.

References:

[1] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pp. 9–14, IEEE, 2010.

[2] Paul Scovanner, Saad Ali, Mubarak Shah, A 3-dimensional SIFT descriptor and its application to action recognition, in: Proceedings of the International Conference on Multimedia (MultiMedia’07), Augsburg, Germany, September 2007, pp. 357–360.

[3] L. Xia, C.-C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on,pp. 20–27, June.

[4] Ren, L. Shakhnarovich G., Hodgins K., Pfiste , and Viola P.”Learning silhouette features for control of human motion”. *ACM Trans. Graph. 24*(4), 1303–1331, 2005.

[5] Morrison, P.; JuJia Zou, "An effective skeletonization method based on adaptive selection of contour points," *Information Technology and Applications, ICITA 2005. Third International Conference on* , vol.1, no., pp. 644-649 vol.1, 4-7 July 2005

[6] M. Corporation, “Kinect for windows.” http://www.microsoft.com/en-us/kinectforwindows/, cited April 2013

[7] Niu F and Mottaleb M. “View invariant human activity recognition based on shape and motion features.” In *International Symposium on Multimedia Software Engineering ISMSE*, pages 546 – 556, Dec 2004.

[8] Niu F., Abdel-Mottaleb M., “HMM-Based Segmentation and Recognition of Human Activities from Video Sequences,” *IEEE Intern. Conf. on Multimedia and Expo (ICME)*, Amsterdam, Netherlands, pp.804-807, July, 2005.

[9] Yilmaz, A., Shah, M.: Actions As Objects: A Novel Action Representation. In: IEEE CVPR, San Diego, IEEE Computer Society Press, Los Alamitos (2005)

[10] Haritaoglu I, Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE transaction on Pattern Analysis and Machine Intelligence*, 22(8):809 –830, 2000.

[11] Vignola, J., Lalonde, J.-F. and Bergevin, R., Progressive human skeleton fitting. In: Proceedings of the 16th Vision Interface Conference, Halifax, Canada. pp. 35-42. 2003.

[12] Laptev I, and Lindeberg T. Space-time interest points, In ICCV, p. 432-439, 2003

[13] Haritaoglu I., Harwood, and L. Davis. Who,when, where, what: A real time system for detecting and tracking people. In *Proceedings of the 3th Face and Gesture Recognition Conf.*, pages 222–227, 1998.

[14] ChenHsuan-Sheng, Hua-Tsung Chen, Yi-Wen Chen and Suh-Yin Lee,“Human Action Recognition Using Star Skeleton,” in *Proc. 4th ACM international workshop on Video surveillance and sensor networks*, 2006, pp. 171-178.

[15] Althloothi S., Mahoor M., and Voyles R., “Fitting distal limb segments for accurate skeletonization in human action recognition”, Journal of Ambient Intelligence and Smart Environments, 2012.

[16] Gorelick L., Blank M., Shechtman E., Irani M., Basri R., “Actions as Space-Time Shapes,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247 – 2253, Nov. 2007.

[17] Dempster, A. P., Laird, N. M., and Rubin, D. B., “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B(Methodological)*, vol. 39, NO. 1, pp.1-38, 1997.

[18] Wang, L., Ning, H., Hu., W.: Fusion of Static and Dynamic Body Biometrics for Gait Recognition. In: International Conference on Computer Vision, Nice, France, pp. 1449–1454 (2003).

[19] Efros, A.A., Berg, A.C., Mori, G., Malik, J. Recognizing Action at a Distance. In: ICCV 2003. IEEE International Conference on Computer Vision, Nice, France, pp. 726–733.IEEE Computer Society Press, Los Alamitos (2003)

[20] David G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision (IJCV) 60 (2)(2004) 91–110.

[21] Poppe R. Vision-based human action recognition: A survey. Image and Vision Computing 28 976–990, 2010

[22] Moeslund T., HiltonA., and Krger V. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104:90–126, 2006.

[23] Gavrila D. M.. The visual analysis of human movement: A survey. Computer Vision and Image Understanding, 73:82–98, 1999.

[24] Aggarwal J. K. and CaiQ.. Human motion analysis: A review. Computer Vision and Image Understanding, 73:90–102, 1999.

[25] Althloothi S., Mahoor M., Zhang X., and Voyles R., 2014. Human activity recognition using multi-features and multiple kernel learning. Pattern recognition, 47(5), pp.1800-1812.

[26] Liang Wang, David Suter, Informative shape representations for human action recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR’06), vol. 2, Kowloon Tong, Hong Kong, August 2006, pp.1266–1269.

[27] Tran D**.** and SorokinA. Human activity recognition with metric learning. In European Conference on Computer Vision, 2008.

[28] Yu E. and Aggarwa J. K. l. Detection of fence climbing from monocular video. In *International Conference on PatternRecognition*, pages 375–378, Hong Kong, 2006.

[29] Elden Yu and J. K. Aggarwal, "Human Action Recognition with Extremities as Semantic Posture Representation", International Workshop on Semantic Learning and Applications in Multimedia (SLAM, in conjunction with CVPR), Miami, FL, June 2009.

[30] ChunS., HongK., and JungK., 3D star skeleton for fast human posture representation," in Proceedings of World Academy of Science, Engineering and Technology, vol. 34, pp. 273{282, October 2008.

[31] Ming-yu Chen and Alex Hauptmann, " MoSIFT: Reocgnizing Human Actions in Surveillance Videos ". CMU-CS-09-161, Carnegie Mellon University, 2009

[32] Wu, G., Mahoor, M.H., Althloothi, S., Voyles, R., “SIFT-Motion Estimation (SIFT-ME): A New Feature for Human Activity Recognition”, The 2010 International Conference on Image Processing, Computer Vision, and Pattern Recognition, Los Vegas, July 2010

[33] Althloothi, S., Voyles, R., Mahoor, M.H., Wu, G., “2D Human Skeleton Model from Monocular Video for Human Activity Recognition”, The 2010 International Conference on Image Processing, Computer Vision, and Pattern Recognition, Los Vegas, July 2010.

[34] Althloothi, S.,”Human Action Recognition Via FusedKinematic Structure and Surface Representation” (2013). ElectronicTheses and Dissertations. 27.