DEEP LEARNING ARCHITECTURE FOR PEDESTRIAN 3-D LOCALIZATION AND TRACKING USING MULTIPLE CAMERAS

Kikyung Kim, Byeongho Heo, Moonsub Byeon, Jin Young Choi *

Perception and Intelligence Lab
Department of Electrical and Computer Engineering, ASRI
Seoul National University, South Korea
{koreaton, bhheo, msbyeon, jychoi}@snu.ac.kr

ABSTRACT

In this paper, we propose a novel deep-learning architecture for accurate 3-D localization and tracking of a pedestrian using multiple cameras. The deep-learning network is composed of two networks: detection network and localization network. The detection network yields the pedestrian detections and the localization network estimates the ground position of a pedestrian within its detection box. In addition, an attentional pass filter is introduced to effectively connect the two networks. Using the detection proposals and their 2-D grounding positions obtained from the two networks, multi-camera multi-target 3-D localization and tracking algorithm is developed through min-cost network flow approach. In the experiments, it is shown that the proposed method improves the performance of 3-D localization and tracking.

Index Terms— 3-D localization, pedestrian detection, multi-camera multi-target tracking, deep-learning

1. INTRODUCTION

Tracking-by-detection framework in multi-camera surveillance has been sucessfully achieved in recent years. In a tracking-by-detection framework, pedestrian detection and multi-camera multi-target tracking (MCMTT) are combined to find and track pedestrians. In MCMTT, 3-D localization is essential to associate detections, because most MCMTT methods [1, 2, 3, 4, 5, 6, 7, 8] try to solve the problem in 3-D space. Nevertheless, most studies overlook the error in estimating the 3-D position of an object precisely.

In general, the 3-D position is estimated by the projection of a 2-D pedestrian ground position (2-D PGP) into 3-D space using a camera matrix. In the projection, bottom center points of 2-D detections are usually used as a 2-D PGP estimate. However, this estimate frequently causes 3-D localization errors. Figure 1(a) shows the results of the conventional pedestrian detectors (LDCF [9]). In Figure 1(a), the 2-D PGP varies depending on the pedestrian's posture. This means that it is difficult to get a precise 2-D PGP without considering the posture of the person in the bounding box. In addition, distortion in the image is caused by camera lens refraction or change in angle of view. Figure 1(b) shows that the 2-D PGP can be changed by an image distortion. From the above observations, we can infer that it is necessary to find 2-D PGP considering the appearance in the bounding box.

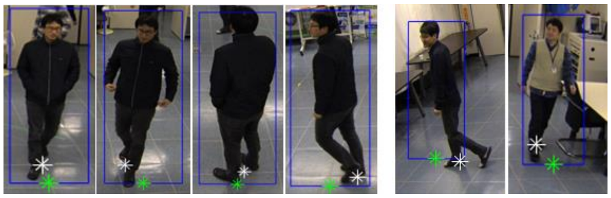
(a) Pose variation (b) Image distortion

Fig. 1. Left : Compared to the actual pedestrian 2-D ground position (white), the one given by the conventional detections (green) varies depending on the posture. Right : Image distortion misleads the estimation of the pedestrian 2-D ground position.

For most MCMTT methods, the tracking input is a set of bounding boxes given by a pedestrian detector. However, the detector may generate a wrong 2-D PGP due to the issues mentioned above. Then most existing MCMTT methods have no chance to correct the 3-D localization error caused by the wrong 2-D PGPs. 3-D localization accuracy has been improved by smoothing the tracking result as a post-processing [1, 2, 4] function. However, the method does not solve a fundamental problem, causing localization errors. To decrease localization error and improve tracking performance, it is required to estimate an accurate 2-D PGP.

Inspired by this investigation, this paper proposes a novel deep-learning network that performs 2-D localization as well as the pedestrian detection to solve the MCMTT problem in 3-D space. As shown in Figure 2, the proposed deep-learning network (DL-net) is designed to be composed of two net-

*This work was supported by the IT R&D program of MOTIE/KEIT. [10041610] and the Brain Korea 21 Plus Project.

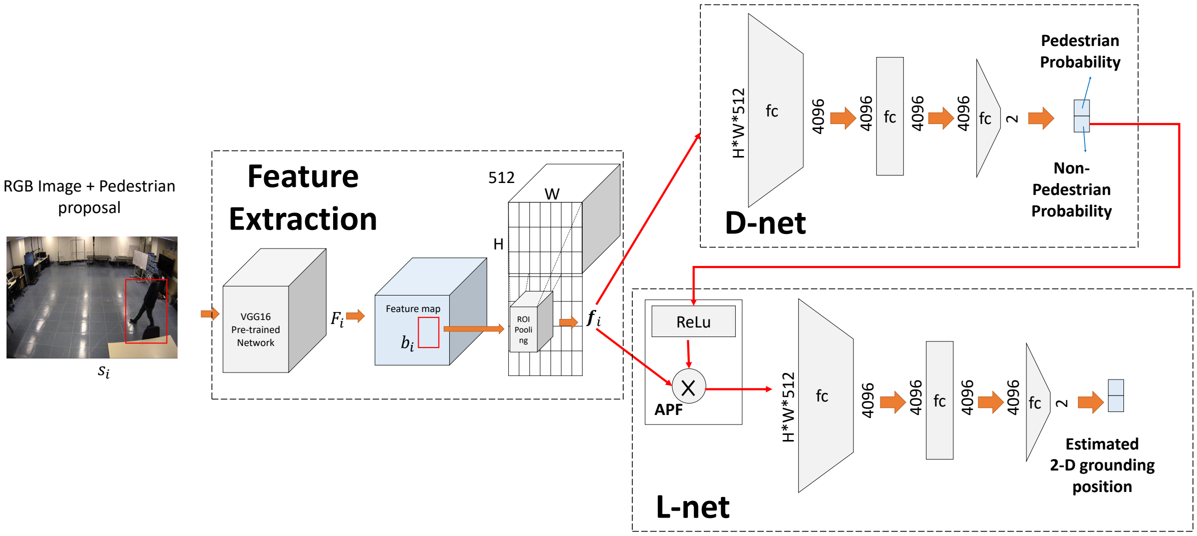

Fig. 2. **Architecture of deep-learning network (DL-net) composed of detection and localization network.** Input image and pedestrian proposals are fed to the input of DL-net. D-net yields detection scores and L-net yields a grounding position from the feature of input proposal. Attentional pass filter (APF) only delivers proposals that are likely to be pedestrians.

works: a detection network (D-net) and a localization network (L-net). D-net yields a detection score of a pedestrian from the proposals given by the object proposal method. L-net estimates the 2-D PGP from the pedestrian's posture in the input bounding box. In addition, an attentional pass filter (APF) is introduced to pass a detection candidate that may be a pedestrian. Connecting D-net to L-net, APF can reduce the load to compute feature of a detection candidate that might not be a pedestrian. From the output of DL-net, the MCMTT problem is developed through the min-cost network flow approach followed by [4]. Accurate 3-D localization and tracking can be achieved from the accurate 2-D PGPs given by the proposed network. The improvement of 3-D localization and tracking is verified through the experiments by showing localization error and public tracking performance measures.

2. NETWORK DESIGN

To detect pedestrians and perform localization at the same time, we propose two loss functions in the proposed network: detection loss function and localization loss function. Detection loss learns whether an input candidate box is a pedestrian or not. Localization loss learns grounding positions which indicates the point where the pedestrian supports the ground. We observe that the box regression loss in previous studies [10] has a limitation to learn a well-localized point. The proposed localization loss can learn an accurate grounding positions from training data. The attentional pass filter (APF) is introuduced to connect detection network (D-net) and localization network (L-net) effectively. APF delivers only the features related with the pedestrian presumed by D-net to L-net. As a result, D-net and L-net are efficiently connected through APF without computational redundancy.

Detection Network (D-net). A training set is defined by $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ...\}$, where $\mathbf{s}_i = \{I_i, c_i, t_i, b_i\}$. I_i refers to an input image, c_i the camera, and t_i the time index. $b_i = (x_i, y_i, w_i, h_i)$ is a detection box with the left top coordinates (x_i, y_i), width w_i and height h_i. A training sample $\mathbf{s}_i$ is fed to VGG-16 network [11] to generate feature map X_i from I_i. To crop specific box region from F_i, ROI-pooling layer takes the input as F_i and b_i. An output of ROI-pooling layer [10] is $\mathbf{f}_i$ which means pooled ROI feature vector from F_i corresponding to b_i. $\mathbf{f}_i$ is fed to D-net to generate a discrete probability p_i which is the output of D-net. p_i is a 2-dimensional vector, where non-pedestrian probability $p_{i,0}$ and pedestrian probability $p_{i,1}$.

Specifically, D-net has the following structure: FC(H × W × 512, 4096)-ReLu-DR-FC(4096, 4096)-DR-ReLu-FC(4096, 2), followed by fast rcnn fully-connected network [10]. FC means a fully connected layer and DR means a drop out layer. Compared with [10], D-net does not perform box regression. Instead of using box regressor, L-net is introduced for an accurate localization.

Localization Network (L-net). Unlike D-net, L-net performs regression task to estimate points indicating 2-D PGP. In the first step of L-net, p_i and $\mathbf{f}_i$ are fed to APF defined by

$$APF(p_i, \mathbf{f}_i) = ReLu(p_{i,1})\mathbf{f}_i. \tag{1}$$

The output of APF is a feature map determined as a pedestrian by D-net. It is inefficient for all inputs to be passed through L-net because most of the detection inputs are not a pedestrian class. APF has a role to prevent low confidence $\mathbf{f}_i$ to be fed to the fully-connected layer. The output of L-net is a 2-dimension vector $\tilde{g}_i = (\tilde{g}_i^x, \tilde{g}_i^y)$. $\tilde{g}_i$ indicates the relative coordinate of a 2-D PGP. And the absolute coordinate g_i is given by $g_i = (\tilde{g}_i^x + x_i, \tilde{g}_i^y + y_i)$. In section 3, we use g_i as an i-th observation of tracking inputs. Compared to box regressor, L-net estimates the 2-D PGP instead of tightening the bounding box. L-net is configured similar with D-net by following structure: APF-FC(H × W × 512, 4096)-ReLu-DR-FC(4096, 4096)-DR-ReLu-FC(4096, 2).

3. MCMTT WITH PROPOSED NETWORK

Track Assignment Formulation. MCMTT in this paper is formulated as an optimization problem to assign the detections to multiple tracks corresponding to pedestrians. Assume that N_c overlapped cameras are set at indoor space. Then N_c images are generated every frame. Let g_i be the 2-D PGP of the detection b_i. Here let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, ...\}$ be the given information for the tracking formulation, where $\mathbf{u}_i = \{b_i, g_i, c_i, t_i\}$, c_i is the camera index and t_i is the time index. With $\mathbf{u}_i$, we define the k-th association set $\mathcal{D}_k$ by

$$\mathcal{D}_k := \{\mathbf{u}_i | \forall \mathbf{u}_i, \mathbf{u}_j \in \mathcal{U}, i \neq j : c_i \neq c_j \wedge t_i = t_j\}, \quad (2)$$

which means a set of 2-D detections at the same time, but in different view, which can be a candidate set of detections from different cameras for a pedestrian. MCMTT is a problem to find a linked set of $\mathcal{D}_k$ associated with each target. In this paper, the tracking problem is formulated by min-cost flow problem similar to [4].

$$\begin{aligned}\mathcal{F}^* = \arg\min_{\mathcal{F}} \sum_k C_k f_k + \sum_k C_{en,k} f_{en,k} \\ + \sum_{k,l} C_{k,l} f_{k,l} + \sum_k C_{ex,k} f_{ex,k}.\end{aligned} \quad (3)$$

Equation (3) implies that the association set $\mathcal{D}_k$ has flow f_k with the cost C_k. $f_{k,l}$ and $C_{k,l}$ are the flow and cost of the temporal edge between $\mathcal{D}_k$ and $\mathcal{D}_l$. $f_{en,k}, C_{en,k}$ and $f_{ex,k}, C_{ex,k}$ are the flow and cost of the source and sink respectively. f is a binary integer value, where $f = 1$ implies that the corresponding edge is a part of the corresponding trajectory and $f = 0$ implies that the edge is not used.

Cost Design. We design new C_k, which means the reconstruction cost of association set $\mathcal{D}_k$.

$$C_k = C(\mathcal{D}_k) = \lambda_{rec} \frac{\sum_{u_i \in \mathcal{D}_k} dist(g_i, G_k; c_i)}{V_k}, \quad (4)$$

where λ_{rec} is a weighting constant. $dist(g_i, G_k; c_i)$ is defined by distance between a 3-D point G_k and the line back-projected from g_i using camera calibration information c_i. V_k is the number of cameras where G_k is visible. G_k is the 3-D vector minimizing the distance between the lines generated from all g_i in $\mathcal{D}_k$. That is,

$$G_k = \arg\min_G \sum_{u_i \in \mathcal{D}_k} dist(g_i, G; c_i). \quad (5)$$

G_k means the virtual 3-D PGP of $\mathcal{D}_k$. $dist(g_i, G_k; c_i)$ is given by

$$\begin{aligned}&dist(g_i, G_k; c_i) = \\ &\frac{||(G_k - \Phi^{c_i}(g_i, z_{min})) \times (\Phi^{c_i}(g_i, z_{max}) - \Phi^{c_i}(g_i, z_{min}))||_2}{||\Phi^{c_i}(g_i, z_{max}) - \Phi^{c_i}(g_i, z_{min})||_2},\end{aligned} \quad (6)$$

where $\Phi^c(g, z)$ is a projection function related to camera c which deliver from the 2-D coordinate g to 3-D coordinate $(\cdot, \cdot, z)$. z_{min} and z_{max} are the constants of the minimum and maximum height of a pedestrian. $dist(g_i, G_k; c_i)$ implies the error between G_k and g_i in 3-D space . The reconstruction error $C(\mathcal{D}_k)$ means the average error of 3-D reconstruction from each cameras. $C(\mathcal{D}_k)$ mainly depends on how accurate g_i is. Most of MCMTT methods, g_i is obtained from b_i by the bottom center of the detection box, i.e.,

$$g_i = (x_i + \frac{w_i}{2}, y_i + h_i). \quad (7)$$

In our works, instead of using Equation (7), L-net yields g_i as

$$g_i = \text{L-Net}(p_i, \mathbf{f}_i). \quad (8)$$

While Equation (7) is a linear mapping from bounding box to 2-D PGP, Equation (8) is non-linear mapping from an image to 2-D PGP. The fact that using image-level feature and non-linear mapping function give a better result than using (7).

Equation (3) is a binary integer programming problem (BIP) like in [4]. We use branch-and-cut procedure to solve Equation (3), which is implemented in the Gurobi optimization library [12].

4. EXPERIMENTS

4.1. Experimental Setting

Dataset. The proposed network was evaluated on PETS 2009 dataset [15], and SNUPIL dataset [2]. In PETS 2009 dataset, we chose S2.L1 scenario for evaluation. The camera calibration information was given using Tsai camera calibration model [16].

Network Training. First we fine-tuned D-net to caltech pedestrian benchmark dataset [17]. We used pre-trained model [11] for feature extraction network. Then we followed the training setting and network parameters to train D-net like in [10]. After that, we trained L-net with detection proposals and corresponding 2-D PGPs in PETS 2009 and SNUPIL. In PETS 2009, we used 6, 8 views for training and 1, 5, 7 views for test. In SNUPIL, we used 4 views for training and 1, 2,

Dataset	Detector	Cameras	MOTA [%]	MOTP [%]	MT	PT	FM	IDS
PETS 2009 S2.L1	HOG-SVM [13]	1,5,7	77.16	41.75	19	4	28	12
	DPM [14]	1,5,7	97.34	59.57	23	0	2	4
	D-Net	1,5,7	90.36	80.13	23	0	2	4
	DL-Net	1,5,7	**97.59**	**80.37**	23	0	**1**	**3**
SNUPIL	DPM [14]	1,2,3	53.48	69.90	4	6	11	18
	D-Net	1,2,3	74.38	78.63	8	2	10	16
	DL-Net	1,2,3	**81.09**	**82.92**	8	2	**4**	**5**

Table 1. Tracking performance evaluation with different detection methods in PETS 2009 and SNUPIL

Method	CAM 1	CAM 5	CAM7
HOG-SVM [13]	20.20	25.28	27.50
DPM [14]	9.92	13.87	14.43
D-net	8.17	11.16	13.69
DL-Net	**7.07**	**10.01**	**11.59**

Table 2. Quantitative results on the SNUPIL dataset. The error shown in the table is an euclidean distance between grounding position and ground truth. We used a pixel unit and 0.25 intersection of union (IOU) constant to determine the boxes corresponding to ground truth.

3 views for test. We used log loss for softmax classification for D-net and L-1 loss for regression task L-net. The detail setting for using multi-task loss is followed by [10].

Ground Truth. We generated about 30,000 samples for training L-net in PETS 2009 and 50,000 in SNUPIL. We generated 3-D ground truth trajectory using the pedestrian locations of each camera views. 3-D ground truth trajectory was estimated by minimizing the reconstruction error of 2-D ground truth based on Equation (6). Ground truth of 2-D PGP is generated every 5 frames in both dataset.

Measure. To evaluate the accuracy of PGP, we used the average of the Euclidean distance to ground truth as a measure. Specifically, we compute the Euclidean distance between the output of L-net and the ground truth. The detection box is determined based on 0.5 threshold intersection of union (IOU) based on the ground truth box. To evaluate tracking performance, we used the widespread CLEAR measures used in [18], called Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP). And like in [19], we also used the metrics of Identity Switches (IDS), Track Fragments (FM), Mostly Tracked (MT), Partialy Tracked (PT). We evaluate tracking performance every 5 frames in both dataset.

4.2. Experimental Results

Localization Accuracy. The quantitative results regarding the accuracy of pedestrian localization are depicted in Table 2. In most cases, DL-net shows the best performance. Additionally, we evaluate D-net, removing L-net from DL-net. D-net is similar to a fast rcnn network [10]. As shown in Table 1, L-net has the positive effect of improving 2-D PGP accuracy by adding it to D-net.

MCMTT Performance. The quantitative evaluation of the tracking performance is depicted in Table 1. The overall tracking performance of the proposed method is better than those associated with other detection results. In PETS 2009, the proposed method shows a much better performance than other algorithms. In MOTP, in particular, representing the accuracy on 3-D localization is greatly improved. This implies that the accurately estimated 2-D PGP has a positive effect on the generation of the accurate position in the 3-D trajectory. As shown in Table 2, the tracking result with DPM shows a high MOTA, while D-net shows a high MOTP. DPM shows a high recall performance in the detection of the pedestrian, and D-net shows a low localization error. In addition, DL-net shows a higher MOTA performance than D-net. This shows that an accurate 2-D PGP can decrease the ambiguity between pedestrians who are close together. Since DPM missed distorted pedestrians, DPM in SNUPIL does not yield a high MOTA as in PETS 2009. However DL-net's 2-D PGPs are robust to image distortion problem. As a result, DL-net shows high MOTA and MOTP in SNUPIL.

5. CONCLUSION

In this paper, we have proposed a deep-learning network that simultaneously detects pedestrians and estimates their ground positions. Compared to previous tracking methods, ours has focused on the importance of pedestrian localization accuracy. Unlike the method of box regression, the proposed method directly gives a 2-D ground position from the detection box to improve localization accuracy. It is meaningful that enhancing the localization accuracy of detections significantly improves the performance of MOTP, representing both tracking and localization in 3-D space.

6. REFERENCES

[1] Moonsub Byeon, Songhwai Oh, Kikyung Kim, Haan-Ju Yoo, and Jin Young Choi, "Efficient spatio-temporal data association using multidimensional assignment in multi-camera multi-target tracking.," in *BMVC*, 2015, pp. 68–1.

[2] Haanju Yoo, Kikyung Kim, Moonsub Byeon, Younghan Jeon, and Jin Young Choi, "Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[3] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua, "Multiple object tracking using k-shortest paths optimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1806–1819, 2011.

[4] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3650–3657.

[5] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[6] Saad M Khan and Mubarak Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 505–519, 2009.

[7] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn, "Branch-and-price global optimization for multi-view multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1987–1994.

[8] H. Possegger, S. Sternig, T. Mauthner, P. M. Roth, and H. Bischof, "Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[9] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[10] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] "Gurobi optimization," `http://www.gurobi.com/downloads/gurobi-optimizer/`.

[13] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[14] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*. IEEE, 2008, pp. 1–8.

[15] Anna Ellis, Ali Shahrokni, and James Michael Ferryman, "Pets2009 and winter-pets 2009 results: A combined evaluation," in *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*. IEEE, 2009, pp. 1–8.

[16] Roger Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[17] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence*, vol. 34, 2012.

[18] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.

[19] Yuan Li, Chang Huang, and Ram Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 2953–2960.