2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

Real-Time Full-Body Human Gender Recognition in (RGB)-D Data

Timm Linder Sven Wehner Kai O. Arras

Abstract— **Understanding social context is an important skill for robots that share a space with humans. In this paper, we address the problem of recognizing gender, a key piece of information when interacting with people and understanding human social relations and rules. Unlike previous work which typically considered faces or frontal body views in image data, we address the problem of recognizing gender in RGB-D data from side and back views as well. We present a large, gender-balanced, annotated, multi-perspective RGB-D dataset with full-body views of over a hundred different persons captured with both the Kinect v1 and Kinect v2 sensor. We then learn and compare several classifiers on the Kinect v2 data using a HOG baseline, two state-of-the-art deep-learning methods, and a recent tessellation-based learning approach. Originally developed for person detection in 3D data, the latter is able to learn the best selection, location and scale of a set of simple point cloud features. We show that for gender recognition, it outperforms the other approaches for both standing and walking people while being very efficient to compute with classification rates up to 150 Hz.**

I. Introduction

Knowledge about humans, their social relations and normative rules is important for interactive robots to provide effective and user-friendly services. Recognizing human gender is a key ability to this end. The problem has traditionally been addressed in the computer vision and surveillance communities whose approaches use visual appearances of faces [1]–[4], frontal upper-body views [5], [6], or full-body views [7]–[9]. Image data have, however, drawbacks particularly in robotics applications: they provide appearance cues only and strongly depend on proper illumination conditions which may change frequently and drastically when cameras are deployed on mobile robots. RGB-D data, on the other hand, are generally less sensitive to ambient conditions and provide 3D range data that allow for the extraction of geometric cues as well. Thus, we adopt RGB-D and 3D data for the purpose of full-body gender recognition in this paper and make the following contributions:

- We propose a novel gender recognition method based upon a depth-based tessellation learning approach. This is an extension of our work on people detection using a top-down classifier [10]. Our method, while characterizing 3D objects using simple geometric point cloud features, in addition to the selection of local features and thresholds, also learns the scale and location at which these features are computed. In our experiments, the approach outperforms a HOG baseline clearly and two state-of-the-art deep-learning methods for RGB-D object recognition by a small margin on Kinect v2 data while being easier to learn and faster to compute.
- We present a large, gender-balanced, annotated, multi-perspective RGB-D dataset with full-body views of over a hundred different persons captured with both the Kinect v1 and v2 sensor. The subjects stand and walk at different distances and relative angles to the sensors and have given their consent for the use of their images for research purposes.

T. Linder, S. Wehner, and K. O. Arras are with the Social Robotics Lab, Dept. of Computer Science, University of Freiburg, Germany. http://srl.informatik.uni-freiburg.de, {linder,arras}@cs.uni-freiburg.de. This work has been partly supported by the European Commission under contract number FP7-ICT-600877 (SPENCER).

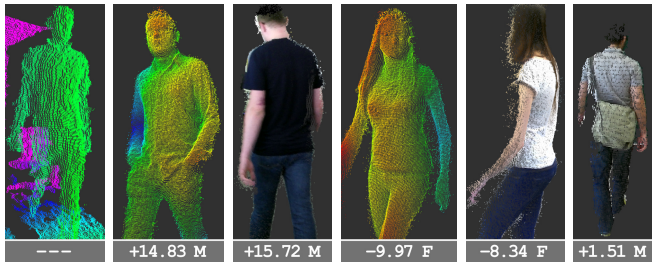


Fig. 1. 3D and RGB-D data from our human attribute dataset with corresponding gender classification results. The leftmost image has been taken with a Kinect v1 sensor, the remaining ones with a Kinect v2. The numbers below each image are confidences of the boosted classifier, the sign indicates the predicted gender label (M for male, F for female).

II. Related work

The problem of gender recognition has traditionally been addressed with visual appearances of frontal faces, mostly using local binary pattern (LBP) features [1], [5], sometimes in combination with raw pixel values, haar-like features [3] or HOG descriptors [2], [6]. The best methods achieve above 90% accuracy on large datasets with several thousands of training images. [3], [4] review different classification algorithms, with SVM performing best and boosting methods following slightly behind at lower computational cost. In addition to facial features, some approaches also extract information from hair and clothing, *e. g.* [5], leading to even higher classification rates (96%) at the cost of speed (less than 1 Hz).

Generalizing the task to full-body views is challenging due to the high variety in human appearances, poses, and distances from the sensor, considering *e. g.* the case of rear views when no face is visible. [7] use *poselets*, that represent small parts of the body under a specific pose, to recognize various attributes including gender. They achieve around 82.4% accuracy on a large database with different poses and viewpoints. [8] examine different combinations of

978-1-4799-6922-7/15/$31.00 ©2015 IEEE

appearance-, shape- and color-based features. Their overall accuracy is 80% on frontal views only. Lastly, [9] employ a convolutional neural network (CNN) on color images, which automatically learns the most informative features during training. They achieve an accuracy of 80.4% on a dataset of 924 full-body images, including rear and frontal views. The reader is also referred to the survey in [11] where the authors compare the results of different visual face-based, body-based and gait-based gender recognition techniques.

Looking at methods that take 3D information into account, [12] recognize gender from a large set (2484 persons) of 360° full-body high-resolution laser scans created with an expensive stationary scanner. The approach requires several costly steps including hole-filling, mesh smoothing and normal computation and assumes no clothing. [13] compute skeleton- and surface-based geometric features (such as torso-to-hip distance) on 3D point clouds captured with a Kinect v1 for the purpose of people re-identification. First preliminary results on RGB-D gender recognition are presented in [14] but the authors apparently focus only on frontal images where the upper body is fully visible. Their method relies on the availability of a skeleton tracking algorithm.

Unlike these methods, we present a full-body gender recognition algorithm using consumer-grade RGB-D sensors that predicts gender with an average accuracy of 91% for standing persons and 87% including walking people over a large number of different body orientations and distances to the sensor. Our method solely relies on 3D point clouds, is very fast to compute (150 Hz) and requires only minimal preprocessing of the data.

Datasets: While there exists a large number of color image-based databases for gender recognition [15]–[19], there are only few suitable full-body RGB-D datasets. Datasets such as our own RGB-D people dataset [20] for person tracking or the BIWI RGBD-ID dataset [21] contain too few individuals or are insufficiently gender-balanced to train a reliable gender classifier. The IIT RGB-D person re-identification dataset [13] includes 79 people walking down a hallway mostly in frontal or rear view. In the TUM gait from audio, image and depth database [22], people also walk mostly into the same direction. None of these datasets include data captured with the new Kinect v2 sensor, which provides a significantly improved depth resolution.

Thus, the gender- and age-annotated dataset presented here (see Sec. IV for details) is the only one to include multiple complex walking patterns so as to have many views of people at various distances to the sensor and relative orientations. It also contains a close-up sequence that is valuable for human-robot interaction. Unlike other RGB-D datasets, our dataset has more participants (118 persons), is largely gender-balanced, and uses both the Kinect v1 and Kinect v2 sensors. This makes it the largest and most complete dataset for the purpose of human attribute recognition in RGB-D.

III. Our method

We propose a novel method for gender classification based upon our existing work on 3D object characterization for the task of person detection in 3D range data [10]. The method takes a bottom-up top-down approach where we classify object detection hypotheses from a bottom-up classifier using a learned top-down model. The bottom-up classifier can either be a simple region-of-interest (ROI) detector or a more sophisticated detector, typically tuned for higher recall. Here, we focus on the top-down method and assume to have a simple ROI detector which extracts candidate person detections from the scene in the form of 3D point clouds.

The top-down method characterizes the point cloud by a set of features computed on the measurements within axis-aligned voxels on the 3D objects and uses AdaBoost to create a strong classifier with the best features and voxels. What is special about this method is that the boosted classifier not only selects the best features and thresholds, but also the best combination of voxels on which these features have found to be informative. Thus, the classifier also learns the best scales and locations of features on the 3D object for the classificiation task at hand. This allows the robust and stable description of complex articulated shapes, as will be demonstrated in the experiments.

A. Tessellation Generation

We assume persons to fit into a fixed-size bounding volume $\mathcal{B}$, centered around the median in x and y of the point cloud. The size of $\mathcal{B}$ can either be fixed and taken from the maximum expected object size or learned from a training set as in [10].

We subdivide the volume into voxels which leads to the question of how a volume can be tessellated into a collection of smaller volumes, a problem well known as tiling in computational geometry. For the sake of simplicity, we consider only axis-parallel voxels which reduces the complexity of the problem but still leaves an infinite number of tessellations of $\mathcal{B}$. Thus, we define a set of proportion constraints $\mathcal{C}$ to exclude extreme aspect ratios of voxels and a list of increments $\mathbf{s}$ by which voxels will be enlarged. Each element $\mathbf{c} = (w, d, h) \in \mathcal{C}$ is a width-depth-height triplet with multipliers of the respective voxel dimension.

The resulting procedure, Algorithm 1, generates all possible voxel sizes subject to $\mathcal{C}$ and $\mathbf{s}$. Defining the remainder after ceiling-division $\texttt{rem}(a, b)$ as $|a - \lceil \frac{a}{b} \rceil b|$, the algorithm tests whether voxels can fill a volume $\mathcal{B}$ without gaps and subdivides $\mathcal{B}$ into a regular grid. The function $\texttt{Tess}(\mathcal{B}, w, d, h, \Delta_w, \Delta_d, \Delta_h)$ produces a regular face-to-face tessellation of $\mathcal{B}$ with voxels of size (w, d, h) and offset $(\Delta_w, \Delta_d, \Delta_h)$ to also allow voxels that overlap each other. The algorithm generates gapless subdivisions of $\mathcal{B}$ that are complete in that no tessellation is missing under the given constraints. In contrast to [10], we also allow slightly protruding voxels with a tolerance θ.

As constraints we choose scaling factors $\mathbf{s} = (0.1, 0.2, ..., 0.8)\ [m]$ and proportions $\mathcal{C}$ being the set of all permutations of $\{\{1,1,1\}, \{1,1,1.25\}, \{1,1,2\}, \{1,1,2.5\}, \{1,1,3\}, \{1,1,4\}, \{1,1,5\}, \{1,1,6\}, \{1,1,8\}, \{1,1,10\}, \{2,2,3\}, \{4,4,2\}, \{4,4,3\}\}$. These constraints extend the ones considered in [10] to account for more detail in the

Algorithm 1: Compute all axis-parallel tessellations $\mathcal{T}$ of a volume $\mathcal{B}$.

Input: Bounding volume $\mathcal{B}$ of size $w_\mathcal{B} \times d_\mathcal{B} \times h_\mathcal{B}$, set of voxel proportion constraints $\mathcal{C}$, list of voxel scaling factors $\mathbf{s}$, protrusion tolerance θ. **Output:** Set of all possible tessellations $\mathcal{T}$

$\mathcal{T} \leftarrow \{\}$
foreach $s_j \in \mathbf{s}$ **do**
  foreach $\mathbf{c}_k = (w_k, d_k, h_k) \in \mathcal{C}$ **do**
    $w = s_j \cdot w_k;\ d = s_j \cdot d_k;\ h = s_j \cdot h_k$
    if $\texttt{rem}(w_\mathcal{B}, w) < \theta \wedge \texttt{rem}(d_\mathcal{B}, d) < \theta \wedge \texttt{rem}(h_\mathcal{B}, h) < \theta$ **then**
      $\mathcal{T} \leftarrow \mathcal{T} \cup \texttt{Tess}(\mathcal{B}, w, d, h, 0, 0, 0)$
      $\mathcal{T} \leftarrow \mathcal{T} \cup \texttt{Tess}(\mathcal{B}, w, d, h, \frac{w}{2}, \frac{d}{2}, \frac{h}{2})$
    end
  end
end
return $\mathcal{T}$

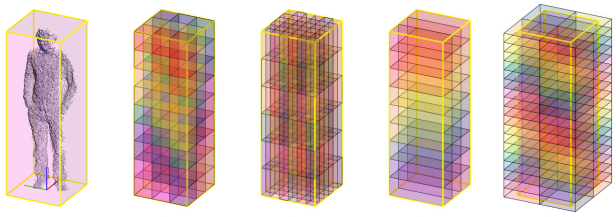


Fig. 2. Left: person candidate point cloud, centered around the median in x and y. The other pictures show example tessellations of the bounding volume $\mathcal{B}$ generated using our tessellation algorithm. We also allow protruding voxels, shown in the rightmost picture.

data and lead to 134 valid tessellations, examples are shown in Fig. 2.

B. Classifier Training

Let $\mathcal{T}_j$ be the jth valid tessellation and $\mathcal{V}_j^i$ its ith voxel. Then, for each $\mathcal{V}_j^i$ of all generated $\mathcal{T}_j$'s, we determine the set $\mathcal{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of points inside the voxel's volume. With the goal to describe shape properties locally, we then compute a set of nine simple 3D point cloud features $f_1, ..., f_9$ that characterize geometrical and statistical properties of $\mathcal{P}$, see Table I. Most of them can be computed very efficiently from the points' scatter matrix via eigenvalue decomposition and none of them require estimation of the surface normals.

Training samples are formed by stacking the features of all voxel point clouds of all tessellations into one large feature vector and associating the corresponding ground truth gender label. We train an AdaBoost classifier with n_{weak} decision stumps as weak learners. After training, the final model is given by the collection of all voxels in which *at least one feature* has been selected. The resulting strong classifier achieves a double objective, it selects the best features ('best' quantified by the AdaBoost voting weights) and selects the optimal subdivision $\mathcal{T}_{opt}$ of $\mathcal{B}$ for the classification task at hand. The method can select an arbitrary number of features in each voxel – a large number, for instance, means that the voxel contains a particularly salient local shape – and may also select a mixture of voxels from *different* tessellations. Note that the approach uses 3D point cloud data only.

Description	Expression
Number of points	The point count of $\mathcal{P}$ denoted as n. $f_1 = n$
Density	Captures the normalized point density w.r.t. the entire point cloud: $f_2 = \frac{n}{N_\mathcal{B}}$
Sphericity	Captures the level of sphericity from the ratio of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ extracted from the scatter matrix of $\mathcal{P}$. $f_3 = 3\frac{\lambda_3}{\sum_i \lambda_i}$ where $\lambda_1 > \lambda_2 > \lambda_3$
Flatness	Measures the degree of planarity from the eigenvalues. $f_4 = 2\frac{\lambda_2 - \lambda_3}{\sum_i \lambda_i}$
Linearity	Captures the level of linearity from the eigenvalues. $f_5 = \frac{\lambda_1 - \lambda_2}{\sum_i \lambda_i}$
Standard deviation w.r.t. centroid	Measures the compactness of points in $\mathcal{P}$, $f_6 = \sqrt{\frac{1}{n-1}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})^2}$ where $\bar{\mathbf{x}}$ is the centroid.
Kurtosis w.r.t. centroid	Captures the peakedness of points in $\mathcal{P}$, fourth centralized moment of the data distribution in $\mathcal{P}$. $f_7 = \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})^4 / f_6$.
Average deviation from median	Alternative measure of compactness. $f_8 = \frac{1}{n}\sum_i \|\mathbf{x}_i - \tilde{\mathbf{x}}\|$ where $\tilde{\mathbf{x}}$ is the vector of independent medians $\tilde{\mathbf{x}} = (\tilde{x}, \tilde{y}, \tilde{z})$.
Normalized residual planarity	Alternative measure of flatness. Squared error sum of a plane fitted into $\mathcal{P}$ normalized by n. $f_9 = \sum_i^n (a\,x_i + b\,y_i + c\,z_i + d)^2$ where a, b, c, d are the parameters of the plane derived from the eigenvalues of the scatter matrix.

TABLE I
POINT CLOUD FEATURES

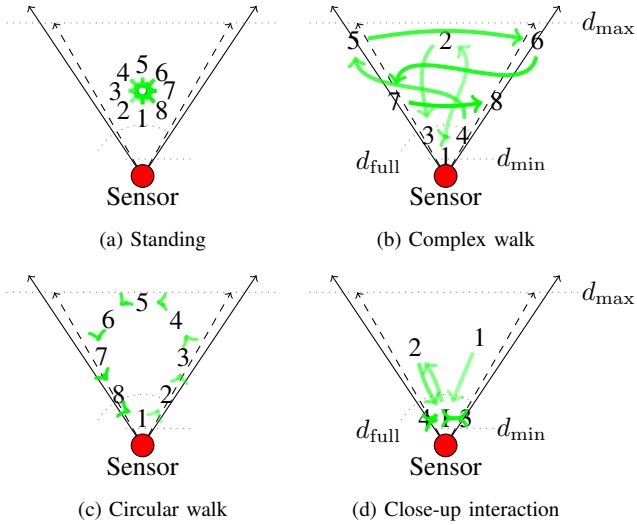
IV. Human Attributes Dataset

In this section we present our human attribute dataset, which is motivated, as mentioned above, by the limitations of existing datasets for the task of full-body, multi-perspective gender classification from RGB-D data. We acquired and annotated an RGB-D dataset of 118 persons (54 male, 64 female) under controlled conditions. The mean age is 27 years ($\sigma = 8.7$), the age of the youngest participant is 4 and the oldest 66 years. The data has been collected at 15 Hz in three different indoor locations under controlled lighting conditions and annotated with gender and age. The subjects performed several standing and walking patterns that have been designed to cover all relative orientations and the full RGB-D sensor range between $0.5m$ to $4.5m$. The sensors used are the ASUS Xtion Pro Live (internally similar to Kinect v1) and the Kinect v2[1] so as to make it possible for researchers to study effects of different RGB-D data qualities. The sensors were stacked on top of each other and recorded all sequences at the same time. We did not notice any cross-talk effects between the sensors which is likely due to the different measurement principles, structured light for Kinect v1, time-of-flight for Kinect v2. They were mounted at around $1.5m$ height and tilted slightly downwards, approximately replicating the setup of a typical mobile robot or handheld device.

For each participant, we recorded four different sequences. In sequence 1 the subject is standing at around $2.5m$ distance from the sensor and rotates clockwise in $45°$ steps (no continuous data capture, 1 image per step). Sequence 2

[1]About 75% of the data were recorded using a developer preview version of the Kinect v2 sensor. We did not observe any notable difference in data quality compared to the final version, apart from the $4.5m$ range limitation.

Fig. 3. Example images from our RGB-D human attribute dataset. The first two images are part of sequence 1 (static poses), images 3–5 are included in sequences 2 and 3 (walking patterns), and the last image is part of sequence 4 (close-up interaction).



(a) Standing (b) Complex walk

(c) Circular walk (d) Close-up interaction

Fig. 4. Standing and walking patterns that people in our dataset performed. (a) Standing pose in 8 different orientations, at around $2.5m$ distance to the sensor. (b) A complex walking pattern, designed to capture a large variety of poses and distances to the sensor. (c) Circular walking pattern in both directions. (d) A close-up human-robot interaction pattern. $d_{\min}$ and $d_{\max}$ denote the Kinect v2's minimum and maximum depth sensor distance and d_{full} the distance above which a person is fully visible in our setup.

(typical length ∼370 frames) consists of a video of the person performing a complex walking pattern so as to capture various distances from the sensor and relative orientations. In sequence 3 (∼300 frames), the person walks on a circle that covers almost the entire view frustum, in both clockwise and anti-clockwise direction. Finally, sequence 4 (∼280 frames) simulates a close-up interaction with a robot, where the subject steps back, forth and sideways in front of the sensor as if he/she is physically interacting with the robot's touch screen or manipulator. The sequence is thought to be a relevant benchmark for human-robot interaction that contains many vertical and horizontal occlusions of people as well as cases of missing out-of-range depth data. Fig. 4 visualizes the patterns in sequences 1–4 and Fig. 3 shows some example color frames from our dataset.

We further post-processed the data (see Fig. 5). We computed foreground segmentation masks and point cloud surface normals which were required for the deep-learning classifier described below. As a result, our dataset's format is largely compatible with the University of Washington's (UW) RGB-D object dataset [23]. To compute the foreground mask, we applied a depth-based ROI extraction method that, similar to [24], projects a height map onto the ground plane

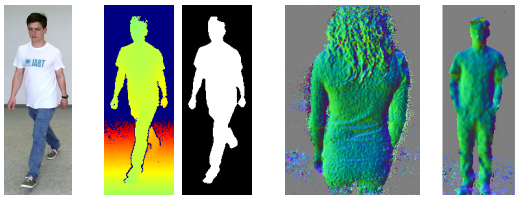


Fig. 5. Post-processing of the data. Left: Initial color image and colorized depth image of a person from sequence 2, recorded with the Kinect v2. Middle: foreground segmentation mask. Right: Point cloud normals for a person close to the sensor (sequence 4, distance <$1.0m$) and a subject in a static pose (sequence 1, distance ca. $2.5m$).

and finds a local maximum. We then center a cylinder at the maximum and mark all points inside the cylinder and above ground as foreground. Normals are computed using a k-NN method with $k = 25$.

In total, the dataset contains around 131,800 RGB-D frames which results in approximately 300 GB of data. 105 participants (47 male, 58 female) have given their consent for sharing their images with other researchers, to which we will grant access to this data upon request at our website.

V. Experiments and Results

In the experiments, we investigate the ability of the tessellation learning approach to recognize gender of the subjects in our dataset using data from the Kinect v2 sensor. We compare our approach, which relies solely on 3D range data, with two state-of-the-art deep-learning methods for RGB-D object recognition and an RGB-D histogram of oriented gradients (HOG) approach. Concretely, the considered methods are:

- Histogram of oriented gradients (HOG) and histogram of oriented depths (HOD), computed on the RGB and D image, respectively. HOG is a successful and widely used descriptor for person detection in image data, HOD has equally been used for this purpose in depth data [20]. The two feature vectors are concatenated and then fed into a linear SVM. We evaluate two window sizes, 32×64 and 64×128, and use the HOG and SVM C++ implementations provided by the OpenCV library.
- Convolutional-recursive neural networks (CRNN). CRNNs are a recent deep-learning method by Socher et al. [25] for RGB-D object recognition. We use the Matlab code provided at the authors' website, which we had to modify to use distinct training/test splits on a per-person basis, and to reduce memory consumption. In addition to the softmax classification layer as proposed in [25], we also used an alternative linear SVM classification stage. The results in this section were obtained using the SVM classifier, as experiments showed no significant difference between the two classifiers for the task at hand.
- Hierarchical matching pursuit (HMP) by Bo et al. [26] is another state-of-the-art deep learning method. The approach uses sparse coding techniques and has, like the CRNN method, also been specifically designed for RGB-D object recognition. In addition to the foreground masks required by CRNN, HMP also requires point

Sequence	(1) Standing	(1)–(3) +Walking	(1)–(4) +Interact.	(1)–(4) $d > 0.8m$
HOG32	84.78%	70.55%	70.70%	70.55%
HOG64	86.27%	74.10%	74.36%	74.33%
CRNN	86.64%	83.46%	83.73%	84.10%
HMP	88.07%	85.28%	**86.10%**	85.42%
Ours	**91.07%**	**86.41%**	85.29%	**87.47%**

Fig. 6. Average gender classification accuracy: HOG with window sizes 32×64 and 64×128, CRNN, HMP, and our proposed tessellation learning classifier. For the top-down classifier, which only operates on the depth modality, we use $n_{\text{weak}} = 500$. The other methods use both color and depth. The last column shows results if we ignore all test samples closer than $0.8m$ distance to the sensor.

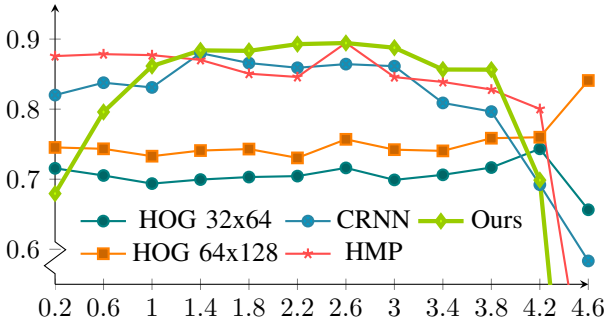


Fig. 7. Average accuracy of each classifier as a function of person-to-sensor distance in meters, obtained on the full dataset containing sequences (1) to (4). The sharp decline at around $4.5m$ distance is due to the maximum range limit of the developer preview of the Kinect v2 sensor.

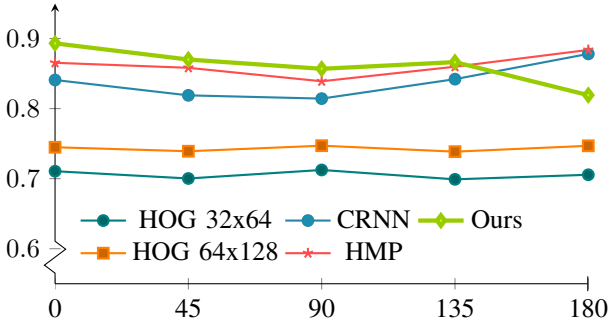


Fig. 8. Average gender classification accuracy of each classifier as a function of person orientation in degrees (same parameters as in Fig. 7). 0° means that the person is looking into the RGB-D sensor.

cloud normals as an input which we provided as part of the post-processing steps of our data (see Sec. IV). Here, we also use the Matlab implementation provided by the authors, which had to be modified slightly to process our training and test splits.

For training these classifiers, we form the training set by concatenating all four sequences of all 118 persons and split it into equally-sized training and test splits on a per-person basis. By never including the same person instance in both training and test set, we want to ensure that the classifiers do not learn individual person appearances. For each classifier, we perform at least 10 runs of repeated random sub-sampling validation with a training/test set split ratio of 1:1 to ensure that there is always sufficient person variety in the training set. As sequences 2 to 4 are recorded with 15 fps which gives a large number of locally similar frames, we subsample every 5$^{\text{th}}$ frame (for CRNN & HMP every 20$^{\text{th}}$) to keep learning times within reasonable bounds.

We evaluate our tessellation learning approach for both $n_{\text{weak}} = 100$ and $n_{\text{weak}} = 500$ weak classifiers. The method is implemented in C++ in two variants, a regular non-optimized version that runs on a single core and an optimized version that uses OpenMP to parallelize feature calculations on the CPU. The two variants will be considered when evaluating inference performances of the different approaches. We use the AdaBoost C++ implementation by OpenCV. Our code will be made publicly available as a ROS package upon publication of this paper.

A. *Classification Accuracy Results*

Figure 6 shows the average classification accuracies (over multiple validation runs) for the different sequences in the dataset: for the eight standing poses of sequence 1, the standing poses and the two walking patterns (seq. 1–3), the entire dataset (seq. 1–4) and the entire dataset except test samples with persons closer than $0.8m$ to the sensor, below which we observed noise and missing data artefacts produced by the Kinect v2 sensor. It can be seen that all methods achieve more than 80% accuracy under the rather controlled conditions of sequence 1. As soon as people start walking, and the shapes of people become more diverse and articulated, motion blur starts to play a role and the HOG-based methods drop to around 70%. CRNN and HMP both perform fairly solid across all sequences, with HMP in all cases being about 2% better than CRNN. A similar observation was also made by [25] on the UW RGB-D object dataset. See also Fig. 1 which shows example classification results along with their input images.

Despite not using any color information from the RGB image, our proposed tessellation learning method generally outperforms all other methods on the task of full-body gender classification. Only for close-range subjects in the interaction sequence 4, HMP is slightly more accurate. Examples for which our methods fails under close-range conditions are shown in Fig. 9.

We further analyze these results with respect to relative distance and body orientation.

Impact of distance to sensor: Fig. 7 shows the classification accuracy against person distance to sensor. HOG, CRNN and HMP deliver relatively constant results in terms of accuracy across the available RGB-D sensor range. At very close-range ($< 0.6m$), our tessellation approach breaks down as the point clouds provided by the sensor extend beyond the near clipping plane, such that the shape of the person becomes very hard to distinguish (Fig. 9). Here, the other methods could be in advantage because they can fall back onto the RGB image, which – at this close distance – is potentially of very high resolution. At $> 4.2m$ distance, parts of the point cloud start to vanish at the far clipping

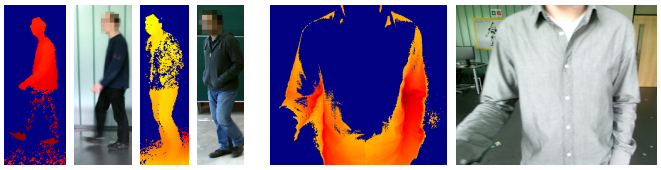


Fig. 9. Example RGB-D images where our tessellation learning approach fails to classify the person's gender correctly. Problems mainly occur at the sensor's near and far clipping planes, and with certain types of black clothing. The RGB image is shown just for illustration purposes, and is not used by our method.

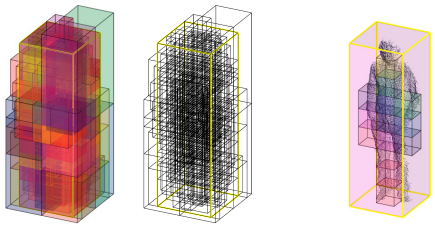


Fig. 10. Learned best tessellations using 500 weak classifiers (left two pictures), trained on the full dataset. Right: Learned regular tessellation using only cube-shaped voxels of side length $0.2m$, trained on seq. (1).

Tessellation type	$\mathbf{n_{weak}}$ 100	500
Regular tessellation $0.1m$	75.93%	76.49%
Regular tessellation $0.2m$	83.19%	84.16%
Regular tessellation $0.3m$	81.54%	82.96%
Learned tessellation	**89.75%**	**91.07%**

Fig. 11. Comparison of gender classification accuracy of the tessellation learning approach using the learned tessellations with mixed-size voxels (selected by AdaBoost from all voxels across all generated tessellations), against the same method with only a set of fixed-size voxels with side length $0.1m$, $0.2m$ and $0.3m$.

plane. The drop in performance of the tessellation learning approach is thus due to limitations of the depth sensor rather than the method itself.

Impact of person orientation: For sequence 1 of the dataset, we can directly determine person's orientations in $45°$ steps from the frame number. For the other sequences, we track the center of the person blob using a Kalman filter-based nearest-neighbor tracker with constant velocity motion model, and then determine the orientation from the track velocity estimates. As can be seen in Fig. 8, gender recognition is rather stable across relative view angles for all methods. The deep-learning methods excel on rear views of a person $(180°)$, whereas for our tessellation learning method, rear views appear to be slightly more difficult than frontal or side views.

B. Resulting Learned Tessellation

With the encouraging results of the previous subsection, we seek to better understand the learned tessellation. First, we examine if it actually makes sense to learn a volume subdivision over the object as a combination of voxels from different tessellations versus a predefined regular tessellation. We compare the learning approach with a grid of cube-shaped voxels of fixed size. We use the same nine point cloud features and train an AdaBoost classifier with the same number of weak classifiers on sequence 1 of our dataset. Fig. 11 shows the classification accuracy over several grid sizes. It can be seen that the learned tessellation performs clearly better than the best tessellation using cube-shaped voxels of size $0.2m$.

Fig. 10 (left) shows the resulting tessellation learned by our method using 500 weak classifiers on the full dataset. The most commonly used features in descending order of frequency (in brackets) are $f_3\,(94)$, $f_5\,(84)$, $f_2\,(64)$, $f_6\,(64)$, $f_9\,(61)$, $f_4\,(50)$, $f_8\,(50)$, $f_7\,(20)$, $f_1\,(13)$. From the wireframe representation, it can be seen that the highest concentration of voxels is located at above waist height and around the upper body, indicating that these regions contain more relevant information than *e. g.* the legs. A similar observation can be made when allowing only fixed-size cube-shaped voxels of size $0.2m$ with $n_{\text{weak}} = 100$ (Fig. 10, right).

C. Runtime Performance Results

Runtime performance is key in applications of robots in real-world human environments, particularly because they may be surrounded by several persons at the same time that all need to be classified. We therefore analyze the time it takes for all methods to predict the gender class given an RGB-D image. For the comparison, we assume that the RGB-D data have already been pre-processed, i.e. the person has been detected and cropped out from the RGB-D image or point cloud, second, point cloud normals have been calculated (required for the HMP method), and third, the sub-cloud containing the person has been transformed into the origin of a local coordinate frame (required for our method). We measure the time it takes to train the classifiers on sequence 1 of the dataset that contains the eight standing poses, and the time it takes to classify a single person, averaged over a large number of frames.

We anticipate that the comparison is not fully fair at this point because CRNN and HMP are implemented in Matlab whereas all other methods are implemented in C++. We believe, however, that the comparison is still able to reveal a trend that we may see in the light of Matlab code running $10\times$ to $100\times$ slower than C++ code. The computer used is a regular desktop PC with Intel Core i7-2600 CPU.

As expected, the runtimes for learning and prediction in Fig. 12 show that the deep-learning methods are computationally most expensive in both training and testing. The HMP approach, although it uses much higher-dimensional feature vectors at the classification stage than CRNN (188,300 vs 32,768) still performs about $2.5\times$ faster during testing. Our tessellation learning approach is very fast. The non-optimized implementation achieves more than 40 Hz for classification, the parallelized variant on four CPU cores even 150 Hz. This is still clearly the fastest method even if we assume a conservative speed up factor of 100 for a C++ implementation of CRNN and HMP.

	Training full seq. (1)	**Testing** single frame
HOG 32x64	$<$ 1 min	90 ms
HOG 64x128	4 min	120 ms
CRNN (Matlab)	420 min	7500 ms
HMP (Matlab)	35 min	3000 ms
Ours, 1 thread	–	24.0 ms
Ours, 4 threads	17 min	6.7 ms

Fig. 12. Average training and testing durations for the different methods. The tessellation learning approach is clearly the fastest classification method even if we assume a conservative speed up factor of 100 for a C++ implementation of CRNN and HMP.

VI. Conclusion

In this paper, we presented a novel tessellation learning method for gender classification in 3D point clouds that achieves up to 91% accuracy on standing people and 87% when including walking people without using any color information. It outperforms an RGB-D HOG baseline by a wide margin and two state-of-the-art deep-learning methods for RGB-D object recognition by a small margin while being very fast to compute. The approach, based upon our previous work on people detection [10], characterizes objects by a selection of the best local shape features and thresholds, as well as the best scales and locations at which these features are computed on the 3D object. Our results underline the importance of scale when characterizing objects – a rather under-explored problem in the context of interest point-based recognition methods as pointed out in [27].

The result also indicates that much gender-relevant information is contained in the depth image and that RGB data may even distract the model from proper classification. However, more systematic experiments are needed to support this conjecture.

We also presented a new large-scale annotated RGB-D dataset for human attributes with full-body views of 118 persons captured with both first- and second-generation Kinect sensors.

In future work we plan to also incorporate color features and evaluate the robustness of tessellation learning for the general task of RGB-D object recognition. On-going work is also concerned with the integration of this approach onto a real-world service robot in crowded environments.

References

[1] H.-C. Lian and B.-L. Lu, "Multi-view gender classification using local binary patterns and support vector machines," in *Advances in Neural Networks (ISNN 2006)*, ser. LNCS, 2006, vol. 3972.

[2] L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognition Letters*, vol. 31, no. 11, 2010.

[3] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Transactions on Pattern Analysis & Machine Intell.*, vol. 30, no. 3, 2008.

[4] J. Bekios-Calfa, J. Buenaposada, and L. Baumela, "Revisiting linear discriminant techniques in gender recognition," *IEEE Trans. on Pattern Analysis & Machine Intell.*, vol. 33, no. 4, 2011.

[5] B. Li, X.-C. Lian, and B.-L. Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 76, no. 1, 2012.

[6] M. Castrillón-Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Improving gender classification accuracy in the wild," in *Proceedings of the 18th Iberoamerican Congress on Pattern Recognition (CIARP 2013)*, ser. LNCS, vol. 8259, 2013.

[7] L. D. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *IEEE Int. Conf. on Comp. Vis. (ICCV)*, 2011.

[8] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," in *9th IEEE Int. Workshop on Visual Surveillance, ICCV 2009 Workshops*, 2009.

[9] C. B. Ng, Y. H. Tay, and B.-M. Goi, "A convolutional neural network for pedestrian gender recognition," in *Int. Symposium on Neural Networks (ISNN'13)*, ser. LNCS, vol. 7951, 2013.

[10] L. Spinello, M. Luber, and K. O. Arras, "Tracking people in 3D using a bottom-up top-down detector," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'11)*, Shanghai, China, 2011.

[11] C. Ng, Y. Tay, and B.-M. Goi, "Recognizing human gender in computer vision: A survey," in *PRICAI 2012: Trends in Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7458.

[12] J. Tang, X. Liu, H. Cheng, and K. Robinette, "Gender recognition using 3-d human body shapes," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 41, no. 6, 2011.

[13] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *ECCV 2012 Workshops and Demonstrations*, ser. LNCS. Springer Berlin Heidelberg, 2012, vol. 7583.

[14] R. C. Luo and X. Wu, "Real-time gender recognition based on 3d human body shape for human-robot interaction," in *ACM/IEEE International Conference on Human-robot Interaction (HRI'14), Late Breaking Reports Track*, 2014.

[15] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1997.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[17] K. J. Ricanek and T. Tesafaye, "MORPH: a longitudinal image database of normal adult age-progression," in *Int. Conf. on Automatic Face and Gesture Recognition (FGR'06)*, 2006.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[19] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[20] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, San Francisco, USA, 2011.

[21] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool, "One-shot person re-identification with a consumer depth camera," in *Person Re-Identification*, ser. Advances in Computer Vision and Pattern Recognition. Springer, 2014.

[22] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation*, vol. 25, 2013.

[23] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Int. Conf. on Robotics & Automation (ICRA)*, 2011.

[24] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *Int. Conf. on Robotics & Automation (ICRA)*, Hong Kong, China, 2014.

[25] R. Socher, B. Huval, B. P. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Neural Information Processing Systems (NIPS'12)*, 2012.

[26] L. Bo, X. Ren, and D. Fox, "Learning hierarchical sparse features for RGB-(D) object recognition," *International Journal of Robotics Research*, vol. 33, no. 4, 2014.

[27] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3d object recognition in cluttered scenes with local surface features: A survey," *IEEE Trans. on Pattern Analysis & Machine Intell.*, 2014, in press.