WNet
Joint multiple head detection and head pose estimation from a spectator crowd image

Yasir Jan, Ferdous Sohel, Mohd Fairuz Shiratuddin & Kok Wai Wong
Murdoch University, School of Engineering and Information Technology
{Y.Jan, F.Sohel, F.Siratuddin, K.Wong}@murdoch.edu.au

Introduction
Research in crowd analysis is focused on various aspects such as crowd counting, body detection, head detection (HD), and head pose estimation (HPE). Previous HD and HPE techniques have various limitations:
- Some techniques perform single task only, i.e. either HD or HPE.
- Some techniques perform joint HD and HPE but for single head images only.
- Some techniques perform HD on multiple heads, but perform HPE on single cropped heads only.

This paper proposes a technique which performs joint task of HD and HPE of multiple low resolution heads in a crowd image.

Contributions
1. To the best of our knowledge, the proposed technique is the first CNN based technique to perform joint HD and HPE of multiple heads.
2. It does not need each head image to be cropped and passed through the network individually for HPE.
3. It accurately performs HD and HPE of more number of heads using lesser number of images.

Method
The proposed technique is based on image to image transformation. It first converts a cluttered crowd image into a simplified image and then extracts information of head bounding box and head pose.

The generated HRM image will have all the non head pixel values suppressed to 0. In CCH image the head centers are marked with 10 sized white pixels, and the rest of the head region is single color coded depending on the head pose.

Color codes: Head pose right, left, front, down and away are coded with colors R, G, B, RG, and RB respectively.

Grouping is done in following sub-steps:
- White pixels are located in the output image, which identifies the head centers.
- Noise is removed by averaging the centers which lie within the average head size distance.
- All the colored pixels lying within the average head size distance are grouped together as one head.
- The location of the colored pixels gives the head bounding box region limits.
- The majority color of the pixels within the region identifies the head pose.

Dataset
The Spectator Hockey (S-HOCK) dataset [2] is used for the experiments. Total videos = 25 (5 camera views ×15 different matches).

Single video features:
- Duration = 1 sec, 590 annotated frames, Resolution = 1280 × 720
- Data split: 20%, 0% and 11 videos are used for training, validation and testing respectively.

Experiments
Only one frame out of 10 consecutive frames are selected for experiments to reduce redundancy.

Training frames: 1850 frames sliced into 20 subframes (each Total 37200 subframes).

Testing frames: 10721 frames sliced into 20 subframes (Total 214420 subframes).

The proposed WNet architecture consists of 2 UNets, each one of them is trained separately.

UNet block 1: Training is done for 30 epochs using 3720 image pairs.

UNet block 2: Training is done for 10 epochs using 3720 image pairs and 3720 horizontally flipped image pairs.

Results
Output is tested for multiple HD IoUs (IoU > 0.3, 0.4, 0.5, 0.6, 0.7). The precision, recall and F1 score of head bounding boxes for varying IoU is calculated. It achieves more than 0.6 precision, recall and F1 score for HD (IoU > 0.5).

HPE performed in benchmarks [2] and proposed WNet have different protocols.

Benchmarks HPE:
- Tested on 34949 single head cropped images.
- WNet HPE: Tested on fewer number of frames (1023 frames / 20460 subframes), but more number of heads i.e. 15095 which is 4.4 times more than the benchmark number of heads (34949).

Analysis
HPE results are discussed.
- Head count of accurate HPE using WNet is greater than benchmarks.
- Benchmarks test single head at a time, while WNet tests multiple heads in a single pass.
- WNet uses lesser testing data frames and estimates higher number of accurate head poses.

Conclusion
- The proposed WNet architecture can perform joint multiple HD and HPE.
- Results show that the proposed technique uses lesser number of frames and accurately estimates pose of more number of heads.

Future Work
- Results could be improved using head segmented datasets.
- Color codes could be correlated to the head pose angle, to extend the range.

References