LEARNING-BASED HUMAN DETECTION APPLIED TO RGB-D IMAGES

Patrisia Sherryl Santoso
Electronics Dept, Electrical and Computer Engr. College
National Chiao Tung University, Hsinchu, Taiwan
Email:Patrisia0593@gmail.com

Hsueh-Ming Hang
Electronics Dept, Electrical and Computer Engr. College
National Chiao Tung University, Hsinchu, Taiwan
Email: hmhang@mail.nctu.edu.tw

Abstract - **Accurate human detection is still a challenging topic due to complicated environments in the real world. In addition, the RGB-D cameras are becoming popular at reasonable price, such as Microsoft Kinect sensor, which provides both RGB and depth data. The depth information often helpful for detection. We adopt the R-CNN method in this paper, which combines the Selective Search technique to generate region proposals and the CNNs (Convolutional Neural Networks) to learn features. A depth map encoding technique (HHA) is adopted to match the CNNs format for learning features. The HHA and RGB images are our inputs. We propose several algorithms to combine their information in constructing various human detectors. Our information fusion structures include CNN, SVM together with PCA for features reduction. More accurate human detection results are shown with the aid of depth information.**

Keywords – Human Detection, CNNs, depth map, HHA depth encoding, RGB-D fusion

I. INTRODUCTION

Human detection is an important task in automotive safety, intelligent robotics and video surveillance. However, human detection faces challenges due to variation of human clothes and pose, occlusion, illumination changes and unique backgrounds in the real-world environment.

Recently, the depth sensors are becoming popular at reasonable price. Indeed, camera and range finders are preferred sensors for this detection task. For example, one of the prominent sensors is the Microsoft Kinect sensor, which provides the RGB and depth data. Recent studies show that the detection performance can be improved using RGB-D data captured by Kinect sensors.

In human detection, one key element is to identify the robust features. Well-known features such as HOG [1], LBP [2], DPM [3], etc are proposed by researchers. However, these features are hand-craft, pre-defined features. They may not match the target features of final identification. Typical algorithms use only low and mid-level information from the image and neglect the higher semantic information. However, this higher semantic information is critical in producing robust features in human detection. Recently, numerous researchers use the deep learning methods such as Convolutional Neural Networks (CNNs) to extract robust features. It is widely used because it can automatically learn and generate goal-matching features from the given input images.

Except for extracting robust features, another key point for human detection is localization. Localization is placing appropriate-size windows at correct locations, which contain one or more humans. The sliding window is a traditional method that finds potential location in an image. However, sliding window is computationally expensive since it searches for every possible location at many different scales inside an image. Here, instead of using a sliding window, a method called region proposals such as Selective Search [4] is used to generate a number of good quality regions (boxes) based on a segmentation technique.

Our algorithm in this paper for people detection contain following elements.

- We adopt R-CNN method [5] and add the depth information together to design a human detector.
- We adopted the HHA encoding method [6] to convert the depth map to match the CNNs format and use it as selective search input to generate the robust region proposals.
- We use CNNs as a feature extractor to extract features from the RGB data and the HHA data.
- We use the SVM machine to classify regions
- We use PCA to reduce feature vector dimension to reduce redundant features and speed up the SVM training time.

II. R-CNN FOR HUMAN DETECTION

In this section, we will describe how we apply the R-CNN approach to do human detection. It has five processes: region proposals generation, region proposals labelling, region proposals filtering, features extraction, and regions classification.

A. Region proposals generation

We adopt the selective search [4] method to generate region proposals on the EPFL dataset [7]. It generates proposals by taking RGB as input. Then applies oversegmentation [8] to create superpixel regions. Next, it groups the superpixel regions hierarchically and stops until it becomes one whole image.

B. Region Proposal labelling

In order to train CNN, we need to provide labels for each proposal. We label all region proposals as positive candidates, if a region overlaps with a ground-truth box with a ratio, OR (Overlap Ratio), $\geqslant 0.5$. The rest are labelled as negatives.

C. Region Proposal filtering

Due to many variant sizes of region proposals generated by the selective search [4], we like to eliminate some unlikely candidates. Indeed, some of them are too large and some too small. Hence, we filter (remove) regions to avoid high false positive probability. Figure 1 shows the region proposals result after this filtering process.

D. Features extraction

Before features extraction, we fine-tune the CNNs (CaffeNet) weights using 100 images (9,229 region proposals). Since the CaffeNet is already trained by 1,2 million of ImageNet Database, so we only need to fine-tune layer 6 to 8 and changed the output of layer 8 from 1000 to two classes (Human and Non-human).

Figure 1 Region proposals after filtering process

The output of "Human" is a probability denoted as P_1. Similarly, the output of "Non-human" is denoted as P_0. The final decision is follows:

$$\text{If } (P_1 > P_0 + \delta)\text{, Input is "Human"}$$
$$\text{Otherwise, Input is "Non-human"} \quad (1)$$

Where δ is a control parameter. Unless particularly stated, δ is typically 0. However, to achieve a higher or lower detection rate, we may need to set δ to non-zero values. For SVM, in the training process, we assign weights to the two classes, W_0 for "Non-human" and W_1 for the "Human". Normally, we assign equal (=1) to both data sets. However, to increase or decrease DR, we assign different weights to two classes.

We extract features using the fine-tuned CNNs. Then, we retrieve feature vectors from layer 7, whose dimension is 4096 D for each proposal.

E. *Region Classification*

We classify regions using two schemes: CNNs and SVM machine. For CNNs, we directly use layer 8 as described in Section 2.4 (eq. 1). For SVM, using the extracted features and labels we train SVM. Then, the trained SVM is used to classify the test data. In this study, our SVM machine comes from LIBSVM [9]. We use C-SVC and the Radial basis function as the SVM kernel.

III. R-CNN ON DEPTH MAP

From Figure 1 we can see that the region proposal results show that many boxes are unreliable. For example, one box contains several persons and some boxes contain only a part of the human body. Therefore, in this section we use the depth maps to improve results further.

We convert the depth map into HHA [6], so that it matches the input format of the CNNs scheme, which was originally designed for RGB.

We use the depth HHA data as the selective search input to generate region proposals. Then, the same training processes in Section 2 are applied. Figure 2 shows an example of region proposals generated using the depth map HHA data as the selective search input. As seen from Figure 2, the depth map HHA data produce more reliable (correct) region proposals. That is, the results are more correct to detect human.

IV. LEARNING FEATURES FROM RGBD IMAGES

We combine both RGB and depth HHA information to produce the final decision. There are several ways to fuse the information from these two sources. (1) CNN classifier: we use CNN to process each data type and then examine their output probabilities and make the final decision. (2) SVM classifier: We combine the features from both of them in the process of SVM decision. At the input stage, either RGB or depth HHA data can be used to generate region proposals. Together, there are a number of variations (schemes).

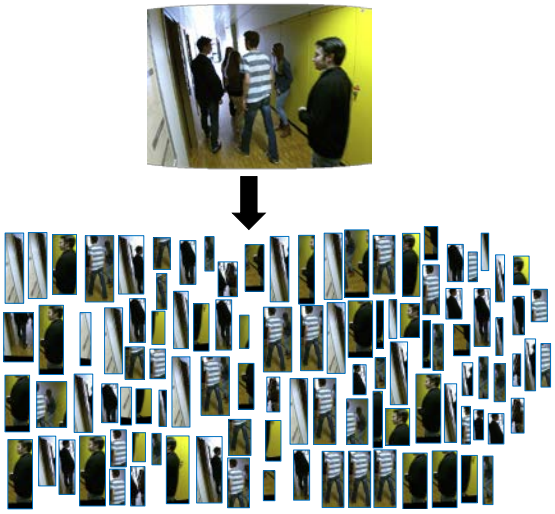

Figure 2 Region Proposals generated by using depth HHA

A. *Outputs fusion with CNNs classifier*

Figure 3 illustrates how we combine the outputs (probabilities) of each CNN classifier and make the final decision. After passing each region proposal candidate through the CNN, we first obtain a probability for each proposal whether it is a human or non-human. Next, we use the probability vector from either or both RGB and HHA data to make decision. For example, one region (bounding box) at the RGB input has a probability 0.5 for human (and non-human). In parallel, apply the same bounding box to the HHA image and produce an output probability 0.65 for human and 0.35 for non-human. Adding the probability vector values together, we obtain a score of 1.15 for human and 0.85 for non-human. The final classifier decides that this test region to be human because the final score (probability) of human is higher than the non-human score. In this example, the initial bounding box (region proposal) may come from either RGB data or depth data, and the results are quite different.

B. *Features fusion with SVM classifier*

Figure 4 shows how we combine the RGB and the HHA features to train the SVM classifier. Similar to the previous section, we first apply the bounding boxes on RGB data and the depth HHA data separately. Next, we extract features trained by CNN based on all region proposals (and the ground truth). Then, using these feature vectors to train the SVM classifier. We thus have a feature set derived from the RGB data and a separate feature set derived from the depth HHA data.

In order to reduce the dimension of feature vectors, we apply PCA (Principal Component Analysis) to the feature vectors. We design two different schemes, pre-PCA and post-PCA.

For pre-PCA, PCA is applied to the feature extraction of RGB and HHA, separately. Next, we combine the reduced-features of RGB and HHA data together to train SVM. Each original CNN produces a 4096D feature vector. After applying PCA to the original feature vectors, each feature vector is reduce to 2048D. Next we concatenate the reduced-features and produce the 4096D vectors of the fused information (RGB and HHA).

For the post-PCA system, PCA is applied to RGB data and HHA features together.

We first mix the features of RGB and HHA together to produce an 8092D feature vector. We apply PCA to this huge the 8092D feature vector and reduce it to 4096D. And use the mixed 4096D vectors to train SVM.

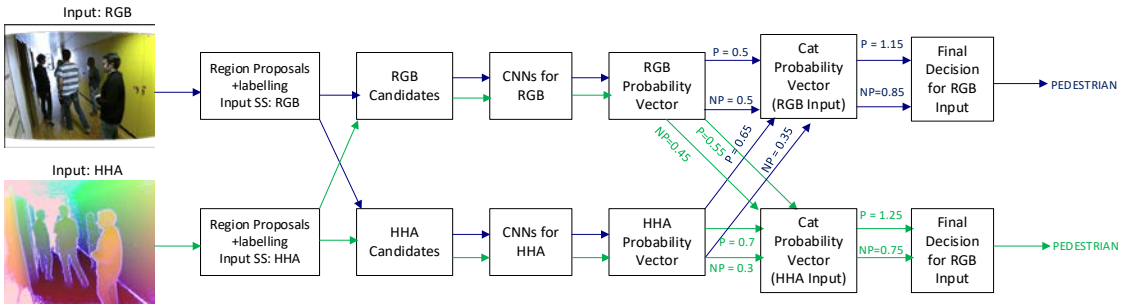

Figure 3 Diagram of Fusion schemes for CNNs Classifier

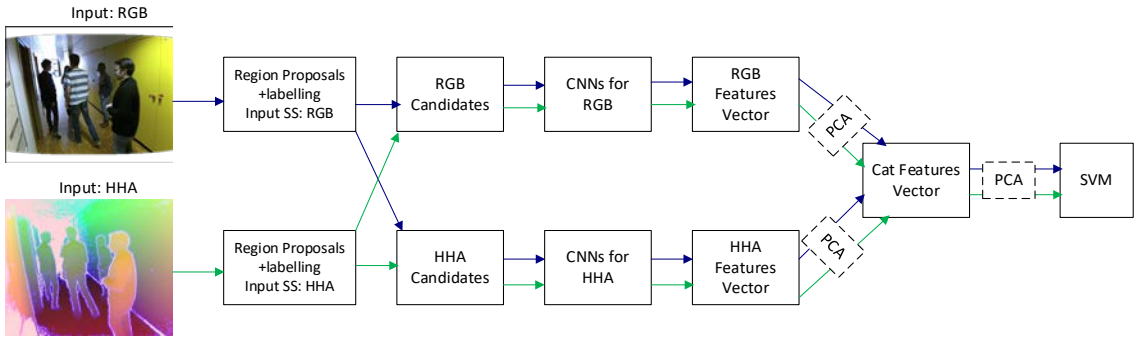

Figure 4 Diagram of feature fusion for the SVM classifier

V. EXPERIMENTAL RESULTS

Our CNNs experiments are performed on NVIDIA GPU GeForce GTX Titan X. We use random 10 Images that different from training sets to test our method. The number of region proposals from this test sets are 866 regions that generated by RGB and 702 regions by HHA. The accuracy of our system is measured using two indicators: Human Detection Rate (DR) and False Positive Per Frame (FPPF) [10]. Each prediction categorized as positive if OR>0.5 and negative otherwise.

A. Results on CNNs classifier

Table 1 shows the accuracy indices, of which the bounding boxes generated using RGB data; and Table 2 is using HHA boxes. The HHA region proposals (boxes) produce better result than the RGB region proposals. The RGB column shows the results of using RGB data only, and HHA column shows the HHA data only. The last column is the fused (combined probabilities) results. The fused results are not better in this case.

Table 1 Accuracy results using CNNs classifier with RGB region proposals

RGB	RGB	HHA	Fused
FPPF	28.69	23.27	26.35
DR	69.92	63.95	68.06

Table 2 Accuracy results using CNNs classifier with HHA region proposals

HHA	RGB	HHA	Fused
FPPF	20.19	18.37	20.06
DR	86.79	86.35	85.21

In order to choose the best scheme, we change the parameter values in the process to make the comparison more clearly. That is, we alter the δ value in eq. (1) as shown in Table 3.

Table 3 DR and FPPF pairs based on multiple δ (CNN)

δ	CNN HHA-RGB		CNN HHA-HHA	
	DR	FPPF	DR	FPPF
-0.9	82.54	17.26	75.67	14.32
-0.7	84.22	18.37	79.76	16.22
-0.5	85.40	18.99	82.67	17.31
-0.3	85.58	19.76	84.09	18.17
-0.1	86.48	20.10	85.66	18.27
0	86.70	20.10	86.3	18.3
0.1	87.22	20.32	87.07	18.56
0.3	88.00	20.98	87.94	19.24
0.5	88.00	21.81	89.03	19.66
0.7	88.63	22.99	90.12	20.78
0.9	89.75	24.97	91.78	22.19

Figure 5 shows a plot figure DR/FPPF using multiple δ as shown in Table 3. From figure, the best result is using HHA to generate the region proposal and applying them to the HHA data.

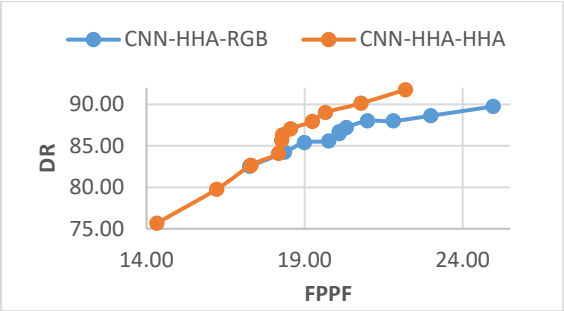

Figure 5 Comparison of CNN-HHA best schemes with different δ

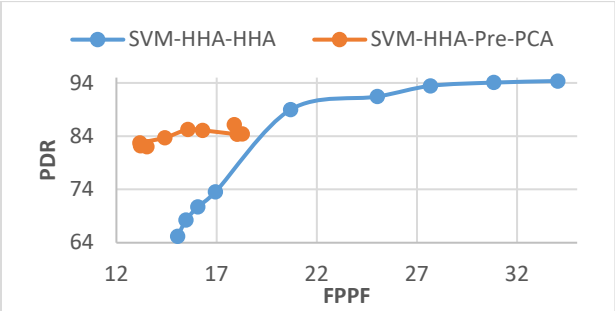

Figure 6 Comparison of SVM-HHA best schemes with adjusted weights

B. *Results* on SVM classifier

Table 4 shows the accuracy results using SVM of which bounding boxes are generated by the RGB data; and Table 5 are generated by HHA data. HHA boxes produce better results because it generates regions that are more robust those generated based on RGB data (as described in Section 3). Table 6 shows a running time on SVM machine and using PCA can make detection run faster.

Table 4 Accuracy results using SVM classifier with RGB region proposals

RGB	RGB	HHA	Fused	Pre-PCA	Post-PCA
FPPF	25.71	28.19	26.90	26.00	22.85
DR	64.26	70.42	67.56	47.75	59.10

Table 5 Accuracy results using SVM classifier with HHA region proposals

HHA	RGB	HHA	Fused	Pre-PCA	Post-PCA
FPPF	20.69	20.77	20.28	15.56	17.14
DR	87.29	89.06	85.55	85.27	84.23

In order to choose the best scheme, we vary the weights in SVM training. W_0 is attached to non-human set, and W_1 is attached to the human set. The results of two more promising configurations are shown in Table 6.

Table 6 DR and FPPF pairs based on multiple weight on SVM

W_0	W_1	SVM HHA-HHA		SVM HHA-PrePCA	
		DR	FPPF	DR	FPPF
9	1	65.19	15.06	82.02	13.52
7	1	68.26	15.47	82.21	13.20
5	1	70.71	16.06	82.73	13.17
3	1	73.54	16.95	83.72	14.42
1	1	89.00	20.70	85.27	15.56
1	3	91.44	25.02	85.11	16.31
1	5	93.44	27.67	84.35	18.03
1	7	94.07	30.84	84.40	18.29

Figure 6 shows a plot figure DR/FPPF using different W values (Table 6). From this figure, the SVM on HHA scheme (HHA box) has a larger DR, but the SVM on the combined Pre-PCA (HHA box) has a smaller FPPF. For example, if we pick up a similar FPPF value at 15.5%, the DR for the SVM-HHA-HHA is approximately 70%, while SVM-HHA-Pre-PCA is 85%. However, for some reason, the DR value of SVM-HHA-Pre-PCA saturates at round 85% in our experiment, and cannot go higher.

VI. CONCLUSIONS

In this paper, we investigated the application of the R-CNN method to human detection. We use it to extract robust features for the more accurate detector design. Furthermore, we use RGB-D data as inputs. Thus, the HHA depth encoding technique is adopted to convert the depth map into the HHA format enables to match the CNN format. Experiments show that the depth HHA data generate more reliable region proposals using the selective search algorithm. Next, we use HHA proposals (boxes) on the RGB and HHA images; then extract the image features using CNNs. To make the final human or non-human decision, we can examine the output probability of individual data or their combined probability values. In the CNN classifier configurations, the RGB-data only seem to produce the best result.

We can use the CNN derived features to train the SVM machine and then, the human vs non-human decision is made by SVM. We also use PCA to reduce feature space (dimension). After extensive simulation and comparison, we conclude that the SVM classifier with pre-PCA leads to a higher detection rate at low False Positive Per Frame (FPPF) probabilities.

VII. ACKNOWLEDGEMENT

This work was supported in part by the MOST, Taiwan under Grant MOST 104-2221-E-009 -069 -MY3 and by the Aim for the Top University Project of National Chiao Tung University, Taiwan.

VIII. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," CVPR, pp. 886-893, 2005.

[2] Y. Mu, S. Yan., Y. Liu, T. S. Huang and B. Zhou, "Discriminative local binary patterns for human detection in personal album," *CVPR,* pp. 1-8, 2008.

[3] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach.,* vol. 32, no. 9, pp. 1627-1645, 2010.

[4] J. R. Uijilings, K. van de Sande, T. Gever and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision,* vol. 104, no. 2, pp. 154-171, 2013.

[5] R. B. Girshick, J. Donahue, T. Darrel and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR,* pp. 580-587, 2014.

[6] S. Gupta, R. Girshick, R. Arbelaez and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," *ECCV,* pp. 345-360, 2014.

[7] T. Bagautdinov, F. Fleuret and P. Fua, "Probability Occupancy Maps for Occluded Depth Images," *CVPR,* 2015.

[8] P. F. Felzenswalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *IJCV,* vol. 59, no. 1,3,4,5,7, pp. 167-181, 2004.

[9] C. C. Chang and C. J. Lin., "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 27:1--27:27, 2011.

[10] Z. Wang, S. Yoon, S. J. Xie, Y. Lu and D. S. Park, "A high accuracy pedestrian detection system combining a cascade AdaBoost detector and random vector functional-link net," The Scientific World Journal, no. 105089. doi:10.1155/2014/105089, 2014.

[11] E. Shelhamer, J. Donahue, J. Long, Y. Jia and R. Girshick, "DIY Deep Learning for Vision: a Hands-on Tutorial with Caffe [Powerpoint slides]," 2014. [Online]. Available: https://docs.google.com/presentation/d/1UeKXVgRvvxg9OUdh_UiC5G71UMscNPlvArsWER41PsU/preview?slide=id.g67c2fe267_379_19. [Accessed June 2016].

[12] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.

[13] A. Babenko, A. Slesarev, A. Chigorin and V. S. Lempitsky, "Neural codes for image retrieval," *ECCV,* 2014.

[14] S. Gupta, P. Arbelaez and J. Malik, "Perceptual organization and recognition of indoor scenes," *CVPR,* pp. 564-571, 2013.

[15] Y. Jia, et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," *ACM int'l conf. on Multimedia*, Nov. 2014.