ABSTRACT
Pedestrian detection based on the framelet features in noisy depth images is investigated in this paper. For capturing the local features and attenuating the effects of noise in depth images, a features optimization model is proposed to adaptively select the framelet features for classification. The selected framelet features extracted by the model and SVM with a linear kernel is adopted as the feature and classifier, respectively. The proposed framelet features under a tight and redundant system can preserve the shape information while reducing the impact of noise. Experimental results also show that the proposed method based on framelet features can achieve a great improvement in noisy depth images, and the improvement is over one order of magnitude than HDD and HOG.

Index Terms—Pedestrian detection, framelet, adaptive selection features.

1. INTRODUCTION
Pedestrian detection in images is an importance topic and is greatly needed in robotics, surveillance, etc. However, its challenging is due to the variations of illumination, human posture, clothing, complex background and occlusion. To overcome the shortcomings of intensity images in pedestrian detection, approaches based on depth images are gaining increasing attentions from computer vision researchers.

Most current approaches are based on intensity images (color or gray images). In 2005, Dalal and Triggs proposed a method using Histograms of Oriented Gradients (HOG) as the feature descriptor and SVM as the classifier [1]. Many newer methods are inspired by HOG which is a milestone contribution to pedestrian detection. Due to the high computational costs of HOG in real-time detection, Zhu et al. adopted AdaBoost and Cascade to accelerate HOG feature descriptor selection for pedestrian detection [2]. To solve the problem of partial occlusion, Wang et al. [3] combined HOG and LBP (Local Binary Pattern) as feature descriptor and achieved good performance. To promote the research on pedestrian detection, Dollár et al. [4] created a large realistic pedestrian detection dataset and evaluated 16 different state-of-the-art methods. It is shown that “no single feature has been shown to outperform HOG”, and “nearly all modern detectors employ some form of gradient histograms” [4].

The recently developed depth cameras can reduce the negative impact of bad illumination and complex background on pedestrian detection because the pixel values are not sensitive to light intensity and colors. Time-of-flight (TOF) cameras and Microsoft Kinect [5] are two kinds of popular depth cameras. The TOF camera resolves distance based on measuring the time-of-flight [6] of a light signal between the camera and the subject, and Kinect measures distance based on stereo triangulation with the help of infrared patterns. Kinect has a higher resolution and measures the distance more accurately. But the unique advantage of the TOF camera is its high frame rate with over 100 FPS [7]. So TOF camera can capture fast human motion that other sensors can not do.

Many recent pedestrian detection methods in depth images are also based on HOG [8], or inspired by HOG [9, 10]. However, noise in the depth images of TOF camera causes the gradient features to be not computed accurately, and the common smooth filters affect the sharp edges in an image while reducing the noise. In order to overcome this problem, we propose a framelet-based method that can reduce the impact of noise by automatically selecting the local framelet features. Under a tight and redundant system, the image local features are extracted by the framelet-based optimal model. The structure (sharp edges) in the depth images can be preserved meanwhile the noise is depressed. So it can extract robust features which can achieve a better detection rate.

The outline of the paper is as follows. A framelet-based method is proposed in Section 2. Experiments and analysis are presented in Section 3. Finally Section 4 concludes the paper.

2. FRAMELET FEATURES OPTIMIZATION MODEL
2.1. Gradient Feature Extraction
HOG as a local feature descriptor is effective to detect the pedestrian in depth images. Inspired by the HOG feature descriptor, the histogram of depth difference (HDD) was proposed and successful applied to depth images captured by time-of-flight (TOF) camera [10]. One of the key steps of the
HDD descriptor is to detect local variations of depth images. Suppose \( f(i, j) \) be the value at the \( i \)th row and \( j \)th column of the depth image \( f \), the gradients in the horizontal and vertical directions are respectively defined as

\[
\Delta_h(i, j) = f(i + 1, j) - f(i - 1, j),
\]
\[
\Delta_v(i, j) = f(i, j + 1) - f(i, j - 1). \tag{1}
\]

\( \Delta_h(i, j) \) and \( \Delta_v(i, j) \) are two first-order difference operators in horizontal and vertical directions, respectively.

For conveniently denoting the two dimensional framelet system, let the matrices \( T_L \) and \( T_1, T_2, T_3, T_4, T_5, T_6, T_7 \), and \( T_8 \) respectively represent the low-pass \( \tau_{0,0} \) and framelet operators \( \tau_{0,1}, \tau_{0,2} \) , \( \tau_{1,0}, \tau_{1,1}, \tau_{1,2}, \tau_{2,0}, \tau_{2,1} \) and \( \tau_{2,2} \), and the framelet coefficients \( \hat{f} = T_H f \) with \( T_H = [T^T_1, T^T_2, \ldots, T^T_8] \).

The framelet coefficients shown in the first row of Fig. 2 are generated by the framelet operators \( \tau_{0,1} \) and \( \tau_{1,0} \) from noisy depth images in Fig. 2. It is obvious that the coefficients are seriously contaminated by noises. That is to say, the noisy framelet coefficients can not accurately denote the local features of depth images, and we need to remove the noise in framelet coefficients and attenuate its effect on pedestrian detection. Thus, we need to propose a framelet feature optimization model for selecting useful framelet coefficients. Due to prior sparsity of framelet coefficients, the feature selecting optimization model for depth images is proposed as

\[
\hat{g} = \arg \min_g \left\{ \frac{1}{2} \| g - \hat{f} \|^2 + \| \Gamma \hat{f} \|_1 \right\}, \tag{2}
\]

where \( \frac{1}{2} \| g - \hat{f} \|^2 \) is the fidelity term, \( \| \Gamma \hat{f} \|_1 \) is the regularization term, and regularization parameter \( \Gamma \) defines a diagonal matrix as

\[
\Gamma := \text{diag}(\cdots, \gamma_t, \cdots), \quad \gamma_t \geq 0.
\]

Assume \( g_{\ell,i,j} \) and \( \hat{f}_{\ell,i,j} \) be the coefficients of the \( \ell \)th sub-band at \( i \)th row and \( j \)th column of \( g \) and \( \hat{f} \), respectively. Due to separability of framelet feature optimal model in Eq. (2) for every entry, it can be decomposed into an optimal model of every \( g_{\ell,i,j} \) as

\[
\hat{g}_{\ell,i,j} = \arg \min_{g_{\ell,i,j}} \left\{ \frac{1}{2} \left| g_{\ell,i,j} - \hat{f}_{\ell,i,j} \right|^2 + \gamma_{\ell,i,j} \left| \hat{f}_{\ell,i,j} \right| \right\} , \tag{3}
\]

where \( \gamma_{\ell,i,j} \) is the corresponding parameter in \( \Gamma \). The well known soft thresholding operator [13] is the optimal result for Eqn. (3) as

\[
\hat{g}_{\ell,i,j} = \text{sign}(\hat{f}_{\ell,i,j}) \max\{|\hat{f}_{\ell,i,j}| - \gamma_{\ell,i,j}, 0\}. \tag{4}
\]

The thresholding parameter \( \gamma_{\ell,i,j} \) determines whether the framelet feature \( \hat{f}_{\ell,i,j} \) is selected for pedestrian detection.

We assume that the difference between original coefficient \( g_{\ell,i,j} \) and observed coefficient \( \hat{f}_{\ell,i,j} \) follows a white zero-mean Gaussian distribution and the coefficient \( g_{\ell,i,j} \) is modeled as a Laplace distribution. Using the MAP technique, the \( \gamma_{\ell,i,j} \) is estimated by

\[
\gamma_{\ell,i,j} = \frac{\sqrt{2} \sigma_{\ell,i,j}^2}{\sigma_{\ell,i,j}}, \tag{5}
\]

where \( \sigma_{\ell,i,j}^2 \) is noise variance of the \( \ell \)th subband coefficients and \( \sigma_{\ell,i,j}^2 \) is the local signal variance [11, 12].
Fig. 2. The comparisons with and without the framelet features optimization model. The horizontal gradients are shown in the 1st and 3rd columns, and the vertical gradients are shown in the 2nd and 4th columns. The 1st and the 2nd rows present the gradients computed without and with the optimization model, respectively. It can be seen that the noise is depressed.

The parameter $\gamma_{f,i,j}$ is determined by the $\sigma_2^2$ and $\sigma_1^2$, $\gamma$. Due to the mean of each subband framelet coefficients being zero, the noise variance of each framelet subband $\sigma_2^2$ can be estimated by the method proposed in [14]. The local signal variance $\sigma_1^2_{f,i,j}$ is given by [15]

$$\sigma_1^2_{f,i,j} = \max \left\{ \sum_{k \in \mathcal{R}(f_{i,j})} \frac{\sqrt{2} |f_{i,j}[k]|}{|\mathcal{R}(f_{i,j})|} - \sigma_2^2, 10^{-6} \right\},$$

where the index set $\mathcal{R}(f_{i,j})$ is the neighborhood of the coefficient $f_{i,j}$ in the corresponding subband and $|\mathcal{R}(f_{i,j})|$ is the cardinality of the set $\mathcal{R}(f_{i,j})$. The number $10^{-6}$ is used only for numerical convenience. In this paper, a window with size of $3 \times 3$ for $\mathcal{R}(f_{i,j})$ is used in experiments.

A framelet features detection algorithm is presented as follow.

**Algorithm 1 (Framelet Features Selection)**

1. Calculate $\tilde{f} = T_{\text{off}}f$;
2. Calculate every $\gamma_{f,i,j}$ in Equ. (5);
3. Get $\tilde{g}_{f,i,j}$ by the thresholding operator in Equ. (4);

As the HDD method [10], the framelet features $\tilde{g}_1$ and $\tilde{g}_3$ are selected for experiment, respectively representing the local variations in horizontal and vertical directions. The depth difference of HDD denotes by a magnitude $\mathcal{M}(i,j) = (\tilde{g}_{f,i,j}^2 + \tilde{g}_{f,i,j}^2)^{1/2}$ and an orientation $\theta(i,j) = \arctan(\tilde{g}_{f,i,j}, \tilde{g}_{f,i,j})$. Each framelet subband is divided into overlapped blocks with four cells of size of $8 \times 8$ pixels, and any two adjacent blocks overlap half size of block. For a depth image with size of $64 \times 128$ pixels, there are $15 \times 7 = 105$ blocks for features detection. A histogram with orientation bins is used to describe statistical features of $\mathcal{M}(i,j)$ and $\theta(i,j)$ for each block, and the bins are uniformly spaced over $[0^\circ, 360^\circ]$. The local geometrical structures of each block is described by the two components $\mathcal{M}(i,j)$ and $\theta(i,j)$, and the histogram of these components at each block will be fed into a SVM classifier for pedestrian detection.

### 3. EXPERIMENTS

A large dataset has been collected\(^1\) to evaluate the performance of methods. The dataset is captured by a TOF camera with its type SwissRanger\(^TM\) SR4000 [6] designed by Mesa Imaging AG, and the distances between objects and the camera rang from 0 to 5 meters. The depth dataset is divided into training and testing sets. The training set contains 3160 positive and 14199 negative samples, and the testing set includes 1477 positive and 56802 negative samples.

As shown in Fig. 1, the noisy depth images are obvious with the artifacts in the smooth area. These artifacts will exist in the framelet coefficients detected by vertical and horizontal framelet operators as shown in the first row of Fig.2. To reduce the impact of noise, we need to determine which framelet features are efficient to describe the local geometric structures for classification. Thus, the automatic framelet features selection is proposed in Algorithm 1, and this algorithm is practical and attractive in real application of pedestrian detection. The framelet features shown in the second row of Fig. 2 are selected by the sparsity formula in Equ. (5), which show that noisy artifacts are removed and the edges are still preserved, comparing with the noisy framelet features in the first row of Fig. 2. To some degree, the efficient framelet features can be selected by our Algorithm 1.

In experiments, we utilize a popular classifier, linear SVMs [16], as the learning paradigm. To evaluate different kinds of features, Detection Error Tradeoff (DET) curves [1] are employed, which describe the miss rates against false positives per window (FPFW). Three different kinds of feature extraction method on depth images are evaluated. They are HOG [1], HDD [10] and the proposed framelet features method. The orientation bins for HOG are spaced over $[0^\circ, 180^\circ]$ and there are 9 bins as suggested in [1]. The bins for HDD are a little different, and are spaced over $[0^\circ, 360^\circ]$.

\(^1\)The dataset is available at http://yushiqi.cn/research/deepdataset
It is shown in [10] that 18 bins the orientation space is divided into for HDD performs best in depth image based pedestrian detection.

The experimental results are shown in Fig. 3. It is clearly shown that the proposed framelet feature extraction achieves a great improvement. The proposed method decreases the miss rate from HOG’s 7.9% and HDD’s 3.8% to 0.34% at FPPW=10^{-4}, and its improvement is over one order of magnitude. This shows that the framelet-based feature extraction method is robust to depth image noise and achieves the best results.

4. CONCLUSIONS

An adaptive features detection method is proposed to automatically select framelet features for pedestrian detection in noisy depth images. The selected framelet features can be effective to represent the local features and reduce effects of the noise. The proposed features detection method is adaptive, feasible and effective, and reduce the miss rate for pedestrian detection in noisy depth images.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (Grant No. 61202158), the Guangdong Natural Science Foundation (Grant No. S2012040006740), Shenzhen Fundamental Research Program (Grant No. JCYJ201105130443A, CXZZ20120618112909681, JCYJ20120613112757342), Natural Science Foundation of SZU (Grant No. 201211) and Open Projects Program of National Laboratory of Pattern Recognition.

5. REFERENCES