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.Shiratuddin, K.Wong}@murdoch.edu.au

Murdoch UNIVERSITY

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 proposed pipeline utilizes two of the UNet [1] blocks for image conversion, and follows the following steps.

1. A UNet converts a crowd image into a head-region-masked (HRM) image.
2. Another UNet generates a color coded head (CCH) image from grayscaled HRM image.
3. Based on the colors in CCH image and Euclidean distance, the head centers, head bounding box regions and their poses are identified.

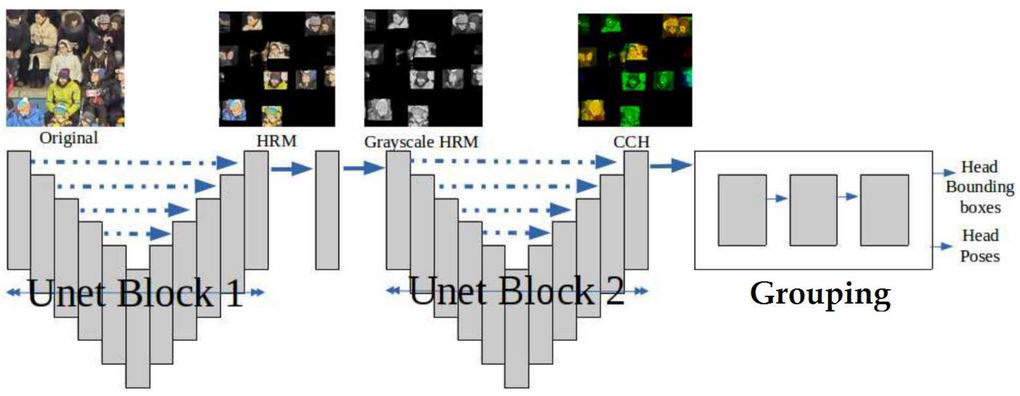

Figure 1: WNet Pipeline.

The generated HRM image will have all the non head pixel values suppressed to 0. In CCH image the head centers are marked with 3×3 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 GB channels 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 : 75 (5 camera views $\times$ 15 different matches).
Single video features : Duration 31 sec, 930 annotated frames, Resolution 1280×1024
Data split : 2, 2 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 : 183 frames sliced into 20 subframes each (Total 3720 subframes).
Testing frames : 1023 frames sliced into 20 subframes each (Total 20460 subframes).
Testing heads : Total count is 155085, with average width and height of 36.46 and 40.69 pixels.

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.

Figure 2: S-HOCK training input and output.

Results

Output is tested for multiple HD IoUs (IoU $> 0.3, 0.4, 0.5, 0.6, 0.7$). The precision, recall and f1score 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).

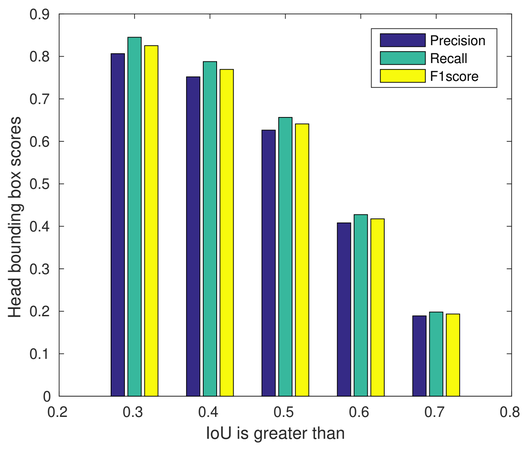

Figure 3: Head bounding box results.

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

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. 155085 which is ≈ 4.4 times more than the benchmark number of heads (34949).

Method	Avg. Accuracy 1	Accurate HPE headcount	Method Type
Orozco	0.368	≈ 12861	Only single HPE
WArCo	0.376	≈ 13140	Only single HPE
CNN [2]	0.346	≈ 12092	Only single HPE
SAE [2]	0.348	≈ 12162	Only single HPE
WNet (IoU = 0.3)	0.321	$\approx$ **39825**	**Joint multiple HD and HPE**
WNet (IoU = 0.4)	0.323	$\approx$ **35064**	**Joint multiple HD and HPE**
WNet (IoU = 0.5)	0.325	$\approx$ **30241**	**Joint multiple HD and HPE**
WNet (IoU = 0.6)	0.337	$\approx$ **20905**	**Joint multiple HD and HPE**
WNet (IoU = 0.7)	0.34	≈ 10545	Joint multiple HD and HPE

Table 1: HPE accuracy and headcount.

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

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[2] Francesco Setti, Davide Conigliaro, Paolo Rota, Chiara Bassetti, Nicola Conci, Nicu Sebe, and Marco Cristani. The s-hock dataset: A new benchmark for spectator crowd analysis. *Computer Vision and Image Understanding*, 159(Supplement C):47 – 58, 2017. Computer Vision in Sports.