INTEGRATING STEREO VISION WITH A CNN TRACKER FOR A PERSON-FOLLOWING ROBOT

Sahdev R.*, Chen B. X.* and Tsotsos J. K.

Dept. of Electrical Engineering and Computer Science and Center for Vision Research

York University , Toronto, Canada

{sahdev, baoxchen, tsotsos}@cse.yorku.ca

** denotes equal contribution*

YORK UNIVERSITÉ UNIVERSITY · Centre for Vision Research · NCFRN NSERC Canadian Field Robotics Network · RCCRT Réseau canadien CRSNG pour la robotique de terrain · Canada Research Chairs / Chaires de recherche du Canada · Canada

Abstract

A Stereo vision based CNN Tracker for a person following robot is introduced. The robot follows the human in real time in a dynamic environment using a CNN which is trained online from scratch. Our Robot can follow the human under challenging situations - pose changes, illumination changes, appearance changes, wearing/removing a jacket/backpack, exchanging jackets, etc. The robot can follow the target even when the target is transiently not seen in the scene. A stereo dataset is also built.

Introduction

- Person Following robots have many application such as autonomous carts in grocery stores, personal guides in hospitals, or airports for autonomous suitcases.
- A CNN based tracker is proposed [2] which runs at 20fps and is trained online from scratch; the robot follows the target (human) in dynamic environments.
- Our approach generalizes to not only humans but to any object following task. Object needs to be known *a priori.*
- A novel extensive stereo dataset is built for the task of person following robots.
- A Proportional Integral Derivative (PID) Controller is used to follow the target by the robot.

Fig. 1. Robot being used for person following behavior

Approach

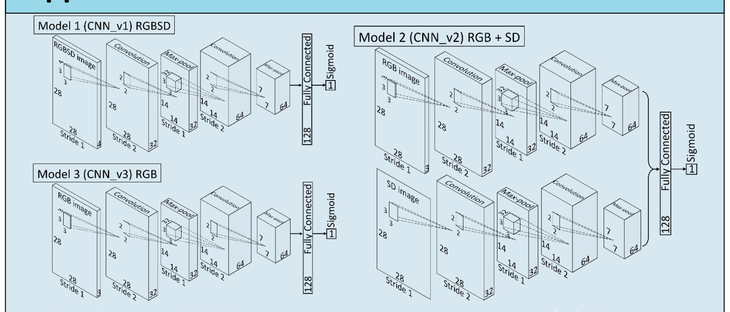

Fig. 2. Three CNN Models: Model 1 takes a 4 channel RGBSD image as input, Model 2 takes a RGB and SD image as input. Model 3 takes only a RGB image.

- 3 different CNN models are proposed to track the given target. (Fig. 2). Each CNN is trained from scratch. The kernels are initialized with random weights.
- Our CNNs do not require any pre-training on any existing dataset.
- *Initial Training Set Selection*: User needs to stand at a pre-specified distance in a bounding box for the training to begin. During the first few seconds, the initial training is done. Human is the positive class and the background patches form the negative class.
- *Test Set Selection:* After initialization, we filter out the subsequent patches based on depth as shown in Fig. 3. Patches with the highest response is the target and form the test set.
- *Update CNN Tracker Step:* The CNN Tracker is updated by adding the most recent positive class patches which form the positive class pool. Patches around the target form negative class.

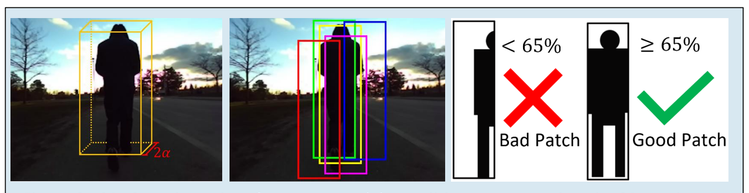

Fig. 3. 3D Search region for test set. (a) candidate test patches in 2D region (based on a sliding window approach), (b) search region with respect to depth, (c) pixels in black are with ± α *meters* from the previous depth. If black pixels are less than 65% of the patch, the patch is bad, else, it is a good patch . The number 65% is chosen experimentally as this covers the human body completely in most of the cases. According to (c), the red and blue patches in (a) are bad patches, the green, pink, and yellow patches are good patches

- Localization and Target pose estimated to compute local path of the target. (Fig. 4a) Local path replicated when robot cannot see the target (Fig. 4b)

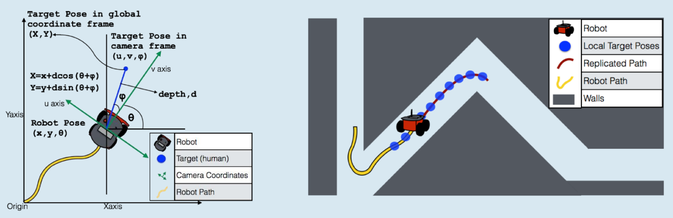

Fig. 4. (a) Target Pose Estimation (b) Target Path Replication

Our Dataset

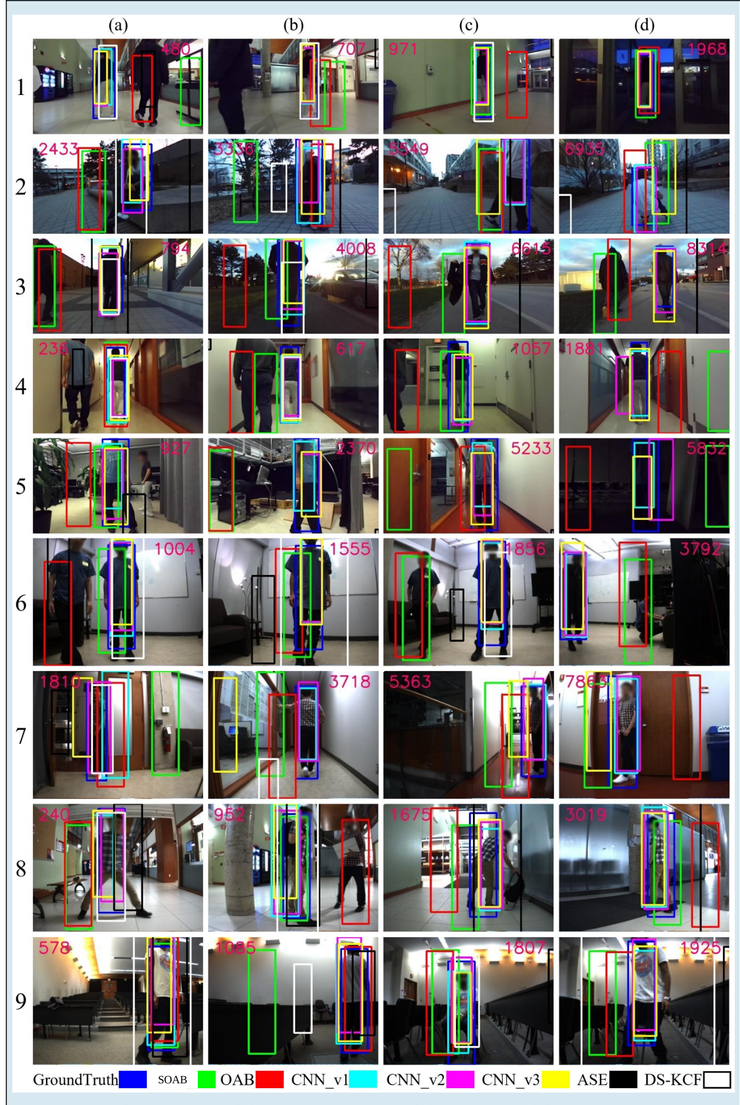

Fig. 5. (1): Hallway 2; (2) Outdoor walking; (3): Sidewalk; (4): Corridor Corners; (5): Lab and Seminar; (6): Same Clothes 1; (7): Corridor Long; (8): Hallway; (9): Lecture Hall (SOAB [1], OAB [5], ASE [3], DS-KCF [4])

- Dataset built at Lassonde Building at York University.
- Our Dataset has 11 different places (as shown in Fig. 5). 9 Indoor and 2 outdoor places
- 11 long sequences exist with each sequence having atleast 2000 frames and up-to 12000 frames
- Dataset built using 2 different stereo cameras Point Grey Bumblebee 2 Stereo Camera and ZED stereo camera.
- Frame rate: 20 fps and captured at a resolution of 640 by 480 using the Bumblebee stereo camera and 672 by 376 using the ZED Stereo camera

Experimental Results

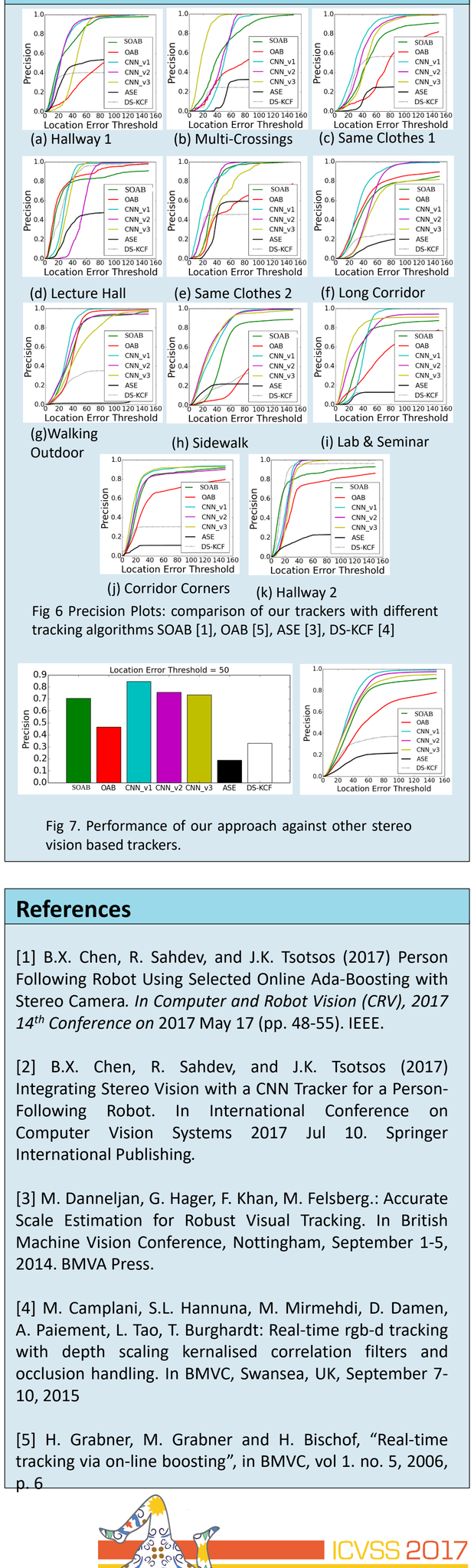

Fig 6 Precision Plots: comparison of our trackers with different tracking algorithms SOAB [1], OAB [5], ASE [3], DS-KCF [4]

Fig 7. Performance of our approach against other stereo vision based trackers.

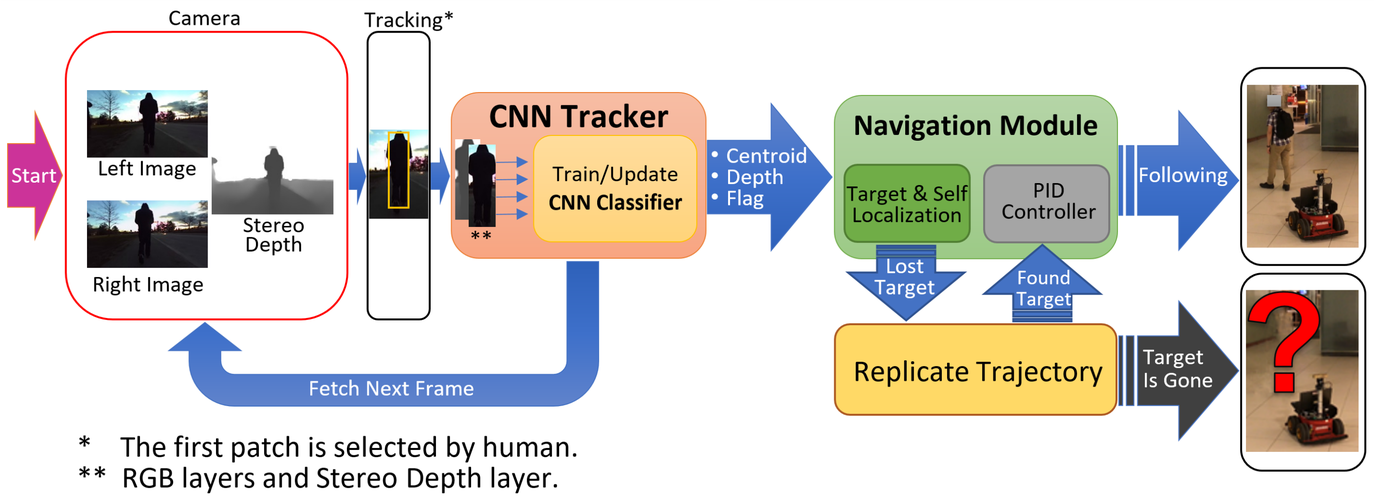

Fig. 8. System design for our approach using Convolutional Neural Network (Tracking Module) and Path Replication/PID controller (Navigation Module)

References

[1] B.X. Chen, R. Sahdev, and J.K. Tsotsos (2017) Person Following Robot Using Selected Online Ada-Boosting with Stereo Camera. *In Computer and Robot Vision (CRV), 2017 14th Conference on* 2017 May 17 (pp. 48-55). IEEE.

[2] B.X. Chen, R. Sahdev, and J.K. Tsotsos (2017) Integrating Stereo Vision with a CNN Tracker for a Person-Following Robot. In International Conference on Computer Vision Systems 2017 Jul 10. Springer International Publishing.

[3] M. Danneljan, G. Hager, F. Khan, M. Felsberg.: Accurate Scale Estimation for Robust Visual Tracking. In British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press.

[4] M. Camplani, S.L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, T. Burghardt: Real-time rgb-d tracking with depth scaling kernalised correlation filters and occlusion handling. In BMVC, Swansea, UK, September 7-10, 2015

[5] H. Grabner, M. Grabner and H. Bischof, "Real-time tracking via on-line boosting", in BMVC, vol 1. no. 5, 2006, p. 6

ICVSS 2017 · Sicily ~ 09-15 July · International Computer Vision Summer School