Deep Learning for Autonomous Vehicles

Course project - Milestone 3 - Report

Students :

Sylvain Pietropaolo
Yann Martinson
André Gomes

June 5, 2020

Contents

Introduction

After the implementation of pedestrian detector and re-identification algorithms in the 2 previous milestones, the third and last step is to combine the two in order to track peoples on a video. The algorithm presented in this report takes a video as an input and gives the same video as output with labelled bounding boxes on the different pedestrians. First, this report will expose how the 2 previous algorithms have been combined. Then, the results will be presented to finally end with the limitations and the principal difficulties encountered.

1 Method overview

As a summary, the 2 previous milestones gave :

1. An algorithm to detect multiple pedestrians on an image and to place a bounding box around each of them.
2. An algorithm trained to extract the suitable features from images of pedestrians and to distinguish them between each others based on these features.

The first algorithm takes an image as an input and gives the positions of all the bounding boxes delimiting the pedestrians from the other objects. The second algorithm takes a list of images, corresponding to the pedestrians, and extracts some particular features.

The goal of the presented algorithm is to combine the two previous algorithms to identify all the pedestrians of a video and to re-identify them in order to be able to track all the persons during all the time they are present in the video.

As the 2 algorithms take images as inputs, the first step is to extract all the frames of the video. Then, frame after frame, the bounding boxes around pedestrians are extracted and the features of each of them computed. Using the features extracted, the bounding boxes are related between each frame to track each person individually. The different steps of the algorithm through the firsts frames are detailed below :

(i) Frame 1 :

   (1) Use the first algorithm to extract the pedestrians bounding boxes.
   (2) Use the second algorithm to extract their features.
   (3) Store the features of each bounding boxes in the query list.

(ii) Frame 2 :

   (1) Use the first algorithm to extract the pedestrian bounding boxes.
   (2) Use the second algorithm to extract their features.
   (3) Stock the features of each bounding boxes in the gallery list.
   (4) Compare the features of the images in the query to the features of the images in the gallery with a score matrix. This score matrix has a score relating each image on the query with each image on the gallery.
   (5) Label the persons in the gallery according to this scoring matrix, taking the label of the query features best associated to them.

(6) Update the features of the re-identified query with the features of the corresponding gallery image.

(7) Add the features of the non-labelled images in the gallery to the query list, giving them a new label.

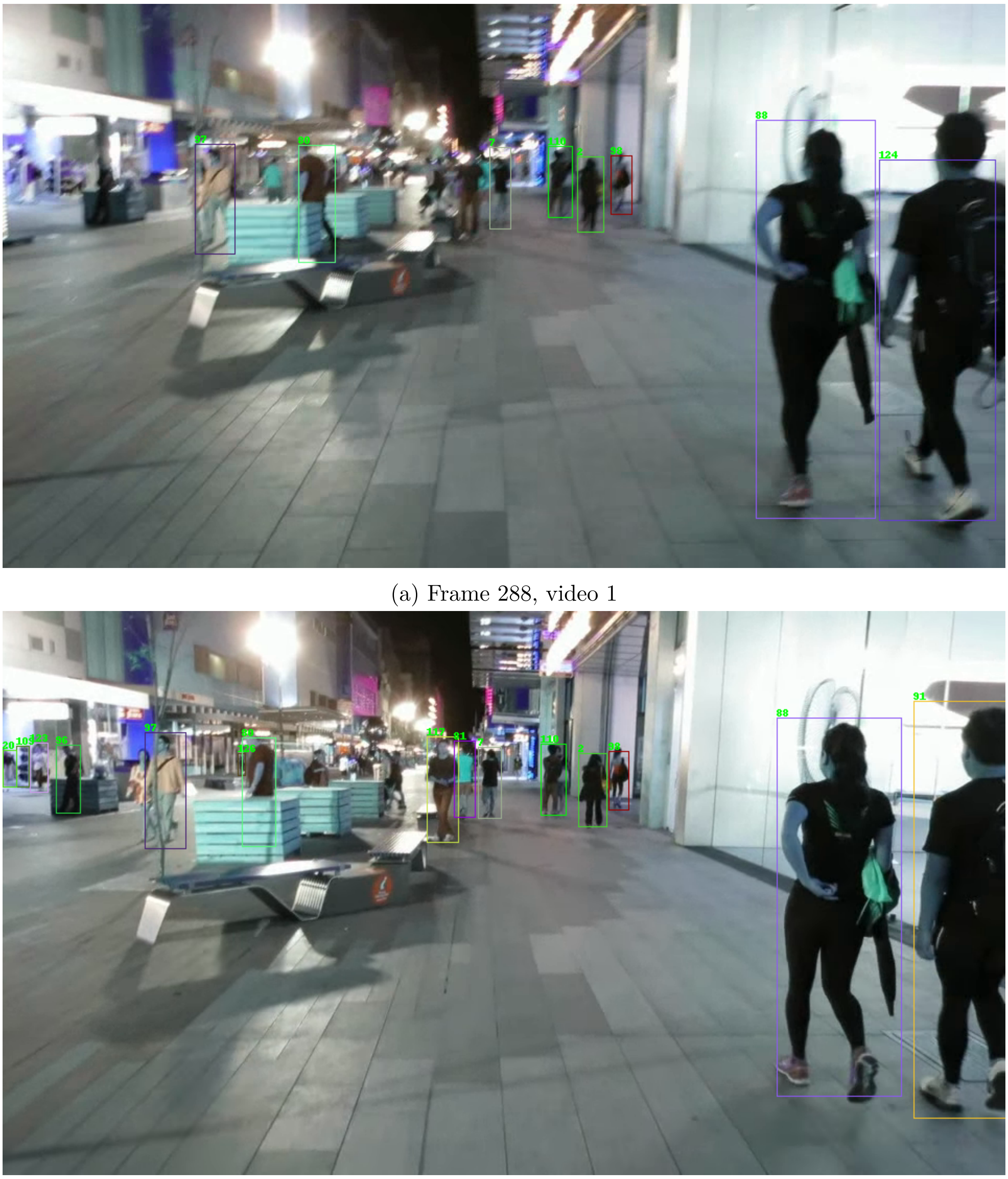
(iii) Frame n : Do the same steps as (ii) with the updated query list. Repeat this until the last frame.

For clarity let's precise how the re-identification (step ii. 4.) is performed. The algorithm of the milestone 2 extracts the features in order to differentiate the different pedestrians but do not classify them by itself. It is thus necessary to introduce a way to compare the features of two different images and to interpret this comparison to re-identify the pedestrians along the different frames. For this purpose, a score matrix is computed by the matrix product of the query features with the transpose of the gallery features. The row and column indices of the matrix give a "feature score" between each query and gallery image. The maximum score of the matrix is then found and if it is higher than a hand-set threshold, the gallery image is considered to be the same person as the associated query. The corresponding line and column of the matrix are removed and the maximum score is found again. Hence once an image in the gallery is associated to one in the query, the image on the query can no longer be associated with any remaining image in gallery still not labelled. This steps are repeated until there is no score higher than the threshold. To be noted that, for a gallery image to be assigned to a query, the center of the 2 bounding boxes need to be less than a certain threshold. Indeed, as the frames come from a video, for a person to be the same from one frame to another, it seems reasonable to expect that the boxes are close. If the maximum score is for 2 boxes far from one another, then the score is considered as "wrong" and set to zero. This provides prior knowledge to the re-identification algorithm and helps to get correct labelling.

As a summary, the re-identification is done by setting by hand a threshold for the score. If the score is higher than this threshold the gallery is considered to be the same person as the query.

2 Results

The described algorithm has been tested on 3 different videos. One common problem arising with all the videos is due to the first algorithm from Milestone 1. The model used in the algorithm extracts too many bounding boxes, sometimes several for the same person or on objects not representing a pedestrian. As a result, the frames of the videos are saturated with 'wrong' bounding boxes. For all these identified bounding boxes, the model however provides a score on how confident it is with the prediction. Hence, applying a threshold on that score reduces the wrongly identified bounding boxes and helps to select only the relevant one. There is nevertheless a trade-off when choosing this threshold. It either removes too many bounding boxes and some pedestrians are no longer identified, or it removes too few of them and there is still non relevant bounding boxes. Finally, it is preferred to fix a threshold removing too many bounding boxes to improve the clarity of the video. Some frames extracted from the output videos are shown below to have an idea of the results.

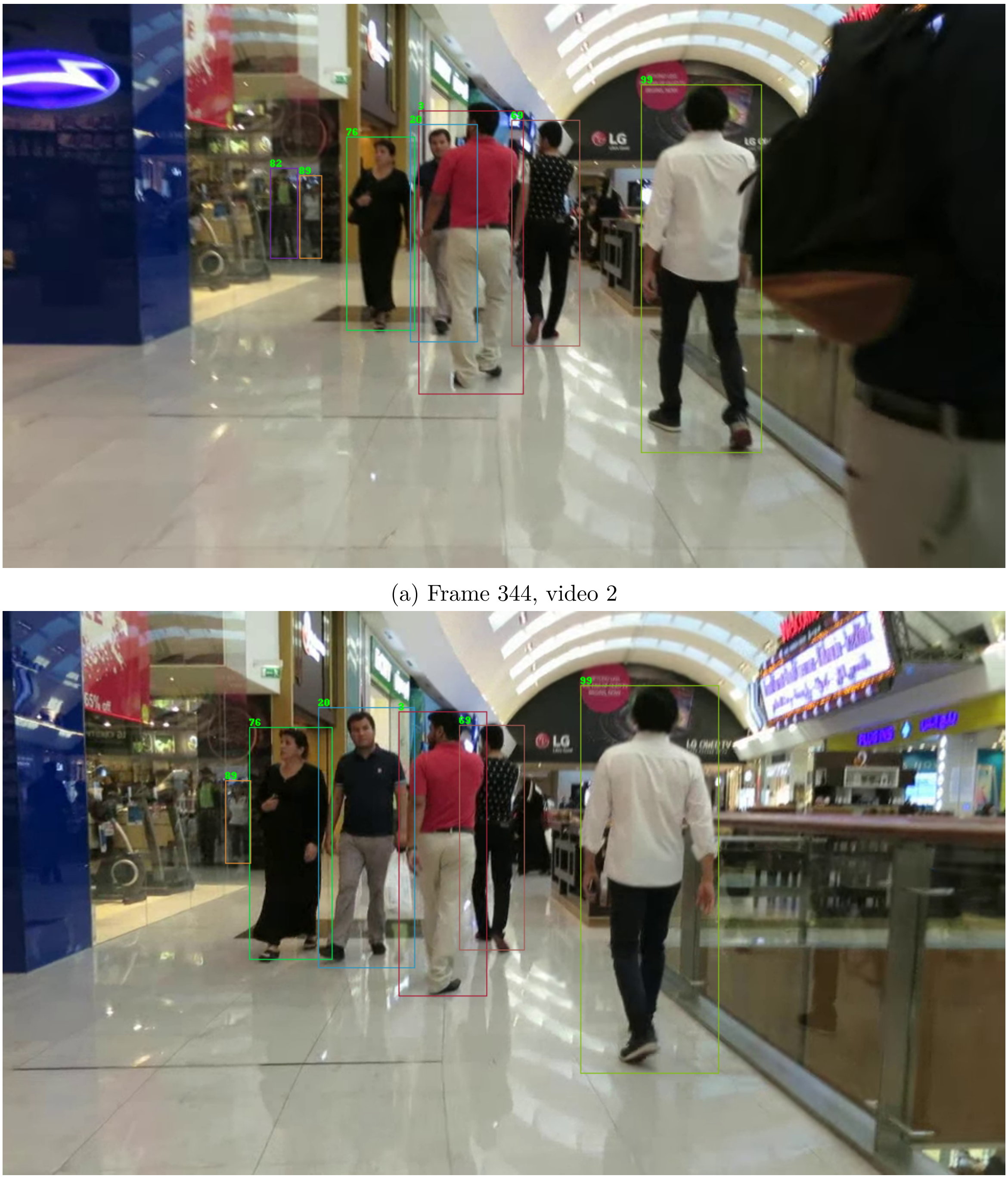


(a) Frame 288, video 1

(b) Frame 306, video 1

Figure 1: Frame comparison video 1

For the first video, we see that the re-identification works quite well. However, a couple issues appear, when for example the frame is blurry, as in frame 288, the algorithm does not re-identify the person on the far right in his height, but detects a pedestrian and gives him a new label. A couple frames later, it re-identifies him with the first label, as shown with the number on the bounding boy which is lower than the one in frame 288.

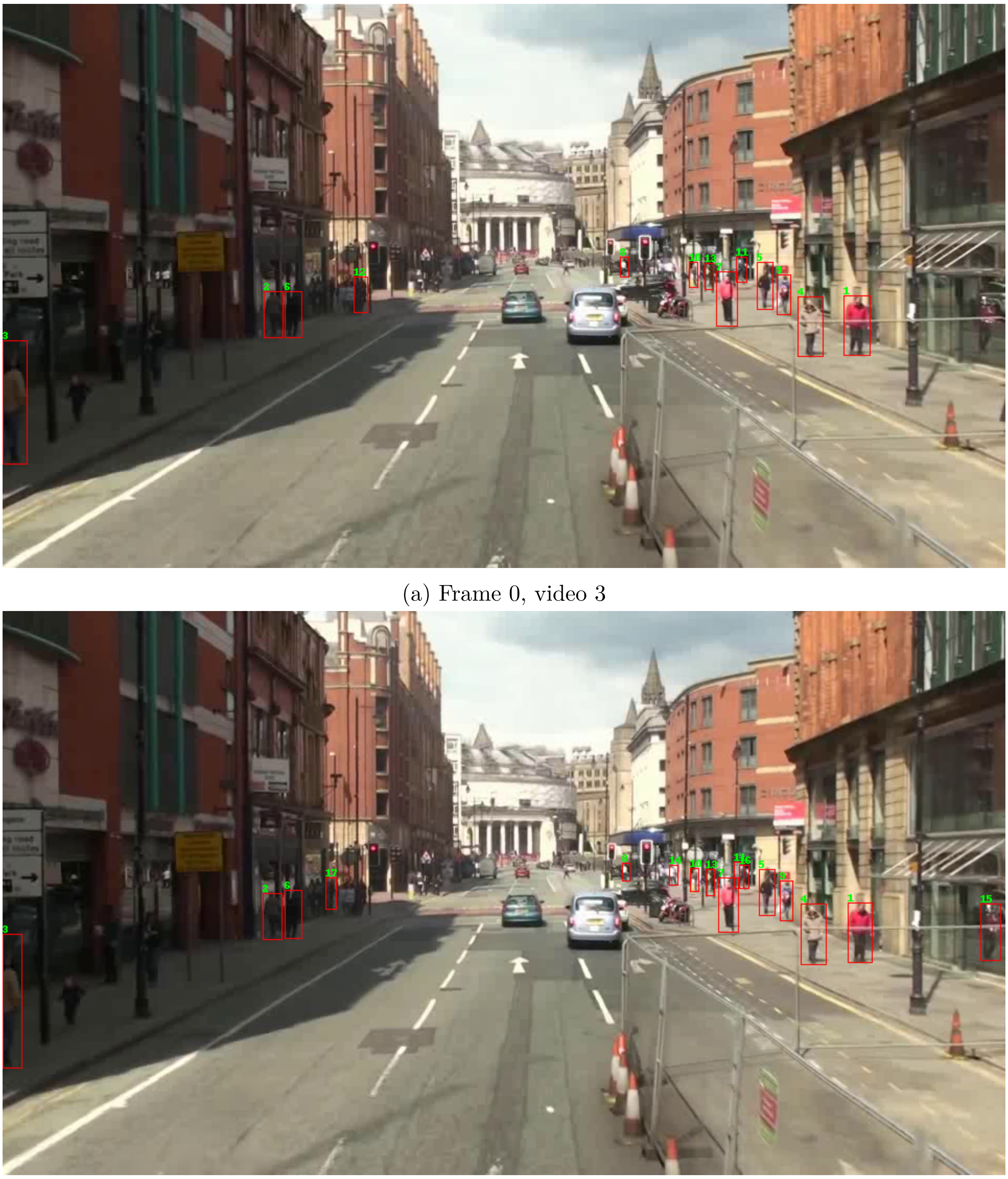


(a) Frame 344, video 2

(b) Frame 363, video 2

Figure 2: Frame comparison video 2

For the second video, we see that the re-identification works quite well. However, a couple issues appear, such as the person with label 82, in the far back left, disappearing on the frame 363. Also, the apparition of the person on the right side on frame 344 is not yet noticed.



(a) Frame 0, video 3

(b) Frame 1, video 3

Figure 3: Successive frames of the 3rd video

As one can see on the previous figures, the pedestrians on the background do not very often appear inside a box because of the threshold on bounding boxes confidence. Additionally, the re-identification seems to have some problems to re-identify little bounding boxes or occluded pedestrians. But for clearly identified persons it seems to work well.

3 Limitations and difficulties

As always in machine learning, an algorithm has to be tuned to the particular problem it is facing. This can be done by adapting the structure of the algorithm or by tuning some of its parameters. In the present case, these parameters represent mainly thresholds, but it could also be by changing starting point, rates etc.. The evaluation of these parameters is carried through some scoring method, such as benchmarks or accuracy. In this last milestone, it is only possible to evaluate the impact of the parameters visually. With videos ranging between 600 and 900 frames it would indeed by too tedious to evaluate the accuracy of the final algorithm manually.

Additionally, this algorithm depends a lot of the 2 previous milestones. And if the re-identification algorithm seems to work properly, more difficulties are encountered with the algorithm extracting the bounding boxes. Indeed, the one used here and trained on the ECP dataset extracts too many bounding boxes which are not relevant every time. For example it puts several bounding boxes on a single pedestrian and thus confuses the re-id algorithm. As explained before, it forces us to set a threshold to decrease the number of bounding boxes identified on each frame. But if this allows to have a single box by pedestrian it also removes some boxes and some persons thus disappear between different frames from the algorithm point of view.

Another difficulty encountered is the setting of the threshold for the re-identification. As every hard-coded parameter, it makes the algorithm less adapted to different configurations of light, size and scale for example.

Conclusion

The two algorithms, one for object detection and the other one for person re-identification, are combined into a single and last algorithm that can take any video as input and apply the consecutive steps on each frame to detect and re-identify pedestrians. The results show that this has been implemented successfully. However, we see clear limitations of the proposed algorithm, mostly due to the trained model for extracting the bounding boxes around the pedestrians. Additionally, the processing of each frame is computationally high, especially for extracting the bounding boxes and computing their respective features. By having the full video at first, it is easier to compute each and stack them back into a video at the end. For a real-time implementation, a more optimised algorithm would be required along with better trained models for person identification.