I-074

Understanding Concurrent Activities of Human in Daily Lives by Hierarchical Interpretation of Each Body Part

ロクマン ジュアンダ　　金子　正秀
Juanda Lokman†　Masahide Kaneko†

1. Introduction

Interpretation of human motion could be the most challenging problems in computer vision research area, because human body as non-rigid articulated object can yield many different poses from a simple movement/gesture to complex and complicated acrobatic motion. Human motion analysis has been applied to the wide spectrum of applications, such as man-machine interaction, surveillance, choreography, sports, medical science, content-based retrieval, video conferencing, etc.

Besides as non-rigid articulated objects, some parts of human body are usually occluded (self-occlusion or even occluded by other objects) while being seen from a single point of view. In daily lives, humans also usually wear different clothes (gender, seasons, fashions). Clothing yields another problem in computer vision, particularly for tracking and labeling human body parts.

Moeslund and Granum [1] divided a system for analyzing human body motion into initialization, tracking, pose estimation, and recognition. In this paper we only concern with recognition/interpretation of human motion based on angular pose of the major body parts (head, torso, shoulder, upper and lower arms, thigh and calf of the legs), while we assume that pose sequences for any activities are obtained from synthetic animated images.

Two typical approaches for interpretation of human motion exist depending on whether the knowledge (model) about the object is a priori known (model-based) or not (motion-based). In model-based recognition, people use motion information to construct geometry structure of the object for recognition, while motion-based recognition directly uses the motion information for recognition.

The major drawback of motion-based recognition is view-dependent. It depends on a specific point of view or with a slightly distinct angle. In contrast, geometric model for recognition may not depend on the point of view as long as the geometry structure of the model can be constructed. And also human activity can be perceived as a sequence of poses. This leads to model-based recognition.

Occlusion is the one of the main problems in motion analysis. Single camera or multiple camera schemes [6] were used to capture human motion or figure to resolve the problem of occlusion. But this did not guarantee the occlusion will not happen (e.g. for the case human is sitting behind a table).

Human can recognize or interpret the action of body parts that can only be seen. Besides that, human action does not always involve all the body parts, for example walking mainly involves only legs, pointing or waving only employs hand(s), etc. Human can perform multiple actions simultaneously as well, i.e., human can do walking while waving, and making a phone call with a mobile phone. Based on these facts and that human activity can be viewed as a sequence of poses [2], we propose a customizable model-based approach for human motion interpretation in daily lives.

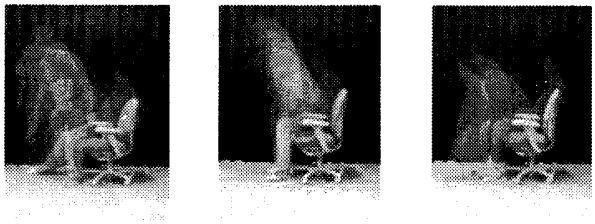


Fig.1: Variation of sitting.

In this paper, we also consider the variation of activities shown in Fig. 1, i.e., sitting on the chair (can be attained from standing pose as well as kneeling/reclining pose), raising hand (there are many ways to raise the hand).

In section 2 we describe the stick model that we use in our approach, section 3 addresses the body part analysis and interpretation, and section 4 will talk about the hierarchical interpretation scheme. And finally in section 5 we will show our experimental results.

2. Stick Model

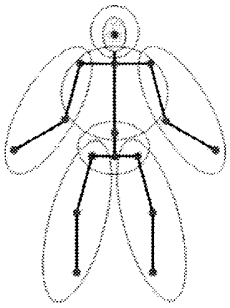


Fig. 2: Stick model and its clustering.

Figure 2 shows the 3D stick model that we use in this approach. There are 11 joints on this model and their degrees of freedom (DOF) are as follows:

- Neck (3 DOF)
- Shoulder (3DOF)
- Elbow (1 DOF)
- Upper Torso (2 DOF)
- Center Torso (1 DOF)
- Pelvis (2 DOF)
- Hip (3 DOF)
- Knee (1 DOF)

† 電気通信大学 大学院電気通信学研究科
The University of Electro-Communications

We assume that using this stick model and synthetic animated images from this stick model, we can get any pose of human activities. From this sequence of poses, the sequence of angular pose of each joint can be obtained, and later these angular parameters are going to be used in the analysis and interpretation of human motion. Suppose that each degree of freedom can rotate 360 degrees (forward/backward, lateral or rotation). Pose for an activity can slightly different from one person to another, even repetition of the activity from the same person as well. Thus, instead of using all the possible degrees $-180° < \mu \leq 180°$, we quantize the degree into several levels. In our case the interval Δ is 15°, thus, finally only 24 posible states of angular pose are considered.

3. Body Part Analysis and Interpretation

Usually in any kind of recognition including activity recognition, once the features (sequence of poses) are obtained, one has to compare the test sequence to a model. But, as we know that there are many variations in motion to perform just a single action, it is impossible to train the system for the entire variation.

Figure 3 shows the variation of pose for Raising Hand motion. The variations depend on the relative motion of the limbs and the speed. But if we see the movement of the pose as vector, each path is sum of small vectors. Thus give the same result (vector summation) for all paths. Now, each movement of the limbs, torso, etc. can be viewed as vector displacement. As a vector, the movement does not depend on how it moves but only the end of the movement relative to the initial position. Let's only consider joint movement (rotation) in forward or backward direction (a plane which spans by the vertical direction and facing direction).

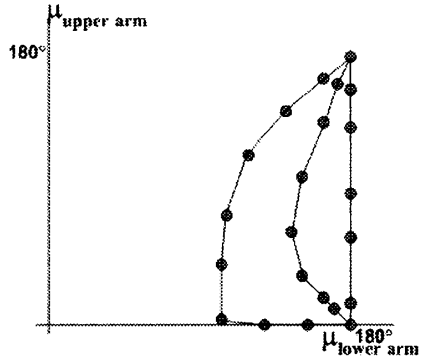


Fig. 3: Pose variation of raising hand.

Each posture can be accomplished by moving forward or backward. Each orientation from each joint has the weight table as shown in Table 1. In this table, we cluster again 24 possible states of pose into 12 clusters. Thus each orientation has 12 most likely postures (D_m). Each posture can be carried out by two directions, for example state s=3 for D_8 can be achieved from s = 1 or s = 2 as well as from s = 4~6. Using Table 1, we try to accumulate the pose that being through by the sequence. Values in Table 1 are used to make the accumulation as positive value if being through. The weight value of Table 1 and the following equations are used to calculate the most likely posture D_m (m = 0,...,11) from the sequence of pose.

$$D_m^{(JOINT)} = \sum_i \delta\, T_m(x_i)\big|_{(JOINT)}; m = 0,...,11 \quad (1)$$

$$\delta = \begin{cases} x_i - x_{i-1}, & i > 1 \\ 0, & i = 1 \end{cases} \quad (2)$$

Where x_i is a sequence of poses, T is table (Table 1) lookup as function of x_i.

Because it cannot have all possible variation samples to train the system, it needs to incorporate knowledge of how to accomplish the action.

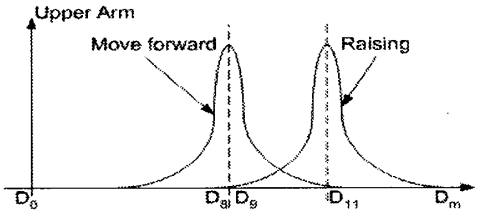


Fig. 4: Example of knowledge representation of micro actions as a graph.

Knowledge representation for each orientation of joint and each basic motion (part of action that can be used to build more complex action) can be written as a graph (see Fig. 4) in which value/weight of the graph represents the most required pose for that action. For example in Fig. 4 we can see that the most likely pose for raising the upper arm is depicted with normal distribution function.

4. Hierarchical Interpretation

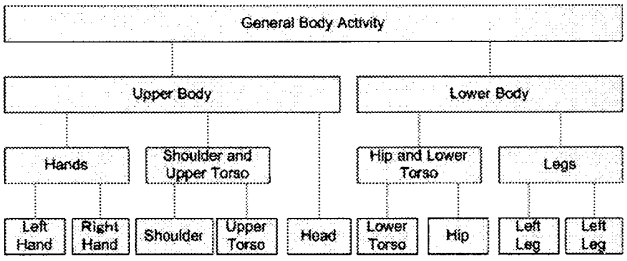


Fig. 5: Hierarchical interpretation.

As shown in Fig. 5, we analyze each body part separately at the low level, and combine them at higher level, thus make multilevel interpretation. For example, at the lowest level, we analyze and interpret only left/right hand, and then we combine and interpret both hands, and again at next level we combine them with upper body, and at the top most level as whole body motion interpretation. But we don't only focusing on general activities (such as walking, jumping, etc.) as a single output at the top level interpretation, each level of interpretation will give a recognition output as well. If there is no meaning for the motion of that body part, no output will be produced, but it will pass to the next higher level.

5. Experiments

We did an experiment for interpretation of some basic movements. The input of pose sequence is obtained from

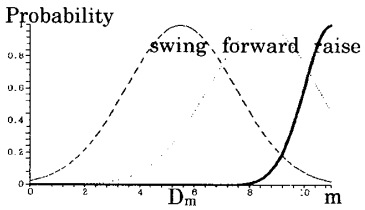
own-developed software by manually changing the pose of the human figure (character). From the pose sequences we already obtained, we trained and tested them for interpretation. If we apply the motion capturing technique, we can obtain a pose sequence from real moving images. However we have skipped this process here to concentrate on the interpretation of movement.

Our proposed method is based on the pose angle of each activity and the same activity cannot have the absolute pose angle, thus we represent the pose angular probability for each action as normal distribution as shown in Fig. 6. Normal distribution function (with $1/(2\pi)^{1/2}\sigma$ omitted) for each body part and kind of basic motion are shown in Figure 6(a, b, c). We also use this distribution as weighting factor to calculate the most likely posture D_m. We show the distribution as continuous one but actually we used discrete value of distribution.

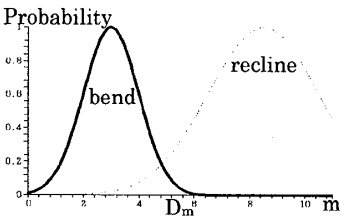
We calculate the most probable action for each body part (hand, torso, leg) and each basic motion (swing, forward, raise, etc.) using the following equation:

$$A_k^{(body\ parts)} = \sum_{(JOINT)} \sum_{m} D_m^{(JOINT)} N_k(\mu, \sigma) \tag{3}$$

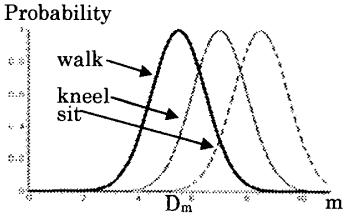
Where A is a discriminant function for basic motion k of each body part and m is the most likely posture that obtained from the pose sequence. N is weight distribution with mean and standard deviation as shown in Fig. 6. Each body part involves several joints that relate to the part, for example a hand includes elbow joint that links the upper-lower arm and shoulder joint that links upper arm to the shoulder.



(a) Distribution for hand



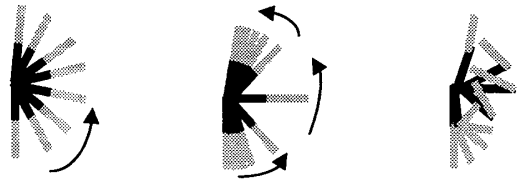
(b) Distribution for torso



(c) Distribution for leg

Fig. 6: Example of weight distribution for each macro action.

In Fig. 7, we show the proposed method to recognize several variant of the same activity with different speed and orientation (e.g. raising hand). We will show that the method is speed invariant. The black stick represents an upper arm while the grey stick represents a lower arm. Figure 7(a) shows the action of raising (right) hand with constant speed. Figure 7(b) shows the action of raising hand with variable speed (slow, fast and then slow again). Figure 7(c) shows the same action with variation of the limb movement. With these three examples, we try to show the proposed method works well as speed invariant and regardless of limb formation to achieve hand raising.



(a) (b) (c)

Fig. 7: Example of Raising Hand with different speed and variation of the limb.

Table 2 shows the result of recognition for three different raising hand motions shown in Fig. 7. Value in Table 2 is value of calculating a discriminant function according to Eq. (3). Each pose is calculated among the other basic motions. From Table 2, we can see only pose that relates to the action has the higher score. It means all the test pose sequences are recognized as the same "raising hand" activity.

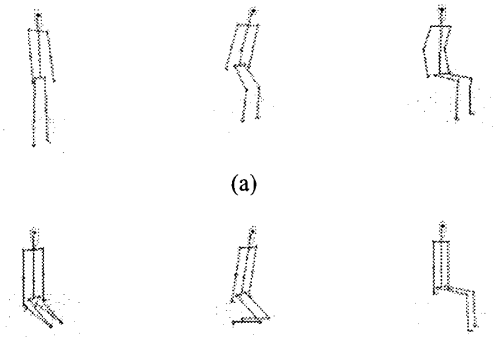
Table 2: Result of several hand raising.

	Raise	Forward	Swing
Case (a)	**300.89**	221.71	-77.19
Case (b)	**294.51**	138.17	-119.48
Case (c)	**373.94**	287.10	-0.74

Table 4 shows the result of recognition for several examples of activities such as raising hand, bowing, picking up, walking, sitting and kneeling. Some of the examples only show the activity of one part of body e.g. hand, torso or leg, and some show the activity that includes two parts of body e.g. "pick up" activity involves hand and torso. Walking or sitting can be viewed as an activity that involves hand and leg or just leg. In our multilevel interpretation framework, it will be viewed as activity of hand and leg as well as only leg, because human can do other activity simultaneously while walking. Table 4 shows the test data are correctly classified according to its basic motion of each body part. "Pick up" activity involves two body parts (hand and torso) as shown in Table 4, and this activity is accomplished with hand moving forward and bending of torso.

Figure 8 shows the proposed method works well on the variation of action regardless of the initial pose. For example "Sitting" activity, usually sitting is viewed as an action from standing pose to sitting down the chair. But as we know "sitting" activity can be started from any kind of previous poses e.g. standing as well as from kneeling, reclining, etc. to sitting down (on the chair). Figure 8 shows two examples of sitting activity

from two cases of initial poses (standing and kneeling). The sequence of pose in case (a) is that from standing position to sitting, while (b) shows that from kneeling to sitting. Table 3 shows the result of recognition for both cases. It shows that our method can recognize both of the activity as "sitting" regardless of the previous pose. This is because our methods analyze the activity based on the human pose particularly the last pose of the sequence (Goal of the activity can be seen on several last pose of human body).



(a)

(b)

Fig. 8: Motion variation of sitting activity from standing and kneeling as initial pose.

Table 3: Result of sitting activity from standing and kneeling.

	Walking	Running	Sitting	Kneeling
Case (a)	-65.66	130.2	**521.01**	266.92
Case (b)	123.93	234.62	**295.61**	-19.92

6. Conclusion

Human can perform very complex and several activities simultaneously. The proposed method works well for activity recognition regardless of speed and variation of motion. Our method to interpret human activity by each part of body can be used to recognize several human activities of each body part that can be seen. In the future, we will extend our method for interpreting several human activities that are performed simultaneously in daily lives scene.

Reference

[1] T. B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding*, Vol. 81, No. 3, pp. 231-268, March 2001.

[2] J. Ben-Arie, Z. Wang, P. Pandit and S. Rajaram, "Human Activity Recognition Using Multidimensional Indexing," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 8, pp. 1091-1104, August 2002.

[3] C. Cedras and M. Shah, "Motion-based Recognition: A Survey," *IEEE Proceeding on Image and Visual Computing*, Vol. 13, No. 2pp. 129-155, March 1995.

[4] M. Shah and R. Jain, "Motion-based Recognition," *Kluwer Academic Publisher*, Vol. 9, 1997.

[5] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, Vol. 73, No. 3, pp. 428-440, March 1999.

[6] Q. Cai and J. K. Aggarwal, "Tracking Human Motion Using Multiple Cameras," *Proceeding of Intl. Conf. on Pattern Recognition*, Vienna, Austria, pp. 68-72, August 1996.

Table 1: General weight table for one orientation of each joint.

D_m \ s	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
0	Sign(δ).2	-1										
1	1	Sign(δ).2	-1									
2	1		Sign(δ).2	-1								
3	1			Sign(δ).2	-1							
4	1				Sign(δ).2	-1						
5	1					Sign(δ).2						
6	0						Sign(δ).2	-1				
7							1	Sign(δ).2	-1			
8							1		Sign(δ).2	-1		
9							1			Sign(δ).2	-1	
10							1				Sign(δ).2	-1
11							1					Sign(δ).2

Table 4: Result of recognition for several activities.

Micro Action / Test Data	Hand			Torso		Leg		
	Raise	Forward	Swing	Bend	Recline	Walking	Sitting	Kneeling
Raise Hand	**300.89**	222.71	-77.19					
Bow				**169.60**	2.004			
Pick Up	194.36	**338.34**	47.38	**217.24**	18.77			
Walk	33.31	189.61	**337.73**			**808.95**	255.75	250.06
Sit						-65.65	**521.01**	266.92
Kneel						-83.20	365.77	**433.57**