This document provides the supplementary information about the annotation for PETS2009 CC and S2L2 datasets. These datasets and the camera calibration data can be found in [1]. There are 8 camera views (C1-C8) in the PETS2009 City Centre (CC) dataset. There are 7 camera views, which correspond to C1, C3-C8 of the CC dataset and are renamed as C1-C7, in the PETS2009 S2L1 dataset. Here we will use the original camera view IDs in the CC dataset.

There are 795 frames in the PETS2009 datasets. Since there is content discontinuity at frame 395, many researchers used the first half of these frames for training and the second half for testing. There are totally 200 frames annotated by us. The ground truths of one frame in every two frames, from frame 392 to 790 of the dataset, are extracted according to camera views C1, C2, C5, C6 and C8.

The ground truth data at each frame is saved in an XML file. The file name represents the frame number in the dataset. In the XML file, the tag <object> </object> represents one annotated pedestrian in that frame. For each object, the ground truths of height and location are shown with tags <height>, </height>, and <top>, </top>, respectively. To be specific, the data of height is the ratio of actual height of the person to the average height 1700 mm of the pedestrians. The information of the location in the top view is shown by tags <x>, </x>, and <y>, </y>, which shows the image coordinates of the pedestrian in a virtual top view of resolution 1000 x 1000 pixels. The upper left corner of the top view image is considered as the origin of the image coordinate system.

The top view coordinate (x, y) in the annotation can be converted to the world coordinate (X_w, Y_w) by using following equations:

$$\begin{cases} X_w = (x - 500) \times 50 \\ Y_w = (y - 500) \times 50 \end{cases}$$

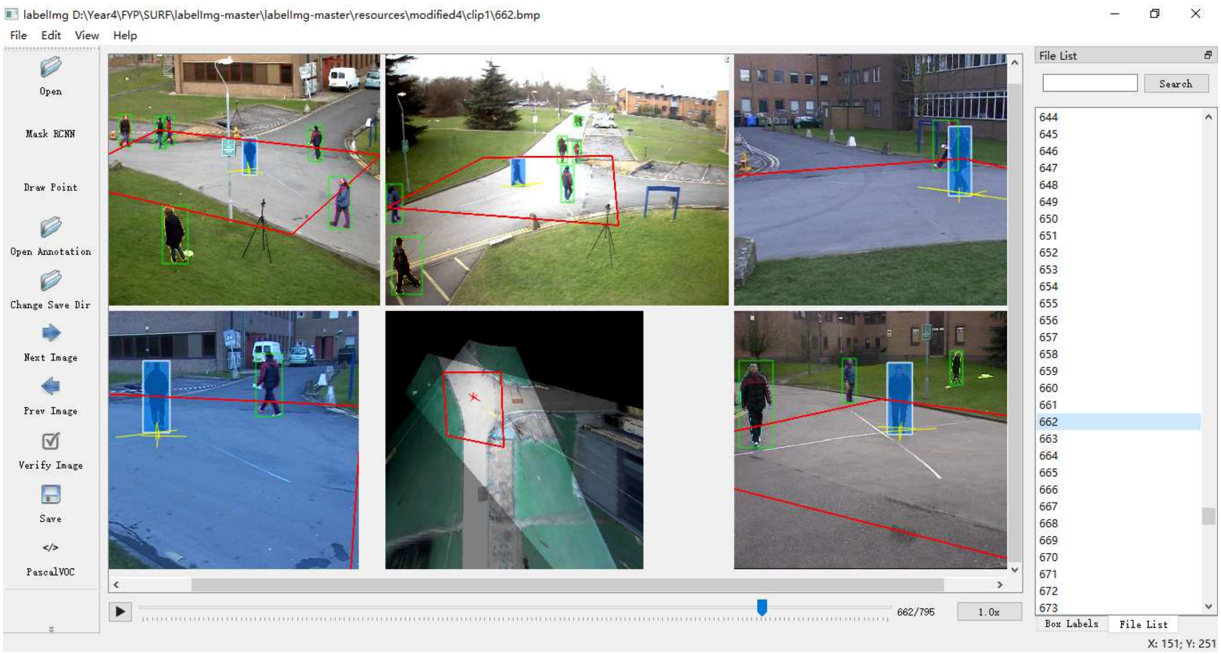
The world coordinate (X_w, Y_w) can also be converted back to the top-view image coordinate (x, y) as follows:

$$\begin{cases} x = X_w \div 50 + 500 \\ y = Y_w \div 50 + 500 \end{cases}$$

The unit of the world coordinate is millimetre (mm). The unit of the top-view image is a pixel which corresponds to a size of 50 mm in the world coordinate system. The origin point of the world coordinate corresponds to the centre of the top-view image, since the size of the top-view image is 1000×1000 pixels.

There is a synchronization problem between camera views C1-C4 and C5-C8. The frame numbers in the annotation files are the original ones in the PETS2009 dataset. In reading the images of views C5-C8, the frame number should be incremented by 1 at frames 482-579, by 2 at frames 580-654 and by 3 at frames 655-790. There are no changes in views C1-C4.

The ground-truth data of PETS2009 were extracted by Ms. Tingyu Lin by using a multi-camera interactive tool. This software package was developed by Tingyu Lin herself with Python [2]. Only the pedestrians within the area of interest (AOI, the red quadrangle) were annotated. The user interface and quality of the ground truth annotation are shown as follows.



labelImg D:\Year4\FYP\SURF\labelImg-master\labelImg-master\resources\modified4\clip1\662.bmp

Figure 1. The user interface of the annotation tool.

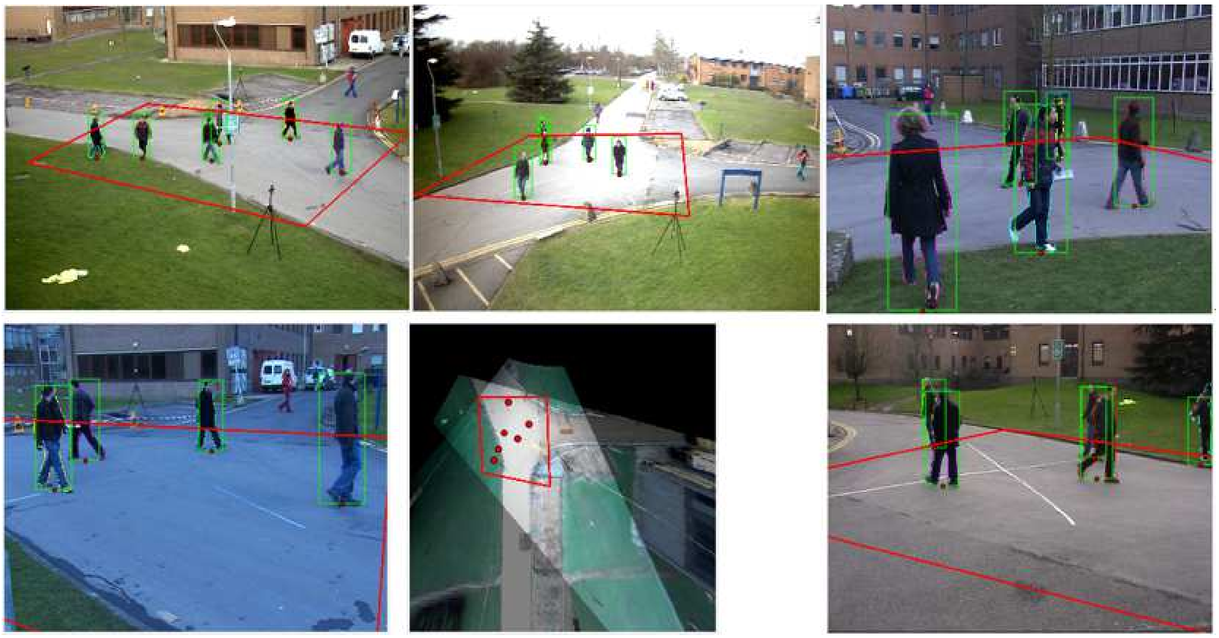


Figure 2. The annotation data visualized in the individual camera views and a synthetic top view.

For any further enquiry, please contact Dr. Ming Xu (Email: ming.xu@xjtlu.edu.cn) at Xi'an Jiaotong-Liverpool University.

References

[1] PETS2009 Dataset, http://www.cvg.reading.ac.uk/PETS2009.

[2] Tingyu Lin, An Annotation Tool for Deep Multi-Camera People Detection, B.Eng. Dissertation, Xi'an Jiaotong-Liverpool University, June 2020.