Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 124 (2017) 167–172

Procedia Computer Science

www.elsevier.com/locate/procedia

4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

Tracking People by Detection Using CNN Features

Dina Chahyati*, Mohamad Ivan Fanany, Aniati Murni Arymurthy

Machine Learning and Computer Vision Laboratory, Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia

Abstract

Multiple people tracking is an important task for surveillance. Recently, tracking by detection methods had emerged as immediate effect of deep learning remarkable achievements in object detection. In this paper, we use Faster-RCNN for detection and compare two methods for object association. The first method is simple Euclidean distance and the second is more complicated Siamese neural network. The experiment result show that simple Euclidean distance gives promising result as object association method, but it depends heavily on the robustness of detection process on individual frames

Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017.

Keywords: Multiple Object Tracking; Faster-RCNN; Convolutional Neural Network; Siamese Neural Network

1. Introduction

Multiple object tracking has been an interesting field of research due to its challenges and importance. It is the main aim of surveillance system and video understanding. Recently, tracking by detection methods had emerged as immediate effect of deep learning remarkable achievements in object detection. Multiple people or pedestrian tracking also takes advantages from this progress, but still remains as hard task since not all challenges are solved. These challenges includes pose, size and shape change as well as occlusions.

Previous research in multiple people tracking has shown promising results by using convolutional neural network (CNN) features. Leal-Taixe, Ferrer and Schindler[1] used Siamese CNN, modified linear programming and gradient boosting classifier on MOT Challenge dataset[2]. Zhu, Porikli and Li[3], used region-based CNN known as Faster RCNN[4] and SSVM as classifier in dataset PETS[5]. Faster-RCNN features are previously best known for object detection, as it was the main approach used by top-rank object detector in Large Scale Visual recognition Challenge[6] (ILSVRC) 2016.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: dina.chahyati@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017
10.1016/j.procs.2017.12.143

The challenge in ILSVRC 2016 was to detect objects that falls into 200 categories. If we were to use tracking by detection method, then CNN features should perform quite well since it has succeeded in detecting various objects.

Our current research is find meaningful information in surveillance videos, such as how many man or woman enter a scene, towards which directions, wearing what cloths and bringing what items. To achieve this goal, we need to track multiple person and their characteristics. For the time being, the characteristics that we explore is their gender. We have trained Faster-RCNN[12] to recognize man or woman in a scene, but the result still needs to be improved. In many cases, a person is detected as man in one frame and then as woman in the next frame. We assumed that if we can detect every person's trajectories than we can use that information to better classify the person's gender in overall scenes.

Since we used Faster-RCNN features for gender detection, we want know if we are able to reuse those features for tracking. The task then become objects association task, since we need to associate the same person in consecutive video frames. In this paper, we compare two methods for object association. The first method is simple Euclidean distance and the second is more complicated Siamese neural network. We use our own dataset because it is already gender annotated.

2. Related works

Two main components of tracking by detection framework is the detected objects and the method used to associate objects in consecutive frames. Objects in this experiment is detected by Faster RCNN, and thus the features of each objects are generated by the Faster-RCNN network. We then use this feature to associate one person in frame *n* to the same person in frame *n*+1. The easiest way to associate them is to compare the Euclidean distance between object features. More complicated way is to use Siamese Neural Network to find objects that are similar to each other. In this chapter, we will briefly presents Faster-RCNN and Siamese neural network as the building block of our experiments. We also describe the tracking evaluation metric used in this paper.

2.1. Faster RCNN

Faster-RCNN framework[4] is known as one of the top person detectors in ILSVRC 2016 Challenge[6]. It is build using Caffe and the code that we used is available online[7]. The idea of Faster-RCNN is to apply CNN network on input image to create feature maps, and then pass the feature map to the Region Proposal Network (RPN). The feature maps corresponding to the region proposals are then passed to a classifier to detect what object it contained, as shown in Fig. 1.

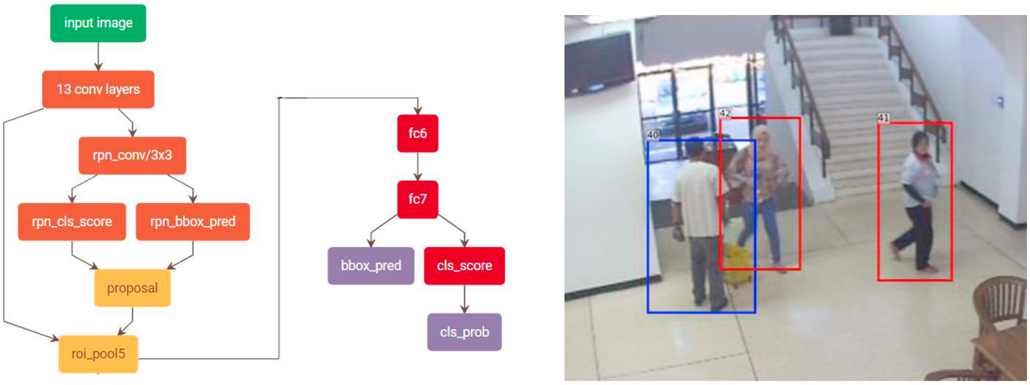

Fig. 1. (a) Faster-RCNN building block; (b) example of detected person.

The CNN network in Faster-RCNN can take any architecture provided in Caffe Model Zoo such as ZF, AlexNet, GoogleNet, VGG-net, etc. In this experiment we used VGG16, introduced by Visual Geometry Group[8], which has thirteen convolutional layers with various sizes (64, 128, 256, 512), two fully connected layer (called 'fc6' and 'fc7') and one softmax layer as classifier. The CNN features that we refer in this paper corresponds to the output of 'fc7' layer which size is 4096. VGG16 is chosen for its simple yet powerful architecture.

The RPN is inserted between the 13th convolutional layer and the 'fc6' layer in VGG16 network. RPN is trained to classify region as containing or not-containing object. First it created nine anchor windows with different size (52,72,96,100,136, 184, 192, 264,376) and different ratios (1:1, 1:2, 2:1). The anchors are shifted in all area in the image, thus creates about 6000 boxes in the image. Convolutional features in the boxes are then passed as input for RPN. Each box is then given a score. High score indicates object (as opposed to background). Then nms (non-maximum suppression) are applied with certain threshold to combine intersecting boxes. RPN returns top 300 region/boxes proposal. These 300 proposal are then classified into real world objects such as TV monitor, chair, person, etc. Nms is also applied in this step. Out of 300 region proposal generated by RPN, only few of them are classified as real objects. Most of them falls to "background" class.

Faster-RCNN that is able to classify 20 classes of PASCAL dataset (person, chair, sofa, bottle, etc) in an image is available for public use[7]. Since we want to know the gender of each person on the set, we used transfer learning[9] to classify person into man or woman class. The transfer learning is done by initializing the network with 20-classes detector Faster-RCNN parameters, and then re-train the network with our own data to detect only two classes (man and woman).

2.2. Siamese Neural Network

Siamese Neural Network (SNN)[10], as the name explains, consists of twin functions G_W which share the same set of parameters W, and a cost module that generates the distance $D_{w=}\|G_w(x_1) - G_w(x_2)\|$, as shown in Fig.2. The input of SNN is a pair of images (x_1, x_2) and a label Y. The label Y may be $Y = 0$ (similar) or $Y = 1$ (dissimilar). A loss function L combines D_W with label Y to produce the scalar loss $L_S = (0.5)(D_w)^2$ or $L_D = (0.5)\{\max(0, m - D_w)\}^2$. L is called the contrastive loss function

$$L(W, Y, x_1, x_2) = (1 - Y)(0.5)(D_w)^2 + (Y)(0.5)\{\max(0, m - D_w)\}^2$$

where $m > 0$ is a margin that defines a radius around $G_w(x)$. Dissimilar pairs contribute to the loss function only if their distance is within this radius[8]. Parameter W is updated using stochastic gradient.

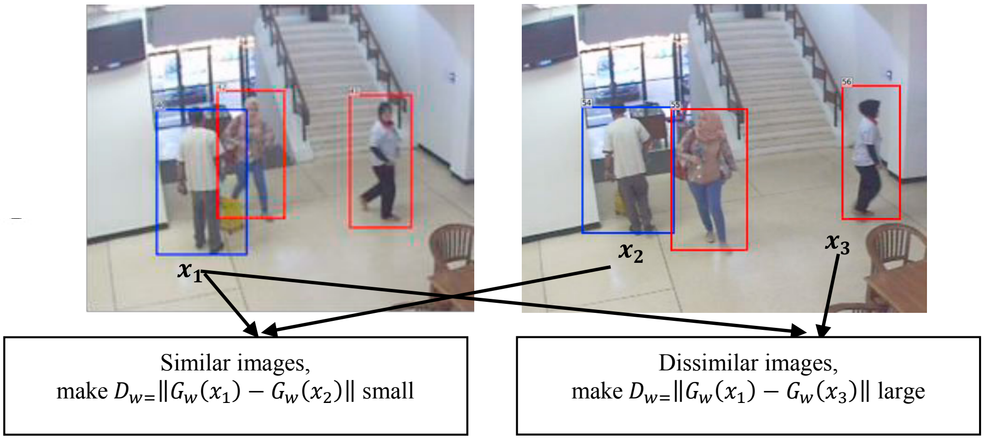
x_1
x_2
x_3
Similar images,
make $D_{w=}\|G_w(x_1) - G_w(x_2)\|$ small
Dissimilar images,
make $D_{w=}\|G_w(x_1) - G_w(x_3)\|$ large

Fig. 2. Visualization of SNN object association between frames

2.3. Multiple Object Tracking (MOT) evaluation metric

One of the most popular tracking evaluation is MOT evaluation metric[2]. We use 12 tracking metrics of MOT:

- Recall: percentage of correctly detected targets, compared to the ground truth
- Precision: percentage of correctly detected targets, compared to all detected objects
- FAR: number of false alarms per frame, number of false positives divided by total number of frames
- GT: number of ground truth trajectories

- MT: mostly tracked trajectories, more than 80% of ground truth trajectory length is tracked
- PT: partially tracked trajectories, between 20% - 80% of ground truth trajectory length is tracked
- ML: mostly lost trajectories, less than 20% of ground truth trajectory length is tracked
- FP: number of false positives
- FN: number of missed targets
- IDs: number of ID switches. ID switches happens when a person is detected as different person due to missed association or is it was occluded by other object
- MOTA: multi-object tracking accuracy in [0,100], MOTA = 1 – error, where error is defined as the total number of FN+FP+IDs compared to the total number of ground truth objects.

3. Method

In this paper, we compare the result of tracking using simple minimum Euclidean distance and more complex SNN between each person in consecutive frames. Each person is represented by 4096 Faster-RCNN feature vector. For the Euclidean distance method, we only need to calculate the distance between two feature vectors of each input pair. A pair would be considered the same person if their Euclidean distance is minimum. This is due to the assumption that convolutional features of similar objects generated by Faster-RCNN should be quite similar compared to features of dissimilar objects.

Unlike minimum Euclidean distance, SNN needs to be trained. We used two third data for training the SNN. Taking the 4096 feature vector as an input, the SNN base network consists of three fully connected layer with 255 node each. We use ReLu as activation functions in all layers.

4. Dataset

We use surveillance image dataset from our campus area. There are 14 scenes taken from the same area in different time. The resolution is 640 x 480 pixel. Each scene contains 4-8 trajectories from 21 – 89 frames. There are several movements such as moving from left to right, leaving the camera, moving towards the camera, and staying, as seen in Fig.3. Scene 11 is the most complicated because it contains 8 person (trajectories) with many occlusion. Although detection result of Faster-RCNN shows promising result, but there are still errors as shown in Fig. 3. Incorrect detections includes false detections and incorrect bounding boxes. The overall detection accuracy for independent frames in our dataset[10] is 0.75.

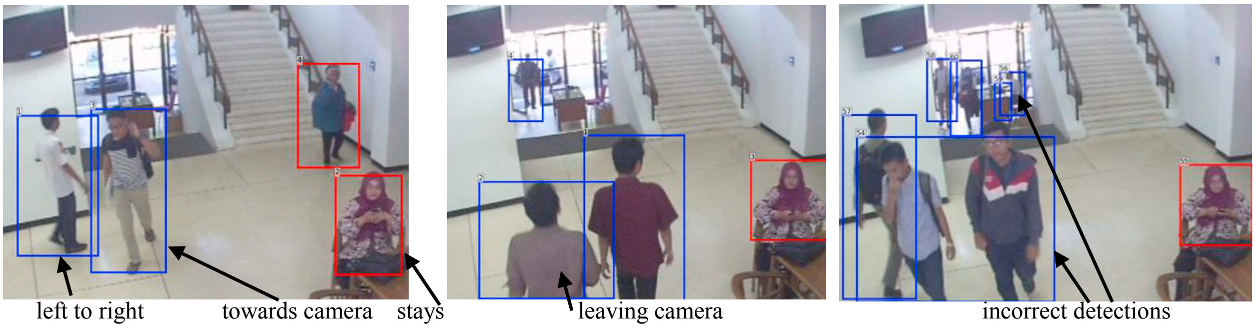

Fig. 3. Screenshot of the dataset

5. Result and Discussion

The result is our experiment is shown in Table 1. It is quite interesting to see that simple minimum Euclidean distance association performs well compared to SNN in most scenes. SNN suffers much errors due to higher number of ID switch, as consistently shown in all scenes. ID switch happen when a person in frame t was tracked as a new person because it was considered to be dissimilar to all person in $t-1$. For SNN, we set the threshold of similarity to 0.5.

Table 1. Tracking Result

Scene	#Frames	Type	Recall	Prec	FAR	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA
1	78	Ecdist	90.86	95.72	0.1	5	5	0	0	8	18	16	11	**78.68**
		SNN	82.23	95.29	0.1	5	4	1	0	8	35	54	16	50.76
2	52	Ecdist	95.45	92.31	0.13	4	4	0	0	7	4	7	2	**79.55**
		SNN	89.77	94.05	0.1	4	3	1	0	5	9	14	4	68.18
3	29	Ecdist	93.15	93.15	0.17	4	4	0	0	5	5	8	4	**75.34**
		SNN	75.34	91.67	0.17	4	2	2	0	5	18	16	6	46.58
4	30	Ecdist	92.19	96.72	0.07	4	3	1	0	2	5	2	1	**85.94**
		SNN	76.56	96.08	0.07	4	2	2	0	2	15	8	3	60.94
5	50	Ecdist	74.82	99.05	0.02	4	2	2	0	1	35	6	12	**69.78**
		SNN	80.58	99.12	0.02	4	2	2	0	1	27	42	14	49.64
6	38	Ecdist	80.95	95.51	0.11	4	2	2	0	4	20	6	11	**71.43**
		SNN	80	95.45	0.11	4	2	2	0	4	21	31	12	46.67
7	59	Ecdist	64	96.97	0.07	6	1	5	0	4	72	4	20	**60.00**
		SNN	78.5	97.52	0.07	6	2	4	0	4	43	49	23	52.00
8	30	Ecdist	82.35	97.67	0.03	4	1	3	0	1	9	1	3	**78.43**
		SNN	84.31	95.56	0.07	4	1	3	0	2	8	14	1	52.94
9	40	Ecdist	76.05	99.22	0.03	5	2	3	0	1	40	5	11	**72.46**
		SNN	81.44	97.14	0.1	5	3	2	0	4	31	32	14	59.88
10	22	Ecdist	49.48	90.57	0.23	6	2	3	1	5	49	3	7	**41.24**
		SNN	69.07	91.78	0.27	6	3	2	1	6	30	23	12	39.18
11	89	Ecdist	39.3	86.81	0.27	8	0	8	0	24	244	22	50	27.86
		SNN	59.95	91.63	0.25	8	0	8	0	22	161	65	53	**38.31**
12	42	Ecdist	62.44	91.79	0.26	7	3	4	0	11	74	23	16	45.18
		SNN	71.07	90.91	0.33	7	1	6	0	14	57	34	21	**46.70**
13	21	Ecdist	75	86.44	0.38	4	2	2	0	8	17	23	6	**29.41**
		SNN	73.53	86.21	0.38	4	2	2	0	8	18	28	10	20.59
14	56	Ecdist	49.7	81.19	0.34	6	0	5	1	19	83	13	27	30.30
		SNN	66.67	82.71	0.41	6	0	6	0	23	55	36	29	**30.91**

It is also worth noted that the recall number is not the same for SNN and Euclidean distance. Calculating recall in tracking is not simply by comparing the bounding box of detected person to the ground truth, but also comparing their ID (or trajectories). If the association between consecutive frames is not correct then they are not considered to be a correct detection, and directly counted as missed target (FN). Recall also shows how good the detection works. Suppose there are 10 person in a scene ground truth and only 7 is detected by Faster-RCNN. Then the maximum recall number is 70%, depending on the association correctness. Scene 10 and 11 suffers low tracking accuracy because of poor detection result. As mentioned earlier, there are many occlusion in these scenes that the detection process cannot handle.

6. Conclusion and Future Works

The experiment result show that simple Euclidean distance gives promising result as object association method, but it depends heavily on the robustness of detection process on individual frames. SNN does not outperform Euclidean distance, but we need to keep in mind that there are a lot of parameters tuning that needs to be further explored. This issue will be the focus of our next research.

References

[1]L. Leal-Taixe, T.C. Ferrer, K. Schindler, *Learning by Tracking: Siamese CNN for Robust Target Association*, https://arxiv.org/pdf/1604.07866v3.pdf

[2] A. Milan, L. Leal-Taixe, I. Reid, K.Schindler, *MOT16: A Benchmark for Multi-Object Tracking*, https://arxiv.org/pdf/1603.00831v2.pdf

[3] G. Zhu, F. Porikli, H.Li, *Robust Visual Tracking with Deep Convolutional Neural Network based Object Proposals on PETS*, CVPR 2016.

[4] S.Ren, K.He, R.Girshick, J.Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PP, Issue: 99, 2016

[5] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, *PETS 2016*. http://www.pets2016.net/

[6] Large Scale Visual Recognition Challenge 2016, http://image-net.org/challenges/LSVRC/2016/results

[7] Faster-RCNN code, http://github.com/rbgirshick/py-faster-rcnn

[8] K. Simonyan, A. Zisserman,*Very Deep Convolutional Networks for Large-Scale Image Detection*, https://arxiv.org/pdf/1409.1556

[9] S.J. Pan, G.Yang, *A Survey on Transfer Learning*, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, Issue 10, 2009

[10] R. Headsell, S.Chopra, Y. LeCun, *Dimensionality Reduction by Learning an Invariant Mapping,* CVPR 2006.

[11] Y. LeCun, *Learning Hierarchies of Invariant Features*, http://helper.ipam.ucla.edu/publications/gss2012/gss2-12_10739.pdf

[12]. D. Chahyati, M.I. Fanany, A.M. Arymurthy, *Man Woman Detection in Surveillance Images,* ICoICT 2017.