Local Shape Context Based Real-time Endpoint Body Part Detection and Identification from Depth Images

Zhenning Li and Dana Kulić
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
z237li@uwaterloo.ca, dkulic@ece.uwaterloo.ca

***Abstract*—For many human-robot interaction applications, accurate localization of the human, and in particular the endpoints such as the head, hands and feet, is crucial. In this paper, we propose a new Local Shape Context Descriptor specifically for describing the shape features of the endpoint body parts. The descriptor is computed from edge images obtained from depth data generated by a time-of-flight sensor. The proposed descriptor encodes the distance from a reference point to the nearest edges in uniformly sampled radial directions. Based on this descriptor, a new type of interest point is defined, and a hierarchical algorithm for searching good interest points is developed. The interest points are then classified as head, feet, hands and others based on learned models. The system is computationally efficient, and capable of handling large variations in translation, rotation, scaling and deformation of the body parts. The system is tested using videos containing a variety of motions from a publicly available dataset, and is shown to be capable of detecting and identifying endpoint body parts accurately at very high speed.**

***Keywords*-local shape context; endpoint body part; depth image; human motion capture; gesture recognition;**

I. Introduction

Human body movement conveys extensive information regarding the human's actions and intentions. Computer vision based human motion capture and gesture recognition extract this information in a non-contact way, enabling a more natural and intuitive interaction between humans and machines.

To ensure accurate localization of the human posture, a fast method to locate the head, hands and feet from sensor data is very important. In marker-less full body human motion capture, the localization of these extremity body parts provides a very strong cue for the inference of the full body configuration [1], [2]. Knowledge of their locations reduces the search space significantly, helping to deal with the biggest difficulty in full body tracking: the high dimensionality problem. When tracking of the target is lost, the locations of these body parts also provide important information in helping the tracker to recover from incorrect configurations [3], [4]. In addition to being used as cues for whole body tracking, the detected locations can be used directly for gesture recognition to enable gesture based human-machine interaction. Because this class of body parts are the endpoints of a human body and have special importance, we name them as ”endpoint body parts”.

In detecting the locations of the endpoint body parts, processing speed is crucial. Whether the end goal is motion capture, gesture recognition or other applications, the endpoint locations are usually used as input information for the higher-level tracking or recognition, and this higher-level processing is also often required to be at frame rate. In this paper, we will propose an accurate and very efficient method to detect and identify these body parts, as well as their 3D orientations.

If we take a close look at the endpoint body parts, we can discover that they actually have very distinguishable 2D shape features. On the one hand, a common feature shared by all the endpoint body parts is that they are connected only to one other body part, and this makes their interior contour shapes open only on one side. On the other hand, different endpoint body part categories (head, hand and foot) have their own characteristic shape features that are distinguishable from the others. By applying a good shape descriptor on an image with clear body part shapes, we can unify the detection task and identification task within one framework.

The first question is how to obtain images with high quality body part shapes, described by the contour edges of the body parts. These edges can be both on the boundary of the human silhouette and inside the silhouette. When using traditional color or monochrome cameras, the shape edges are often cluttered by other edges, such as the edges due to clothing pattern. Even the high quality segmentation of the human silhouette is a very difficult task. Moveover, the environment conditions can affect the quality of the images significantly, for instance lighting changes. To avoid these issues, we propose an approach that uses the depth images generated from a Time-of-flight (TOF) sensor, which preserves most of the shape edges, while removing most of the cluttering edges.

The next question is what a good shape descriptor is for this application. Inspired by the shape context concept proposed by Belongie et. al. [5], we propose a novel Local Shape Context (LSC) Descriptor specifically for describing the shape features of body parts. This descriptor describes the local shape of different body parts with respect to a given reference point on the human silhouette, and is shown to be very effective at discriminating and classifying endpoint body parts, as well as being computationally efficient. A new type of interest point is defined based on the LSC Descriptor, which is called Interest Reference Point (IR Point). A hierarchical IR Point selection algorithm is designed to further conserve computational resources. The detected endpoint body parts are then classified according to learned models of each class based on the LSC feature. The proposed technique is tested using a publicly available dataset [6] and achieves around 85 percent precision and recall for each class at a speed of 100Hz on a Quad Core PC.

The paper is structured as follows. In Section II, existing approaches for body part detection are reviewed. Next, we propose our LSC Descriptor based approach for localizing the body parts in Section III. We show the experimental results in Section IV, and provide both a quantitative and qualitative analysis of the system performance. Further discussion can be found in Section V, and the paper is concluded in Section VI.

II. RELATED WORK

Significant effort has been devoted to the detection and identification of body parts. Some researchers consider body part identification as a standalone problem, while others integrate it as a subsystem in body motion recovery. The most commonly used approach is based on the assumption of skin color, such as in [7] and [8]. The body parts are first detected using skin color segmentation, and then additional heuristics are used to obtain a robust result. Obviously, these methods make a strong assumption about the appearance of the body part to be detected, which is not necessarily satisfied in a real application.

An interesting approach that explores more general features of body parts is proposed by Haritaoglu et al. [9], who develop a body part labeling system using silhouettes only. The system is based on two observations: 1. The head, hands, elbows, feet and knees are more likely to be found on the silhouette boundary. 2. The human body in any given posture has a topology structure which constrains the relative locations of body parts. They first classify the human motion into predefined posture categories, and then apply convex hull analysis to label the body parts according to the prior knowledge of that posture. However, this method is constrained to the known posture classes. Also, when the body parts to be labeled are not on the silhouette boundary, the algorithm will fail.

Wu et al. propose an edgelet detector in [10], which makes use of the silhouette oriented information. An edgelet is a short segment of a line or curve, and a set of edgelets can represent the shape of a body part. The system is trained using the edgelet features for the head and shoulder contour, torso contour and legs contour, and these body parts are searched for within an image. A joint likelihood is then formed by combining the responses from the body part detectors, and used to detect the full body. However, in this work, the postures are constrained to walking and standing, limiting the generality. Considering additional postures, e.g. squatting, will result in edgelet features that can be very different. Furthermore, between straight standing and deep squatting, there are a myriad of continuous postures with varying edgelet features. If the system is trained using a greater range of postures, the detector performance is likely to degrade.

Recently, depth sensors that can produce high quality depth images have gained increasing attention. Based on the depth images obtained from a TOF sensor, Plagemann et al. developed an endpoint body parts identification and localization system [11]. They define a new type of interest point in 3D, a geodesic extremum on the surface mesh of a human subject. This definition is based on the fact that the geodesic distances from the endpoint body parts to the center of the body are usually the longest. After the interest points are found, a boosted patch classifier is applied to the patches around each interest point for identification. This system explores an essential feature of the endpoint body parts, but forming the surface mesh and searching for interest points are computationally demanding, requiring 60 ms to process each frame.

Compared to the previous work, our proposed method is more general and more efficient. It makes no assumption of specific color or specific motion, and achieves a frame rate of 100Hz with accurate detection and classification.

III. APPROACH

A. Local Shape Context

In [5], Belongie et al. propose a shape context descriptor for measuring shape similarity. The idea is that among the points on the shape contour, the vectors from a reference point to all the other points contain rich global shape information with respect to that point. Assuming the edges that describe the shape have already been found and sampled into a number of points, a histogram that statistically summarizes the distribution of all these vectors is defined as the shape context for the reference point. The shape of the entire contour is described by the shape contexts of all the sample points. This descriptor is shown to be very effective in object recognition, and it is inherently invariable to translation and small deformation. In addition, it can also be made invariant to rotation and scaling. Later, Mori et al. applied the shape context descriptor to estimate human body configurations [12], but the need for exemplars severely limits its application. Also, no processing speed is reported.

When considering body part detection and identification, however, the original shape context descriptor is not suitable. The biggest problem is that this descriptor encapsulates the global shape information at each point, but since a human body is a highly articulated object, the global shape and the location of the endpoint body parts can vary tremendously. Even if a localized, window-based version of the shape context descriptor is used, it is difficult to exclude extraneous edges arising from adjacent body parts in a window. Furthermore, although the calculation of the shape context is not time consuming, the matching of the points can be very slow. The total cost of the matching for every point to every point must be minimized by comparing 2D histograms, and the computation cost can be high.

Inspired by the shape context concept, we propose a new shape descriptor for body part detection and identification, which we name the Local Shape Context (LSC) Descriptor. This descriptor is defined with respect to a reference point within the human silhouette, and it describes the local shape around that point. A LSC Descriptor is composed of a set of vectors from the reference point to the points on the nearest edges in radial sampled directions (Fig. 1). If the edge is not found within a certain range in one direction, the distance will be made equal to that range. A corresponding LSC feature is defined as one vector consisting of all the lengths from the reference point to the nearest edge along each radial line in clockwise order. It does not matter which direction is used as the starting point.

This descriptor contains rich shape information about the local area, while excluding the extraneous edges of other body parts effectively. More importantly, this descriptor is concise. Compared to the previous shape context, the LSC Descriptor of one single point is sufficient to describe a local shape, if the reference point is well selected and the directions are sampled densely enough. Obviously, this descriptor is invariant to translation. In the following, we will show the effectiveness of applying this descriptor for detecting and identifying endpoint body parts, and discuss the ways to make it invariant to rotation, scaling and deformation.

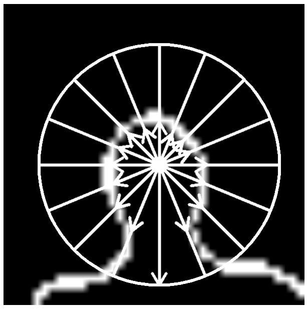
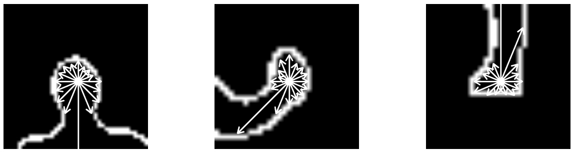


Figure 1. LSC Descriptor Illustrated on a Head

B. Hierarchical Interest Reference Point Selection

As described in the previous section, an LSC Descriptor is defined with respect to a reference point. A naive approach would be to sample through the entire human silhouette and compute the LSC Descriptor for each sampled point, and classify them into different categories. However, in this case, in order to ensure a good detection performance, the sampling would have to be dense globally. In addition, the computation cost of the classification would also be high, because every sampled point would need to be evaluated. To avoid these problems, a two layered hierarchical Interest Reference (IR)Point selection algorithm is used to improve efficiency.

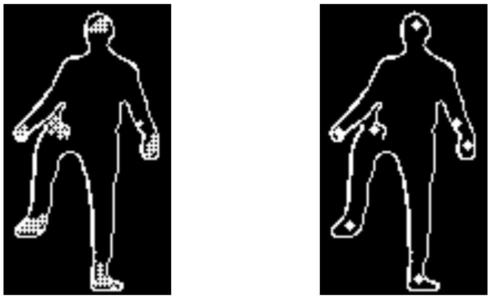
Based on the LSC concept, we define two tiers of IR Points. Tier 1 IR Points are selected based on the fact that endpoint body parts have common shape features distinguishing them from most of the non-endpoint body parts. At a point close to an endpoint body part inside the body, there is a convex edge contour which is open only on one side within certain angle range (Fig. 2). This means that, when we try to detect edges in a set of sampled directions, edges will be detected in most directions, but will not be detected in several continuous directions within a narrow range. We develop an algorithm to identify such points, these are selected as Tier 1 IR points. Tier 2 IR Points are then formed by clustering Tier 1 points according to Euclidian distance in the 2D plane.



2.1: Head 2.2: Hand 2.3: Foot

Figure 2. LSC Descriptor at Endpoint Body Parts

1) Tier 1 Interest Reference Point Selection: We first sample over the human silhouette uniformly to identify Tier 1 IR points. Because we just need to know whether there is an edge in a sampled direction, instead of using distance, we only use a binary value to indicate whether an edge is found within the search range or not. To further simplify the calculation, the radial sampling density is sparse, and the search range is relatively small. Note that 4 connection should be used if the edges could be thin. The LSC feature is thus a vector of binary values, which is actually a ring rather than a line. This means the last element in this array is connected to the first one. On this ring, we detect and count the falling edges, which are the changes from 1 to 0. Reference points with more than one falling edge are definitely not a candidate for an endpoint body part, so they are removed. Then, for the points with only one falling edge, we count the continuous 0s between the falling edge and the raising edge. The points with more than 3 continuous 0s are discarded. The remaining sampling points are the Tier 1 IR Points, shown in Fig. 3.1. Notice that they are located mainly at the endpoint body parts, but some are also located at other locations which meet the criterion. In the figure, the human subject lifts up his right leg, so a convex contour is formed near the hip, resulting in Tier 1 IR points. Tier 1 IR Points are also found on the left forearm, because the special orientation of the forearm, the slightly bent wrist and the sparse radial sampling of the descriptor.


3.1: Tier1

3.2: Tier2

Figure 3. Result for Tier 1 IR Points and Tier 2 IR Points

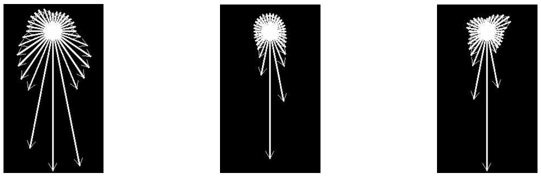
2) Tier2 Interest Reference Point Formation: Once Tier 1 IR Points are identified, clustering is performed to form Tier 2 IR Points. Since some endpoint body parts may not be detectable, and some other body parts may appear to be similar to endpoint body parts, the number of clusters is unknown. Thus the commonly used K means clustering is not suitable [13]. We use a hierarchical clustering algorithm in our system, specifically, the agglomerative clustering algorithm [14]. The clustering starts by assigning each point to an individual cluster. At each step, the two closest clusters are merged to form a new one. This continues until all the points are in one cluster. For every merging, the newly generated cluster is the parent of two previous ones, leading to the formation of a binary tree. At each merging, the distance between the two merged clusters is recorded as the precision level of the new cluster. After setting up the tree, given a certain precision, we can traverse the tree to search for the clustering at the required level. The definition of the distance between two clusters must be defined, and we use the median average-linkage, for the purpose of discarding the outliers.

After clustering is performed according to a given precision, which is a distance of 10 pixels in our case, the center of the cluster is calculated by averaging the locations of the points in the cluster. This is a Tier 2 IR Point (Fig. 3.2). For each Tier 2 IR point, the LSC feature is calculated at a finer level, with more radial sampled directions and a larger searching range. The distance in each direction from the reference point to the nearest edge is calculated to form the LSC feature. From the LSC feature, the orientation is computed as the direction of the largest distance value, or the direction of the middle one if there are multiple continuous maximum distances. To illustrate this concept, in Fig. 2, the orientation of the head is straight down, that of the hand is towards bottom left, and that of the foot is straight up. This orientation is very useful in aligning body parts with the model when classifying. By performing the alignment, the variance caused by rotation can be removed.

3) Searching Range Adaptation: Note that the searching range should not be a constant, but should be adaptive to the depth changes. That means when a reference point is closer to the TOF sensor, the searching range should be larger, and vice versa. So before calculating the LSC feature for a given reference point, the depth value is first read and an adapted searching range is calculated according to the projection model.

C. Endpoint Body Part Identification

1) Identification Using a Template Model: The next task is to classify the Tier 2 IR Points into the four categories, namely head, hand, foot and other. A learning based classification approach is used. The simplest model is a deterministic template model. Assuming that we know the initial posture of the human subject in each clip, then we can extract a template for each endpoint body part at the beginning of the detection, as shown in Fig. 4. After the Tier 2 IR Points are identified in a given frame, we first rotate the IR points to the template orientation by aligning the orientation element of the template and of the IR point, and then compute the distances from the LSC features of the Tier 2 IR Points to the templates. Based on these distances, a distance matrix is computed between the templates and the IR points. Note that we have two other templates for each hand and foot, which are symmetric. Because the radial sampling density for generating the template is the same as for Tier 2 IR Points, the dimensions are the same. Thus Euclidian distance is used to calculate the distance. The best match is found by a greedy algorithm, which always looks for the smallest distance for the next matching. More complicated algorithms to find better matches can also be considered, such as using the Hungary Algorithm to minimize the total distance of the matches [15].


4.1: Head

4.2: Hand

4.3: Foot

Figure 4. Deterministic Template for Each Endpoint Body Part

2) Identification using a Multivariate Gaussian Model: The deterministic template model does not account for the deformation of the 2D shapes of the body parts. However, the shapes may vary significantly due to two reasons: the change of view angle, and the actual change of the endpoint shape, such as hand closing and opening. The former case can apply to all the categories, but the latter is only obvious for the hands and feet. A straightforward improvement of the identification is to use a probabilistic model instead of a deterministic template. We model the LSC feature for each category using a multivariate Gaussian model, by specifying a mean and a covariance matrix among the directions of the feature. The mean and the covariance matrix are calculated using the LSC features obtained from 27 manually selected frames which cover a variety of the possible shapes of each of the endpoint body parts. Given a Tier 2 IR Point, the probability that describes the likelihood that the IR point corresponds to one of the categories is calculated. The same greedy algorithm is applied for finding the best match. Using the probabilistic model significantly improves the robustness to shape variations in the hands and feet.

3) 3D Position and Orientation Computation: After identification, the 2D locations of the endpoint body parts are obtained. Also, the orientations are described in 2D. However, it is not difficult to calculate the corresponding values in 3D. For locations, this is the inverse problem of recovering the depth images. Combined with the sensed depth value and the given projection model, we are able to transform the location to 3D. For the orientation, we compute the depth at two points along the 2D orientation line and calculate the vector between them.

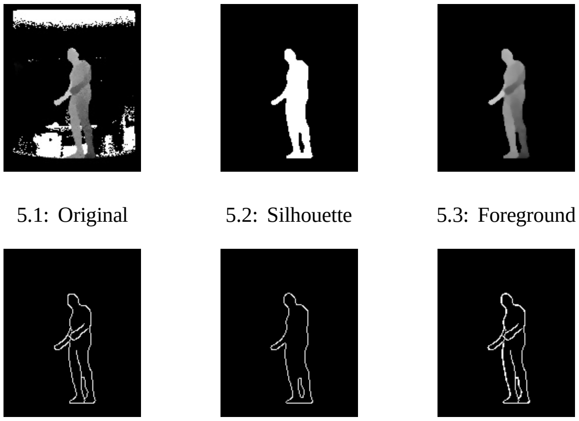
4) Local Sliding Window: For calculating the LSC feature for a given body part, the location of the reference point matters. The classification result is improved if the reference point for evaluating an endpoint body part candidate can be close to that of the model template. When generating the models, the reference points are approximately at the center of each endpoint body part. Because Tier 2 IR Points are calculated by averaging Tier 1 IR Points, the location should already be close to the center of each body part, but there is no guarantee. To improve the IR point positioning, we apply a local sliding window around each Tier 2 IR Point, and extend the LSC Descriptor of one Tier 2 IR Point as the LSC Descriptors of a set of sampling points around that IR Point.

D. Processing Steps

The depth dataset we use is publicly available, originally generated by Ganapathi et. al. using a Swissranger SR4000 TOF camera by MESA Imaging AG, with a resolution of 144 by 176 [4]. The dataset contains 28 clips of a variety of different motions, along with the ground truth data obtained from a marker based motion capture system. The depth is represented by a 3D point cloud with coordinates described in the eye coordinate frame. Our first task is to recover the depth image using the projection matrix which is provided along with the dataset. A typical recovered depth image is shown in Fig. 5.1.

After the depth images are obtained, the depth values are thresholded to generate the silhouette (Fig. 5.2). To denoise the sensing data, we remove the background and zero depth points, and smooth the resulting image using a median filter. After that, the image is thresholded again, and a clean binary human silhouette image is obtained. We then mask the smoothed depth image using the silhouette image, to obtain the smoothed foreground depth image (Fig. 5.3).

The next task is to extract the shape of the human body, which is described by the shape edges. The shape edges contain the entire contour edge, as well as the edges inside the contour due to the depth discontinuity. The Canny edge detector [16] is applied twice, to both the binary silhouette image (Fig. 5.4) and the smoothed foreground depth image (Fig. 5.5), but with different thresholds. The results from the two pipelines are then integrated by an ”OR” operation and generate the final edge image as shown in Fig. 5.6. We do this because we want to ensure the contour edge is continuous, and the inner shape edges are controllable by adjusting the Canny thresholds separately.



5.1: Original 5.2: Silhouette 5.3: Foreground

5.4: Edge1 5.5: Edge2 5.6: Edge

Figure 5. Image Processing

After depth data preprocessing, the approach described in Sec. III is implemented to achieve the endpoint body part detection and identification task. The complete algorithm is summarized in Fig. 6.

IV. EXPERIMENTAL RESULTS

A. Parameter Settings

During the experiment, there are several important parameters to set. When selecting Tier 1 IR Points, the initial point sampling is carried out over a uniform grid with sampling points at every 4 column by 4 row point. The radial sampling

```
for frame number = first to last do
    Step1: read and process the depth image (Sec. III-D)
    Step2: sample and select Tier 1 IR Points (Sec. III-B1)
    Step3: cluster Tier 1 IR Points to form Tier 2 IR Points (Sec. III-B2)
    Step4: classify Tier 2 IR Points (Sec. III-C)
    Step5: evaluate the result using ground truth data (Sec. IV-B)
end for
```

Figure 6. Algorithm Pseudocode

is carried out in 16 directions and the searching range is set to 9 pixels before adaptation. For the LSC feature calculation of Tier 2 IR Points, the radial sampling density is set to double the Tier 1 IR Points density (i.e., 32 directions), and the searching ranges before adaptation are also doubled. The range of the local sliding window is set to be within a 9 by 9 square around each Tier 2 IR Point. Parameter selection is discussed further in Section IV-B.

B. Quantitative Results

The system is tested using all 28 clips contained in the dataset [6], including simple and complicated human motions. The motions include arm waving, leg raising, squatting, turning around and playing baseball motions. Furthermore, in some clips, when the human subject approaches too close to the sensor, the silhouette becomes incomplete in the depth image. Using the parameter settings specified above, the system requires 10 ms on average to detect and identify endpoint body parts through all 7900 frames.

We evaluate the obtained results quantitatively, using the ground truth data obtained from a marker-based motion capture system. For each endpoint body part category, we identify true positives, false positives and false negatives. Given a classified Tier 2 IR Point, we calculate the distance from this point to the ground truth position of the corresponding body part in 3D. If the distance is within 20 cm, we identify this as a true positive, otherwise it is a false positive. When a body part is not detected in one frame, this is identified as a false negative. By doing this, we assume that all the five endpoint body parts are detectable in every frame, but there are actually a few exceptions. As a result, the identified false negative rate will be slightly higher than its actual value.

We then compute the precision and recall for each classifier, which are widely used measures for evaluating classifier performance. The precision is calculated as the number of true positives divided by the sum of true positives and false positives, and the recall is calculated as the number of true positives divided by the sum of true positives and false negatives. The results using the parameter settings described in Section IV-A are shown in Table I.

Table I
PRECISION AND RECALL

	Head	Hand	Foot
Precision	0.87	0.80	0.99
Recall	0.91	0.82	0.91

Table II
CONFUSION MATRIX (COLUMNS ARE ACTUAL, ROWS ARE PREDICTED, NORMALIZED BY THE ACTUAL NUMBER)

	Head	Hand	Foot
Head	6343(99.09%)	7(0.07%)	1(0.01%)
Hand	58(0.91%)	10630(99.13%)	801(5.38%)
Foot	0(0%)	86(0.8%)	14089(94.61%)

A confusion matrix is also computed to further characterize the classifier performance. Using the calculation of the false positive, if a Tier 2 IR Point A is identified as an endpoint body part, say head, but the identification is incorrect, we compute the distances from A to the ground truth of all the other body parts, i.e. hands and feet. If any of these distances are smaller than 20cm, e.g. from A to right hand, we identify this as a confusion between head and hand. The confusion matrix is shown in Table II.

We also varied the parameters to see how the performance is affected. First we altered the searching range for Tier 1 IR Points, while maintaining the searching range for Tier 2 IR Points as doubling the Tier 1 range. The results show that as the searching range decreases, head detection suffers most. This is because typically the head is the largest of the three body part categories. When the range is increased to 11, the performance also worsens, because during the detection of Tier 1 IR Points, fewer sampled points have directions along which edges are not detected. Next we altered the sampling density of the points used to generate Tier 1 IR candidates. We found that as the density drops from every 2 points to every 4 points, the performance does not decrease appreciably, but the speed becomes much faster. This is because much of the processing time for each frame is taken in clustering when the number of Tier 1 IR Points is large, and denser sampling increases the number of Tier 1 IR Points. Finally, we tested the effect of adding the local sliding window. Before adding the sliding window, the precision and recall are shown in Table III. We can see that the precision is not affected much, but the recall is much lower than when a local sliding window is used.

Table III
PRECISION AND RECALL WITHOUT LOCAL SLIDING WINDOW

	Head	Hand	Foot
Precision	0.86	0.83	0.98
Recall	0.58	0.75	0.43

C. *Qualitative Analysis*

Fig. 7 shows examples of correct and incorrect detections. Comparing Fig. 7.2 with Fig. 3.2, we can see that although Tier 2 IR Points are also detected on the right thigh and the left forearm, after identification using the LSC feature, those two reference points are successfully discarded. Fig. 7.3 shows that although the 2D shape for the feet changes significantly, the system can still successfully detect and recognize them because of the probabilistic model. Fig. 7.4 demonstrates the effect of the depth-adapted searching range; even though the head is closer to the sensor and appears to be bigger, the search range is adapted so that the head is still correctly detected. Fig. 7.5 and Fig. 7.6 demonstrate that even when the shape edges for endpoint body parts are inside the contour, such as the left hand in both frames, due to our edge detection method, they can still be correctly identified in these frames.

There are also some typical incorrect situations. In Fig. 7.7, the feet cannot be identified, due to the fact that they are too close to the sensor and thus after foreground segmentation, this portion of the human body is lost. Even though the feet cannot be detected, the detection of the hands and head is unaffected. In Fig. 7.9, the right foot is not detected, because the view angle relative to this foot has changed significantly, and this foot appearance is not included in the limited training data. This can be easily improved by using a more comprehensive data set to train the system. A similar problem is seen in Fig. 7.11, where the left foot looks very unlike any of the training data. But surprisingly, in Fig. 7.11 the right foot is still detected although the person turns around and the right foot is partially occluded by the left foot. Fig. 7.9 shows the situation when the right hand is too close to the head and appears to be connected in the depth image, in this case both of these two body parts cannot be detected. This difficulty also exists in other depth image based methods, such as [11]. Although Fig. 7.10 is a similar situation to Fig. 7.5, the shape edges for the hands cannot be detected because the arms are too close to the torso and thus no depth edge is found. If we use a more accurate depth sensor with better resolution, this problem can be alleviated. Making the thresholding of the Canny detector adaptive can also improve performance. In Fig. 7.12, the system confuses the left hand and the head, because of the shape changes due to occlusion and deformation.

V. DISCUSSION

As seen from the experimental results, the proposed approach is effective and efficient at detecting and identifying endpoint body parts from depth images. The newly proposed LSC descriptor contains rich shape information about body parts, and is very easy to compute. It is invariant to translation, and through the orientation computation, adapted searching range, use of probabilistic models, it can be made at least partially invariant to rotation, scaling and deformation. Combined with the hierarchical IR Point selection algorithm, the system is able to localize the endpoint body parts at a very fast rate with high precision and recall rates. Further more, the system does not take any temporal information into account, so the output of our system can be seen as raw sensing data, to be used as input for solving tracking problems.

However, there are also several factors which degrade performance. First, our detection and identification is shape based, so when the shape of one body part looks very similar to another, this approach may fail in distinguishing them. Secondly, the quality of the depth images affects the performance significantly. Especially when the endpoint body parts are inside the body silhouette contour, the detection is completely based on the depth discontinuity. If the sensor is very noisy or the resolution is too low, the shape edges inside the contour cannot be found. Thirdly, depth image based body part detection suffers when the body part to be detected is too close to other body parts. A potential solution to this problem is to also incorporate other cues to compensate for the edge losses. For example, if the shape edges are very clear in color image, they could also be combined into the final shape edge image. Finally, since our approach is completely shape based, it does not incorporate any other knowledge, but other sources of information can be very helpful. For example, if temporal information is used to filter the detection result, many confusion errors will be eliminated. Also the knowledge of the location of the entire human body and of the human motion can also help to eliminate mislabeling.

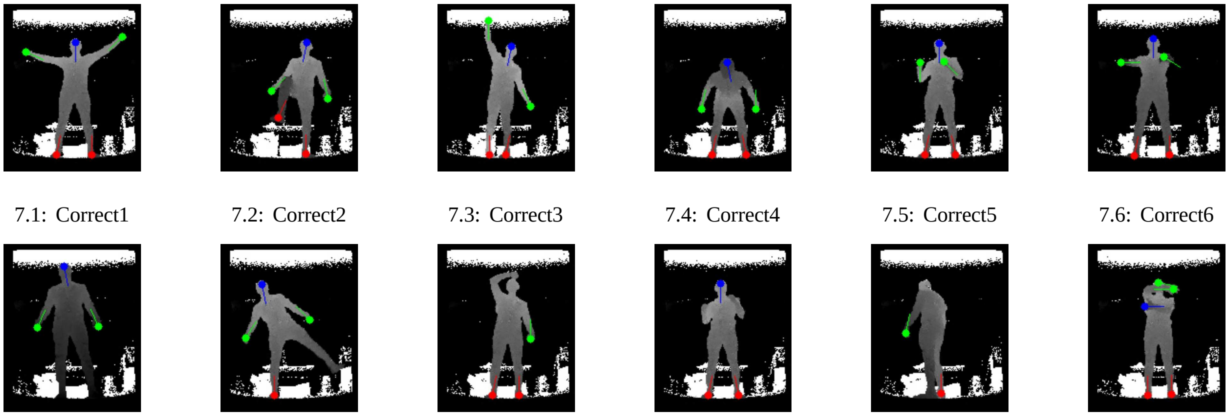
VI. CONCLUSIONS

For both motion capture and gesture recognition applications, localizing the endpoint body parts is very important. In this paper, a new method for fast detection and classification of endpoint body parts was proposed, based on a novel image feature describing the local shape near the body parts of interest. An efficient algorithm was developed for locating and classifying LSCs, using edge images from depth data. The proposed algorithm is computationally efficient and shows excellent classification performance at 100Hz, with a high precision and recall.

In future work, we will improve the system by enhancing the shape edge extraction and applying more sophisticated classification methods. We will also consider incorporating edge information obtained from a color camera to improve the edge image. Finally, the output of this system will be used to facilitate full body motion tracking [1].

ACKNOWLEDGMENT

The authors gratefully acknowledge V. Ganapathi et al. for making the dataset used in this paper publicly available [6].



7.1: Correct1 7.2: Correct2 7.3: Correct3 7.4: Correct4 7.5: Correct5 7.6: Correct6

7.7: Incorrect1 7.8: Incorrect2 7.9: Incorrect3 7.10: Incorrect4 7.11: Incorrect5 7.12: Incorrect6

Figure 7. Selected Correct and Incorrect Frames

References

[1] Z. Li and D. Kulić, "A stereo camera based full body human motion capture system using a partitioned particle filter," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3428–3434, IEEE, 2010.

[2] Z. Li and D. Kulić, "Particle filter based human motion tracking," in *Proceedings of the IEEE International Conference on Control, Automation, Robotics and Vision*, 2010. To Appear.

[3] P. Azad, T. Asfour, and R. Dillmann, "Robust real-time stereo-based markerless human motion capture," in *8th IEEE-RAS International Conference on Humanoid Robots, 2008. Humanoids 2008*, pp. 700–707, 2008.

[4] V. Ganapathi, C. Plagemann, S. Thrun, and D. Koller, "Real time motion capture using a single time-of-flight camera," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA, USA), June 2010.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 509–522, 2002.

[6] V. Ganapathi and C. Plagemann. Dataset used in this paper: http://ai.stanford.edu/ varung/cvpr10.

[7] P. Azad, A. Ude, T. Asfour, and R. Dillmann, "Stereo-based markerless human motion capture for humanoid robot systems," in *2007 IEEE International Conference on Robotics and Automation*, pp. 3951–3956, 2007.

[8] M. Siddiqui and G. Medioni, "Real time limb tracking with adaptive model selection," *Pattern Recognition*, vol. 4, pp. 770–773, 2006.

[9] I. Haritaoglu, D. Harwood, and L. Davis, "Ghost: A human body part labeling system using silhouettes," in *International Conference on Pattern Recognition*, vol. 14, pp. 77–82, Citeseer, 1998.

[10] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1-Volume 01*, p. 97, IEEE Computer Society, 2005.

[11] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time Identification and Localization of Body Parts from Depth Images," in *IEEE Int. Conference on Robotics and Automation (ICRA), Anchorage, Alaska, USA*, 2010.

[12] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," *Computer VisionECCV 2002*, pp. 150–180, 2002.

[13] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, p. 14, California, USA, 1967.

[14] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[15] G. Carpaneto, S. Martello, and P. Toth, "Algorithms and codes for the assignment problem," *Annals of Operations Research*, vol. 13, no. 1, pp. 191–223, 1988.

[16] J. Canny, "A computational approach to edge detection," *Readings in computer vision: issues, problems, principles, and paradigms*, vol. 184, 1987.