2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Real-time Identification and Localization of Body Parts from Depth Images

Christian Plagemann[*] Varun Ganapathi[*] Daphne Koller Sebastian Thrun

Artificial Intelligence Laboratory,
Stanford University, Stanford, CA 94305, USA.
{plagem, varung, koller, thrun}@stanford.edu

***Abstract*— We deal with the problem of detecting and identifying body parts in depth images at video frame rates. Our solution involves a novel interest point detector for mesh and range data that is particularly well suited for analyzing human shape. The interest points, which are based on identifying geodesic extrema on the surface mesh, coincide with salient points of the body, which can be classified as, e.g., hand, foot or head using local shape descriptors. Our approach also provides a natural way of estimating a 3D orientation vector for a given interest point. This can be used to normalize the local shape descriptors to simplify the classification problem as well as to directly estimate the orientation of body parts in space.**

Experiments involving ground truth labels acquired via an active motion capture system show that our interest points in conjunction with a boosted patch classifier are significantly better in detecting body parts in depth images than state-of-the-art sliding-window based detectors.

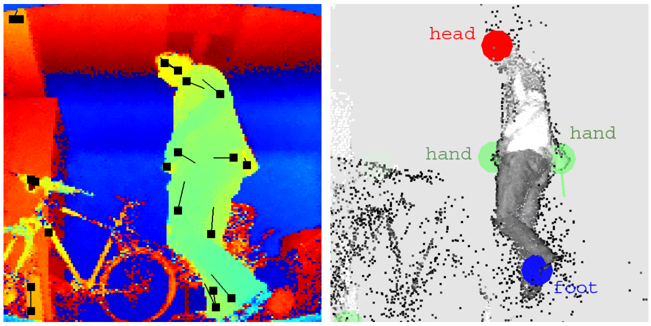

Fig. 1. The proposed system identifies and localizes human body parts in 3D space from single depth images. The left image shows a color-coded depth image and detected interest points overlayed in black. The 3D point cloud representation of the same data as well as the body part detections are shown on the right.

I. INTRODUCTION

In this paper, we present a fast and robust system for detecting, identifying and localizing human body parts in depth images. The output of our system can either be used directly, e.g., to infer human gestures, or it can be used as preprocessing for other algorithms, such as full-body tracking or surveillance. As one potential application domain for the developed technology, human-robot interaction (HRI) relies to a large degree on non-verbal communication. Apart from communicating through speech and language, humans use their whole body as an effective medium for communication. For instance, humans often demonstrate tasks rather than describing them verbally and point in order to refer to objects in their environment. Our algorithm was developed specifically for fast operation at video frame rates, since natural communication requires a low-latency action-reaction cycle. The presented system requires 60 ms per camera frame to estimate the 3D locations of hands, feet and head as well as their 3D orientations.

The task of understanding human shape and motion from sensor information, such as camera images, is hard because the human body is highly variable – both between subjects and within subjects. The human skeleton has many degrees of freedom and the skeleton itself is hidden under several flexible layers of body, skin and clothes. In addition, the appearance of a scene under natural lighting conditions is highly variable, which has been a major hurdle for vision-based approaches to human motion analysis. In this work, we take advantage of recent advances in sensor technology to remove the visual appearance of a scene as one source of ambiguity. In particular, we use a time-of-flight sensor to acquire depth images as the primary data source.

Time-of-flight sensors are a rapidly improving technology that can provide dense depth measurements at every point in the scene at high frame rates. They can be seen as the two dimensional extension to the laser range scanners, which have become standard in applications such as robot localization. The range data provided by these cameras both facilitates segmentation of the human body from background structure and it can also disambiguate visually similar poses.

The contributions of this paper are two-fold: First, we present a novel interest point detector for range data. It is fast to compute and robust against the typical sources of noise found in depth image data. Moreover, the detector provides estimates for the 3D orientation of foreground structure around the interest point locations. Our second contribution is an integrated system that detects human body parts from depth images in real-time using a boosted classifier that assigns local shape descriptors extracted at interest point locations to body part classes.

[*] C. Plagemann and V. Ganapathi contributed equally to this work. Additional material and videos can be found at
http://stanford.edu/ ˜ plagem/publ/icra10.

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

II. RELATED WORK

The automatic analysis of human shape and motion from sensor data is a strongly researched topic in several areas. Moeslund et al. [1] illustrate this impressively for the field of computer vision by reviewing more than 350 papers from this field. There are a multitude of approaches and algorithms on this topic. In this paper, we focus on the task of accurately detecting parts of the human body, such as the head, hands and feet, from a single depth image.

For intensity cameras, many papers have explored improvements on the basic paradigm of finding interesting points in the image and computing descriptors for the local patches around them. These detectors identify points in the image for which the signal changes two-dimensionally, that is, at corners, junctions and blobs. Several authors have attempted to port computer vision algorithms directly to three dimensional data to develop detectors for locally interesting points. Steder *et al.* [2], for example, apply vision-based interest point detectors on depth images. Ruhnke *et al.* [3] applied the Harris corner detector [4] to the depth image in order to construct 3D object models in an unsupervised fashion from partial views. Recently, Zou *et al.* [5] applied the concepts of the SIFT detector [6] to arbitrary point clouds by finding interest points that are local extrema of the DOG function (difference of Gaussians) over a locally defined curved surface. In a similar fashion, Hu *et al.* [7] consider spectral geometric features on triangle meshes. Nickel *et al.* [8] recognize pointing gestures using stereo vision. They primarily use the color images to identify the location of the face and hands using standard vision techniques, and use the depth sensor only to determine the final pointing direction. Ido *et al.* [9] enable a robot to identify the hands of a human by finding small patches that are the closest to the robot. While suitable for their application, this makes strong assumptions about the human pose, which we do not make here.

With the exception of the last, a commonality in the above work is that interest points are identified using local information. Because the time-of-flight sensor is currently noisier and lower resolution than intensity cameras, the local curvature can vary widely due to noise. We instead define interesting points by using the nature of the matrix of pairwise geodesic distances between all identified foreground depth measurements.

III. PROBLEM SETTING AND OVERVIEW

The task is to extract information about human body parts, such as their visibility or their location and orientation in space, from depth images. We define a depth image as a set $R = \{r_{i,j}\}$, $i = 1,\ldots,n$, $j = 1,\ldots,m$, $r_{i,j} \in \mathbb{R}$ of distance measurements $r_{i,j}$ relative to a camera location $\in \mathbb{R}^3$ and orientation $\theta \in \mathrm{SO}(3)$.

In a preprocessing step, the distance measurements are transformed into a 3D point cloud using the known camera calibration and truncated according to a bounding box. The specific sensor used in our experiments is the Swissranger SR4000 [10] Time-of-flight camera by MESA Imaging AG, Switzerland, but most other range sensors such as laser range finders or dense stereo reconstruction systems would be applicable as well. The SR4000 sensor yields 176x144 pixels at 30 frames per second. In our experiments, we rotated the sensor by 90° to better match the dimensions of different recorded human subjects.

Our goal is to estimate for every depth image a set $\mathscr{D} = \{(c_d, \mathrm{x}_d, \theta_d)\}_{d=1}^{D}$ of body part detections including class assignments $c_d \in \{\texttt{head}, \texttt{hand}, \texttt{foot}\}$, part locations $\mathrm{x}_d \in \mathbb{R}^3$ and optionally their orientations $\theta_d \in \mathrm{SO}(3)$. Ideally, these detections should become available at a high frame rate so that dependent higher-level functions, such as a module for human-robot interaction, do not suffer from time delays.

As detailed in the following, we take a bottom-up approach to this problem by identifying and classifying potential body part location directly from the range data – as opposed to, for instance, first fitting a full-body model of the human subject and inferring the body part configurations in a top-down fashion.

We provide an algorithm for identifying a novel type of interesting point based on geodesic distance between vertices in a mesh. This particular interest point detector has the added advantage of providing a stable estimate of local pose, which can be used to normalize image patches prior to feature extraction.

Our detection pipeline consists of the following steps:

1) Construct a set of connected surfaces meshes from the point cloud
2) Identify interest points on each of these meshes
3) Extract local descriptors for the interest points
4) Classify the descriptors to body part classes
5) Sort patches by classifier confidence

We now discuss the most important steps in detail, that is, interest point extraction and patch classification. The remaining steps are described briefly at the beginning of the experimental section where we describe the experimental setup.

IV. INTEREST POINT DETECTION

In this section, we present a novel interest point detector for range data. The task is to select vertices from an arbitrary surface mesh embedded in the 3D Euclidean space that are invariant to mesh deformations, translations, rotations and noise. As a visual example, consider several depth images of a moving person or an arbitrary articulated object. Our goal is to extract sets of interest points from these images that correspond to the same points on the observed body, even though the underlying body articulations differ. This property of an interest point detector is termed *stability*. On the other hand, we would also like to maximize the *repeatability* of the detector, that is, the likelihood of detecting a once-detected interest point again should be high.

Our algorithm is inspired by the insight that geodesic distances on a surface mesh are largely invariant to mesh deformations and/or rigid transformations. More visually, the distance from the left hand of a person to the right hand along the body surface is relatively unaffected by her/his posture.

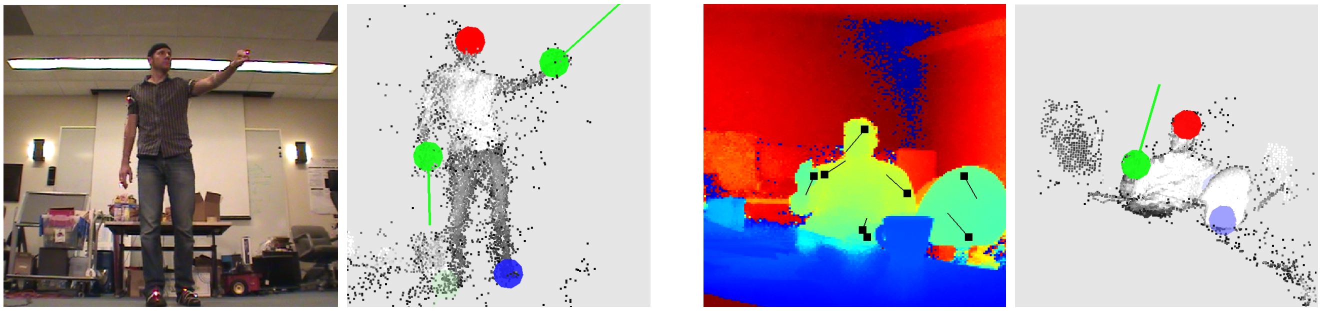

Fig. 2. Exemplary results. Left: A test situation including active markers for evaluation. The second image shows the acquired 3D point cloud colored by the intensity return values and the body part detections *head* (red), *hand* (green, including orientation) and *foot* (blue). Right: A person sitting behind a desk in a cluttered office environment, facing the camera (raw color-coded depth image as recorded by the camera and the point cloud visualization).

On a more local level, this also holds for the case in which the body surface is represented by several non-connected meshes due to segmentation and occlusion effects.

The proposed interest points are termed AGEX, Accumulative Geodesic EXtrema. They are computed by incrementally maximizing geodesic distances on the surface mesh. Specifically, we define for a given mesh $\mathscr{M}$ and $k >= 1$ the sets AGEX$_k(\mathscr{M})$ recursively as follows:

- For all $k > 1$, AGEX$_k(\mathscr{M})$ extends AGEX$_{k-1}(\mathscr{M})$ by a vertex $\mathrm{p} \in \mathscr{M}$ that maximizes the geodesic distance to all vertices contained in AGEX$_{k-1}(\mathscr{M})$.
- AGEX$_1(\mathscr{M})$ contains exactly one vertex: the geodesic centroid of the mesh.

Note that we chose to initialize AGEX$_1$ as the set containing only the geodesic centroid of the mesh for simplicity of implementation. Alternatively, it is relatively straightforward to start the process at AGEX$_2$ and to define this set as containing the endpoints of the longest path on the mesh surface.

AGEX$_k(\mathscr{M})$ can be computed efficiently in an iterative way using Dijkstra's algorithm (see, e.g., [11] or standard text books on graph algorithms). We consider the graph that corresponds to the acquired surface mesh: mesh points become the vertices of the graph and edges are created for neighboring mesh points. Points are considered neighbors, if (a) the corresponding pixels in the depth image are neighbors and (b) their distance in 3D scene space does not exceed a threshold $\delta_{\text{connected}}$. We now consider the geodesic centroid v_c of the mesh (i.e., the starting situation AGEX$_1(\mathscr{M}) = \{v_c\}$) and calculate the shortest paths to all other vertices following Dijkstra's algorithm. The vertex v_s which is found to have the longest of these shortest paths is added to the interest point set to yield AGEX$_2(\mathscr{M}) =$ AGEX$_1(\mathscr{M}) \cup v_s$. We now add a zero-cost edge between the v_c and v_s and update the shortest paths. This is repeated until k vertices have been added to the interest point set. The runtime of Dijkstra's algorithm per iteration is $O(|E| + |V| \log |V|)$, where $|E|$ is the number of edges in the graph and $|V|$ the number of vertices. In our case, in which we consider the 8-neighborhood of depth image pixels and need to run the shortest path algorithm k times, this results in $O(k \cdot (8n + n \log n))$, where n is the number of vertices. Given that the number of vertices is bounded by the relatively low number of available depth image pixels and one is typically interested in less then $k = 20$ interest points, the computational cost for computing AGEX$_k(\mathscr{M})$ is effectively very low.

A. *Estimating the Interest Point Orientation*

We can assign an orientation to each extracted interest point k_i in a natural way by "tracing back" the shortest path that lead to its detection. Concretely, we mark the graph vertex corresponding to the interest point and consider the matrix of path costs produced by Dijkstra's algorithm. We successively follow the shortest incoming edges to neighboring vertices until a maximal geodesic distance of δ_{orient} to a point p is reached. The resulting vector $\mathrm{o}_i := \mathrm{k}_i - \mathrm{p}$ is treated as the 3D orientation of the interest point k_i.

Figure 3 visualizes local depth image patches extracted at interest point locations normalized by their estimated orientation (i.e., the patches have been rotated such that the orientation vector points downwards). It can be seen that this normalization step brings the main features of each class to a close alignment, which in turn, makes learning and classification significantly easier.

B. *Properties of the AGEX Interest Points*

Note the following properties of the proposed interest points:

- The detection algorithm presented above is executed for all subsets of connected meshes and, thus, can deal with situations in which a mesh falls into several parts, e.g., due to occlusions in the depth image. As a result, additional interest points are found at the occlusion boundaries.
- The resulting interest point set approximates a uniform distribution over the mesh surface and it grows incrementally with k until it eventually contains all mesh points.
- AGEX$_k(\mathscr{M})$ is not a uniquely identified set of points. If several points $p \in \mathscr{M}$ maximize the geodesic distance to all points in AGEX$_{k-1}(\mathscr{M})$, a random one of them is chosen to extend the set.
- For identifying human body parts, the order in which the interest points are detected typically corresponds to their importance, that is, points close to the hands, feet

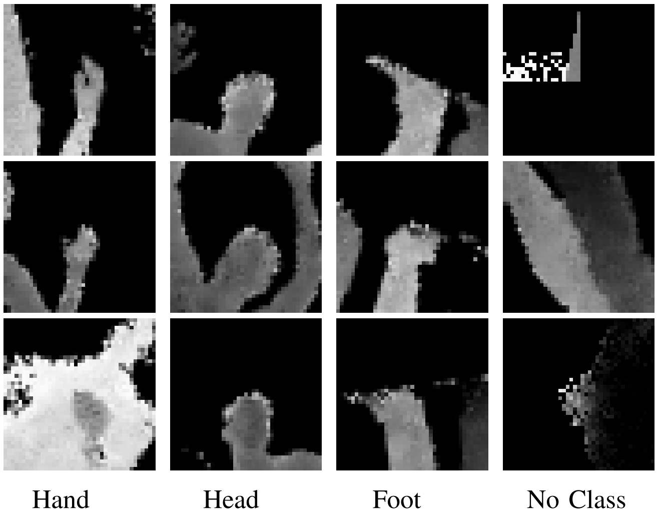

Fig. 3. Typical examples of local descriptor patches for the different body part classes. All patches have been automatically centered at their interest point locations and normalized by orientation using the respective orientation estimates.

and the head are found first, before other surface points are added to the set.

- The orientation estimates are particularly useful to normalize local shape descriptors by their orientation. This operation leads to better invariance properties of the descriptors and, thus, to a simpler classification problem.

V. IDENTIFICATION OF PARTS

The algorithm described in the previous section provides a set of interest points, which we consider as the hypothesis space for potential body part locations. In this section we describe how to learn and apply classifiers for local patch descriptors that assign body part labels to the interest points.

A popular alternative to the explicit pre-selection of potential location for detections is the so-called sliding window approach. Here, a set of learned patch classifiers is applied to the entire image (or a higher dimensional state space) in a partially overlapping fashion. The local maxima of the classifier responses are then taken as detection results.

In the current literature there is no clear preference for either the interest point-based detection approach or the sliding-window class of detectors. The preferred choice seems to depend mainly on the particular task and data set that is to be solved. As our experimental evaluation shows, the AGEX interest points outperform the sliding window approach by a large margin in our application. This is due to the fact that (a) there is a large overlap between the sets of AGEX points and the true locations of the considered body parts – which considerably constrains the hypothesis space and (b) the normalization of the patch descriptors by the estimated orientation drastically reduces the intra-class variability that the classifier has to learn and represent. Both aspects serve to improve the generalization performance of the algorithm.

A. Local Shape Descriptor, Classifier and Learning

As local descriptors at interest points we consider 41x41 depth image patches surrounding the point, rotated w.r.t. the estimated orientation vector. The result is a set of patches from the image that are likely to be centered on salient parts of the human body. Figure 3 shows typical examples of local descriptors for the different classes considered in this paper.

To assign body part labels to the extracted descriptor patches, we follow a supervised learning approach inspired by Torralba *et al.* [12], which is implemented in the STAIR Vision library [13]. Given a labeled training set of image patches, the approach constructs a random dictionary of local image structures and learns a boosted classifier using the dictionary entries as basic features. This requires a training set of positive examples, i.e., patches centered on parts of interest points and a large quantity of negative examples, i.e,. patches of the background or other parts not of interest.

We obtain training data by simultaneously recording with an active marker-based motion capture system, specifically the PhaseSpace system. This system consists of four cameras and several visible light LED markers placed on parts that we wish to detect. The LEDs in the PhaseSpace system encode their unique id. Thus, the output of the motion capture system is the three-dimensional location of distinct markers with sub-millimeter accuracy. We register coordinate systems of the depth sensor and the PhaseSpace system by manually corresponding measured data with marker locations. With sufficient pairs of corresponding 3D points, we can compute the rigid transformation between the coordinate systems in closed form using a singular value decomposition (SVD). We experimentally verified that the markers are not visible in the depth camera images, which is to be expected since the SR4000 sensor only responds to surfaces that reflect the specific infrared light it emits.

VI. EXPERIMENTAL EVALUATION

The algorithm was evaluated on several sequences involving a variety of movements of the entire body, three different human subjects and several environments including clutter. The goal of this evaluation is two-fold. First, we evaluate the overall detection and classification performance of our integrated system and discuss typical situations that lead to accurate or inaccurate detections. In the second part, which is detailed in Sec. VI-A, we compare our approach to the state-of-the-art alternative for body part detection.

In the experiments, the raw depth sensor data was preprocessed by (i) de-noising the image and removing so-called mixed pixels at depth discontinuities and (ii) removing points that fall outside a bounding box. *Mixed pixels* are depth measurements that effectively average the depth value of a foreground object and the background due to the temporal integration of raw measurements by time-of-flight cameras. Furthermore, we perform agglomerative clustering on the 3D point cloud to produce a set of surface meshes as input to the AGEX interest point detection algorithm. Video material is available at `http://stanford.edu/~plagem/publ/icra10`

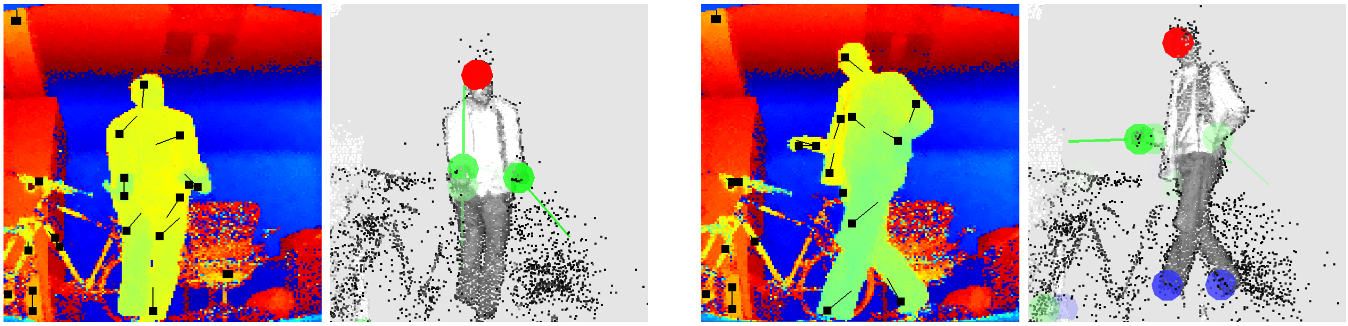

Fig. 4. Left: The algorithm detects both hands well, even though they point directly towards the camera and do not stand out in the silhouette of the person. Right: The left hand of the person is detected with a low confidence only, since the orientation vector of the corresponding interest point is not aligned with the body part.

Figures 2 shows the range of situations considered. The two left-most diagrams exemplify the laboratory conditions under which the true part locations are known via recorded active marker positions and on the right, a significantly more complex scene including a person sitting behind a desk in a cluttered office environment is depicted.

Figure 5 gives the precision/recall curves as well as the confusion matrix for the body part classifiers evaluated at the extracted AGEX interest points for a test sequence. The training set consists of 789 recorded frames from a different sequence, resulting in 6312 patches extracted at interest point locations. The time requirements for learning all classifiers on a standard quad-core desktop PC were approx. 240 minutes per class. The confusion matrix was computed for a classifier threshold of 0.75, that is, patches causing classification confidences above 0.75 were counted as positive classifier output.

Each point on the precision/recall curves corresponds to a specific threshold opposed on the classifier confidence. Naturally, at high confidences, the precision takes a high value and the recall a low one. As the confidence threshold lowers, recall increases. Each cell of the confusion matrix gives the frequency at which a given classifier responds to patches from a given (true) class.

As can be seen from the diagrams, the head of the test subject is classified nearly perfectly across all frames. The hands and the feet of the subject are detected well, but have higher error rates in general. One interesting question is how many false negatives are caused simply by the fact that we restrict our consideration to AGEX points. We evaluated this on the test set and found that 98% of the patches containing the head were in AGEX$_5$. The respective numbers for hands and feet were 82% and 79% respectively. In experiments, we found that the state-of-the-art alternative to our approach – the sliding window detector – shows lower performance, even though it considers all possible image locations. We give quantitative results of a comparison in Sec. VI-A

Figures 1, 2, 4 and 6 show exemplary results of the part detector in different realistic situations. Figure 1 and the left part of Fig. 2 show body configurations that are hard to analyze using a regular intensity camera since the hands take up only few pixels in the image space and their appearance is highly ambiguous. Figures 2 and 6 show instances in which the 3D orientation vectors of the interest points are highly correlated with the pointing directions of the hands. The right-most diagram in Fig. 2 shows a situation in which this is not the case. Classification performance can decrease when the orientation estimate of the interest point has high uncertainty. This can occur, for instance, when the arm is pointed directly at the camera so that only the hand is visible, and no part of the arm itself. Then the hand will form its own disconnected interest point. As a result, almost any orientation of the hand is possible, which presents difficulty for the patch classifier. This could be corrected by identifying cases where the orientation is uncertain and using some other means for orientation normalization prior to patch extraction.

Some reduction in recall is due to not detecting interest points at the appropriate locations. Consider, for example, the left foot in the left image in Fig. 2. The interest point detector assumes that interesting points occur at geodesic extrema. When this assumption is violated, performance can suffer. For instance, when the hand touches the thigh, the first appropriate interest point for the hand may be contained in AGEX$_{1000}$ rather than already in AGEX$_5$ – in which instance the interest point detector has degraded to selecting a large part of the image.

A. Comparison to the Sliding-window Approach

We compared our algorithm against the state-of-the-art approach to body-part detection, that is, the sliding-window detector (SW). Both algorithms were implemented to use the

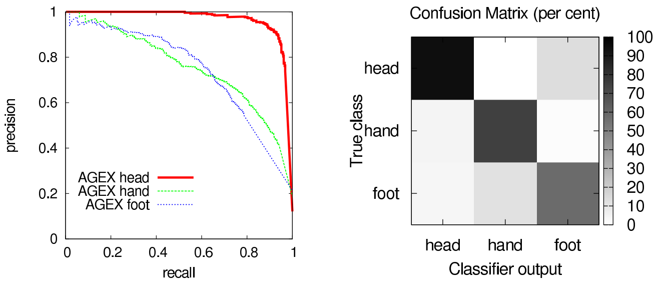

Fig. 5. Precision/recall curves and confusion matrix for the body part classifiers at the AGEX interest points extracted from a real test sequence.

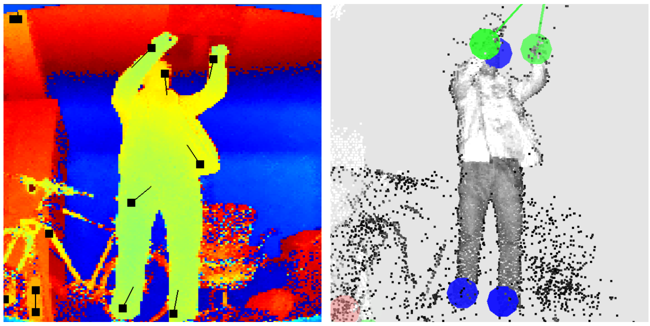

Fig. 6. Increased classifier confusion due to self occlusion. The head receives a high confidence from the *foot*-classifier since the local descriptor patch is influenced by the occluding right arm.

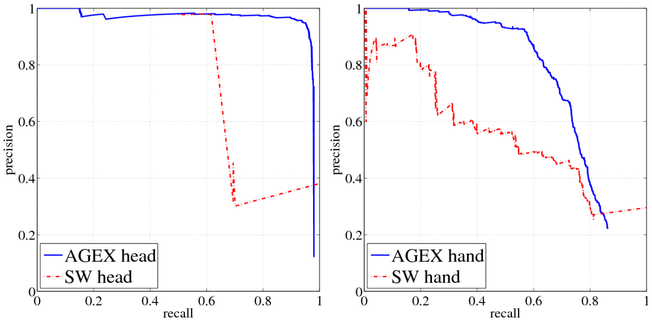

Fig. 7. Precision/recall curves for the sliding window detector (red, dotted) and AGEX (blue, solid) for the classes *hand* and *head*.

same classification algorithm for image patches as described in Sec. V-A. The sliding window algorithm was trained using the standard paradigm of extracting overlapping patches from the training images, with all patches containing parts of interest as positive examples and a large number of random patches in the scene that either contain other parts or the background as negative examples. At test time, SW classifies all possible patches in the image and detections are filtered using non-maximal suppression.

Figure. 7 shows the precision/recall curves of the two algorithms for the body parts *head* and *hand*. The free parameter of the curves is a threshold applied to the classifier confidence. We use the standard metric of considering a detection to be correct when the area intersection of the detected bounding box and the true bounding box divided by the union of the areas is greater than 50%. As the graphs show, both algorithms perform well at identifying and localizing the head of a human. Our detector offers significantly higher precision as the recall increases, whereas the standard detector rapidly drops off. The head usually has a consistent orientation during standard movements of a human, and therefore the benefit of our algorithm mostly presents itself as increased recall. When identifying hands, AGEX performs significantly better across the range of recall levels.

We also evaluated the false negative rate caused by the restriction to the AGEX interest point set and found that 402/407, or 98% of the patches containing the head were in AGEX$_5$, which is consistent with the previous experiment. The respective number for hands was 735/851 (86.4%). Thus, our maximum recall is reduced, but the precision increases significantly by allowing the classifiers to only learn the distinction between types of normalized AGEX patches. At the same time, the AGEX algorithm uses fewer negative training examples and makes better use of them. At test time, rather than considering all possible windows and scales, AGEX only needs to consider five patches centered around points in AGEX$_5$, which also represents a significant decrease in computational complexity.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we defined the AGEX interest point detector and provided an efficient algorithm for executing it on depth data. We evaluated the efficacy of this detector in the context of real-time human part detection using learned patch classifiers and found that it significantly increases precision and recall while simultaneously decreasing computational complexity. This work serves to provide a solid foundation of local detectors which can be incorporated in more complicated algorithms that use more global inference algorithms to incorporate temporal and articulated model cues.

REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, Dec. 2006.

[2] B. Steder, G. Grisetti, M. Van Loock, and W. Burgard, “Robust online model-based object detection from range images,” in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, St. Louis, MO, USA, Oct. 2009.

[3] M. Ruhnke, B. Steder, G. Grisetti, and W. Burgard, “Unsupervised learning of 3d object models from partial views,” in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2009.

[4] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proc. of The Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[5] G. Zou, J. Hua, M. Dong, and H. Qin, “Surface matching with salient keypoints in geodesic scale space,” *Computer Animation and Virtual Worlds*, vol. 19, 2008.

[6] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[7] J. Hu and J. Hua, “Salient spectral geometric features for shape matching and retrieval,” *The Visual Computer*, vol. 25, no. 5, pp. 667–675, 2009.

[8] K. Nickel and R. Stiefelhagen, “Visual recognition of pointing gestures for human–robot interaction,” *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[9] J. Ido, Y. Matsumoto, T. Ogasawara, and R. Nisimura, “Humanoid with interaction ability using vision and speech information,” in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 1316–1321.

[10] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger, and N. Blanc, “An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger),” in *Proc. SPIE*, vol. 5249, 2004, pp. 534–545.

[11] B. Golden, “Shortest-path algorithms: A comparison,” *Operations Research*, vol. 24, no. 6, pp. 1164–1168, 1976.

[12] A. Torralba, K. Murphy, and W. Freeman, “Sharing features: efficient boosting procedures for multiclass object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004.

[13] S. Gould, O. Russakovsky, I. Goodfellow, P. Baumstarck, A. Ng, and D. Koller, “The stair vision library (v2.3),” http://ai.stanford.edu/~sgould/svl, 2009.