Essential Body-Joint and Atomic Action Detection For Human Activity Recognition Using Longest Common Subsequence Algorithm

Sou-Young Jin and Ho-Jin Choi

Dept. of Computer Science, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea (South)

Abstract. We present an effective algorithm to detect essential body-joints and their corresponding atomic actions from a series of human activity data for efficient human activity recognition/classification. Our human activity data is captured by a RGB-D camera, i.e. Kinect, where human skeletons are detected and provided by the Kinect SDK. Unique in our approach is the novel encoding that can effectively convert skeleton data into a symbolic sequence representation which allows us to detect the essential atomic actions of different human activities through longest common subsequence extraction. Our experimental results show that, through atomic action detection, we can recognize human activity that consists of complicated actions. In addition, since our approach is "simple", our human activity recognition algorithm can be performed in real-time.

1 Introduction

Human activities are complicated. It often consists of many different atomic actions such as hand stretching, leg lifting, and/or head moving. Human activity recognition, on the other hand, tries to recognize high level semantic meaning of human actions such as walking, running, jumping. Majority of previous approaches [1–14] recognize human activity by considering different part of human actions as a whole to build a human activity classifier/model. Essentially, such approaches often fail when a human performs complicated actions that are largely deviated from the "normal" activities in their training data.

In this paper, in contrary to previous approaches that use a whole body actions to recognize human activities, we propose an alternative approach that recognizes human activity by detecting essential atomic actions of different human activity classes in training data. Our approach is motivated by the observation that only certain part of atomic actions of a human is sufficient to recognize a human activity. For example, to recognize a "walking" activity, we only require to recognize the body movement and the atomic actions of legs where actions of upper body of a human is irrelevant. This observation also applies to many common human activities such as hand waving, jumping and sitting.

To demonstrate our idea on human activity recognition, we develop an algorithm that works on the human skeleton data. The human skeleton data can

be captured by a RGB-D camera and/or by any other motion capturing system that can gather human skeleton data. For our convenience, we use the Kinect [15] and its SDK to get human skeleton data [16]. We collect many skeleton data of different activity classes, i.e. boxing, jumping, hand waving, sitting/standing. Each training example is labeled manually with the class label. Our goal is to detect the atomic actions within each class such that we can recognize activities by just detecting similar atomic actions in testing video. We assume essential atomic actions of each class are the actions that are repeated by themselves frequently (with variations) in training data.

Our algorithm starts by converting human skeleton data into symbolic sequence representation. The symbolic sequence representation can tolerate inter-class variations of atomic actions, which allows us to recognize human activity in a robust manner. In addition, since we assume essential atomic actions are repeated actions in training data, we convert our problem into a problem that finds the longest common subsequences in training data that are repeated by themselves within and among different training examples within the same activity class. In testing phase, the learnt longest common subsequences will be compared with the symbolic sequence representation of testing data to classify if a testing example belongs to an activity class defined in training phase. Since our recognition algorithm only involves a very simple operation for common subsequent string detection, our algorithm can recognize human activity in real-time. Our experimental results also show that our approach out-performs previous approaches that use the whole body actions for activity recognition.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 describes our algorithm in details. Section 4 demonstrates the effectiveness of our approach through experiments. Section 5 concludes our paper.

2 Related Works

There is a great deal of work targeting activity recognition. We discuss examples most relevant to our approach that recognize human activity using human pose information and refer the reader to the following recent articles by Poppe [17] and Aggarwal and Ryoo [18] for more thorough surveys.

With an adventure of Microsoft Kinect [15], state-of-the-art pose estimation algorithm [16] can predict human pose in terms of 3D body-joint positions from a single depth image. Such information is very useful in human activity recognition. In Sung et al. [19], they use the pose information from Kinect to extract features such as body pose features, hand position and motion information for activity recognition. Their body pose features compose of 10 torso related body-joint orientations, such as relative position between foot/hand and torso, relative position between head and hip, and relative positions between head and torso. For each consecutive frame in a video, they calculate the changes of these body-joint angles and train an activity recognizer using hierarchical maximum-entropy Markov models for each labeled action.

Similar to Sung et al. [19], Tran et al. [20] also use Kinect data for activity recognition. They introduced a polar space which measures the distance and angle orientations of different body part from the center of a body. A 2D histogram is created to capture the frequency of body parts being observed at each different quantized locations. A classifier is trained using the 2D histogram as features to recognize different human activity. While both Sung et al.[19] and Tran et al. [20] demonstrated some successes in human activity recognition using Kinect data, these two approaches consider human activity as a whole body actions. When there is an "abnormal" activity, such as a person is waving hand during walking, these approaches often produce less than satisfactory results.

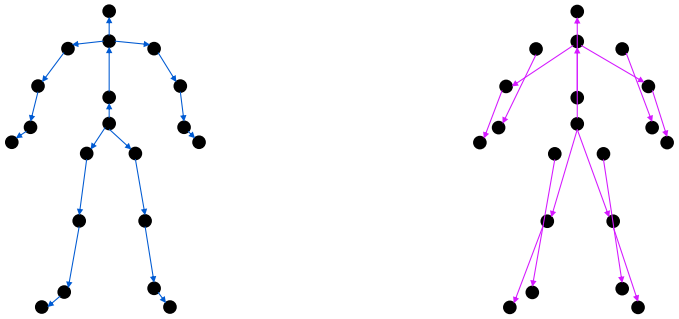
Meanwhile, there are also approaches that try to recognize human activity by considering the importance of different body parts. Ryoo and Aggarwal [21] proposes a description-based approach for activity recognition. Their basic idea is to use a context-free grammar (CFG) to represent and encode the hierarchical structure of human activity. In each frame, the system extracts poses and gestures for each body part: head, upper-body, lower-body, and hand position. If all of required poses and gestures are recognized and the time intervals for them satisfy the relationships described in the representation, the system deduces the action is occurred. Unlike other previous works, Ryoo and Aggarwal treat body parts separately so that poses and gestures are recognized in each body part. However, in their training process, it requires manual encoding on poses or gestures for each body part as well as the time intervals for each atomic action using their CFG syntax.

Chakraborty et al. [22] propose the ensemble of body-part detectors using hidden Markov model. Similar to our motivation, they observe that not all body-parts contribute equally to all action classes. Thus, their approach is especially robust in distinguishing between similar actions since it only considers the certain body-parts that has major contribution to the actions. Nevertheless, their approach still requires manual labeling to characterize different body-parts for different human activity class.

Comparing our work with previous works in [21] and [22], our approach only requires manual label of different activity class. It is automatic in detecting essential body-joints and atomic actions that are necessary for activity recognition. Moreover, our symbolic sequence representation is robust in body-joint angle orientation and it is efficient in activity recognition after training.

3 Algorithm

In this section, we present our algorithm for activity recognition from human skeleton data. We will first present our representation on how to convert skeleton data into our symbolic sequence representation. Next, we will present our main algorithm on how to learn essential body-joint and atomic action from training data using the Longest Common Subsequence (LCS) algorithm [23, 24]. Finally, we describe how we can detect human activity via sequence matching.



(a)1st order (b)2nd order

Fig. 1. The skeleton data from Kinect SDK contains 19 different body-joint vectors. Each body-joint vector captures the vector orientation from one joint to another joint. On top of the first order neighborhood provided by the SDK, we defined a second order neighborhood that connects adjacent vectors. The body-joint angle of second order vector can be easily computed from the body-joint angles of the first order vector. In total, we have 19 first order vectors and 14 second order vectors which capture human pose information at time t.

3.1 From Skeleton to Symbolic Representation

The skeleton data from Kinect SDK has 19 different body-joint vectors and the center location of human body as illustrated in Figure 1(a). These body-joint vectors encode the vector orientations of different body parts from one body-joint to another body-joint in 3D world-coordinate. Since human actions usually contain different amount of variations, using the original body-joint angles for training can be sensitive. Here, we propose to convert these body-joint angles into a symbolic representation by quantizing the angles from continuous domain into a discrete symbol as illustrated in Figure 2. For each body-joint angle in 3D world coordinate, we project the angle onto xy-plane and yz-plane respectively. The projected angles are then quantized into eight discrete symbols. Hence, we quantize body-joint angles in 3D world coordinate into 64 discrete symbols which is robust enough to encode human action while tolerates small variations across different examples of the same activity. Figure 3 shows an example on how to encode the body-joints angles of left shoulder, left elbow, left wrist, and left hand, into symbols for the left hand waving action at time t and time $t+1$.

Besides body-joint angles, we also need to consider the change of human body location as it also provides very useful information in activity recognition. Similar to body-joint angle, we measure the location difference of "hip center" body-joint between the current and the next frame and then convert the movement vector into our symbolic representation. Figure 4 illustrates our process. We quantize the location difference similar to body-joint angle. After that, we further quantize the magnitude of movement vector to "no movement", "small movement" and "large movement" by two thresholds. The thresholds were set empirically. In

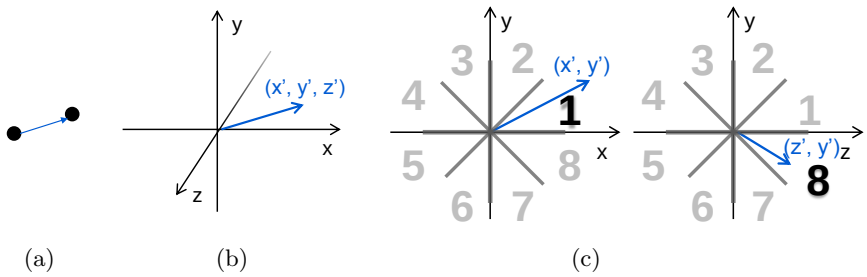
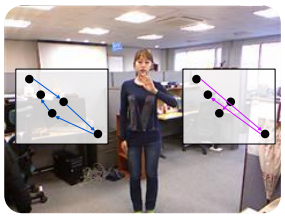
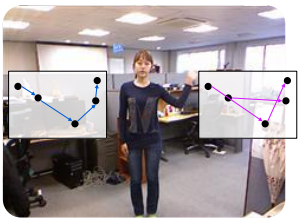


Fig. 2. We convert a continuous angle in 3D space into discrete symbol according the projected angles in xy-plane and yz-plane.





	shoulder	elbow	wrist	hand
1st order adjacent vector	**76**	**75**	**44**	**44**
2nd order adjacent vector		**75**	**55**	**43**

shoulder	elbow	wrist	hand
77	**76**	**13**	**22**
	76	**84**	**13**

Fig. 3. Body-joint symbols for one item of data containing a "waving" action.

our experiments, we found that setting the first threshold as half of average movement in training data, and the second threshold as three times larger than the first threshold produce good recognition accuracy.

After converting skeleton data into our symbolic representation, we obtain 34 sequences where 19 sequences corresponding to the first order body-joint angles, 14 sequences corresponding to the second order body joint angles, and 1 sequences corresponding to human body movement. Since we are interested in "active" actions, we remove symbols that are identical in consecutive frame as illustrated in Figure 5. The compact representation of motion sequences allows us to focus on only the moving part of human activity for training and testing.

3.2 Essential Body-Joints and Atomic Action Detection

From the symbolic representation of motion sequences, we want to detect the essential body-joints and the atomic action that can be used for activity recognition. We define the essential body-joints to be the body-joint vectors that are necessary and sufficient for recognizing certain activity. For example, to recognize a "walking" activity, we only require the body-joint vectors of hips and

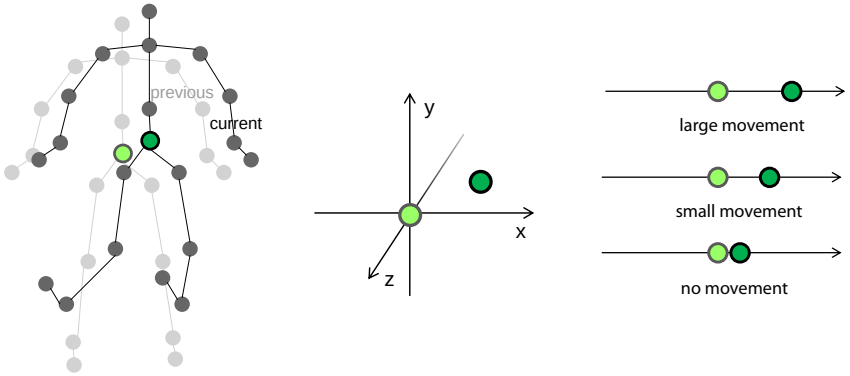


Fig. 4. (a) We compute the movement vector by measuring the location difference of "hip center" body-joint between consecutive frames. (b) The movement vector is converted into symbolic representation according to the movement vector direction. (c) We additionally add another symbol to encode magnitude of the movement vector.

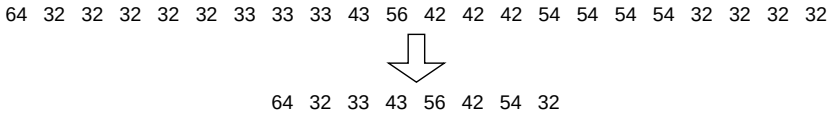


Fig. 5. We remove identical symbols in consecutive frames to adopt a compact representation of motion sequences.

legs and the human body movement vector. The body-joint vectors of head and hands are less important or even irrelevant. This observation also applies to many other human activity that only involves certain parts of body actions but not the whole body actions. We define the atomic action to be the pattern of actions of essential body-joints that can be used to recognize certain activity. Using the "walking" activity as example, this activity involves legs lifting and legs stretching to follow certain sequences. By detecting atomic actions of essential body joints, one can evaluate and recognize if a person has performed certain activity or not.

We adopt the Longest Common Subsequence (LCS) algorithm [23, 24] to serve our purpose. Among the labeled training data, we compare the current longest common subsequence of each body-joint to the symbolic motion sequence of new data of the same body joint:

$$atomic(a,i) = \begin{cases} LCS(atomic(a,i-1), i\text{-}th\ example) & \text{if } i\ > 0 \\ i\text{-}th\ example & \text{if } i\ = 0 \end{cases} . \quad (1)$$

where a is index of body-joint, i is index of training examples, and $atomic(a, i)$ is the atomic action of body-joint a learnt from the first i-th training examples. This process continues until we process all training examples. After performing the LCS algorithm, some body-joints may have no common sub-sequences. We discard these body-joints as we believe these are irrelevant body-joint. Accordingly, we consider a body-joint is essential only if the learnt atomic actions have three or more symbols in its sequences.

During the recognition phase, we build a finite state machine for each atomic action of essential body-joint in each training category. When a new testing sequence comes, the skeleton data can be converted into our symbolic representation in real-time and then verified by the finite state machine. The benefit of finite state machine is that it can effectively filter out consecutive identity symbols in testing caused by slow motion of activity. Hence, our algorithm is robust to the speed variations of human activity.

4 Experimental Results

We evaluate our algorithm in this section using real-world examples. We will first describe our training data. Next, we will evaluate if our algorithm can successfully detect essential body-joints by comparing with ground truth essential body-joints. Finally, we test our learnt essential body-joints and atomic actions to recognize human activity with challenging examples.

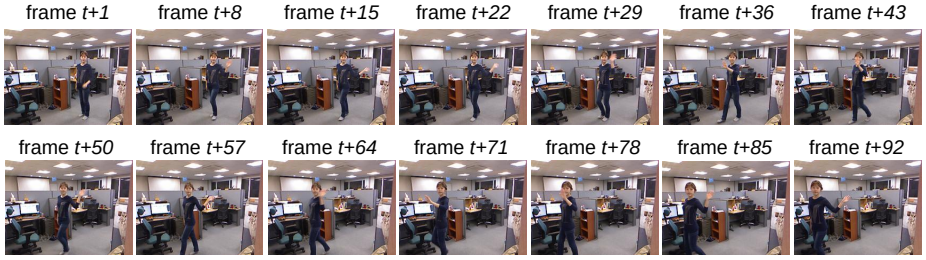
4.1 Training Data

We build our own training data by using Kinect to capture real-world examples. Our training data consists of four different activity categories: "boxing", "jumping", "hand waving" and "sitting/standing". For each activity category, we collect two different type of examples: Single Action (SA) examples and Composite Action (CA) examples. In the single action training data, only the essential body-joints are moving while the other body-joints are static. In composition action training data, not only the essential body-joints are moving but the other body-joints are also free for moving. Figure 6 shows examples of single action and composite action training data. We assign label manually to each training example. For composite action training examples that consist of multiple action categories, multiple labels has been given. Within each category, we collect 16 training examples with different persons performing the same activity.

4.2 Evaluation on the detected essential body-joints

Our first experiment evaluates the effectiveness of our method on essential body-joints detection. We detect the essential body-joints for each category using SA and CA training data separately. Figure 7 shows our detected essential body-joints for different activity category. The skeletons on the left hand side of each category are the results learnt from SA training data, and the skeletons on the

(a) Single Action (SA) example of a "hand waving" action.



(b) Composite Action (CA) example that has both "hand waving" and "walking" actions.

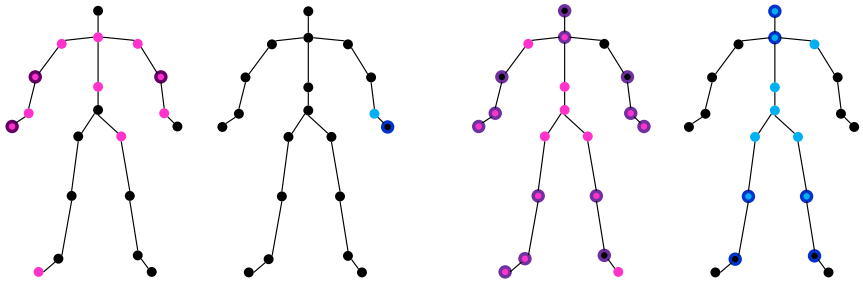
Fig. 6. Examples of our training data.

right hand side of each category are the results learnt from SA training data. The detected essential body-joints were highlighted.

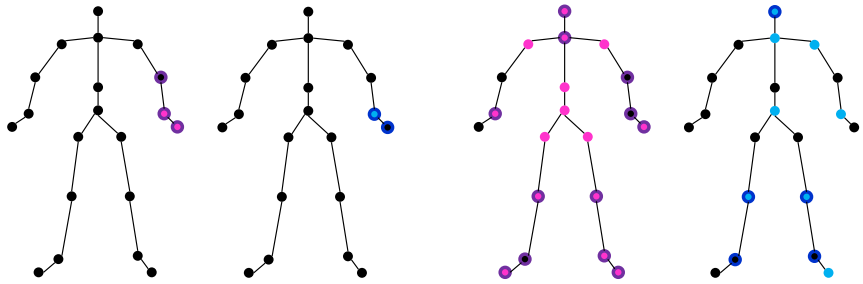
With the SA training data, not only the essential body-joints were detected. Some inessential body-joints were also detected. For example, the body-joints related to two arms should not be detected as essential for a "jumping" action. However, the system identifies some body-joints around arms since the subjects habitually move their arms to perform a "jumping" action. When the system uses the CA training data, the result of essential body-joint detection is better in most action categories. However, although both left and right arms are essential for a "boxing" action, only two body-joints in the right arm - 'wrist right' and 'hand right' - are detected as essential. Though not all essential body-joints are selected for a "boxing" action, the system is able to recognize the action with only these two body-joints. Table 1 shows the learnt atomic action sequences for the "boxing" action and the "hand waving" action. Both wrist and hand were detected as essential body-joints, yet our learnt atomic action sequences are different. Our algorithm can successfully detect different body-joints and atomic actions for different activity category and the "correct" essential body-joints in each activity were included.

4.3 Activity Recognition Accuracy

We evaluate and compare our activity recognition accuracy with approach from Tran et al. [20]. In [20], Tran et al. describe video data by a polar histogram



(a) "boxing" action (b) "jumping" action



(c) "hand waving" action (d) "sitting/standing" action

Fig. 7. Results of essential body-joint detection. Every left skeleton indicates the detected body-joints when the SA dataset is used while every right skeleton is associated with those when the CA dataset is used. A body-joint is filled with a color, if the body-joint's 1st order adjacent vector is selected. In addition, a body-joint's edge is colored if the 2nd order adjacent vector is selected.

Table 1. Comparison of the common sequences between a "boxing" action and a "waving" action. Although the detected essential body-joints are similar, atomic action sequences for each selected body-joint differ from each action.

Body joint		When the system detects essential body joints using CA training data	
Name	Adjacent order	"boxing" action	"hand waving" action
wrist right	1st	**44 55 44**	**33 23**
hand right	2nd	**44 55 44**	**23 33**
wrist right	2nd		**14 34 24**

which each histogram bin represents the frequency of certain body-parts appeared in the histogram bin area. We call this approach a Whole Body-Joint (WBJ) approach since they use all available body-part actions to build their polar histogram for activity recognition. While our approach only uses the detected essential body-joints for activity recognition.

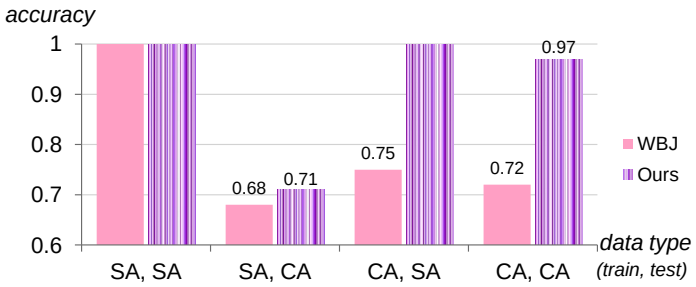


Fig. 8. Activity recognition accuracy. Our approach out-perform WBJ approach when the training/testing data contains composite action.

Similar to previous experiments, we again separate the SA and CA data to train and test the activity recognition accuracy of our approach and the WBJ approach. Note that we adopt the leave-one-out cross validation. Figure 8 compares the average accuracies. In the axis of data type, the first letter is related to the type of training data while the second is associated with that of test data. As we can observe from Figure 8, when training or test data contains composite actions, our approach performs much better than the WBJ approach.

Meanwhile, we have also observed that our approach does not work well when training data is SA but the testing data type is CA. This is because some of inessential body-joints were detected and were used to recognize activity. This result also confirms that essential body-joint detection is necessary for activity recognition. Consequently, our approach performs better when using CA data for training.

Fig. 9. We show seven frames from the CA training data where the ground truth labels are both "hand waving" and "sitting/standing". WBJ approach fails this example, while our approach can still recognize both labels.

Figure 9 shows a challenging testing example. This example contains both "hand waving" and "sitting/standing" activity simultaneously. WBJ approach fails to recognize the labels of this example since their approach uses a whole body actions to recognize activity which is not applicable to this challenging composite action. On the other hand, our approach only uses essential body joints to recognize activity. Both "hand waving" and "sitting/standing" labels were successfully detected using our proposed algorithm.

5 Conclusion

We have presented a simple yet effective algorithm to automatically detect essential body-joints and atomic action sequences for human activity recognition. To handle within class activity variations, we convert continuous skeleton data into discrete symbolic representation which can tolerate small variations of activity performed by different people. After that, we detect essential body-joints and atomic action sequence from training data by extracting the longest common subsequence among the activity examples within the same activity category. In recognition phase, we build a finite state machine for each learnt atomic action sequences of detected essential body-joints to recognize human activity in real-time.

While majority of previous approaches focus on whole body actions of human activity, our approach focuses on the atomic actions of essential body-joints. The experimental results show that our observation that only essential part of body actions is necessary and sufficient for activity recognition is valid. With essential body-joint detection, our approach can outperform recent WBJ activity recognition algorithm on challenging examples which contain composite actions with multiple labels. Our current approach assumes the training data does not contains any noise or outliers, e.g. wrong labels. Such outliers can potentially damage our atomic action detection algorithm. In the future, we shall study how to incorporate probability model into our algorithm to increase the robustness for detecting atomic action sequences. We shall also study how to incorporate some advanced feature selection techniques into our algorithm for essential body-joint detection.

Acknowledgement. This work was supported by the National Research Foundation (NRF) grant (No. 2012-0001001) of Ministry of Education, Science and Technology (MEST) of Korea.

References

1. Wang, L., David Suster, D.: Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
2. Weinland, D., Boyer, E.: Action recognition using exemplar-based embedding. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–7
3. Wang, Y., Huang, K., Tan, T.: Human activity recognition based on r transform. In: Workshop of IEEE Conference on Computer Vision and Pattern Recognition for Visual Surveillance. (2007) 1–8
4. Souvenir, R., Babbs, J.: Learning the viewpoint manifold for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–7
5. Wang, L., Suter, D.: Informative shape representations for human action recognition. In: International Conference on Pattern Recognition. (2006) 1266–1269
6. Huang, F., Xu, G.: Viewpoint insensitive action recognition using envelop shape. In: Asian Conference on Computer Vision. (2007) 477–486

7. Cherla, S., Kulkarni, K., Kale, A., Ramasubramanian, V.: Towards fast, view-invariant human action recognition. In: Workshop of IEEE Conference on Computer Vision and Pattern Recognition for Human Communicative Behaviour Analysis. (2008) 1–8
8. Souvenir, R., Babbs, J.: Learning the viewpoint manifold for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–7
9. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104** (2006) 249–257
10. Huang, W., Wu, Q.M.J.: Human action recognition based on self organizing map. In: IEEE International Conference on Acoustics Speech and Signal Processing. (2010) 2130–2133
11. Ahmad, M., Lee, S.W.: Variable silhouette energy image representations for recognizing human actions. Image and Vision Computing **28** (2010) 814–824
12. Abdelkader, M.F., Abd-Almageed, W., Srivastava, A.: Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. Computer Vision and Image Understanding **115** (2011) 439–455
13. Jia, K., Yeung, D.Y.: Human action recognition using local spatio-temporal discriminant embedding. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
14. Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as spatio-temporal shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2247–2253
15. Microsoft, C.: Kinect for xbox 360 (2010)
16. Shotton, J., Fitzgibbon, A., Cook, M., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011)
17. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28** (2010) 976–990
18. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Computing Surveys **43** (2011) 1–43
19. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: IEEE International Conference on Robotics and Automation. (2012) 842–849
20. Tran, K., Kakadiaris, I., S.K.Shah: Part-based motion descriptor images for human action recognition. Pattern Recognition **45** (2012) 2562–2572
21. Ryoo, M., Aggarwal, J.: Semantic representation and recognition of continued and recursive human activities. International Journal of Computer Vision **82** (2009) 1–24
22. Chakraborty, B., Bagdanov, A.D., Gonzalez, J., Roca, X.: Human action recognition using an ensemble of body-part detectors. Expert System (2011)
23. Hirschberg, D.S.: Algorithms for the longest common subsequence problem. Journal of the ACM **24** (1977) 664–675
24. Bergroth, L., Hakonen, H., Raita, T.: A survey of longest common subsequence algorithms. In: 7th International Symposium on String Processing and Information Retrieval. (2000) 39–48