MULTI-GLIMPSE LSTM
WITH COLOR-DEPTH FEATURE FUSION FOR HUMAN DETECTION

Hengduo Li [1], *Jun Liu* [1]*, *Guyue Zhang* [1], *Yuan Gao* [1], *Yirui Wu* [2]

[1] Fudan University, Shanghai, China
[2] Hohai University, Nanjing, China
{lihd14, ljun, guyuezhang13, gaoyuan14}@fudan.edu.cn, wuyirui@hhu.edu.cn

ABSTRACT

With the development of depth cameras such as Kinect and Intel Realsense, RGB-D based human detection receives continuous research attention due to its usage in a variety of applications. In this paper, we propose a new Multi-Glimpse LSTM (MG-LSTM) network, in which multi-scale contextual information is sequentially integrated to promote the human detection performance. Furthermore, we propose a feature fusion strategy based on our MG-LSTM network to better incorporate the RGB and depth information. To the best of our knowledge, this is the first attempt to utilize LSTM structure for RGB-D based human detection. Our method achieves superior performance on two publicly available datasets.

Index Terms— Human detection, RGB-D, LSTM, Feature fusion

1. INTRODUCTION

Human detection has been a hot research area due to its wide usage in video surveillance, self-driving vehicles, human-machine interaction, and robotics. With the development of depth cameras like Kinect and Intel Realsense, various vision-based applications are boosted with the depth information acquired by the devices which are more robust against illumination and texture variations. Among these applications, RGB-D based human detection receives continuous research attention recently.

Recent years have seen a considerable amount of work to solve the RGB-D based human detection problem. Spinello and Arras [1] takes inspiration from Histogram of Oriented Gradients (HOG) [2] and proposes Histogram of Oriented Depths (HOD) to detect people in dense depth data. A reversible jump Markov chain Monte Carlo (RJ-MCMC) particle filtering method was proposed for human detection and tracking on both fixed and moving color-depth cameras [3]. Bagautdinov et al. [4] proposed a generative model to compute the probabilities of presence of potentially occluding pedestrians from a single depth map provided by RGB-D sensors.

Neural networks have shown their strong capability in a variety of fields, such as object recognition [5][6][7][8], activity recognition [9][10][11][12], semantic segmentation [13][14], and RGB based human detection [15]. Very recently, Xue et al. [16] also explored to apply neural networks for RGB-D based human detection and tracking. A deep CNN was used in their method to identify generated proposals. However, they did not consider the utilization of multi-scale multi-part contextual color-depth information, which is often important for reliable human detection. Based on an observation, when detecting and identifying a target, humans tend to catch a wide-range glimpse to get overview knowledge first, then shrink the area of focus gradually until eyes focus exactly on discriminative parts of the target. Owing to the effective utilization of the contextual information among the multiple glimpses, negative factors like occlusion could be more easily handled by human. Therefore, we propose to explicitly utilize the contextual multi-scale multi-part information for RGB-D based human detection.

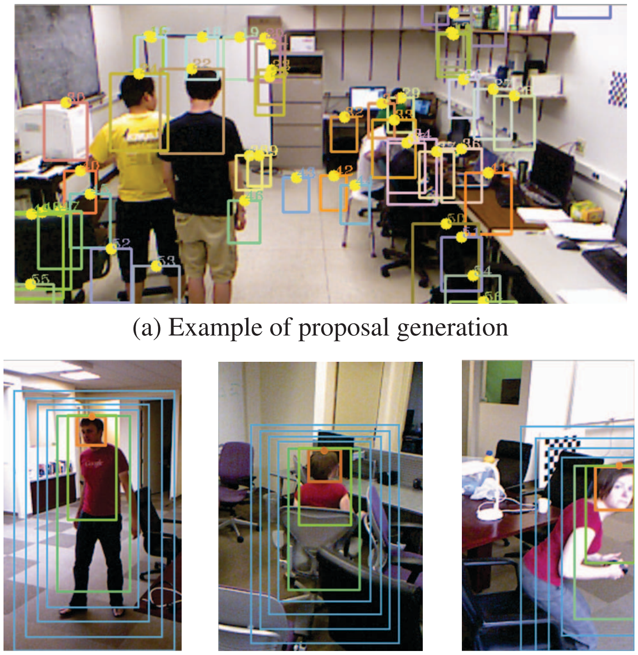
Long short-term memory(LSTM) is a powerful neural network structure which is able to model the dynamics and contextual dependences in sequential information. Consequently, in this paper, we propose a Multi-Glimpse LSTM (MG-LSTM) to model the contextual multi-scale color depth information. Besides, to effectively fuse the two modalities of RGB and depth, we further propose a novel fusion strategy for LSTM. In our fusion framework, two bypass LSTM chains take the multi-scale color depth information respectively and fuse it at the main LSTM chain to generate the prediction. This structure can fuse the data flow of RGB and depth more effectively whilst allows the two bypass chains to remain relatively independent. To the best of our knowledge, this is the first attempt to utilize LSTM network for RGB-D based human detection.

The main contributions of this paper are as follows:

(1) We propose a long-short term memory classification model, Multi-Glimpse LSTM (MG-LSTM), for RGB-D human detection. This model takes context features of multi-scale RGB-D data corresponding to each pre-generated proposal as a sequence of input for classifica-

∗ Corresponding author

978-1-5090-2175-8/17/$31.00 ©2017 IEEE



(a) Example of proposal generation

(b) (c) (d)

Fig. 1: **Examples of proposal generation (a) and multi-scale images clipping (b-d). (displayed in color domain)** Side of clipping windows stops expansion when reaching borders of original image as shown in (d).

tion.

(2) We propose a fusion strategy of LSTM for RGB-D based human detection composed of two bypass LSTM chains receiving extracted features as input and a main prediction-making LSTM chain.

2. MULTI-GLIMPLSE LSTM WITH COLOR-DEPTH FEATURE FUSION

Humans tend to utilize multi-scale visual information contextually when detecting targets. Motivated by this observation, we propose our method which includes three stages: proposal generation, multi-scale multi-part feature extracting and classification. Potential pixels of human head-top are localized as proposals at the first stage. A set of multi-scale RGB-D images of each proposal are clipped and forwarded to pre-trained convolutional neural networks to extract fixed-length feature vectors. Finally, the MG-LSTM network with color-depth feature fusion takes extracted features as input for the binary classification.

2.1. Proposal Generation

For RGB based human detection, various proposal generation methods are adopted including selective search [17], multi-scale combinatorial grouping [18], objectness [19], etc. In RGB-D based human detection, additional depth information offers opportunity for faster and more reliable proposal generation. Recently, Liu et al [20] proposed an ultra-fast proposal generation method, called plausible candidate retriever, for proposal generation in depth image. In this method, every pixel within the depth image is evaluated to judge whether it's a possible location of a human head-top. Plausible head-top pixels are then collected as proposals (candidate human locations). These proposals (candidate human locations) are used for the subsequent human detection procedures.

Compared to other RGB map based approaches, this proposal generation technique proposes much fewer proposals for each image, thus can significantly reduce the time consumption of proposal generation and subsequent processes. According to our experimental results, only around 50 proposals are generated for each image with an extremely small miss rate (0.03). The generated proposals are much fewer than those of Regions with Convolutional Neural Network(R-CNN, around 2,000 region proposals) [15]. Besides, the speed of this process reaches as high as 500 fps. Thus, we use this method for proposal generation. Fig. 1(a) shows a typical result of this method. Readers are referred to [20] for more details of the proposal generation method.

2.2. Multi-Glimplse LSTM

Human uses contextual multi-scale information when detecting and identifying a target. Similarly, our method utilizes clipped multi-scale multi-part color depth images. As shown in Fig. 1(b), for each plausible head-top pixel (candidate human location), a set of multi-scale images are clipped from both color and depth maps. The set of images includes a head-scale, an upperbody-scale, a body-scale and several larger ones. Clipping sizes of head and body are obtained with their real-world sizes, depth value of the plausible head-top pixel and depth camera's intrinsic parameters [20]. Size of nth peripheral image is calculated as: $S_n = S_b * (1+0.3n)$ in which S_b denotes the size of body scale image.

Arranged contextually in large-to-small order, clipped images are further forwarded to pre-trained CNN networks for extracting discriminative features. For color images, the VGG-19 model [6] pre-trained on Imagenet is utilized. For depth images, we use the depth based CNN model trained by Eitel et al. [21], which performs well on object recognition. A sequence of 4,096 dimensional feature vectors is obtained.

LSTM network, with the ability of modeling the contextual dependencies in a sequence of information, is ideal for modeling this sequence of features. Therefore, we propose the Multi-Glimpse LSTM network (MG-LSTM) which takes the sequence of contextual color depth feature vectors as input to make classification. (Fig. 2) The logistic regression classifier completes the binary classification at the last step.

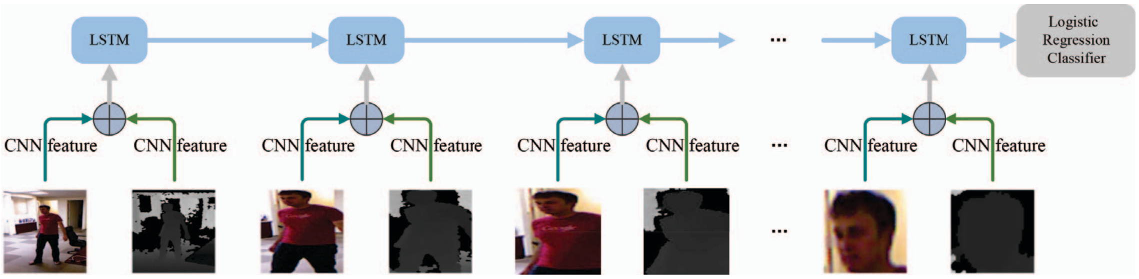


Fig. 2: **Multi-Glimpse LSTM with Color Depth Feature Concatenation.** $\bigoplus$ represents feature concatenation.

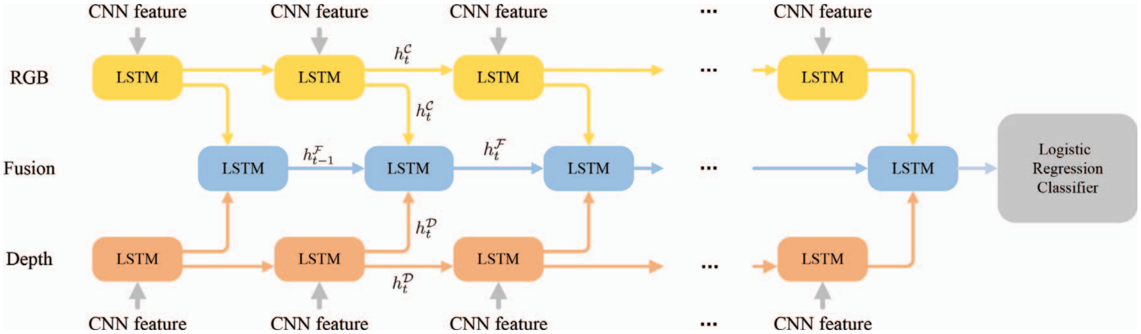


Fig. 3: **Multi-Glimpse LSTM with Color Depth Feature Fusion Strategy.**

The LSTM transition equations are as follows:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ u_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(M \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \right) \tag{1}$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \tag{2}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3}$$

where an input gate i_t, a forget gate f_t, an output gate o_t and formulated input u_t are included. t denotes the glimpse step. σ is the sigmoid function. M is the affine transformation composed by model parameters. x denotes input of concatenated color depth feature vectors. h_t and c_t denote the output state and cell state. $\odot$ indicates element-wise production in the formula.

When modeling the sequence of contextual multi-scale multi-part information, the network gains an overview understanding from the large clipped images firstly, then refinements are added step by step based on the sequential input. Since the multi-scale images within a set are sequential and tightly correlated, a better understanding of the plausible target can be obtained through modeling. Such comprehensive utilization of the contextual information is obviously effective for this binary classification task. Besides, as small-scale images contain human head and upper body which are less deformable whilst large-scale images are more probable to contain intact information, negative impact from occlusion and irregular human poses (Fig. 1(c)(d)) can be reduced.

2.3. Color Depth Feature Fusion

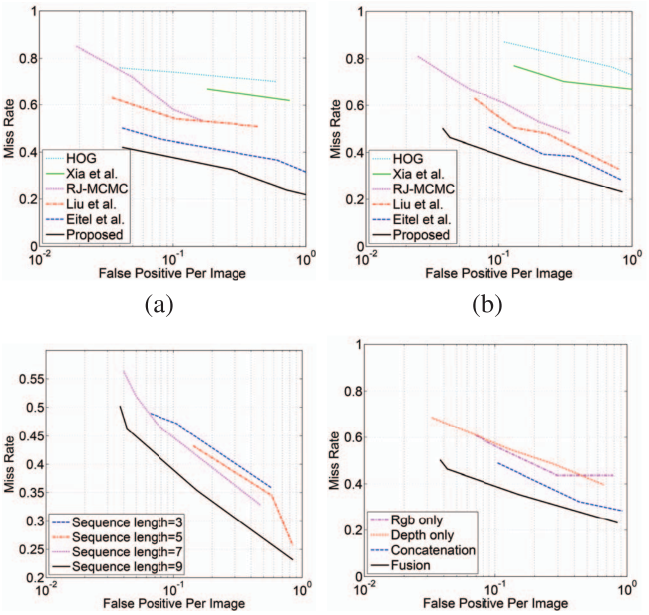
Color and depth are two different while correlating modalities. In Fig. 2, we simply concatenate the two types of CNN features as the input of each step of LSTM. In order to use these two modalities more effectively, in this section, we propose a fusion strategy, in which we merge three chains of LSTM network to achieve color-depth feature fusion. Structure of the fusion network is demonstrated in Fig. 3. Within each step, two bypass LSTM chains take color depth feature vectors respectively and fuse them into the main LSTM chain. The logistic regression classifier connected to the last step of the main chain makes binary classification finally.

In our fusion scheme, the gates of the two bypass chains are formulated in the same way of equation (1-3), yet the gates of the main chain are calculated as follows:

$$\begin{pmatrix} i_t^{\mathcal{F}} \\ f_t^{\mathcal{F}} \\ o_t^{\mathcal{F}} \\ u_t^{\mathcal{F}} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(M^{\mathcal{F}} \begin{pmatrix} h_t^{\mathcal{C}} \\ h_t^{\mathcal{D}} \\ h_{t-1}^{\mathcal{F}} \end{pmatrix} \right) \tag{4}$$

where the subscript $\mathcal{C}$, $\mathcal{D}$, $\mathcal{F}$, denote the Color chain, Depth chain and main Fusion chain respectively.

By applying this color-depth feature fusion strategy, color and depth features can be fused effectively through the data flow compared to simple concatenation. Also, the existence of three chains enables the two modalities remain relatively independent instead of being totally interweaved.



(a)

(b)

(c)

(d)

Fig. 4: **Experimental results.** (a) and (b) show results of different methods on *Office Dataset* and *Mobile Dataset*. (c) and (d) demonstrate the comparison of different glimpse numbers and different fusion methods on *Mobile Dataset*.

3. EXPERIMENT

Datasets. We evaluate the proposed method on two datasets, which were captured with Kinect at 640 × 480 resolution.

(1) *Kinect Office dataset* [3]. This dataset contains 17 video sequences captured in an office. Various human poses such as sitting, standing and walking are included.

(2) *Kinect Mobile dataset* [3]. This dataset contains 18 video sequences collected by a Kinect mounted on a PR2 robot with a horizontal perspective. The robot was moving inside a building, and challenges like illumination variations and complex background are included.

Implementation details. The generated proposals are labeled binarily as human and non-human. The network learns to predict the binary class with the logistic regression classifier, which takes the output state of the last step of MG-LSTM as input. The objective function takes the negative log-likelihood loss to measure the difference between predicted results and true labels. Back propagation through time (BPTT) is used to minimize the objective function. We set learning rate and decay rate as 0.0004 and 0.97 to train our network. One MG-LSTM layer with the neuron size of 256 is used in our network. In our experiment, for Kinect office dataset, we labeled 1,146 positive samples and 22,835 negative samples for training; for Kinect mobile dataset, we labeled 577 / 23,520 for training. The rest of the frames are used for testing. Since the positive-negative sample ratio is rather low in the training set, in each training epoch, we randomly pick a certain number of negative samples out of the entire negative sample set to handle the unbalanced training samples, which makes the positive-negative sample ratio at approximately 1 to 3 within each epoch.

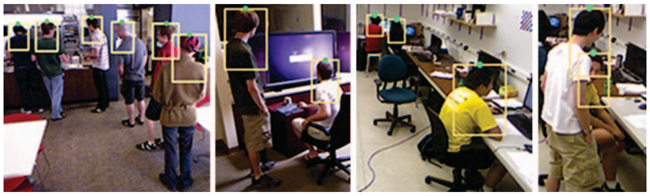


Fig. 5: **Example of detection results.**

Experimental results. We follow the evaluation protocol in [3] and plot the false-positive-per-image (FPPI) vs. missrate curves for evaluation.

To evaluate the performance of our proposed method, we compare it with HOG [2], the depth-based detector proposed by Xia et al. [22], the RJ-MCMC by Choi et al. [3], the Ring-Wedge Mask method proposed by Liu et al. [20] and a deep CNN network proposed by Eitel et al. [21]. Based on the results in Fig. 4(a) and (b), our method obviously achieves the best performance among all compared methods, which indicates the superiority of our MG-LSTM network with color depth feature fusion. Our method is also proved to outperform deep CNN in this task.

Different sequence length (different glimpse numbers) of MG-LSTM are experimented to assess the effectivity of utilizing multi-scale multi-part information. As shown in Fig. 4(c), the network with sequence length 9 outperforms others with shorter sequence length, proving that multi-scale information contributes much to the classification performance.

Besides, in order to evaluate the color-depth feature fusion strategy, we compare our network (MG-LSTM with color depth fusion) with three deep neural networks, i.e. MG-LSTM with color only, MG-LSTM with depth only and MG-LSTM with simple color depth concatenation(see Fig. 2). In Fig. 4(d), our proposed color depth feature fusion strategy clearly shows its effectivity since it exceeds the other three information utilization strategy. Generally, experimental results illustrate marked advantage of our method.

4. CONCLUSION

In this paper, we propose a new Multi-Glimpse LSTM network for RGB-D based human detection. In order to better incorporate the RGB and depth information, we further propose a color-depth feature fusion strategy. The comparative experiments show the superior performance of our method on two RGB-D datasets.

5. REFERENCES

[1] L. Spinello and Kai O Arras, "People detection in rgb-d data," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3838–3843, 2011.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[3] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 1577–1591, 2013.

[4] T. Bagautdinov, F. Fleuret, and P. Fua, "Probability occupancy maps for occluded depth images," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Q. Ke and Y. Li, "Is rotation a nuisance in shape recognition?," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4146–4153, 2014.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[10] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," *ACM SIGMM International Conference on Multimedia Retrieval*, 2016.

[11] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Human interaction prediction using deep temporal features.," *European Conference on Computer Vision. Springer International Publishing*, pp. 403–414, 2016.

[12] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Spatial, structural and temporal feature learning for human interaction prediction," *arXiv preprint*, p. arxiv:1608.05267, 2016.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[14] L.C. Chen, G. Papandreou, I. Kokkinos, and A.L. Yuille K. Murphy, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *International Conference on Learning Representations*, 2015.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[16] H. Xue, Y. Liu, D. Cai, and X. He, "Tracking people in rgbd videos using deep learning and motion clues," *Neurocomputing*, vol. 204, pp. 70–76, 2016.

[17] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, 2013.

[18] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[19] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[20] J. Liu, G. Zhang, Y. Liu, L. Tian, and Y.Q. Chen, "An ultra-fast human detection method for color-depth camera," *J. Vis. Commun. Image R.*, vol. 31, pp. 177–185, 2015.

[21] A. Eitel, J.T. Springenberg, L. Spinello, M.Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[22] L. Xia, C.C. Chen, and J.K. Aggarwal, "Human detection using depth information by kinect," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.