

Algorithms for Tracking and Identifying People in the Image by using Machine Learning Methods

Martin Keršner, Jakub Novák
Czech Technical University in Prague
Prague, Czech Republic

Abstract—Detection, tracking and gender recognition has become fundamental task of surveillance systems. Price of depth sensors is still falling and therefore enables discover new methods based only on depth information. This paper focuses on techniques which can be exploited using depth maps. Detection and tracking make do only with smart usage of histogram. We propose new method of estimation of human body orientation based on derived features from gradient line. Our method nearly achieves the same accuracies as common techniques. However, the best results were reached by method called Histogram Of Oriented Gradients. This feature extraction method and linear SVM classifier obtained 96.4% for estimation of human body orientation and 99% for gender recognition.

Keywords—Depth Map, Human Detection, Estimation of Human Body Orientation, Gradient Line Feature Extraction Method, Gender Recognition.

I. INTRODUCTION

IN THIS paper we deal with wide range of tasks from human body detection to gender recognition, but we mainly focus on estimation of human body orientation. Human body detection is key task for all surveillance systems and since the price of depth sensors is decreasing, detection using depth map creates new opportunities to achieve more accurate results. Estimation of human body orientation can help to watch customers' behavior and react on the base of these observations. Biometric systems recognizing humans only according to their faces still are not so general. Well developed gender recognition classifier could simplify task of face recognition because dimensionality of searched space would get approximately twice smaller or techniques used to face recognition could depend on gender and thus be more specialized.

We employ RGBD information (160×120 pixels) acquired by Kinect. Depth map together with RGB information enriches spatial information about scene and draw near to a way as human can see. It also have several more advantages such as easiness of human body detection or invariance to illumination changes. We largely use depth map, RGB information is only utilized in gender recognition classifier. The depth maps that we experimented with are captured from top-view (fig. 1). White parts are closer to the Kinect and black parts are farther. We can notice there is not full length of human legs. The 3D points, from which the depth map is computed from, were cut in particular distance from Kinect. Therefore, the floor is all black even though it is not perpendicular on Kinect.

The paper is organized as follows: we begin with discussion about related work in section II. Section III describes human

body detection. Section IV gives detail about tracking. In section V we present methods of estimation of human body orientation. Gender recognition is described in section VI. Section VII discusses the experimental results and section VIII concludes the paper.

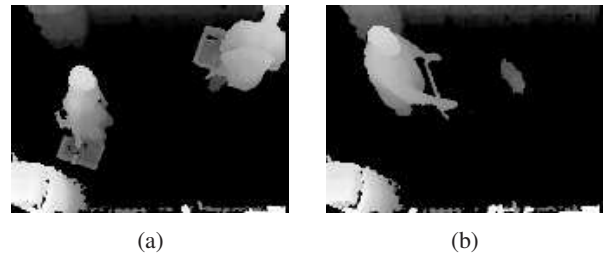


Fig. 1: Examples of input depth map. At the first image (a) there are two people, one is diagonally rotated toward camera and the other one shows only its back. The second image (b) contains a person pushing a shopping cart from left to right of scene.

II. RELATED WORK

Human detection is task, which has already been researched for long time, however it becomes popular again due to advent of depth sensors. Yujie et al. proposed new feature descriptor for human detection called Local Ternary Direction Pattern (LTDP) [1]. LTDP extends Local Binary Pattern and Local Ternary Pattern. Combination of positions of body joints and cropped image parts extracted from RGB image data are inspiring approach of feature extraction from Hao-Jen et al. [2]. Rauter researched head detection from top-view [3] and consequently human detection. His method searches for local maximas and gradient climbing confirms final head candidates.

Wu et al. experimented [4] with different methods to estimate human body orientation and proposed static and motion RGB-D feature extraction method. Other two researches [5], [6] handled estimation of orientation using descriptor called Histogram of Oriented Gradient. They employed linear SVM classifier and SVM Decision Tree respectively.

Gender recognition methods contain many ways how to distinguish between genders. Common approach was proposed in paper [7] using WLD descriptor. Jian-Gang et al. experimented with recognition based on beard detection [8]. Other popular methods were described and compared in paper [9].

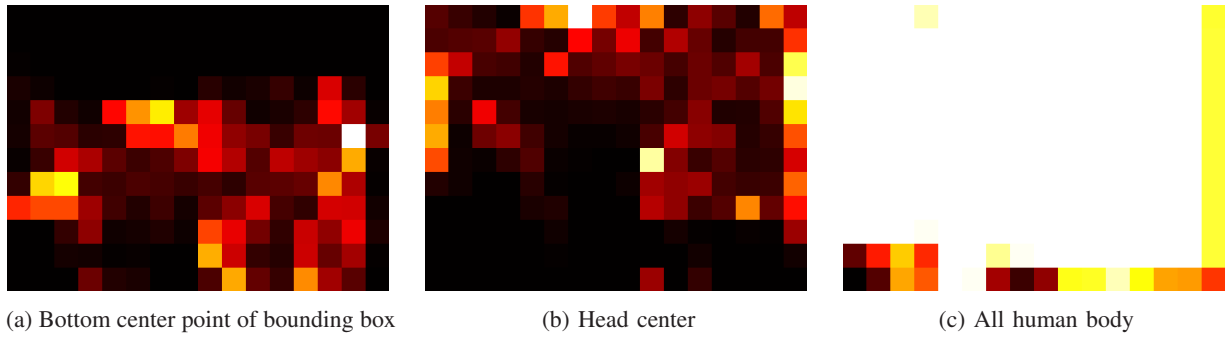


Fig. 2: Representation of heat maps computed from movements of people.

III. HUMAN BODY DETECTION

The human body detection is the first step before we can start to track or identify human. That consists from several preprocessing substeps which need to be executed. Filtering and segmentation can be considered as the main steps of this detection.

Firstly we needed to filter out background. Hence, we created a mask which was computed as maximum of depth values from couple of different scenes with no humans. Our mask computation employed 77 blank scenes. Since we had a mask we could subtract it from input depth map. The output image contains humans and undesirable unfiltered noise.

A. Filtering

Filtering is used to eliminate disruptive noise which can negatively affect performance of later executed algorithms. The main goal is also to preserve original image with its edges and details. The important setting of described methods is the size of neighborhood from which the final values are computed. The filter methods can be split into two parts: linear and non-linear filters.

The simplest linear method is called averaging. The averaging computes the average of neighborhood intensity values for each pixel. Even though this method eliminates the noise, the edges and details disappear. The second method, called Gaussian filter, is also linear filter. Performance of Gaussian filter is slightly better because it uses Gaussian function as kernel. Nevertheless the edges are still not preserved at the image as we would wish. The last method, median filter, is classed as non-linear filter. Median filter computes the median of neighborhood intensity values for each pixel. This filter achieve the best results among these described methods, because it can reduce the noise while preserving the edges, and therefore we have chosen it for our solution.

B. Segmentation

The segmentation is used to divide image into parts in specified way and connect similar parts together. Particular methods differ in complexity and number of segments which can be

distinguished. Our task is to segment remaining background parts from the foreground.

The most naive method is to determine a threshold ad-hoc and separate image into two parts, one with values smaller than threshold and the second one with values higher than threshold. This method could be hardly generalized and threshold would have to be chosen for each image.

The next method is called balanced histogram thresholding [10]. Balance histogram thresholding is also able to divide image data only into two parts. However, the threshold is selected in more sophisticated way. The selection is based on found of a balance between left and right part of histogram. An average of values inside of chosen bin of histogram, where the balance of histogram was found, is then used as a threshold. The accuracy of method is dependent on a number of bins of histogram.

The third method, called Otsu's method [11], also finds only one threshold to divide image data into two parts. Otsu's method is looking for a threshold which would maximize inter-class variance. The algorithm computes inter-class variance for each intensity level and selects that with maximum value.

Another segmentation method, called K -Means [12], is able to divide image in an arbitrary number of parts. Firstly, K pixels are randomly selected to be centroids of clusters. Then the rest of pixels are assigned to the closest clusters and centroids of clusters are recomputed. The algorithm is running until the pixels do not change their clusters.

We propose a method, adaptive thresholding, which handles segmentation task best. A threshold is derived using histogram computed from depth values and method searching for a local minimum. If the first local minimum is not occupying only one bin we select the most right bin of this sequence as the threshold. Afterward we create a binary mask. All values which exceed found threshold are marked as number 1, the rest of them as number 0. The binary mask is then multiplied with input depth map. This method does not only segment background and foreground correctly, but it can even reconstruct damaged parts of depth map, where human covers subtracted background.

IV. TRACKING

Information about human movements is important knowledge from which the part of human behavior can be derived. There is no reason to store and track all depth values of each human. Therefore, we need to determine which significant point is going to be tracked. Since there are not provided depth values of feet, which could seem as the right choice, we have decided for center of head. We consider head detection and subsequently finding center of head as an easy task when we can exploit Depth map. To increase accuracy of detection we introduce a border [3] around all scene, which allows to distinguish between complete and incomplete heads. Size of border at each side was experimentally determined as 10 pixels. The process of tracking starts when human appears in scene, though head is not completely visible. When human crosses borders system starts to store position of head. The closest points between frames are assigned together.

We know about issue, which is not solved yet, when two or more people would touch with their heads. Our head detector would misclassify these heads as one.

This section describes a method of human detection and presents different heat maps of humans' movements created from dataset which we experimented with.

A. Head Detection

Most of described solutions about head detection [3], [13], [14] include some sort of feature extraction and machine learning algorithm. In order to preserve good performance and keep high accuracy of detection, we propose a method based on histogram of depth values. Our solution assumes that the head is the highest point of human body and the size of the head is proportional to the rest of the body. Each value of examined depth map is voted to 1D histogram. According to number of bins, the n -th (in our case penultimate) bin from the end is selected and computed average of values inside. This average indicates a threshold that decides which depth values belong to the head and which do not. Known issues of this algorithm are incomplete figures and raised hands. Since we do not register a human, who does not completely appear in the picture, we do not need to deal with it. The raised hands can cause problems when the hand is touching the head, otherwise we can filter hands out according to size or shape.

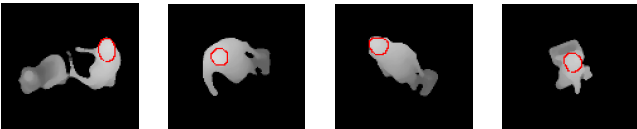


Fig. 3: Examples of well detected human heads.

B. Heat Map

The heat map can be a visualization tool of human movements over time. It can reveal information, which cannot be seen at the first glance. We created three heat maps (HM)

from almost 7 minutes long record where more than 37 people appear. They differ in point at human which we count. The first HM (fig. 2a) is computed from bottom center of bounding box of detected human and the second one (fig. 2b) uses head center of human. The last HM (fig. 2c) employ all depth values of human body.

According to generated heat maps we can notice that people spent more time in bottom right part than in left one. The people neither went there nor they could not because of some barrier. Another observation could be that they came from the right side more often.

V. ESTIMATION OF HUMAN BODY ORIENTATION

After a human is correctly detected we can further use features extraction and machine learning methods to identify additional characteristics.

This section discusses two different approaches how human body orientation can be estimated: common methods to derive features from image data and method with apriori knowledge of appearance of human displayed using depth map.

A. Common Methods

The image in computer interpretation is large matrix with miscellaneous values, there are no relations between them. Feature extraction methods are used to get more significant information about image. Output from these methods has the same size as input, therefore after extraction the output is usually split into blocks and values of each block are voted to histogram. Subsequently these histograms are concatenated together to one feature vector.

Histogram of Oriented Gradients (HOG) is feature descriptor, which was firstly proposed for human detection [15]. The first step of HOG computation is to apply 1D mask $[-1, 0, 1]$ in vertical and horizontal direction. Since derivatives in both directions are obtained, particular orientation of magnitudes can be calculated. Orientations are then voted to cell histograms which contain bins in range from 0° to 180° . Thereafter, cell histograms are connected to blocks. We experimented only with rectangular blocks. Eventually each block can be normalized and thus prevent issues which are related to changes in illumination or shadowing. We employed normalization method called L2-norm (eq. 1).

Local Binary Pattern (LBP) [16] became popular feature extraction methods due to robustness and easiness of computation. The algorithm requires two parameters; size of neighborhood (distance from currently computed pixel) and number of pixels in neighborhood. LBP compares all pixels from neighborhood with its center pixel. If the center pixel is smaller than pixel in the neighborhood then LBP assigns 1 to this pixel, otherwise 0. Concurrently these computed values are multiplied with power of two, where the power is order of pixel in the neighborhood, and finally summed together. One of the advantages of LBP is rotation invariance which enables to describe image still the same.

$$f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \quad (1)$$

The last examined method is called Histogram of Oriented Normal Vectors (HONV) [17]. HONV was designed as feature extraction method specifically to capture characteristics from depth information. The process of computation begins with applying Sobel operator to acquire dx and dy derivatives. The second step is to calculate azimuth and zenith angle from dx and dy derivatives. After that, cells are split and 2D histogram is created for each cell from azimuth and zenith angles. Since there is not proposed any way, how to reduce the size of feature vector, we compute 1D histogram from 2D histogram created at the previous step of algorithm.

B. Gradient Line

Length of feature vector, computed using methods like LBP, HOG or HONV, is too large. It can reach even few thousands of values. We could employ methods for dimensionality reduction to get smaller number of features, however, we propose a method based on gradient line. The idea to use gradient line utilize one property of depth map. Depth values continuously decrease from head as highest point in direction toward to feet. Thus, starting point of gradient line is center of head. In order to obtain the most general gradient line we create line with different orientations (72 lines, from 0° with step 5°). Gradient line is then trimmed at the end where zero values start. If there is anywhere zero value beneath the rest of line, gradient line is not tested anymore. We have experimented with decision rule which determines the most accurate gradient line among all of remaining ones. We tried to make decision based on length of gradient line, sum of depth values beneath the line and ranking (eq. 2). Ranking is computed as the inverse of second power of mean of differences through values which lie on the line. The best result we get when we decide according to the longest line. Since we have selected gradient line, features can be derived. Our method takes into account 6 main features:

- position of head (x-axis, y-axis),
- angle of gradient line,
- length,
- sum of depth values on the line and
- ranking.

Other 12 features are computed as multiplication combinations of main features except the position of head.

$$R = \frac{1}{\bar{\delta}^2} \quad (2)$$

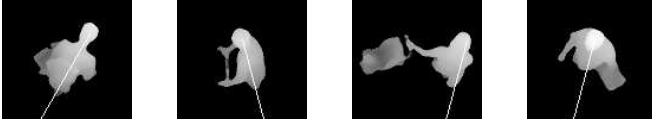


Fig. 4: Examples of found gradient line.

VI. GENDER RECOGNITION

We decided to recognize gender using common techniques, previously discussed in section V-A. Paper [7] has inspired us to use Weber Local Descriptor (WLD) [18]. WLD computes simultaneously two components: differential excitation and orientation. These components are voted into histogram and concatenated eventually. The differential excitation depends on intensity differences between particular pixel and its neighborhood and also on current intensity of pixel. The orientation is computed as gradient orientation of particular intensity of pixel.

Since we have implemented other feature extraction methods we employed them again on gender recognition.

VII. EXPERIMENTS

All experiments were performed using computer with processor Intel® Core™ i3, 4GB RAM running on GNU/Linux 3.13 x86_64. Preprocessing, detection and feature extraction scripts were implemented in GNU Octave 3.8.1¹. Classification algorithms were employed using Machine Learning library for Python 2.7.6² called scikit-learn³. Implementation of all tested methods can be found at GitHub in the account martinkersner. The name of repository is Algorithms-for-Tracking-and-Identifying-People-in-the-Image-by-using-Machine-Learning-Methods.

A. Dataset for Estimation of Human Body Orientation

Dataset consists from 12,385 images of 37+ distinct people. The images are sorted (fig. 5) to 8 classes according to Human Body Orientation. Number of the images in each class is provided by graph below. The images of people were taken from 30fps record. Therefore, in order to prevent doubts about overfitting we are going to perform on data (10 subdatasets) with different ratio of selected samples from the whole dataset.

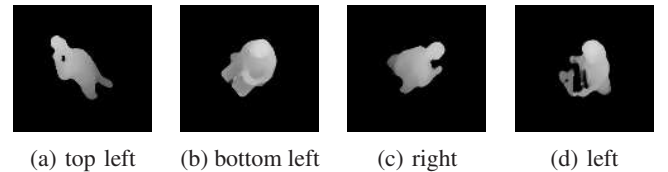


Fig. 5: Samples from dataset used for estimation of human body orientation.

B. Dataset for Gender Recognition

Dataset for gender recognition consists from 290+ images of 5 people (3 male, 2 female). The images contain only front and profile of heads, all mixed together. Each image was converted to gray intensities and size each of them is 71×71 pixels.

¹<http://www.gnu.org/software/octave/>

²<https://www.python.org/>

³<http://scikit-learn.org/stable/>

TABLE I: Accuracies of estimation of human body orientation using HOG with different settings.

Bin	C	B	Accuracy
8	6	3	94.4%
8	8	3	92.8%
9	6	3	95%
9	6	4	95.2%
9	8	3	93.2%

This dataset contain too small amount of samples and therefore we cannot expect from classification algorithms high level of generalization.

C. Estimation of Human Body Orientation

In the research we experimented with three feature extraction methods: Histogram of Oriented Gradients, Local Binary Pattern and Histogram of Oriented Normal Vectors. Linear SVM ($C = 0.001$) was applied to estimate orientation. Training of each dataset employed 10-fold cross validation. Settings of LBP and HONV was the same. Size of block was determined as 10×10 pixels. Features of each block were voted into 10 bins. Size of cell for HOG was 6×6 pixels and size of block was 4×4 cells. Blocks were normalized using L2-norm and histogram were divided into 9 bins.

Experiments were performed (fig. 6a) on 10 subdatasets with various sizes. HONV has distinctly turned out as the least successful method with the highest accuracy 75.2%. LBP achieved accuracies in range from 84.2% to 93.2%. The best results between 84.8% and 96.4% received HOG.

HOG was thoroughly examined. Results are shown in table I, where C indicates size of cell in pixels and B is size of block. All tests were performed with 60% of dataset. Accuracies are alike and thus not so strongly dependant on settings.

Estimation of human body orientation, based on derived features from gradient line, has surprisingly reached high accuracies (fig. 6b), almost as good as HOG achieved, though it contains 200 times less features. The best results between 78.8% and 95.3% were acquired using Random Forest classifier, which outperformed different SVM methods (Linear: $C = 0.5$, Poly: $C = 1$ and Rbf: $C = 1$) and Decision Tree classifiers (Gini and Cross-Entropy; both have 40 as maximum level, 3 as minimal number of samples in leaves). Random Forest classifier utilized 70 estimators. Training of each dataset employed again 10-fold cross validation.

D. Gender Recognition

We applied common feature extraction methods to resolve gender recognition task and achieved more than satisfying results (table II). However, we need to repeat our worries about credibility of these accuracies, because used dataset contain small amount of samples.

The highest accuracy 99% was obtained again with HOG. The setting was utilized as follows: size of cell 6×6 pixels,

TABLE II: Accuracies of different feature extraction methods solving gender recognition task.

Method	Accuracy
HOG	99%
LBP	95.3%
HONV	80.8%
WLD	95.7%

size of block 4×4 , 9 bins and L2-norm. Classification was performed using linear SVM ($C = 0.001$) and 10-fold cross validation.

VIII. CONCLUSION

We have proved that human detection algorithms do not necessarily need to use feature extraction and machine learning methods when we can employ information from depth map. Our proposed solution is properly working until there does not happen special case like touching head with hand.

Estimation of human body orientation was examined by two different ways. The first one utilized Histogram of Oriented Gradient together with linear SVM classifier and obtained 96.4% on our created dataset. The another solution was employing Random Forest and our proposed method based on features derived from gradient line. There were approximately 200 times less features than at HOG method and accuracy 95.3% was nearing to HOG.

Gender recognition task was solved with HOG method and linear SVM.

ACKNOWLEDGMENT

The author would like to thank ČVUT Media Lab and eClub for funding this research within eClub Summer Camp 2014.

REFERENCES

- [1] Y. Shen, Z. Hao, P. Wang, S. Ma, and W. Liu, "A novel human detection approach based on depth map via kinect," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013 IEEE Conference on, June 2013, pp. 535–541.
- [2] H.-J. Wang, Y.-L. Lin, C.-Y. Huang, Y.-L. Hou, and W. Hsu, "Full body human attribute detection in indoor surveillance environment using color-depth information," in *Advanced Video and Signal Based Surveillance (AVSS)*, 2013 10th IEEE International Conference on, Aug 2013, pp. 383–388.
- [3] M. Rauter, "Reliable human detection and tracking in top-view depth images," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013 IEEE Conference on, June 2013, pp. 529–534.
- [4] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li, "Accurate estimation of human body orientation from rgb-d sensors," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1442–1452, Oct 2013.
- [5] K. ngam Panachit and O. S. Guat, "Estimation of human body orientation using histogram of oriented gradients," 2011.
- [6] C. Weinrich, C. Vollmer, and H.-M. Gross, "Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images," in *Intelligent Robots and Systems (IROS)*, 2012 IEEE/RSJ International Conference on, Oct 2012, pp. 2147–2152.

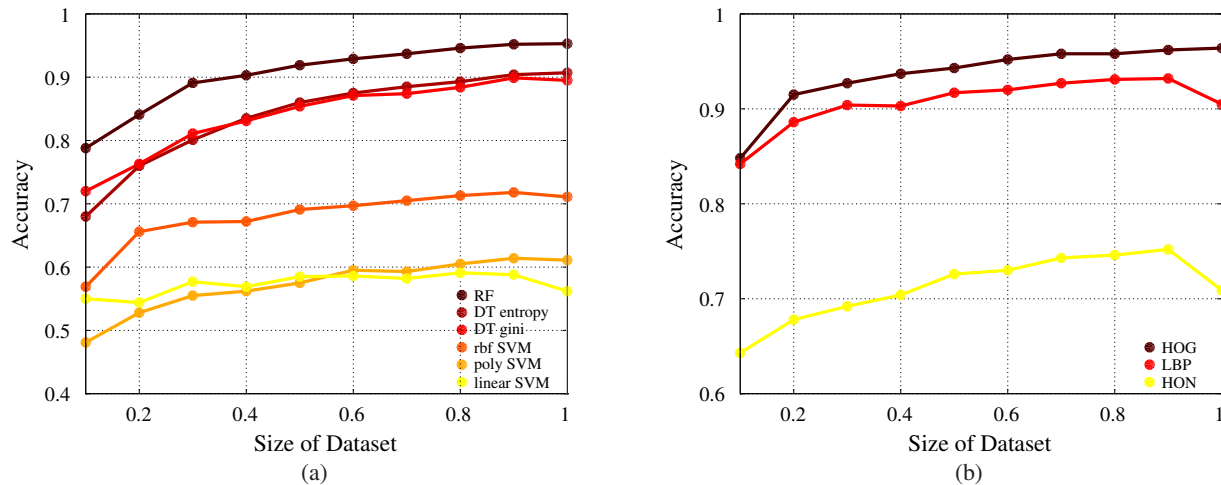


Fig. 6: Estimation of body orientation accuracies computed on subdatasets with different size of samples. The first graph (a) compares common feature extraction techniques. The graph on the right (b) presents results of gradient line method.

- [7] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. Mirza, "Gender recognition from face images with local wld descriptor," in *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, April 2012, pp. 417–420.
- [8] J.-G. Wang and W.-Y. Yau, "Real-time beard detection by combining image decolorization and texture detection with applications to facial gender recognition," in *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2013 IEEE Workshop on*, April 2013, pp. 58–65.
- [9] M. Sakarkaya, F. Yanbol, and Z. Kurt, "Comparison of several classification algorithms for gender recognition from face images," in *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*, June 2012, pp. 97–101.
- [10] A. dos Anjos and H. Shahbazkia, "Bi-level image thresholding - a fast method," in *BIOSIGNALS (2)'08. INSTICC - Institute for Systems and Technologies of Information, Control and Communication*, 2008, pp. 70–76.
- [11] "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp. 62–66, Jan 1979.
- [12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Berkeley, Calif., pp. 281–297, 1967.
- [13] A. T. Nghiem, E. Auvinet, and J. Meunier, "Head detection using kinect camera and its application to fall detection," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, July 2012, pp. 164–169.
- [14] Y. Ishii, H. Hongo, K. Yamamoto, and Y. Niwa, "Face and head detection for a real-time surveillance system," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, Aug 2004, pp. 298–301 Vol.3.
- [15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1.
- [16] I. Oulu, "The local binary pattern approach to texture analysis extensions and applications," 2003.
- [17] S. Tang, X. Wang, X. Lv, T. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Computer Vision ACCV 2012*, ser. Lecture Notes in Computer Science, K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, Eds. Springer Berlin Heidelberg, 2013, vol. 7725, pp. 525–538.
- [18] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "Wld: A robust local image descriptor," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1705–1720, Sept 2010.