Detecting Humans in RGB-D Data with CNNs

Kaiyang Zhou
University of Bristol
kz15291@my.bristol.ac.uk

Adeline Paiement
Swansea University
A.T.M.Paiement@swansea.ac.uk

Majid Mirmehdi
University of Bristol
m.mirmehdi@bristol.ac.uk

Abstract

We address the problem of people detection in RGB-D data where we leverage depth information to develop a region-of-interest (ROI) selection method that provides proposals to two color and depth CNNs. To combine the detections produced by the two CNNs, we propose a novel fusion approach based on the characteristics of depth images. We also present a new depth-encoding scheme, which not only encodes depth images into three channels but also enhances the information for classification. We conduct experiments on a publicly available RGB-D people dataset and show that our approach outperforms the baseline models that only use RGB data.

1 Introduction

RGB-D images encapsulate richer information by providing depth along with color values. In RGB-D human detection, depth information is usually used to reduce the search space [1]. For example, Jafari et al. [2] use depth pixels to extract the regions of interest (ROIs) by classifying each pixel into one of three categories: ground plane, ROIs (objects) and non-ROIs (buildings and walls). Zhang et al. [3] remove the ground plane and ceiling, then propose ROIs based on the density distribution along the depth dimension. In this paper, we also leverage depth information to remove the ground plane, but we further constrain ROIs by exploiting the characteristics of depth.

Convolutional neural networks (CNNs), which exhibit significant powers of discrimination in the color image domain [4], have been successfully applied to people detection. Angelova et al. [5] propose a cascade framework consisting of several CNNs for pedestrian detection where proposals are obtained by a dense sliding window. However, each CNN in each cascade stage is applied repeatedly to each proposal, without sharing computations on convolutions. We also use CNNs for proposal classification, but convolutions are performed only once on the entire image, which is achieved by the ROI-pooling proposed by Girshick [6]. ROI-pooling extracts a fixed-size feature vector for each ROI window in the last set of convolutional feature maps and forwards these feature vectors to the fully connected layers, which actually require fixed-size input.

CNNs have also been applied in RGB-D images for human detection, such as [7, 8]. In [7] however, depth is not included in the classification. Mees et al. [8] develop a mixture of CNNs to extract features independently from different modalities including color, depth and motion, and use a gating network to fuse them for further classification. However, they use the traditional sliding-window approach to generate proposals, which ignores the potential of depth for ROI selection. In this paper, we apply two CNNs to learning features from color and depth images respectively. They perform human detection in ROIs of each modality separately, followed by a fusion of these results to obtain a more reliable detection overall.

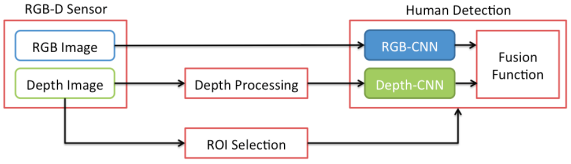


Figure 1: Overview of our human detection system.

Our contributions in this paper are: (a) we develop a fast ROI selection method based on depth to reduce the search space, (b) we propose a CNN-based RGB-D human detector where we design a novel way to fuse detections from RGB and depth images, (c) we propose a fast depth-encoding method, which can produce three-channel depth images that are close to color images in terms of saliency of information. This allows a more effective deployment of depth information by CNNs through the transfer of pre-learnt color features.

2 Proposed Approach

Fig. 1 shows a compact overview of our human detection system. The ROI Selection module exploits the depth information to reduce the search space and generates a set of candidate proposals. The Depth Processing module fills holes in the depth image using a mean-filter, and then it encodes the filled image into three channels. The two networks, RGB-CNN and Depth-CNN, in the Human Detection module, process the RGB and encoded depth images respectively, producing probability scores for the candidate proposals. For each proposal, two probability scores produced by the two networks are combined by a fusion function, leading to a stronger probability that indicates whether the proposal contains a person (upper body) or not. In the end, proposals with high probabilities are kept and then passed through non-maximum suppression, resulting in the final list of non-overlapping windows.

2.1 Depth Processing

A good depth-encoding method needs to satisfy three criteria: (i) it should preserve as much information contained in the original depth as possible, such

as shape, (ii) it should be computationally cheap, and (iii) it should produce an image with characteristics that matches that of a color image, notably in terms of range and contrast. Based on these, we evaluate three existing methods, namely *depth-gray* (DG) and *color-depth* (CD) from [9] and *contrast-enhanced depth-gray* (CE) from [10], and propose a new scheme in this work. In particular, DG normalizes depth pixels to have values ranging from 0 to 255 and then replicates them to three channels. CD maps each normalized pixel to three channels via a reversed jet colormap[1] where the resulting color values are ranged from red (near) through green to blue (far). CE is similar to DG, except that it performs histogram equalization after the normalization and before the replication into the three channels. We further propose a new depth-encoding scheme called *contrast-enhanced color-depth* (CECD), which performs histogram equalization before mapping pixels to the reversed jet colormap. These four schemes are compared in Section 3.

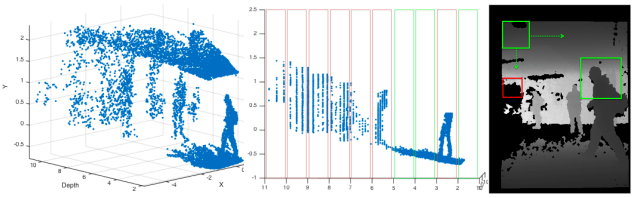
2.2 ROI Selection

The ROI Selection process is composed of three stages, outlined below, which work together to produce candidate proposals. [2]

Ground Plane Detection (GPD) - We project depth pixels to the global 3D world with known depth camera intrinsic parameters (see Fig. 2a). To determine the ground plane accurately in the presence of numerous outliers, we sample the 3D world into a grid of 10×10 cells (see Fig. 2b). Since the ground is usually located in the lower part of a scene, we only sample points from the lower half of the image. For each bin, we compute standard deviation of the points in vertical columns spanning this area (VSTD). The bins with VSTD values larger than an empirical threshold are removed and the rest of the points are fed to a RANSAC-based plane fitting algorithm [11]. This is based on the observation that bins with points largely spread in the vertical direction will contain fewer ground pixels and thus should be filtered out. In the end, pixels close to the plane correspond to non-ROIs and the rest are selected for further ROI selection. Jafari et al. also use an outlier-reduction method before detecting the ground plane in [2]. However, they remove bins with a high density of points, which can lead to the removal of bins that may contain pixels that belong to the ground plane at locations close to the camera (in this case, the points are quite compact).

Scale-Informed ROI Search (SIS) - We slide a window across the remaining pixels. The window width for each pixel is dynamically determined by the depth value which has the benefit of avoiding a time-consuming multi-scale search. In particular, the window width λ at a specific pixel is determined by $\lambda = \frac{fW}{Z}$, where f is the focal length of the depth camera, W is the rough width of a normal human (we used 0.6m), and Z is the depth value (in meters). We only detect the human upper body, so each window is a square bounding box.

Candidate Proposals Filtering (CPF) - To further reduce the number of proposals, we discard those that mainly contain invalid pixels (no depth value) arising from objects with poor reflective properties. To efficiently compute the number of valid pixels in a bounding box, we employ the idea of integral images [12]. A proposal is only selected if its portion of valid pixels is larger than a threshold e.g. one-third. This process can be executed in parallel with the SIS stage to keep the computation fast. Fig. 2c shows an example of valid (green) and invalid (red) proposals.



(a) (b) (c)

Figure 2: ROI selection - see text for explanation.

2.3 Human Detection with CNNs

The architecture of our system is shown in Fig. 3. The blue stream network (RGB-CNN) processes RGB images while the green stream network (Depth-CNN) processes depth images. They are identical in architecture but do not share parameters. We use CaffeNet [13] but change the 1000-way classification layer with a 2-way classification layer to suit our purposes. CaffeNet is essentially a variant of AlexNet [4] where the order of pooling and normalization layers is switched. We replace the max-pooling layer after the 5^{th} convolutional layer with the ROI-pooling layer [6]. Further, we follow [6] to restructure the FC6 and FC7 layers via truncated SVD, resulting in size-reduced weight parameters which helps increase the processing speed. We apply multi-scale detection at test time [6].

Fusion of RGB and Depth Detections - The two networks work independently in each domain to score each proposal. For each proposal, the RGB-CNN produces a probability $P(y = 1|X_c)$ and the Depth-CNN produces a probability $P(y = 1|X_d)$, where $y = 1$ means a window contains a person (upper body), and X_c and X_d represent the proposal region in RGB and depth images respectively. The two probabilities are fused to deduce the final probability $P(y = 1|X_c, X_d)$, which is a stronger evidence for classifying the proposal. We compute the fusion probability as

$$P(y = 1|X_c, X_d) \propto exp((1 - \omega)L(y = 1|X_c) + \omega L(y = 1|X_d)), \quad (1)$$

where $L(\cdot|\cdot)$ is the log likelihood and ω is an adaptive weight dependent on the proposal depth, defined as

$$\omega = \begin{cases} 1 & d <= 1 \\ -\frac{1}{5}(d-1)+1 & 1 < d < 6 \\ 0 & d >= 6 \end{cases}, \quad (2)$$

where d is the depth value in meters and its threshold values in (2) are set empirically. In Eq. (1), if the proposal is close to the camera, we give more weight to the depth information, while if the proposal is far away from the camera, we rely more on RGB information. This is grounded in the fact that the performance

[1] https://uk.mathworks.com/help/matlab/ref/colormap.html

[2] The code for the ROI selection module is available on https://github.com/KaiyangZhou/ROI-Selection

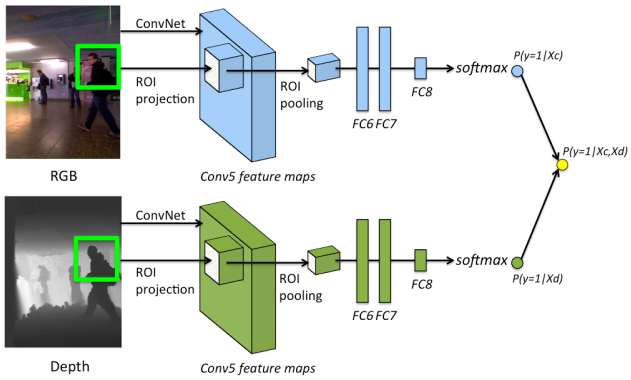


Figure 3: RGB-D human detection CNN architectures. The blue and green streams process RGB and depth images respectively. Each network takes as input a full-size image and the same set of ROIs. The detection probabilities from each network are fused to infer a stronger probability.

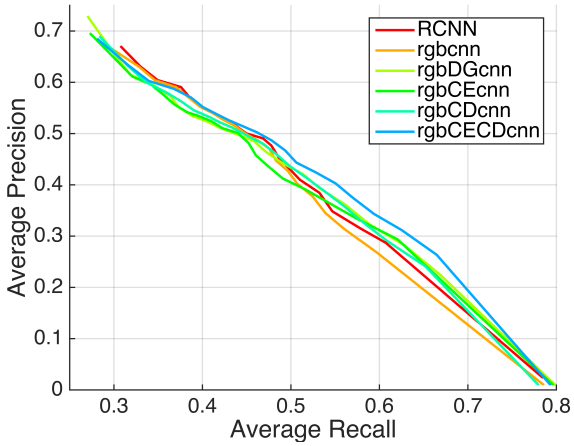


Figure 4: Experimental results of different models and using different depth encodings. The model with the CECD encoding (i.e. rgbCECDcnn) performs the best.

of most depth sensors downgrades as the distance increases. It is also easy to observe in depth images that people closer to the camera have more clear shapes than people who are far away (see e.g. Fig. 2c).

3 Experiments

Dataset - For training data, we collect positive training data provided by SPHERE [14] and negative training data from NUYD2 dataset [15]. This results in 3271 positive images and 7574 negative images. We choose the *RGBD people dataset* [16] to evaluate our approach. This dataset contains three videos each including 1000+ RGB-D images. We found that some annotations were missing in this subset, thus we manually added any missing annotations.[3]

Network Training - We use *Caffe* [13] to train and test our networks with the SGD algorithm (learning rate 0.001, momentum 0.9 and weight decay 0.0005). All the networks are initialized with ImageNet weights, except for the new 2-way classification layer which is initialized with a Gaussian distribution. Fine-tuning is applied to the RGB-CNN (only for fc layers) for 3 epochs where the learning rate is lowered by 1/10 after 2 epochs. The Depth-CNN is fine-tuned with different settings. For each depth-encoding scheme (see Section 2.1), the one producing the best result when combining with the RGB-CNN is selected for later comparison.[4]

Evaluation Methodology - For evaluation, we determine average-precision against average-recall. We adopt the 'no-reward-no-penalty' rule [16] with positive detections at 50% overlap and penalize repeated detections on the same person. We use three models as the baseline: RCNN [17] + our ROI method, our RGB-CNN + our ROI method, and our RGB-CNN + SelectiveSearch [18]. We disable the bounding box regression in RCNN as we found that our proposal method is less sensitive to localization error. We tailor SelectiveSearch particularly for upper body detection by setting the height of each proposal to its width, keeping only the square upper part of elongated proposals, and discarding proposals with width < 50 pixels. The first two baselines let us assess the advantage of using depth information for the detection stage, as well as our proposed color and depth fusion method. The third baseline allows the evaluation of our ROI method comparing to SelectiveSearch used in [17].

Results and Discussion - The experimental results are shown in Fig. 4 where rgbcnn represents our RGB-CNN and rgb$*$cnn with $* \in$ {DG, CE, CD, CECD} represents combined RGB-CNN and Depth-CNN with different depth encodings.

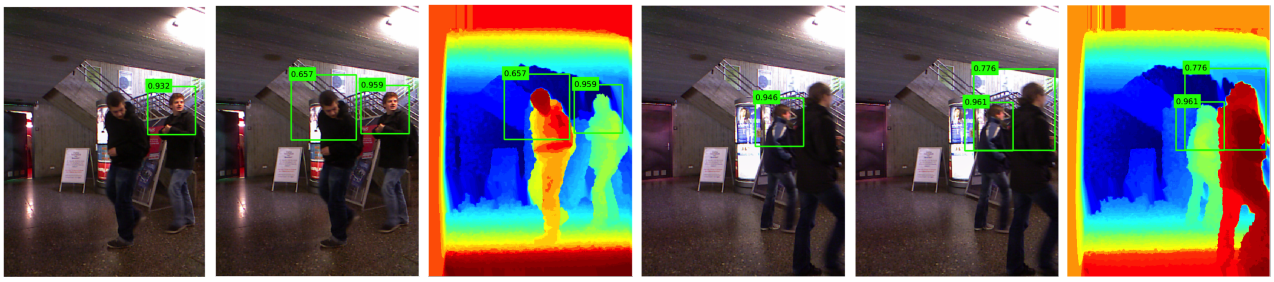
Depth encoding - Overall, the performances of the depth encoding schemes, used with the same pre-trained CNN, can be sorted in the decreasing order as: CECD $>$ CD $>$ DG $>$ CE for a targeted average recall of 0.5. These results show that the CECD-encoding is better at producing new depth images close in characteristics to color images. In other words, by using the CECD-encoding, the fine-tuning of the Depth-CNN requires less effort in adjusting the pre-trained weights.

Combined RGB and depth detection - The rgbCECDcnn outperforms the two baselines, showing that by adding depth the detection accuracy can be improved. Two observations may explain this result: (i) *Depth-CNN is more robust to pose deformations.* The front person in Fig. 5a is a missing detection when using only RGB-CNN. This might be caused by the slightly unusual body pose. By combining with Depth-CNN, this person can be detected (Fig. 5b), as the shape of the person is still a discriminating feature in the encoded depth image (Fig. 5c). (ii) *Depth detections can compensate the absence of RGB detections when color information is not discriminative enough.* When using only RGB-CNN, the front person in Fig. 5d is missed. This is probably because the appearance of the person is somewhat blurred and the skin information is insufficient. Fortunately, in the (encoded) depth domain (Fig. 5f), the strong shape and contrast information can compensate these deficiencies.

Depth based ROI selection - Fig. 6 shows that for

[3]The completed annotation set is available on https://github.com/KaiyangZhou/new-annotations-rgbdpeople

[4]The best fine-tuning for the four Depth-CNNs are: (1) DG-CNN: fc678, epoch=1; (2) CE-CNN: conv345fc678, epoch=3; (3) CD-CNN: conv345fc678, epoch=3; (4) CECD-CNN: fc678, epoch=1.



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

Figure 5: Detection results obtained by rgbcnn (a,d) and rgbCECDcnn (b,c,e,f). It can be seen that the shape/contrast information is well preserved in the CECD depth images (c) and (f).

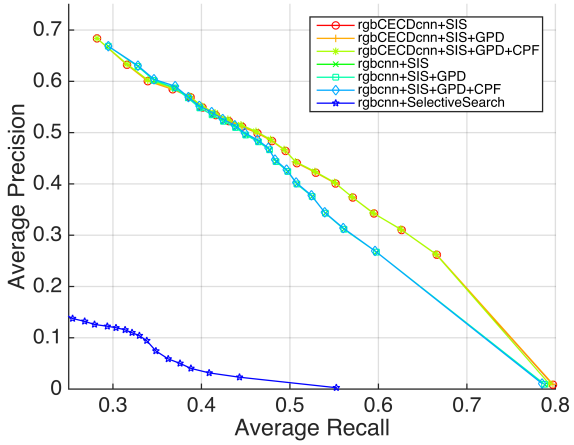


Figure 6: Evaluation of the ROI selection method.

each CNN model, the results of using SIS, SIS+GPD and SIS+GPD+CPF nearly overlap, which strongly suggests that the speed improvement gained by our ROI method does not significantly affect our detection accuracy. The selective search result exposes a severe low-precision problem, with many false positives or repeated detections. Our ROI method did not suffer from these problems, partly because SIS prevents producing large windows on far backgrounds and small windows on close foregrounds. As for speed, although our ROI method produces more proposals (5000 vs. 1500 with tiny proposals being pruned as mentioned), the time spent is two orders of magnitude shorter than that of selective search (14.5ms vs. 2700ms).

4 Conclusion and Future Work

With the infiltration of RGB-D cameras on the one hand and the potential for fast and powerful CNN-based solutions on the other in the vision community, solutions to the problem of people detection in RGB-D data with CNNs are in demand. In this paper, we proposed (1) an effective ROI selection method based purely on depth, (2) a depth-encoding method and (3) a two-stream CNNs framework for people detection, including a novel color-depth fusion approach. We demonstrated that by combining color and depth detections our models outperformed the RGB baselines. In the future, it would be interesting to learn the fusion function parameters from data, which will allow us to perform end-to-end training on the two networks.

References

[1] M. Camplani et al., "Multiple human tracking in rgb-d data: A survey," *IET Computer Vision*, 2016.

[2] O. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *ICRA*, 2014.

[3] H. Zhang, C. Reardon, and L. E. Parker, "Real-time multiple human perception with color-depth cameras on a mobile robot," *IEEE Trans. on Cybernetics*, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[5] A. Angelova et al., "Real-time pedestrian detection with deep network cascades," in *BMVC*, 2015.

[6] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[7] E. Martinson and V. Yalla, "Real-time human detection for robots using CNN with a feature-based layered pre-filter," in *IEEE (RO-MAN)*, 2016.

[8] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *IROS*, 2016.

[9] A. Eitel et al., "Multimodal deep learning for robust RGB-D object recognition," in *IROS*, 2015.

[10] B. Crabbe et al., "Skeleton-free body pose estimation from depth images for movement analysis," in *ICCVW*, 2015.

[11] M. Yang and W. Förstner, "Plane detection in point cloud data," in *MCG*, vol. 1, 2010.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[13] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014.

[14] L. Tao et al., "A comparative home activity monitoring study using visual and inertial sensors," in *HealthCom*, 2015.

[15] N. Silberman et al., "Indoor segmentation and support inference from RGBD images," in *ECCV*, 2012.

[16] L. Spinello and K. Arras, "People detection in RGB-D data," in *IROS*, 2011.

[17] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[18] J. Uijlings et al., "Selective search for object recognition," *IJCV*, 2013.