***Abstract*—Pedestrian detection targets to classify and localize multiple people and return the spatial location of each person from images and videos. Most of the detectors are not ideal for pedestrian detection in crowded scenes. Though Deep Neural Network has a remarkable performance on object detection, the majority of training datasets only label the uncovered parts of objects which makes occlusion detection unlearnable. Moreover, most monocular detectors do not take into account the multi-camera geometry to resolve ambiguities and ignore occlusions. In this work, we present a model to apply multi-view detection based on monocular instance segmentation using CNNs and multi-view geometry to localize, which aims to detect people in crowded scenes. This report shows the usages of pedestrian detection, the challenge for detection, the proposed methodology, and its performance on terrace video sequences.**

***Index Terms*—multi-pedestrian; multi-view; RCNN;**

I. Introduction

PEDESTRIAN detection targets to classify and localize multiple people from images and videos, which is a long-standing problem in computer vision. It has a vast range of applications which can be directly applied in the field of real-world tasks, such as vehicle-assisted driving system, intelligent video surveillance and robots.

Vision-based pedestrian detection technology has been developed over decades, most of them rely on learning algorithms to find the implicit representation of people. There are two difficulties in the design of pedestrian detection system of finding precise features of pedestrians.

1) Inconstant appearance: Pedestrians' appearance varies because of the change of height, pose, sex, cloth and accessory.

2) Occlusion: Pedestrians are moving objects who may be occluded by other objects, especially the occlusion among pedestrians in crowded scenes.

Detecting the pedestrians through multiple cameras within a overlapping field of view, allowing to merge more spatial information, is a reasonable solution for occlusion. This work use homography transformation to fuse information of potential location of pedestrians in a top view to indicate their positions in all camera views.

This work's contribution is: a simple method to apply current state-of-art deep learning detection technology in multi-view detection which is pre-trained by a monocular dataset and provide a more robust detection compared with methods using background subtraction, HOG and SVM.

II. Related work

The traditional paradigms before 2010 were based on variants of feature extraction techniques: HOG (Histograms of Oriented Gradients) and DPM (Deformable Parts Model). HOG extracts features relied on the edge and shape, like color contrast between pedestrians and background, so it is not effectively record the visible information of objects and it is sensitive to noise. As regards the other method, DPM is more advanced, but it need to manually define features and it is not robust to detect objects with rotating, stretching, and changing the visual angle.

Monocular detection works based on CNN, such as YOLO series and RCNN series have made remarkable achievement on COCO dateset, since deep convolutional neural networks are highly robust to lights, color, texture and white balance, and normally have richer dimensional representations of features computed over the input. However, most multi-view detection methods still depends on old and outmoded technologies, such as background subtraction, because deep learning methods rely heavily on training datasets enabling researchers to apply multi-view training on pedestrian detection, but they do not exist.

III. Method for Merging Monocular Detector

The overview of our method is shown as1, it takes two inputs from a monocular detector: masks(foregrounds of pedestrians) and bounding boxes. The truncated mid-lines of bounding boxes are then projected to a top view, where the pedestrians' location candidates are defined. Next RSS (Repulsive Spatial Sparsity Filtering) find local peaks and reduces the duplicates .

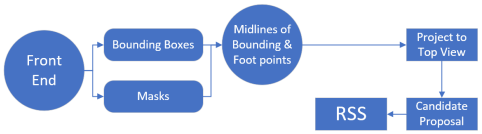


Fig. 1. Method for Merging Monocular Detector

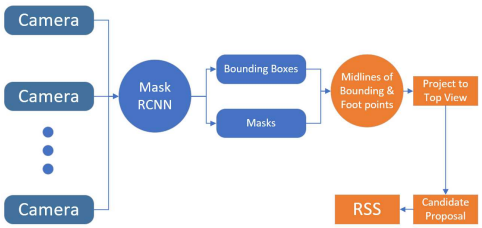


Fig. 2. Overview of the proposed method: using Mask RCNN as front end

A. Candidates Proposal

1) Homography: One difficulty of monocular detection is finding the exact foot positions of pedestrians. The middle bottom point of each bounding box (specifically generated by Mask RCNN in this work) is not the centroid between feet. By default, for eliminating ambiguities, each mid-line of a bounding box is truncated and projected to the top view by homography. A scanner, the horizontal orange line with two gray ends, scans whether it is an occlusion between objects. If either of two ends intersects with the foreground (segmentation), the mid-line should be extended until there is no intersection between the scanner and foreground.

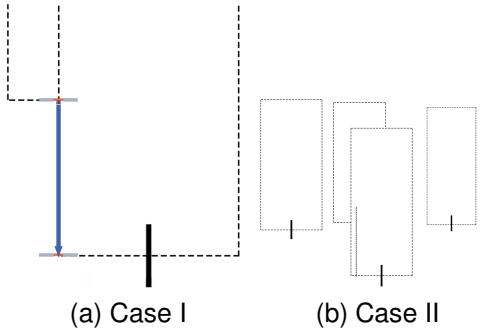
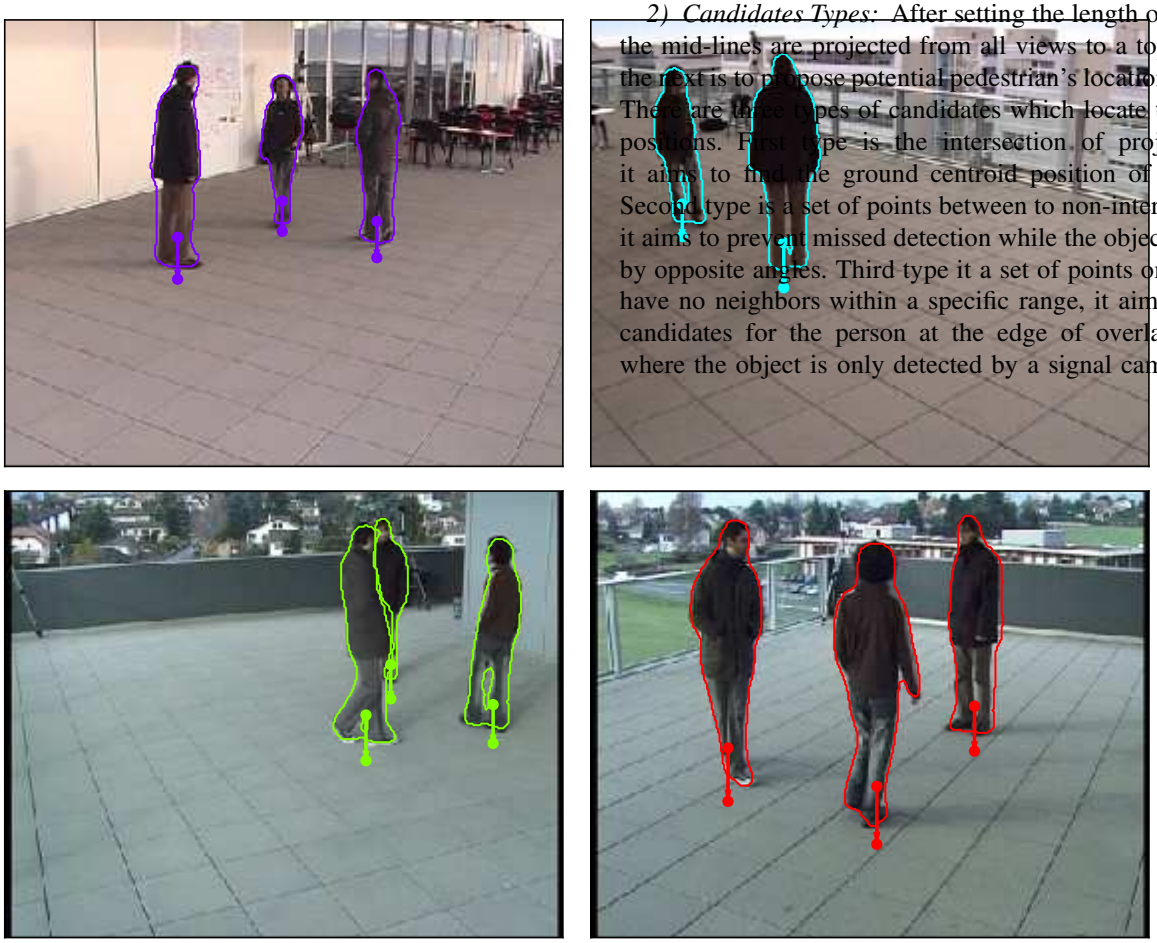
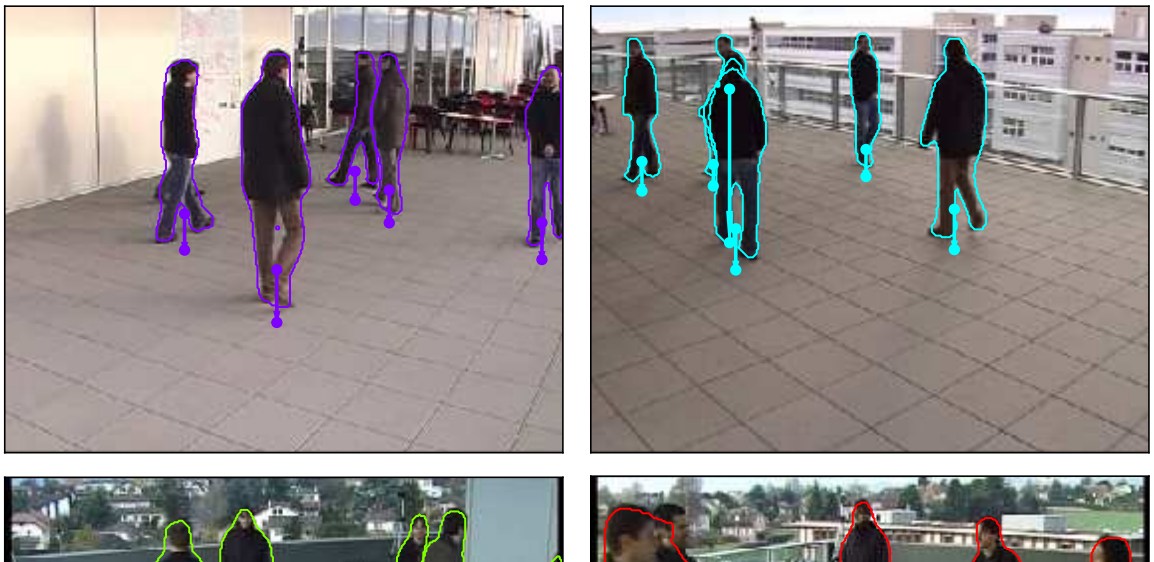


Fig. 3. The truncated mid-lines and masks

2) Candidates Types: After setting the length of truncation, the mid-lines are projected from all views to a top view, then the next is to propose potential pedestrian's location on ground. There are three types of candidates which locate the potential positions. First type is the intersection of projected lines, it aims to find the ground centroid position of pedestrians. Second type is a set of points between to non-intersected lines, it aims to prevent missed detection while the object is detected by opposite angles. Third type it a set of points on a line who have no neighbors within a specific range, it aims to provide candidates for the person at the edge of overlapping field, where the object is only detected by a signal camera.



[t].24



[t].15

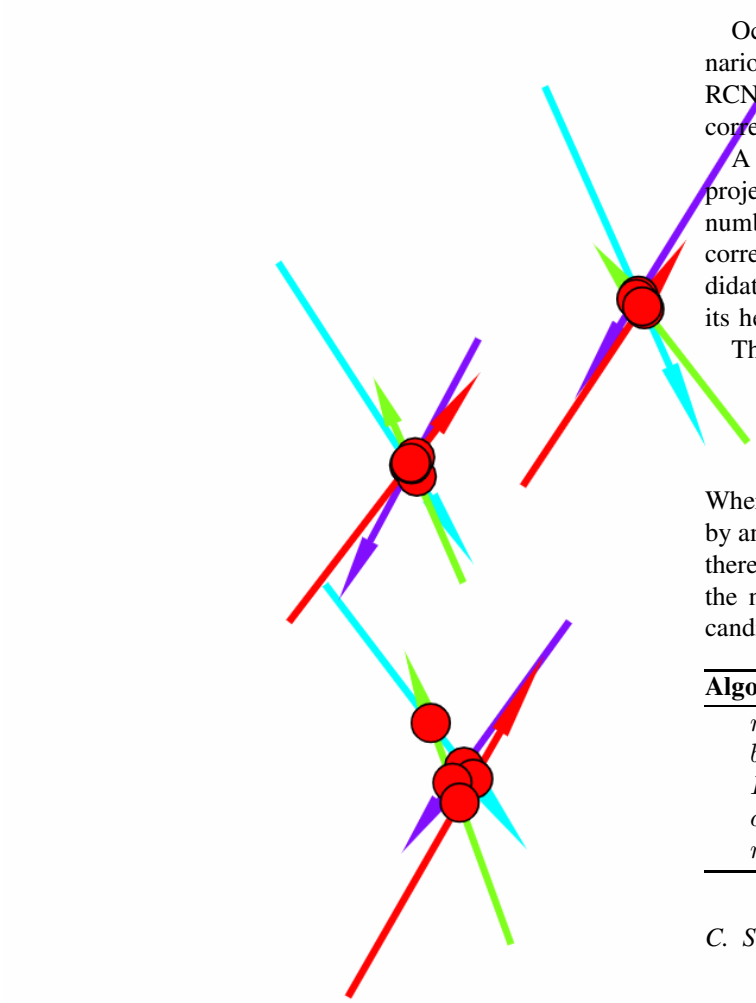
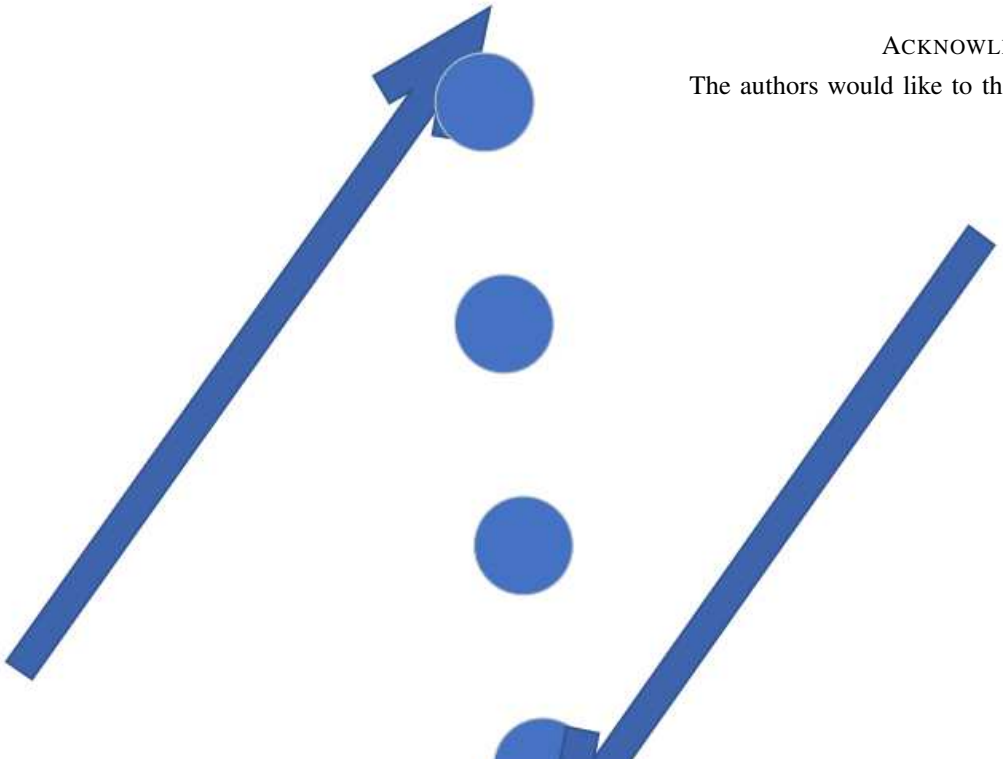


Fig. 5.

[t].15



B. Multi-view Mask Scoring

Occlusion can become a serious problem in complex scenarios with multiple targets. The score generated by Mask RCNN aims to assess classification results, which is not well correlated with multi-view detection.

A likelihood function is applying to candidates after inverse projecting them to camera views from the top view, and numbers of N (N stands for the number of camera views) corresponding bounding boxes are regenerated for each candidate. It is assumed the width of a bounding box is 37.5% its height in this step.

The likelihood function is defined as

$$S_{\text{total}} = \sqrt[n]{\prod_{k=1}^{n} (S_{head}[k] * S_{foot}[k])}$$

Where n stands for how many cameras can detect a pedestrian by an ideal monocular detecting strategy. However, practically, there is no perfect monocular detection. In this work, n is the number of its valid re-projected bounding boxes for the candidate, and it is counted by:

Algorithm 1 Counter

$n = 0$ $segmentation$, $view$ in all_views
$box = Regenerating(candidate, view)$ box in $FOV(view)$ and box is not at $Edge(view)$
$overlapRatio(box, segmentation) > threshold$
$n++$

C. Soft Region Non Maximum suppression

APPENDIX A
PROOF OF THE FIRST ZONKLAR EQUATION

Appendix one text goes here.

APPENDIX B

Appendix two text goes here.

ACKNOWLEDGMENT

The authors would like to thank...