Feature Visualization Based Stacked Convolutional Neural Network for Human Body Detection in a Depth Image

Xiao Liu[1,2,3], Ling Mei[1,2,3], Dakun Yang[1,2,3], Jianhuang Lai[1,2,3], and Xiaohua Xie[1,2,3](✉)

[1] Sun Yat-sen University, Guangzhou 510006, China
xiexiaoh6@mail.sysu.edu.cn
[2] Guangdong Key Laboratory of Information Security Technology, Guangzhou, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

Abstract. Human body detection is a key technology in the fields of biometric recognition, and the detection in a depth image is rather challenging due to serious noise effects and lack of texture information. For addressing this issue, we propose the feature visualization based stacked convolutional neural network (FV-SCNN), which can be trained by a two-layer unsupervised learning. Specifically, the next CNN layer is obtained by optimizing a sparse auto-encoder (SAE) on the reconstructed visualization of the former to capture robust high-level features. Experiments on SZU Depth Pedestrian dataset verify that the proposed method can achieve favorable accuracy for body detection. The key of our method is that the CNN-based feature visualization actually pursues a data-driven processing for a depth map, and significantly alleviates the influences of noise and corruptions on body detection.

Keywords: Human detection · Depth image · Feature visualization Sparse auto-encoder · Convolutional neural network

1 Introduction

Human body detection is a basic task in biometric recognition which can be widely applied in tracking, gait recognition and face anti-spoofing detection [1]. However, earlier detection methods used RGB camera is unavailable in some special cases such as low-lighting scenes. To address this problem, depth cameras have been considered for the human detection. Compared with RGB image, depth image containing the 3D structure of the scene is insensitive to lighting changes. Therefore human body detection in depth image has become an active and attractive research area in the computer vision community [2].

Regarding to human body detection in depth image, most of depth descriptors are similar to those in RGB or gray images [3]. For instance, Wu et al. [4]

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 87–98, 2018.
https://doi.org/10.1007/978-3-030-03335-4_8

proposed Histogram of Depth Difference (HDD) descriptor and Spinello [5] proposed Histograms of Oriented Depths (HOD) descriptor, which were similar to HOG. Yu et al. [6] proposed a Simplified Local Ternary Patterns (SLTP) descriptor, which improved the Local Ternary Patterns (LTP) and apply to human body detection in depth imagery. However, a CNN with only one convolutional layer and one pooling layer is used in [3], which actually expresses a shallow representation of the depth image. In general, a deep representation obtained by a deep CNN is better to express an image than a shallow representation [7]. Therefore, we would like to investigate how to develop Su et al.'s method [3] to a deeper version.

Since it is unreasonable to use original image patches for training all network layers, we employ the feature visualization technology to generate layer-specific images at each network layer, then extract layer-specific image patches to train corresponding SAE. Specifically, we adopt the image representation inverting algorithm [8], which is able to use only information from image representation to reconstruct the image. Figure 1 illustrates the images reconstructed from the features at different convolution layers (conv1–conv5) for a given input image. As shown, the reconstructed images correspond to multi-level semantic abstraction but hold some invariant geometric and photometric information [9]. We call these reconstructed images as layer-specific *high-level images*.

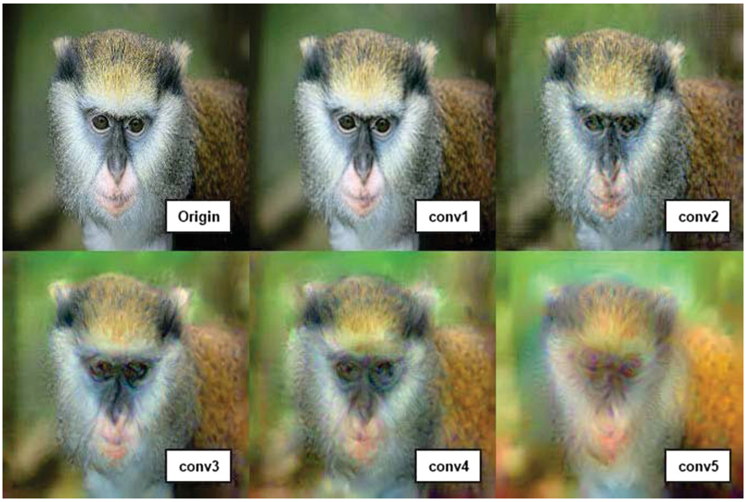


Fig. 1. Illustration of the feature visualization for convolutional neural networks.

Overall, the feature visualization based stacked convolutional neural network (FV-SCNN) is proposed to extract features for human body detection in depth images. FV-SCNN is rather different from Stacked Sparse Autoencoders (SSAE) [10] or Stacked Denoising Autoencoders (SAE) [11]. Specially, the CNN-based feature visualization actually achieves a data-driven processing for depth map, and enables the FV-SCNN to alleviate the influences of noise and corruptions on body detection.

Moreover, sliding window approach is widely used in body detection task [12], but it is rather time-consuming. To address this problem, we follow Su et al.'s method [3] to use the histogram of depth to extract candidate depth planes. Combined with the multi-scale window strategy, our method can not only avoid the time-consuming siding window search, but also generate high-quality candidates for body detection. Compared with [3] that uses k-means algorithm to detect the candidate center, our method is more robust to the noise, corruption, and the non-body parts.

The remainder of the paper is outlined as followed. The detailed introduction of the proposed method is presented in Sect. 2. Experiments are reported in Sect. 3 and the conclusion of the paper is made in Sect. 4.

2 Technical Approach

2.1 The Overview of the FV-SCNN Body Detection Framework

We utilize the FV-SCNN to learn the candidate body centers in depth images, then develop a multi-scale body candidate windows with body centers to locate the body areas.

The proposed FV-SCNN based human body detection framework is shown in Fig. 2. As shown, the proposed model contains two CNNs with each containing one convolutional layer and one pooling layer. In the training module, a large mount of image patches are randomly extracted from original training set (depth images) and used to train a SAE network. In our experiment, the size of image patch is set to 16 by 16. The optimized weights and the bias of the SAE network are employed to construct the filter of the first CNN. After that, the sub-images with fixed aspect ratio (e.g., 120:64) extracted from original training set are resized to 120 × 64 and put into the first CNN, yielding corresponding feature maps. Based on each feature map, a high level image can be reconstructed by using the feature visualization technology [8]. Like the scheme to construct the first CNN, the randomly extracted patches from the first-layer high-level body images are used to train the weights of another SAE for forming the second CNN. After the second CNN is formed, the high-level images are input the second CNN followed by a PCA to produce the final features. The features of different labelled sub-images (body vs. non-body) are further used to train a SVM classifier. Specifically, the sub-image exactly containing a human body is labelled as a body sub-image, otherwise as non-body one.

In the application module, a set of candidate windows (sub-images) are generated for each input image. Each candidate sub-image is resized to 120 × 64 and put into the first CNN. The corresponding high level image reconstructed from the first-layer feature map is further fed into the second CNN to export the features, and further processed by PCA for a dimensionality reduction. The classifier response based on the final features will judge whether current window

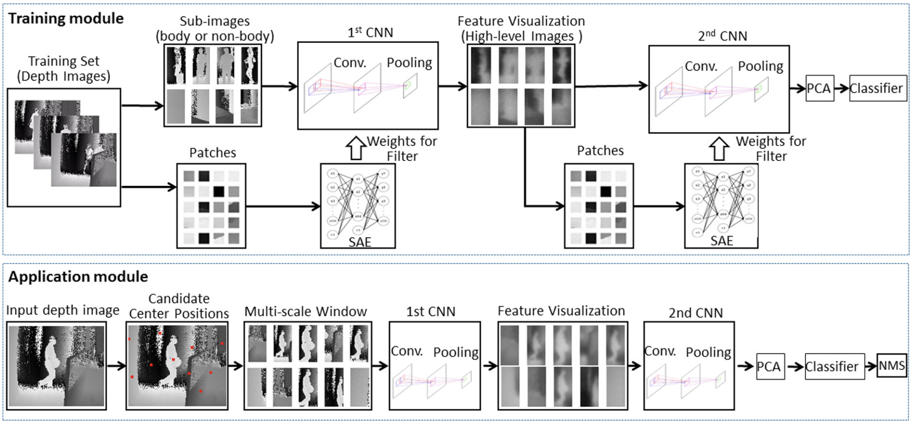


Fig. 2. Illustration of the proposed human body detection framework.

contains a human body. Finally, the non-maximal suppression (NMS) is used to merge the overlapping detected windows and get the final locations.

We demonstrate only two-layer CNNs in our model because a practicable computing cost should be considered for a body detector. Specially, the first CNN-based feature visualization tends to perform a data-driven processing for depth map, and alleviate the influences of noise, corruption, and non-body components on body detection. Actually, our model can be directly extended to more than two layers by utilizing the reconstructed high-level image of a specific CNN as the input of the following CNN. Details about our method are presented in Sect. 3.

2.2 Sparse Auto-Encoder (SAE)

Recently, deep multi-layer neural networks have many levels of non-linearities allowing them to compactly represent highly non-linear and highly-varying functions. Auto-Encoder (AE) is an unsupervised feature learning algorithm which aims to develop better feature representation of input high-dimensional data by finding the correlation among the data. For an AE network, the output vector is equal to the input vector. Training an AE can minimize reconstruction error amounts and obtain the mutual information between input and learnt representation. Intuitively, if a representation allows a good reconstruction of its input, it means that the representation has retained much of the information that was presented in the input. Specifically, the AE is a three-layers neural network with a single hidden layer forming an encoder and a decoder which proposed in [13].

Auto-Encoder (AE) can avoid the labor-intensive and handcraft feature design. When the number of hidden units in AE is less than that of the input units, a compression representation achieved. When the number of hidden units is larger, even more than that of the input units, interesting structure of input data can still be discovered by imposing a sparsity constraint on the hidden units. The Auto-Encoder with only few hidden units activated for a given input is called the Sparse Auto-Encoder (SAE). Specially, the sparsity regularization typically leads to more interpretable features for representing a visual object.

For a SAE network, let $\hat{\rho}_j$ be the mean activation probability in the jth hidden unit, namely $\hat{\rho}_j = (1/m)] \sum_{i=1}^{m} h_j$. Let ρ be the desired probability of being activated. Sparsity is imposed on the network, it is obvious that $\rho \ll 1$. Here Kullback-Leibler (KL) divergence is used to measure the similarity between the desired and actual distributions, as shown in the following equation

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{1}$$

The SAE model can be formulated as the following optimization problem

$$\min_{W,b} \left[\sum_{i=1}^{m} (h_{W,b}(x^{(i)}) - y^{(i)})^2 + \lambda(\|W\|_2^2) + \beta \sum_{j=1}^{k} KL(\rho \| \hat{\rho}_j) \right], \tag{2}$$

where the first term is the reconstruction cost, the second term is a regularization on weight to avoid over-fitting, and the last term enforces the mapping sparsity from the input layer to hidden layer. The parameters λ and β are regularization factors used to make a tradeoff between the reconstruction cost, weight decay and sparsity penalty term. Typically, back-propagation algorithm is used to solve Eq. (2).

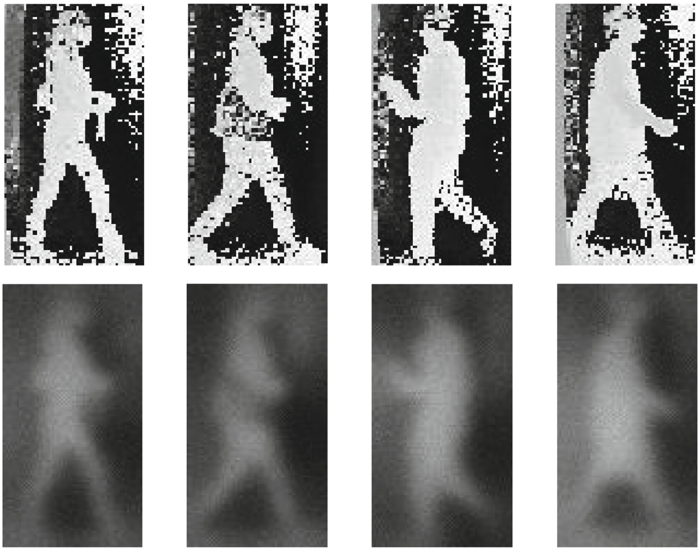


Fig. 3. Examples of feature visualization for depth images. The first row are the original images and the second row are the corresponding high level images reconstructed from CNN features.

2.3 Feature Visualization

In the proposed model, the feature maps of the first CNN are inverted and visualized to generate the high-level images as the input of the second CNN. We adapt Aravindh Mahendran's method [8] to achieve this goal. Given a representation function $\Phi : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^d$ and a representation $\Phi_0 = \Phi(x_0)$ to be inverted, x_0 is the input feature image, reconstruction finds the image $x \in \mathbb{R}^{H\times W\times C}$ that minimizes the objective

$$x^* = \operatorname{argmin}_{x\in\mathbb{R}^{H\times W\times C}} \ell(\Phi(x), \Phi_0) + \lambda\Re(x), \tag{3}$$

where the loss ℓ compares the image representation $\Phi(x)$ to the target one Φ_0 and $\Re : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}$ is a regulariser capturing a natural image prior.

In this paper, as same as [8], we choose the Euclidean distance for the loss function ℓ as follows

$$\ell(\Phi(x), \Phi_0) = \|\Phi(x) - \Phi_0\|^2. \tag{4}$$

For the regulariser $\Re(x)$, it contains two parts which incorporate two image priors, then it can be written as

$$\Re(x) = \Re_\alpha(x) + \Re_{V^\beta}(x), \tag{5}$$

where $\Re_\alpha(x) = \|x\|_\alpha^\alpha$ is the α-norm, which encourages the range of the image to stay within a target interval instead of diverging. Since images are discrete, the total variation (TV) norm is replaced by the finite-difference approximation:

$$\Re_{V^\beta}(x) = \sum_{i,j} \left((x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2\right)^{\frac{\beta}{2}}. \tag{6}$$

Through the above mentioned, the final form of the objective function is

$$\|\Phi(x) - \Phi_0\|_2^2 + \lambda_\alpha\Re_\alpha(x) + \lambda_{V^\beta}\Re_{V^\beta}(x). \tag{7}$$

In this paper, the simple gradient descent procedure is used to optimize the problem of the objective (7). In the iteration process, the parameter x is updated as follows:

$$\begin{aligned} \mu_t &= m\mu_{t-1} - \eta_{t-1}\Delta E(x) \\ x_{t+1} &= x_t + \mu_t, \end{aligned} \tag{8}$$

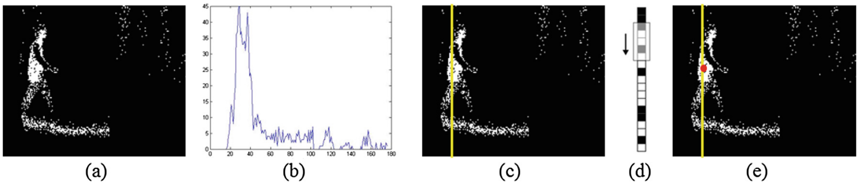
where $E(x) = \ell(\Phi(x), \Phi_0) + \lambda\Re(x)$ is the objective function, $m\mu_{t-1}$ is the momentum with the momentum parameter m, and η_{t-1} is the learning rate.

Some examples of feature visualization are illustrated in Fig. 3. Compared with the contents in original depth images, both noise and non-body components have been cleared up in the high-level images, but essential structures of human body are preserved.

2.4 Localization of Body Candidate

We follow Su et al.'s method [3] to compute the histogram of depth values in a depth image and further to extract a set of candidate depth planes with respect to the local peaks of the histogram. The depth map with respect to each depth plane is converted into a binary image, which indicates whether a pixel belongs to current depth plane (1 for yes and 0 for not). Such a binary image is named as a depth plane mask. For locating the human body, Su et al. [3] apply the k-means clustering on each depth plane mask and regard the clustering center as the body center of candidate. However, the accuracy of this manner could be easily affected by the noise and corruption of the depth map as well as the non-body components.

For more accurately locating the human body center, we propose using the vertical projection method to locate the X-coordinate of the body candidate. Assumes that the human center position is (x_0, y_0). We observe that the x_0-th column of depth map generally contains more points than other columns in current depth plane. Inspired by this, the current depth plane mask is vertically projected onto the horizontal axis. The positions corresponding to the maximum projection value on the horizontal axis are regarded as the X-coordinates of the candidate body centers. For each candidate X-coordinate x_p, we perform an 1-D average filtering on the x_p column of the depth plane mask. In our experiment, the filter length is set to 8. After filtering, the location with respect to the maximum response is taken as the Y-coordinates of the candidate body center. When multiple locations hold the maximum response value, the maximum coordinate and the minimum one among these locations are averaged to be the Y-coordinates of the candidate body center. The proposed candidate body center localization method is illustrated in Fig. 4. At each candidate center, we generate multi-scale windows as the candidate sub-images for the first CNN to get corresponding feature maps, and use the pre-trained classifier to judge the human body.



(a) (b) (c) (d) (e)

Fig. 4. Illustration of candidate body center localization scheme. For each candidate depth plane mask (a), it is vertically projected onto the X-axle (b). The column corresponds to maximum projection value is selected (c), and then is smoothed by a 1-D average filter (d). The location with respect to the maximum filtering response is taken as the candidate body center.

3 Experiments

3.1 Experimental Setting

This section presents experiments on SZU Depth Pedestrian dataset [4,14] to evaluate the proposed method for body detection. We divide the dataset as the principle in [3]. The dataset is captured by a Time-Of-Flight (TOF) camera, only depth images are used in our experiments, the resolution of them is 176 × 144 pixels. The number of training and testing images are 4435 and 4029 respectively. We found that both the training of the feature learning with SAE and softmax classifier need only few training examples, therefore we used 400 training images which extracted 40,000 patches randomly for training the first SAE. The number of neurons in the feature layer of SAE is set as 64. The sub-images are extracted and normalized to the size of 120 × 64. The size of pooling filter in each CNN is set 7×7 and the model finally outputs a 6,720 dimensional feature. The output feature is finally processed by PCA to produce a 1,000 dimensional feature vector and sent to the SVM classifier. For training the classifier, the proportion of positive and negative samples should be 1:6 [3]. We extracted 100 body sub-images and 600 non-body sub-images from the training set. These sub-images are reconstructed to train the second SAE. For body detection application, 300 positive samples (the depth images containing a pedestrian) as well as 300 negative samples are randomly opted for testing.

We first use two experiments to investigate the performance of the proposed candidate localization method, and then compare the FV-SCNN based body detection method with related methods.

3.2 Investigation on the Body Candidate Localization Method

In this experiment, we compare the proposed body candidate center localization method (in Sect. 2.4) with the k-means based method [3]. Both methods work on the same depth plane detected by the histogram analysis method mentioned in [3].

Figure 5 shows a comparison example of candidate center localization on a depth plane in a depth image by using different method, and Fig. 6 illustrates the results on all depth planes. In order to present the superiority of our method in terms of extracting accurate candidate points, we exploits the same classification method to compare the performance of locating candidate points. As shown, the proposed method gets far more accurate center localization result than the K-means method, and generates less candidates. It is notable that the k-means localization method is easily affected by the non-body objects. Further, it is a remarkable fact that k-means cost a lot of computation time. In our experiment, it takes 14.7921 s to generate all the candidate positions for each depth image while our method cost only 0.0046 s. Additionally, k-means based method is sensitive to the clustering initialization.

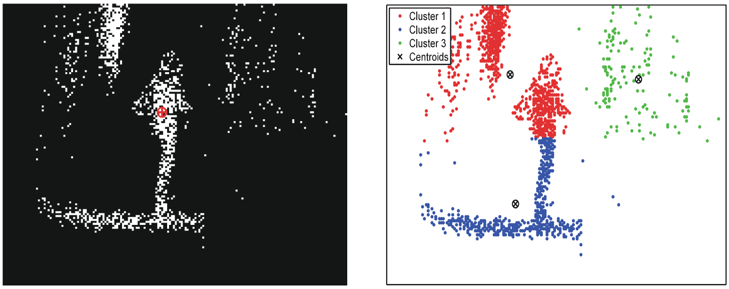


Fig. 5. Illustration of candidate center localization result on a depth plane. Left is the result of our method (shown with red cross) while the right is the result of the k-means based method (shown with black cross). (Color figure online)

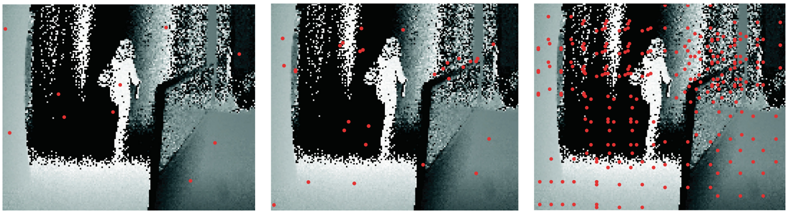


Fig. 6. Illustration of the candidate center localization result on all depth planes. Left is the result of our method and the middle is the result of k-means based method. Right is the result of k-means based method by extending more eight points around each candidate position by 16 pixel step, which is also suggested in [3].

3.3 Investigation on the Feature Visualization Based Stacked Network

In this experiment, we compare the proposed FV-SCNN model with the SAE-CNN method [3] which forms a single-layer CNN by optimizing a SAE. We also implement a specially designed HOG-FV-CNN model, which first performs HOG presentation visualization [15] to obtain a HOG based high-level reconstruction image, and then use a SAE-CNN model to extract the features for classification. Simply speaking, in regard to HOG-FV-CNN, the HOG is used in place of the first CNN in the proposed FV-SCNN model.

The human body detection results by different method are shown in Table 1. In this experiment, the sub-images are directly used for testing. That is, the algorithms do not need to localize the body candidate but only return whether the input sub-image contains a body. The classification accuracy rates are reported. As shown, the performance of HOG-FV-CNN is obviously worse than the FV-SCNN model, even worse than the single-layer SAE-CNN model. The experimental results support that the deeper SAE-CNN network outperforms the single-layer one. Furthermore, the CNN features work better than the hand-designed HOG features in our framework.

Table 1. Human body detection accuracies by different methods.

Methods	SAE-CNN	HOG-FV-CNN	FV-SCNN
Accuracy rate	94.5%	86%	**96.33%**
♯Layer	1	2	2

To deeply investigate the difference between HOG-FV-CNN and FV-SCNN, we illustrate the reconstructed high-level images by these two methods in Fig. 7. As shown, the high-level image generated by FV-SCNN contains the main human structures but suppresses the noise and corruption (especially pay attention to the upper-left part of image). By contrast, the reconstructed high-level image by HOG is much rougher and the body configuration is distorted. The main reason may be that the SAE is learnt by using the patches from the body sub-images, so that the formed CNN responds more prominently on the body parts than the non-body parts. However, as a kind of hand-craft features, the HOG takes responses equally on different parts.

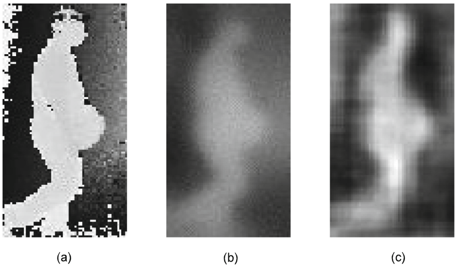


Fig. 7. Examples of high-level images reconstructed by different method. (a) is the original depth image, (b) and (c) are the high-level images reconstructed by FV-SCNN and HOG-FV-CNN, respectively.

3.4 Comparison with State-of-the-Art Methods

We also compared the proposed method with five state-of-the-art depth descriptors for pedestrian detection in depth imagery, including the Histogram of Oriented Depths (HOD) [5], Histogram of Depth Difference (HDD) [4], Relational Depth Similarity Feature (RDSF) [16], Simplified Local Ternary Patterns (SLTP) [6], and SAE-CNN [3]. For a fair comparison, we adopt the same candidate windows and the same classifier (SVM) for different methods. The body detection accuracy is evaluated using the intersection over union (IoU), which is defined as the ratio of intersection to union between the results and ground-truth bounding boxes. When the IoU is larger than 0.5, we treat current result as a correct detection.

The body detection results of different methods are shown in Fig. 8, which plots miss rate against FPPI (False Positives Per Image). Smaller miss rate at a fixed FPPI means more accuracy of the detection. As shown, the proposed FV-SCNN method outperforms all other methods, and SAE-CNN perform better than HOD, HDD, RDSF and SLTP, the superiority becomes more obvious with a higher FPPI because we add the candidate localization method to the proposed method from a systematical standpoint. The result revealing that the feature learning is better than using hand-crafted feature. The proposed FV-SCNN performs better than SAE-CNN, which verifies that a deeper network architecture is more helpful for feature learning, and the proposed feature visualization based network stacking manner is effective.

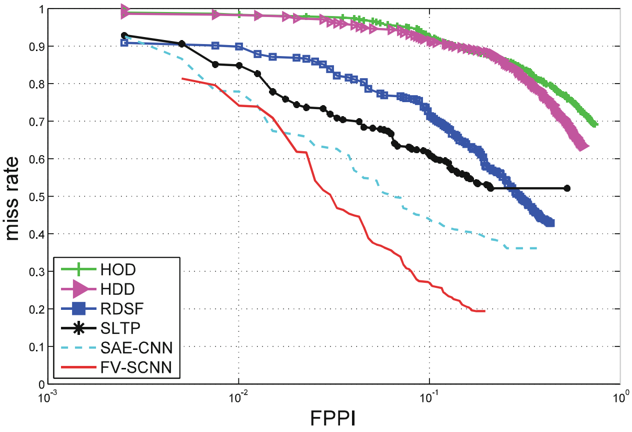


Fig. 8. Comparison of the proposed FV-SCNN with state-of-the-art methods including hand-designed descriptors (HOD, HDD, RDSF) and learning based feature (SAE-CNN).

4 Conclusion and Future Work

This paper presents a feature visualization based stacked convolutional neural network (FV-SCNN), where the feature visualization technology is used for connecting multiple CNN layers. The FV-SCNN can be learned in a layer-wise unsupervised manner by SAE. The FV-SCNN has been demonstrated for human body detection in depth images. Experiments and visualization results reveal that the proposed method significantly alleviates the influences of noise, corruption, and non-body components on body detection. The proposed method also obtains a better body candidate localization result than the traditional methods in body detection. In the future, we would like to apply the FV-SCNN to other visual recognition processing tasks with deeper architectures. We also would like to develop a fine-tuning method for the FV-SCNN, and investigate how to jointly optimize the FV-SCNN and the classifier.

Acknowledgements. This project is supported by the Natural Science Foundation of China (61573387, 61672544), Guangzhou Project (201807010070), and Fundamental Research Funds for the Central Universities (No. 161gpy41).

References

1. Mei, L., Yang, D., Feng, Z., Lai, J.: WLD-TOP based algorithm against face spoofing attacks. Biometric Recognition. LNCS, vol. 9428, pp. 135–142. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25417-3_17
2. Lee, G.-H., Kim, D.-S., Kyung, C.-M.: Advanced human detection using fused information of depth and intensity images. In: Kyung, C.-M. (ed.) Theory and Applications of Smart Cameras. KRS, pp. 265–279. Springer, Dordrecht (2016). https://doi.org/10.1007/978-94-017-9987-4_12
3. Su, S., Liu, Z., Xu, S., Li, S., Ji, R.: Sparse auto-encoder based feature learning for human body detection in depth image. Signal Process. **112**, 43–52 (2015)
4. Wu, S., Yu, S., Chen, W.: An attempt to pedestrian detection in depth images. In: 3rd Chinese Conference on Intelligent Visual Surveillance, pp. 97–100. IEEE Press, Beijing (2011)
5. Spinello, L., Arras, K.-O.: People detection in RGB-D data. In: 2011 International Conference on Intelligent Robots and Systems, pp. 3838–3843. IEEE Press, San Francisco (2011)
6. Yu, S., Wu, S., Wang, L.: SLTP: a fast descriptor for people detection in depth images. In: 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 43–47. IEEE Press, Beijing (2012)
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., et al.: Going deeper with convolutions. In: 2015 CVPR, pp. 1–9. IEEE Press, Boston (2015)
8. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: 2015 CVPR, pp. 5188–5196. IEEE Press, Boston (2015)
9. Mei, L., Chen, Z.-Y., Lai, J.-H.: Geodesic-based probability propagation for efficient optical flow. Electron. Lett. **54**(12), 758–760 (2018). https://doi.org/10.1049/el.2018.0394. Print ISSN: 0013-5194. Online ISSN: 1350-911X
10. Yang, D., Lai, J., Mei, L.: Deep representations based on sparse auto-encoder networks for face spoofing detection. In: You, Z., et al. (eds.) CCBR 2016. LNCS, vol. 9967, pp. 620–627. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46654-5_68
11. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(Dec), 3371–3408 (2010)
12. Uijlings, J., Van De Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
13. Hinton, G.-E., Salakhutdinov, R.-R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
14. Li, Y.-R., Yu, S., Wu, S.: Pedestrian detection in depth images using framelet regularization. In: 2012 IEEE International Conference on Computer Science and Automation Engineering, CSAE, pp. 300–303. IEEE Press (2012)
15. Weinzaepfel, P., Jégou, H., Pérez, P.: Reconstructing an image from its local descriptors. In: 2011 CVPR, pp. 337–344. IEEE Press, Colorado Springs (2011)
16. Ikemura, S., Fujiyoshi, H.: Real-time human detection using relational depth similarity features. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6495, pp. 25–38. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19282-1_3