Exploring Depth Information for Head Detection with Depth Images

Siyuan Chen, Francois Bremond, Hung Nguyen, Hugues Thomas

Exploring Depth Information for Head Detection with Depth Images

Siyuan Chen, Francois Bremond, Hung Nguyen, Hugues THOMAS
INRIA Sophia Antipolis
2004 Route des Lucioles, 06902 Valbonne, FRANCE
siyuanchen@ymail.com, francois.bremond, hung.nguyen, hugues.thomas @inria.fr

Abstract

Head detection may be more demanding than face recognition and pedestrian detection in the scenarios where a face turns away or body parts are occluded in the view of a sensor, but locating people is needed. In this paper, we introduce an efficient head detection approach for single depth images at low computational expense. First, a novel head descriptor is developed and used to classify pixels as head or non-head. We use depth values to guide each window size, to eliminate false positives of head centers, and to cluster head pixels, which significantly reduce the computation costs of searching for appropriate parameters. High head detection performance was achieved in experiments – 90% accuracy for our dataset containing heads with different body postures, head poses, and distances to a Kinect2 sensor, and above 70% precision on a public dataset composed of a few daily activities, which is higher than using a head-shoulder detector with HOG feature for depth images.

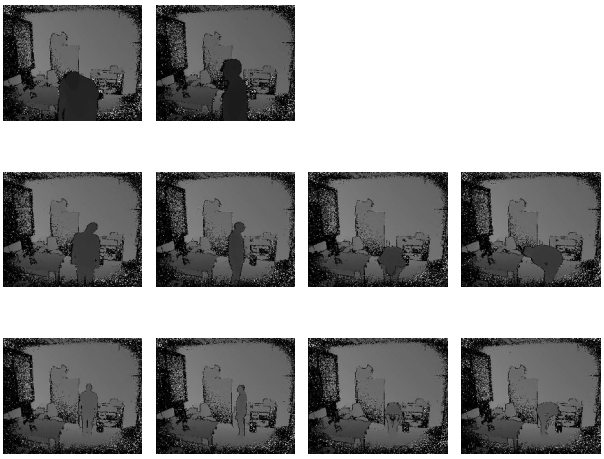

Figure 1. Examples of depth maps of standing (the first two columns) and bending (the 3rd and 4th columns) postures in full 360 ° views to a Kinect2 sensor and with different head poses. Depth maps of the three rows correspond to the distance of one, two and three meters to the Kinect2 sensor.

1. Introduction

Significant research achievements have been made on face recognition and pedestrian detection for applications in surveillance, human tracking, counting people, to name a few. However, they are not applicable when a face is turned away from a camera, in far-field, or few body parts appear in cameras. In these scenarios, robustly locating people is also urgently needed, e.g. in home care centers, to monitor elderly people performing daily activities where face and body occlusions often occur. Thus head detection is highly valuable as an alternative in locating people. Figure 1 shows some occlusion examples in standing and bending cases.

Head detection aims to locate head center robustly in the wild, regardless of body postures and head orientations. Some analytics can benefit from it when head region is required in the first place, e.g. head pose estimation [5] and human gaze estimation [7].

Nevertheless, direct head detection with depth images is not often seen in current research when inexpensive depth cameras become available, allowing us to better handle shape information from depth data than RGB data. Recent studies with depth images include people counting [2][11][17], pedestrian detection [14], predicting body joints position [12], head pose estimation [5], and fall detection [10]. Among them, the methods of head detection are an extension of appearance and feature based methods with RGB data, e.g. patch classification [5] and head-shoulder detector with HOG (Histogram of Oriented Gradients) feature [10]. Their tests are often limited to specific contexts such as facing cameras within one meter or people start falling. Head detection without body and distance constraints has not been addressed.

This paper introduces a novel approach for direct head detection, taking the advantage of the depth information measured from a single Kinect2 camera. As opposed to conventional classification with patch images or omega-shape detector with multi-scaled windows [5] or fixed head

978-1-5090-3811-4/16/$31.00 ©2016 IEEE

size searching and tracking [10], our method is more flexible with head appearance, that is, two-shoulder shape and multi-scaled windows are not necessary. The contribution lies in: 1) utilizing depth information of each pixel to determine different parameters when performing the feature level, probabilistic classifier level and clustering level. This strategy has general implications and may be applied for any other similar parameter requirements; 2) proposing a new head descriptor and filter which not only capture the characteristics of a head and the closest body part around, but also robustly detect people wearing long hair and a hat; 3) creating a new RGBD image dataset for real-world head detection challenges. It contains people in different body postures, standing and bending, in different orientations, forward, backward and profiled, in different distances, around one to four meters away from a Kinect2 camera, and performing different head poses.

2. Related Work

Generally speaking, head detection shares some similarities with face recognition and pedestrian detection. Following their directions, one research line for head detection is to find discriminate features and strong classifying approaches. Haar feature and cascade classifier have shown discriminating power for face recognition [13]. Similarly, HOG feature has demonstrated suitability for pedestrian detection [3]. When applying them for head-shoulder detection in [4], the authors found that HOG feature outperformed Haar and SIFT features. By combining multilevel HOG and LBP (Local Binary Pattern) features, the study [16] significantly improved head detection performance and achieved 89% detection rate at 10^{-4} FPPW (False Positive Per Window). In spite of explicit defined features, inexplicit head features can also be learned with direct input of head patches. This appearance-based method is often seen in head pose estimation where head region is either pre-located or detected along with head pose estimation [9].

Since the advent of low cost RGBD cameras, depth map, which describes measured distances of objects in a scene to a sensing camera, has been treated as an intensity image. As a direct head detection study, [10] detected head with depth data to recognize people for fall detection. In their head detection algorithm, moving objects were firstly detected by background subtraction and clustered by thresholding distance. Then possible head positions were searched throughout contour segments by fitting a circle with a certain head diameter. HOG features were then extracted from the rectangular grids around each head candidate, and went to a head and shoulder classifier for further identification. However, they assumed that head is always on the top of moving objects, and their direct head detection accuracy was not reported for varied body postures and head poses. As another relevant study, [5] used depth image patches to directly estimate head location and orientation at the same time with discriminative random classification and regression forests. Head pose was voted only for positive head patches. In their method, densely extracted patches were required and subjects were facing a camera doing different head poses. From their showed training patch examples, it might be more accurate to say that they are face patches because only the face region was cropped.

Among head detection studies, whether with RGB data or depth maps, most tried to locate a head-shoulder shape, which is one of the distinctions to face detection. It makes sense because head and shoulder are very close body parts and less prone to be occluded in camera views than other parts. However, in some contexts, such as people performing daily activities, both shoulders clearly to be viewed all the time by cameras is not realistic. Therefore, in [10], once head-shoulder shape was found, it would not perform it again. Instead, it began tracking the moving object. Furthermore, HOG and LBP features are useful for describing shape and texture, which might make head-shoulder descriptor sensitive to the omega shape and fail for people wearing a hat. It is also worth noting that to extract these features over a whole image, sliding window needs to be performed at a specific range of scales, the trade-off between detection-error and window metric needs to be considered.

In another research line, efforts are devoted to handle false positives and false negatives given these favorite features and classifiers to improve head detection accuracy. Typical method is to locate ROI (Region of Interest) first by motion segmentation and searching head-like candidates from ROI for head-shoulder classification rather than work on the whole image [10], so that falsely detected heads from environment become impossible. Another improvement were made through head tracking where data association rules with regard to the detected heads and trajectory were employed [15]. Because of these rules, false positives were suppressed and false negatives were recovered.

Although these approaches achieved reasonable results, they inevitably add extra concerns on the robustness of motion segmentation and tracking as both of them are in active research. Meanwhile, they introduce some limitations on head detection, such as requiring a collection of video sequences and head initialization, therefore they are not applicable to still images.

Over recent years, direct head detection did not gain much attention from computer vision community, probably because these difficult cases which are inapplicable to face recognition and pedestrian detection are also challenging for current head detection methods. In this paper, we address these difficult situations. Similar to face recognition with LBP in [1], we use circular windows but not multi-scaled. New head descriptor is developed specifically for

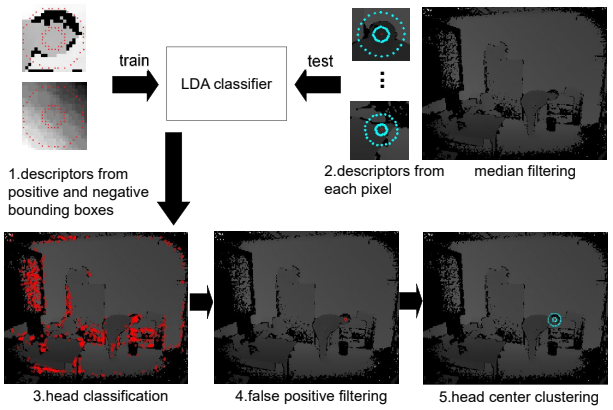

Figure 2. Framework of our proposed head detection approach

depth data to handle shoulder occlusion problems. Depth information also serves for parameter estimation without multi-scaled search, which significantly reduces processing overheads. As far as we know, this is the fist attempt to propose such depth map application.

3. Head Detection with Depth Images

A general framework of our proposed head detection method with depth data is shown in Figure 2. Firstly, a classifier is trained with head descriptors extracted from annotated head regions on depth images after a 3 x 3 median filtering. Secondly, before a new depth image is tested, a 3 x 3 median filtering is also applied and head descriptors are extracted from each pixel. Thirdly, each pixel is classified as head center or non-head-center based on its descriptor. Then a false positive filer is employed to select the most possible head centres of the depth image, followed by head center clustering to determine the final locations. Details of head descriptor with feature arrangement, false positive filtering, and head center clustering are explained below.

3.1. Depth-based head descriptor

For the center of a head, its distance to any other point on the head that can be seen from a camera is only around a few centimeters. Similarly, its distance to any point on the neck or the closest body part such as the back, if the neck is invisible like the bending cases in Figure 1, is also in such a small range but its distance to any other point in the environment, i.e. non-body part, is far larger than that range and also varied. To describe the depth texture and relationship of the head, the closest body part, and its environment, depth information on two circular windows centered at each pixel is used. We use the notation (P,R,C) to indicate P sampling points on a circle of radius R and centered at pixel C, and use $\mathbf{DP}$ and DC to refer the depth values of P points and C respectively. If this pixel C is near the head center, one circular window (R_1) is entirely inside the head and the other (R_2) is completely enclosing the head, as shown in Figure 3 left. The head descriptor is defined as

$$\mathbf{H} = DC - DP_{ij}, i = 1, 2; j = 1, ..., P \quad (1)$$

where DP_{1j} and DP_{2j} are the depth values of the circles inside the head and enclosing the head respectively. These depth differences between the circle center and two circles are concatenated to form a $2P$-dimensional vector representing the descriptor $\mathbf{H}$ at pixel C.

For each pixel, its circular window sizes are determined by its depth value:

$$\begin{aligned} R_1 &= \alpha * f/DC \\ R_2 &= \beta * f/DC \end{aligned} \quad (2)$$

where α and β are constant values, smaller than the low bound and larger than the high bound of statistically measured head size in population respectively. It is different to trying multiple-scaled windows to directly estimate each of varied head size where the processing overhead is relatively high. Meanwhile, since the sampled points on the two circles are not directly from head boundaries, head appearance does not affect the descriptor. Therefore, it can detect a head with a hat on it or with different hair styles. Here f is also a constant value representing the focal length in pixels which can be obtained from geometric calibration [8]. It is worth noting that these pixels whose circular windows (any sampling points) are outside the range of image size or whose depth value is zero due to noise in acquisition are removed from further processing.

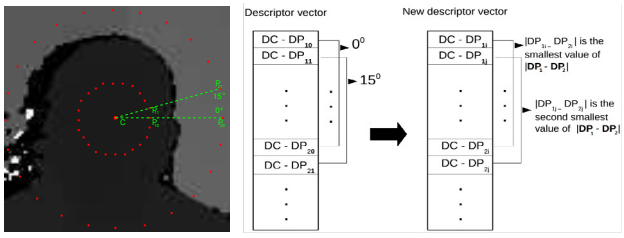

Figure 3. Depth-based head descriptor (left) and feature arrangement (right).

3.2. Feature arrangement

Due to different head poses, body postures and different views of a camera, the closest body part to the head may appear to be any part of the circle R_2. To reduce this confusion during head model learning, we introduce a feature arrangement to align body parts in each descriptor. The body parts will always be placed from the right central point of the circles, at 0° degree as shown in Figure 3 left. That is, the absolute depth difference between DP_{1j} and DP_{2j} is

firstly sorted and then the descriptor $\mathbf{H}$ is reorganized based on its ascending order.

It is based on the assumption that compared to the surroundings, the distance between the head and the closest body part should be smallest. However, it may occur that an object just appears next to the head causing the smallest depth difference. Nevertheless, it will not cause much trouble to the descriptor's discriminative power because few objects have a ring shape surrounding a head. Thus it only affects a small portion of the descriptor's depth values.

3.3. Depth-based false positive filter

Similar to the studies aforementioned [15], there are inevitably many false positives from surroundings after head classification, because the features are hardly unique to head attributes only. Different to face detector using cascaded classifiers [13], a filter is introduced to eliminate false positives. As we mentioned before, the smaller circular window is entirely inside a head region if the circle center is the head center, and there is a margin between the head boundary and all sampling points. If the circle center moves around by a few pixels, this circular window is still inside the head and the other still encloses the head. This makes all these pixels near the circle center detected as head centers. While for most false positives, when one pixel is happened to be detected as a head center, the pixels next to it have low chance to be detected as head centers all, as the red dots shown in Figure 4 left.

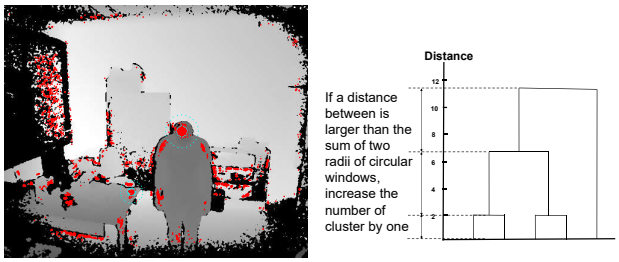

Figure 4. Examples of detected head centers in red over a whole depth map (left), where the true head center has 100% detection rate per window and the pixel on the table corner has lower detection rate per window, both are shown in blue circular windows. The principle of depth-based clustering is described on the right.

According to this observation, we propose the following functions to filter false positives. Suppose after head classification, there are k pixels classified as head centers. L_i is the label for pixel C_i. 1 denotes head center and 0 denotes non-head center. Given a pixel C and its circular window radius R_1, we can obtain a set of pixels $\{\mathcal{P}$=(R_1,C)$\}$ inside this small circular window. We define F as detection rate per circular window :

$$F_i = \begin{cases} \frac{\sum_{n\in\mathcal{P}_i} L_n}{\#\mathcal{P}_i} & if DC_i < D \\ 0 & otherwise \end{cases} \quad (3)$$

where #(·) is cardinality. Then we assign a new label L' for each pixel with this filter:

$$L'_i = \begin{cases} 1 & if F_i = max(\mathbf{F}) \\ 0 & otherwise \end{cases} \quad (4)$$

where only these head candidates who have the highest detection rate per circular window are possible to go to the next procedure. It is assumed that ideally the detection rate near a true head center is 1 and when this condition is met (or detection rates are same but less than 1) by multiple head candidates, they will then be processed in the next clustering stage. Otherwise, only the one with the highest detection rate is chosen and directly taken as the true head center. Meanwhile, we set a condition for a valid depth value D because if it is too far from the camera, the radius of circular window R_1 is too small to discriminate anything while the detection rate can be easily high if it is a false positive.

3.4. Depth-based Clustering

When multiple head candidates survive filtering, the next step is to determine head centers by clustering. Usually it is difficult to know how many heads in images beforehand for the number of clusters. Depth information of each pixel is again used to determine the cluster number. Then each cluster center is taken as the final detected head center.

Hierarchical clustering is employed where a hierarchical cluster tree is created according to the Euclidean distance between every pair of the selected head candidates. Given the depth value of each head candidate, we know their radius R_1 of the smaller circular window. If the cluster center distance between two head candidates is larger than the sum of their radius, the two head candidates is set into two clusters. Therefore, by examining the distance of the clusters at one level and the next in the cluster tree, we can easily find how many levels of clusters whose distances are larger than the sum of their radius to know the number of clusters. This is illustrated in Figure 4 right.

4. Experiments

4.1. Data acquisition and head center labeling

As far as we know, there is few dedicated public depth dataset for head detection. Thus we acquired a dataset containing some difficult head detection cases with a Kinect2 sensor. This dataset consists of 882 depth images obtained from two males and three females – one wearing glasses and one wearing a hat. They performed totally seven sets of different head poses and body postures at different distances. In each set of 126 depth images, one subject firstly stood at about one meter away from the sensor, rotating the head by around $-60\,^\circ$, $0\,^\circ$, and $60\,^\circ$ for pitch, around $-40\,^\circ$, $0\,^\circ$, and $40\,^\circ$ for roll, and around $-75\,^\circ$, $0\,^\circ$, and $75\,^\circ$ for yaw.

The standing posture contains four directions, facing forward, backward to the sensor, left and right in which only one shoulder was visible to the sensor. With this distance, only half body was visible to the sensor. Next the subject stood at about two meters away where the whole body can be seen, performing the same head poses with four standing directions. Then the subject faced the camera and bent the upper body by around $90\,^{\circ}$, performing the same head poses. He/she turned the body to right by $90\,^{\circ}$ and bent again, performing these head poses, and finally turned left by $90\,^{\circ}$ doing the same postures and head poses. Lastly, the subject stood and bent at around three meters away, doing the exactly same head poses and standing directions.

The ground truth of head position on each depth image was manually annotated with a bounding box. So the center of the bounding box is taken as the head center. Later, one non-head bounding box was automatically generated from each depth image by randomly selecting a patch of the same bounding box size outside the labeled head region.

Despite our collected dataset for training and testing, we used a public Kinect2 depth dataset [6] collected by the Cornell University to test the learned head model and compare our algorithm with other available algorithms. As opposed to the designated head poses and body postures in our dataset, this dataset contains a few human daily activities with object interaction in an unsupervised setting.

4.2. Experimental design

LDA(Linear Discriminative Analysis) is employed to classify each pixel as head center or non-head center in the classification procedure. We use leave-one-set-out scheme to train a head classifier with the descriptors extracted from 756 head bounding boxes and 756 non-head bounding boxes. Here half of the bounding box length is taken as radius R_2 and R_1 is one thirds of R_2. For each of the rest sets (126 depth images), the descriptors for each pixel are tested to determine the head candidates for clustering. Centered at each of the final detected head centers, a rectangular is generated whose length is the radius of the larger circular window in order to compare with the annotated head bounding boxes. Dice coefficient is calculated and if it is larger than the threshold of 0.3, the head is correctly detected. Average precision and recall of the seven sets are then computed.

There are a few constant parameters that need to be determined. According to statistics, the smallest and largest head size are around 120 mm and 200 mm. As the Kinect2 sensor was placed about 2 meters high, well above the head of subjects, the head size is smaller in the camera view. We set β to 150 mm, and α is one thirds of β, 50 mm in equation 2 for all datasets. Meanwhile, the value of focus length in pixels f is set to 366 pixels according to [8]. The value of D in equation 4 is set to 4 meters. According to equation 2, it makes R_1 to be 4 pixels, being little informative. The only parameter which is difficult to determine is the number of sampling points on the circular windows in equation 1. We set P to 24, that is, in an interval of $15\,^{\circ}$, and report the head detection results. Then we analyze how P affects the detection performance in Discussion.

Body posture	Distance (approximate)	Precision	Recall
Standing	1 meter	1.00	1.00
Standing	2 meters	1.00	1.00
Standing	3 meters	0.99	0.99
Bending	2 meters	0.88	0.88
Bending	3 meters	0.51	0.51
Overall		0.90	0.90

Table 1. Head detection performance for each body posture and distance category and overall detection performance when P is 24 (an interval of $15\,^{\circ}$).

5. Results

The mean head detection precision and recall for our dataset are presented in Table 1. Please note, if something other than head is incorrectly detected as a head, it creates one false positive and one false negative. To illustrate how our proposed approach being able to detect head in different body postures and head poses, and how distance affects detection performance, we detailed head detection performance in each category besides the overall accuracy.

To demonstrate the influence of the number of sampling points on the two circular windows, we sampled points with different interval of degrees over the circular windows to observe the detection performance, as shown in Figure 5.

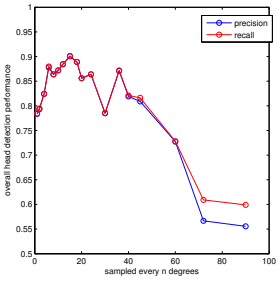

Figure 5. Detection performance vs. number of sampling points

The head detection performance on the dataset of Cornell University is shown in Table 2, where our algorithm is compared to the algorithm in [10]. For this dataset, we used skeleton data as the ground truth.

Another head detection comparison is made between using skeleton, the algorithm in [10], and our algorithm on a benchmark video of Cornell University, where 148 head locations were manually annotated. The results are illustrated in Table 3.

	Nghiem's algorithm [10]	our algorithm
Image number	2785	2785
True positive	1447(51.9%)	1976(70.9%)
False positive	213(7.6%)	301(10.8%)

Table 2. Head detection performance comparison on the Cornell University dataset.

	Skeleton	Nghiem's algorithm [10]	our algorithm
Image number	148	148	148
True positive	110(74.3%)	52(35.1%)	124(83.8%)

Table 3. Performance comparison of three head detection algorithms on a benchmark video of Cornell University.

6. Discussion and Conclusions

High performance was achieved for close-ranged head detection irrespective of varied body postures and head poses in our real-world testing set, above 85% in precision and recall as shown in Table 1. It can attribute to the instinctive usage of depth information. Firstly, a double circular window descriptor successfully captures the depth relationship between a head center, its surroundings and the closest body part. It overcomes the limitation of omega-shape based features when one shoulder is invisible to a camera. Meanwhile, feature arrangement further enhances the depth relationship. It improves the classification accuracy of training data, from around 80% to more than 90%. Secondly, as opposed to the practice of shifting sliding windows by a certain number of pixels, where it assumes that content in windows changes little between the skipped pixels, we utilized the invariance to strengthen the true positive of head center classification on the pixel level. Therefore, from Table 1, 99% of the time the head center was the most dense area if the head was not occluded by a shoulder when standing, disregarding the distance. Thirdly, using depth information to estimate circular window size and the number of clusters works well and significantly reduces the complexity of the algorithms employing multi-scaled windows to scan images and employing parameters tuned for different datasets.

In our approach, the only concerned parameter is the number of sampling points. If it is too big, it not only assigns these points with repetitive depth values and collects more depth information about surroundings in the larger circular window, but also increases processing time, and vice versa. Figure 5 demonstrates that the detection performance can be degraded markedly if the number of sampling points is inappropriately selected. However, performances of P in the range of 18 to 60 (sampled every 6° to 20°) are relatively stable. Here we fix the number of sampling points to 24 across different datasets.

As we mentioned before, there are few studies dedicating their methods for head detection although they are often head related. It is of little interest if we tested their algorithms just for comparison given our knowledge that their detectors were trained for face only or for a limited body rotation while our dataset contains a number of non-face and full body rotation cases. However efforts were made to compare given similar depth data cases. As relative comparisons, [5] has the same processing in terms of frame-by-frame depth images, but used depth image patches. According to their data acquisition, their depth data is similar to the cases in our dataset when subjects are one meter away facing the sensor. The best detection rate is 99% when the window stride is four pixels and the head center error is around 15 ± 22 mm. Although it is difficult to directly compare with the head center errors because we used dice coefficient with a threshold of 0.3, the best head detection performance can be 100%. [12] also shares some common to our approach in terms of using single depth image and per-pixel classification but their objective is to predict 3D position of body joints. As one body part, head was detected with an average precision of 91% with a body rotation of $\pm120^\circ$ to the sensor (distance is unknown), but the specific detection rate for head when a body was in the full 360° scenario was not reported. Their detecting cases are close to the standing cases in our dataset while our detection rate can be above 99%.

With available Kinect2 depth datasets and available head detection algorithms, given trained head models, we found that our algorithm outperforms the head-shoulder detector with HOG feature in [10] and the skeleton method used in [6] on the daily human activities depth dataset, as shown in Table 2 and Table 3.

The main failure cases in our study are when subjects bending and rotating their head in far field. This is likely due to two reasons. Firstly, when subjects bent their body and rotated head away from the sensor, the head is easily occluded by a shoulder, thus the detectable head area is reduced. Secondly, in far field, the accuracy of depth map decreases, especially for reflective and dark area [8]. We find that dark hair becomes visible in depth map and the corresponding distances in the depth map are larger than expected. This phenomenon seldom occurs in the one meter cases. Larger depth values in the hair region can result in the smaller circular windows being larger than the head region. Although they may only account for a few pixels, the detectable head region further shrinks. The consequence is that the detection rate per circular window becomes lower, losing the competition to the objects who share close attributes of the descriptor, such as a table corner and bum.

Future work may involve discriminating non-head regions as this approach is based on the assumption that there are heads in single depth images, and may extend it to crowd scenes. Also the use of Kinect allows to recognize heads within 4 meters, so experimentation of the method using another depth sensor with a wider range may be conducted.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *ECCV*, 2004.

[2] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. D. Bimbo. Real-time people counting from depth imagery of crowded environments. In *AVSS*, 2014.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *ICPR*, 2008.

[5] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *CVPR*, 2011.

[6] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Watch-n-patch: Unsupervised understanding of actions and relations. In *CVPR*, 2015.

[7] K. A. Funes-Mora and J.-M. Odobez. Gaze estimation in the 3d space using rgb-d sensors. *IJCV*, pages 1–23, 2015.

[8] E. Lachat, H. Macher, M. A. Mittet, T. Landes, and P. Grussenmeyer. First experiences with kinect v2 sensor for close range 3d modeling. In *ISPRS*, 2015.

[9] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, pages 607–626, 2009.

[10] A. T. Nghiem, E. Auvinet, and J. meunier. Head detection using kinect camera and its application to fall detection. In *ISSPA*, 2012.

[11] L. D. Pizzo, P. Foggi, A. Greco, and G. Percannella. Counting people by rgb or depth overhead cameras. *PATTERN RECOGN LETT*, 2016.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moor, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[13] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[14] S. Wu. An attempt to pedestrian detection in depth images. In *IVS*, 2011.

[15] R. Xu, Y. Guan, and Y. Huang. Multiple human detection and tracking based on head detection for real-time video surveillance. *Multimed Tools Appl*, 74(3):729–742, 2015.

[16] C. Zeng and H. Ma. Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *ICPR*, 2010.

[17] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li. Water filling: Unsupervised people counting via vertical kinect sensor. In *AVSS*, 2012.