Ground-truthing Large Human Behavior Monitoring Datasets

Tehreem Qasim[1] Robert B. Fisher[2] Naeem Bhatti[1]

1- Department of Electronics, Quaid i Azam University, Pakistan

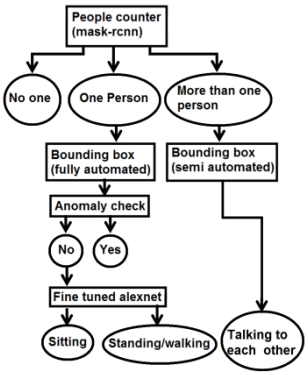
2- School of Informatics University of Edinburgh, UK

Introduction

- Ground-truthing video data is time-consuming
- We exploit: modern detectors and behavior classifiers, plus specialized consistency checks
- Applied to 1 FPS video of small numbers of people
- Reduced labeler clicks by 99+%
- Applicable to behavior monitoring

Overview of method

- Ground-truthing method for large video datasets.
- Mask-rcnn is used as a people counter
- Motion based checks to correct mask-rcnn errors
- Bounding box extraction based on count of people in a frame
- Anomaly detection based on inactivity analysis
- Behavior labeling using a fine-tuned alexnet



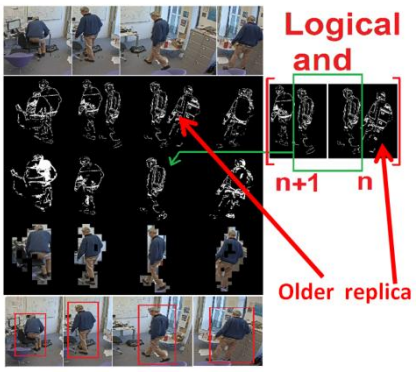
Block diagram of the ground-truthing frame work.

Mask-rcnn based people counter

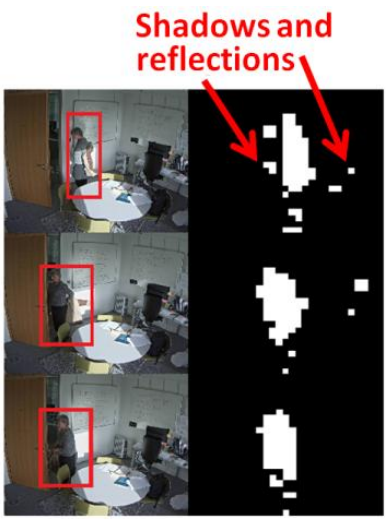
- Initial count of people in video frames obtained from mask-rcnn
- Motion based semi-automated checks to detect and correct errors
- Special tests to correct errors or flag for manual intervention

Bounding box extraction

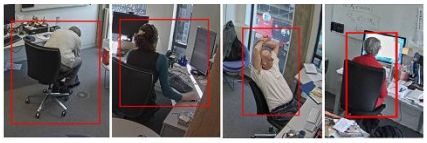
- For single person frames motion (difference of frame gradient) based method to draw bounding boxes



Top row: color frames (day 4, frames 1603-1606)
2nd row: unwanted previous frame replica of a person
3rd row: Logical and operation to get rid of older replica
4th row: Motion cells
5th row: Bounding boxes around motion cells



Shadows and reflection problem solved by retaining the largest connected component of motion cells



- Specialized methods for multi-person frames

Anomaly detection

Behavior labeling

- Label "*talking to each other*" assigned to frames with two or more people
- Single person frame classes: Sitting and standing/ walking
- A fine-tuned alexnet CNN used for classification
- Changes in labels manually verified

Experimental results

- **New office dataset recorded at 1 fps:**

Table 1. People counter results

Office	Manually verified true detected state changes	Manually corrected false detected state changes	Total automatic corrections	Total frames
01	274	213	7,891	236,651
02	19	15	252	54,721
03	99	41	6,251	77,628
04	87	32	8,089	87,715
Total	479	301	22,483	456,715

- Number of clicks required for **people counting** reduced to 0.71% (479 verifications and 301 corrections out of 456,715 frames
- Clicks for **single person bounding boxes** reduced to 0% (for 134,110 single person frames)
- Clicks for **more than one person bounding boxes** reduced to 4.02% (6,178 box initializations out of 72,650 frames requiring total 153,807 box initializations)
- Clicks for behavior labeling reduced to 0.66% (705 verifications and 953 corrections out of 249,955 single person frames)

Conclusions

- Algorithm driven approach reduces 99+% of clicks
- Combination of standard and specialized algorithms
- Applicable to low-frame rate and sparse person-watching video
- Similar reduction in click rate for 15 FPS video

Acknowledgement:
We are thankful to HEC Pakistan for funding the visit of Mr. Tehreem Qasim to the School of Informatics, University of Edinburgh under its IRSIP program.

Office dataset is available at the following link;
http://homepages.inf.ed.ac.uk/rbf/OFFICEDATA/