Fine-Grained Classification of Pedestrians in Video: Benchmark and State of the Art

David Hall and Pietro Perona
California Institute of Technology.

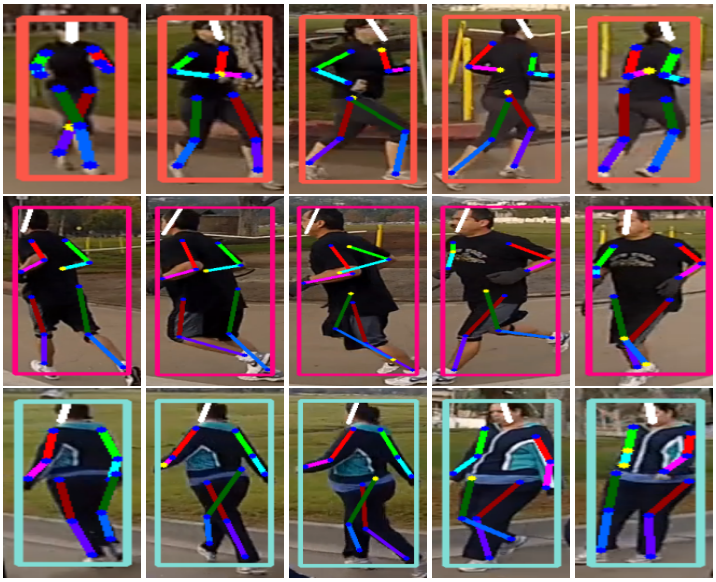

Figure 1: **Three examples from the CRP dataset.** Annotations include a bounding box, tracks, parts, occlusion, sex, age, weight and clothing style.

People are an important component of a machine's environment. Detecting, tracking, and recognising people, interpreting their behaviour and interacting with them is a valuable capability for machines. Using vision to estimate human attributes such as: age, sex, activity, social status, health, pose and motion patterns is useful for interpreting and predicting behaviour. This motivates our interest in fine-grained categorisation of people.

In this work, we introduce a public video dataset—**Caltech Roadside Pedestrians (CRP)**—to further advance the state-of-the-art in fine-grained categorisation of people using the entire human body. This dataset is also useful for benchmarking tracking, detection and pose estimation of pedestrians.

Its novel and distinctive features are:

1. Size (27,454 bounding box and pose labels) – making it suitable for training deep-networks.
2. Natural behaviour – subjects are recorded "in-the-wild" so are unaware, and behave naturally.
3. Viewpoint – Pedestrians are viewed from front, profile, back and everything in between.
4. Moving camera – More general and challenging than surveillance video with static background.
5. Realism – There is a variety of outdoor background and lighting conditions
6. Multi-class subcategories – age, clothing style and body shape.
7. Detailed annotation – bounding boxes, tracks and 14 keypoints with occlusion information; examples can be found in Figure 1. Each bounding box is also labelled with the fine-grained categories of age (5 classes), sex (2 classes), weight (3 classes) and clothing type (4 classes).
8. Availability – All videos and annotations are publicly available

CRP contains seven, twenty-one minute videos. Each video is captured by mounting a rightwards-pointing, GoPro Hero3 camera to the roof of a car. The car then completes three laps of a ring road within a park where there are many walkers and joggers. Each video was recorded on a different day.

The dataset was labelled with bounding boxes, tracks, pose and fine-grained labels. To achieve this, crowdsourcing, using workers from Amazon's Mechanical Turk (MTURK) was used. A summary of the dataset's statistics can be found in Table 1.

Number of Frames Sent to MTURK	38,708
Number of Frames with at least 1 Pedestrian	20,994
Number of Bounding Box Labels	32,457
Number of Pose Labels	27,454
Number of Tracks	4,222

Table 1: Dataset Statistics

A state-of-the-art algorithm for fine-grained classification was tested using the dataset. The results are reported as a useful performance baseline. The dataset is split into a training/validation set containing 4 videos, with the remaining 3 videos forming the test set. Since each video was collected on a unique day, different images of the same person **do not** appear in both the training and testing sets.

The fine-grained categorisation benchmark uses 'pose normalised deep convolutional nets' as proposed by Branson *et al.* [1]. In this framework, features are extracted by applying deep convolutional nets to image regions that are normalised by pose. It has state-of the-art performance on bird species categorisation and we believe that it will generalise to the CRP dataset. Results can be found in Figure 2

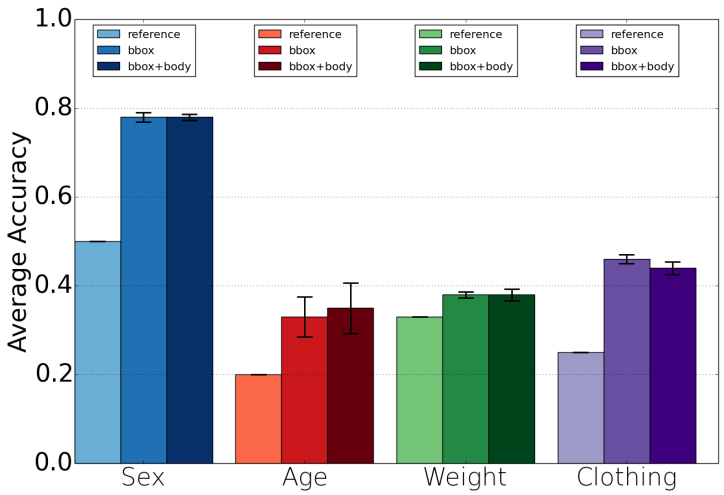

Figure 2: **Fine-grained classification results.** We report the mean average accuracy across 10 different train/test splits, for each of the subcategories in CRP, using the method of [1]. Average accuracy is computed assuming that there is a uniform prior across the classes. The reference value for each subcategory corresponds to chance. The results suggest that CRP is a challenging dataset.

A novel feature of our dataset is the occlusion labelling of the keypoints. Exploiting this information may be the first step towards improving performance for fine-grained classification. Using temporal information is another alternative. Most pedestrians in CRP appear multiple times over large intervals of time. We are planning on adding an identity label for each individual, to make our dataset useful for studying individual re-identification from a moving camera.

[1] Steve Branson, Grant Van Horn, Pietro Perona, and Serge Belongie. Improved Bird Species Recognition Using Pose Normalized Deep Convolutional Nets. In *BMVC*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.