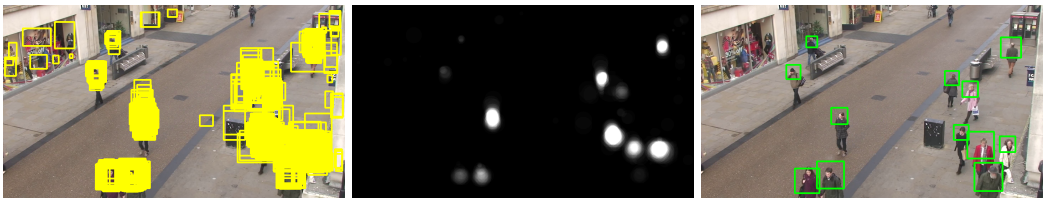
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2016
http://excel.fit.vutbr.cz

Multiple-Person Tracking by Detection

Jakub Vojvoda*



Abstract
Detection and tracking of multiple person is challenging problem mainly due to complexity of scene and large intra-class variations. In this paper, I present a novel on-line method for multiple person tracking based on tracking-by-detection approach. An object tracking component is deployed to increase the performance of the method and decrease the number of detector failures. Furthermore I use a fusion component to associate the responses of the detection and tracking components. The proposed system was evaluated on available datasets and the results shows that it is suitable to use for this task.

Keywords: person detection – body parts detection – responses association — object tracking

Supplementary Material: Demonstration Video

*jakub.vojvoda@gmail.com, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Person detection and tracking is one of the challenging problems in computer vision. Difficulty of this problem is caused mostly by the large variations of scale, appearance, viewpoint, articulation and occlusions. This task is important for many applications, such as surveillance, human-computer interaction, or behavior modeling.

Existing approaches differ in many aspects and with the increasing computational power there are increasing number of approaches using learning or detection-tracking cooperation. However, the person detection and tracking still remains an active research area and recent scientific results [1] and increasing computational power indicate the possibilities and directions of the future research.

In this paper, I present a robust on-line method for detecting and tracking of multiple people in a scene from one static camera. The proposed method is based on a tracking-by-detection approach and cooperation of a detection and tracking part. The detection part is based on a combination of multiple features and models that allow increasing a detection rate while reducing the false positive responses. A problem with partial occlusions is handled by using a body part detection, respectively an upper body detection.

Another component of the tracking system is multi-object tracker. The component is deployed to increase the performance and decrease the number of detection failures in a run-time. In order to control the tracking component, an association of the detection and tracking responses is presented. The association is part of fusion component and is based on a distance metric and percentage of overlap.

The described method was evaluated using the existing datasets and metrics. The results are presented and shows that this approach is suitable for detecting and tracking of multiple people.

2. Related works

During the last years, multiple-person tracking become an active field of research and much effort was put to

solve the problem. Numerous methods for detecting and tracking of people were proposed and differ in accuracy, stability and computational cost.

Histogram of oriented gradient, introduced by Dalal and Triggs [2], is one of the feature descriptors designed for person detection. The basic idea is that the local object appearance can be characterized by distribution of local gradient orientations even without precise knowledge of corresponding gradient positions. An object detection is based on sliding window in which the feature vectors are extracted and then used for object/non-object classification using a linear support vector machine (SVM).

Felzenszwalb et al. [3] enriched the previous model using a star-structured part-based model defined by a root filter and a set of part filters with associated deformation models. Each part of the model captures local appearance properties of the object. An object detection system is based on mixtures of this multiscale deformable part models trained using a discriminative method.

An approach which combines a shape information and a texture information was proposed by Wang et al. [4]. The shape information is described using the HOG features and the texture information using the cell-structured LBP (Local Binary Patterns) features. The occlusions are handled by using global and part detectors and by constructing an occlusion likelihood map which is then segmented by a mean-shift algorithm.

Andriluka et al. [5] extend one of the state of the art detectors to an articulation and limb-based detection approach. They detect approximated articulation of person based on local features that model the appearance of individual body parts. The possible articulation and temporal coherency within a walking cycle is modeled using a hierarchical Gaussian process latent variable model.

Many works deal with the object tracking mainly addressing the requirements for computational cost and precision. A learning method for long-term tracking of single object was proposed by Kalal et al. [6]. The detector localizes an object instance and corrects the tracker if necessary. The learning component initializes the detector and updates it in real-time.

Henriques et al. [7] presented a method for tracking a single object based on correlation filter and HOG (Histogram of Oriented Gradients) features instead of raw pixels. The tracking is formulated as a regression problem for correlation filter learning. A discriminative classifier is trained with sample patches around the object at different scales and translations.

3. Tracking system overview

Detection of multiple people is a difficult problem, but recently proposed methods and approaches shown possibility of detecting people even in crowds and scenes with partial occlusions [3, 8] but a problem with false positive detections still remains. I try to design the detector to cope with this problem and therefore the parts of the detection are based on more than one methods.

The proposed system consists of three main components: detection, tracking and fusion. For the detection, I use three types of different detectors and models: the deformable part detector [3], the detector based on HOG (Histogram of Oriented Gradients) features [2] and detector based on Haar-like features [9]. Scores of detection responses are normalized and associated using a local maxima finding approach. The tracking part is based on kernelized correlation tracker [7]. The fusion component associates the detection and tracking responses and controls the tracker. The overview of the system can be seen in Figure 1 and it will be described in details in the text below.

3.1 People detection

The detection component is based on combination of two approaches: the full body detection and upper body detection. The full body detection provides much information but the detector is not able to handle the articulated deformations of people and partial occlusions. Therefore I use the upper body detector to tackle this problem.

The **full body detector** consists of combination of deformable parts detector [3], HOG detector [2] and Haar [9] detector. The deformable part model was trained using full body images from INRIA dataset and the pre-trained model was used in framework for combined pedestrian detection [10]. The models used within the HOG detector and Haar detector are part of the OpenCV library [1]. I set a low detection threshold to obtain all true positive detections but also a large amount of false positive detections.

The **upper body detector** is based on deformable parts detector [3] and Haar-like features detector [9] which are trained on a head-shoulder images. The pre-trained model for deformable parts detector can be found in framework for pedestrian detection [10] and the upper body model for Haar detector is part of the OpenCV library [1]. The process of detection is the same as in the previous case where the detection responses from all detector are obtained using low threshold value.

[1] http://opencv.org/

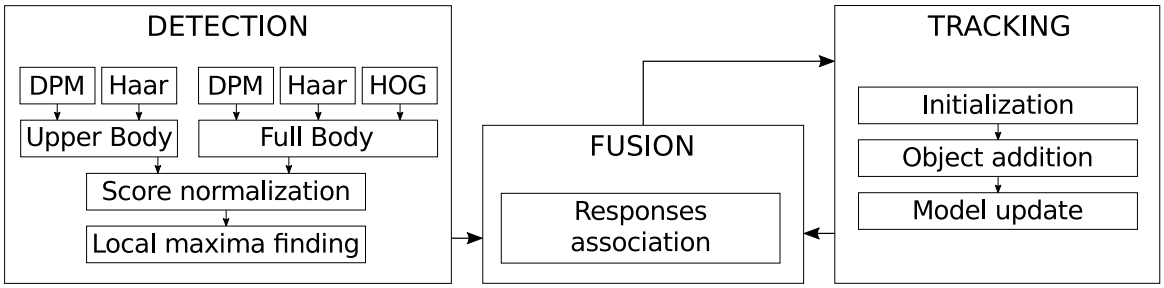


Figure 1. Proposed system overview

To deal with different sizes of the full body and upper body detections, the responses obtained from the full body detector are normalized to 40% of its height. This results to a large number of detection responses, as can be seen in first image of Figure 2.

An experimental analysis of the detection responses shows that the responses with extremely different sizes are most likely false positives. Due to this experiment and the fact that the detection process results in a very large number of detection responses, I filtered out 3% of detections with the smallest and the largest sizes.

Since the output of these individual detectors differ significantly, a **normalization of the scores** is required. The output scores of detection responses are normalized using the standard score approach [10] for each object detector separately as

$$s_0 = \frac{s - \mu}{\sigma}, \tag{1}$$

where μ is mean and σ is standard deviation of detector scores. The normalization results to all score values are positioned around zero value. Then, the obtained responses are normalized to the same value range $\langle 0, 1 \rangle$ as

$$s_{norm} = \frac{s_0 - min}{|max - min|}, \tag{2}$$

where *min* is minimal score and *max* is maximal score of all responses obtained from already processed frames and s_0 is defined in Equation 1.

A higher number of detection responses with a higher sum of scores occur in regions where the people most likely appear. The regions correspond to local maximas in a confidence detection map (see example in Figure 2). The confidence in certain position of the map is computed as a weighted sum of all response scores obtained from full body and upper body detection.

$$\begin{aligned} c &= c_{upper} + c_{full} \\ c_{upper} &= \sum_{i=1}^{N} \sum_{j=1}^{M} w_i s_{ij} \\ c_{full} &= \sum_{k=1}^{P} \sum_{l=1}^{Q} w_k s_{kl} \end{aligned} \tag{3}$$

The calculation is shown in Equation 3 where c is the confidence in a certain position of the confidence map, c_{upper} and c_{full} are confidences of the upper body and full body detectors in the same position, s_{ij} is the score of response j from detector i normalized by Equation 2 and weighted by weight w_i of detector i.

The local maximas in the confidence map can be found using a non-maxima suppression. I use similar approach, but in contrast to the usual non-maxima suppression I search for local maximas using a round window and a smaller neighborhood.

The method of the **local maxima finding** is based on assumption that the local maxima is greater than other values in the same window. If the local maximum is found in the neighborhood centered around the candidate, the confidence is computed as a sum of values in the round window divided by an area of the window. The found local maximum with the confidence above a threshold is then used as a result of the detection.

3.2 Multiple object tracking

I deploy an object tracking component to increase the performance and decrease the number of detector failures. For this purpose the kernelized correlation filter [7] was modified to track multiple objects simultaneously.

The tracker is initialized on the first frame using all regions obtained from the detection. New object to track is added based on the fusion component decision and only if the detection response without corresponding tracking response exists.

[2] image taken from the SUN Database (http://groups.csail.mit.edu/vision/SUN/)

An object is removed from tracking in case the corresponding detection response does not exist for n frames and at the same time, the object is not moving for m frames. Therefore the tracking component has information about the last locations and information about associations of the detection responses with the given tracking response. Finally, in each frame the position and history of tracked objects is updated.

3.3 Fusion of different responses

In order to find new object that are not tracked yet, it is necessary to associate the detection and the tracking responses. The responses are associated using two simple metrics: Euclidean distance and a percentage of overlap.

Let t be a tracking response in location t_c and let d be a detection response in location d_c. The response t and d belongs to one object if

$$\begin{aligned} dist(t,d) &< dist_{max} \quad \text{AND} \\ overlap(t,d) &> overlap_{min}, \end{aligned} \tag{4}$$

where the distance measure and the percentage of overlap is defined as

$$dist(t,d) = \sqrt{(t_{cx}-d_{cx})^2+(t_{cy}-d_{cy})^2} \tag{5}$$

$$overlap(t,d) = \frac{t \cap d}{t \cup d - t \cap d}. \tag{6}$$

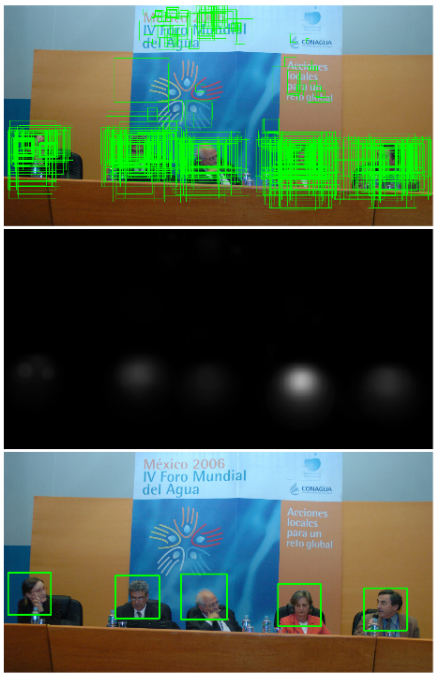


Figure 2. Example of detection steps[2](from up to down): all detection responses, confidence map and results of the described local maxima-finding approach

The detection response d is added to track only if there is no associated tracking response t which satisfies the Equation 4. In practice, this corresponds to finding detection responses which not satisfied the Equation 4 and adding them to the tracking component.

4. Performance evaluation

In pilot experiments, I was interested in accuracy evaluation of the proposed method in order to compare this method with other algorithms. Results of the described method were collected and evaluated using Town Center Dataset [11].

The dataset contains video of the busy town street from one static camera. The video is five minutes long and has 71500 hand labeled full body annotations, with average of 16 people visible at any time.

Figure 3. Sample results obtained for the Town Center dataset using the proposed method where the size of detection responses was adjusted to full body size

For evaluation, I used a criteria of PASCAL VOC challenge where the detection with overlap larger than half with annotation is considered as a true positive. The sample frames are shown in Figure 3 and the evaluation results are shown in Table 1. The precision and recall is defined as

$$precision = \frac{tp}{tp+fp} \tag{7}$$

$$recall = \frac{tp}{tp+fn} \tag{8}$$

where tp, fp and fn are numbers of true positive, false positive and false negative responses. The results of other methods were obtained from the publication [11], in which the dataset was presented.

In the experiments, I used weight w value equal to 1 for each detector (see Equation 3) and the threshold value for obtaining detection equal to 10 (see Section

	precision	recall
Proposed method	**87.9%**	**66.2%**
HOG detector	82.4%	72.3%
Method from [11]	82.0%	79.0%

Table 1. Performance evaluation and the comparison of the proposed approach with other methods

3.1). The maximum distance $dist_{max}$ was set to 30 pixels and minimum overlap $overlap_{min}$ to 0.3% which caused a low number of false detection. By using a lower value for the distance threshold $dist_{max}$ and a larger value for the overlap threshold $overlap_{min}$, a higher recall rate can be obtained but precision rate will drop due to a higher number of false positive responses.

The results shows that the presented method can be used for tracking of multiple people in a scene from one static camera. The method achieved a high precision rate which is caused by a small number of false alarms. The recall rate of the proposed method is inferior to the recall rates of the compared methods due to used fusion thresholds and therefore a larger number of false negative detections. Furthermore, the results showed that by using fusion component even with simple metrics such as the Euclidean distance and the percentage of overlap, it is possible to achieve good results.

5. Conclusion

In this work, I presented the method based on tracking-by-detection approach. Using the different models and object detectors I proposed the person detector which is capable of detecting partially occluded people. I deployed a state of the art object tracker to increase the performance and decrease the detector failures. In order to find new objects that are not tracked yet, the detection and tracking responses are associated using the fusion component. The component controls the tracking by adding new objects and removing them when necessary.

The proposed system was implemented and evaluated on the dataset of busy town street. The method achieved a 87.9% precision rate and a 66.2% recall rate. The results showed that described approach is suitable for multiple-person detecting and tracking in a video sequences.

Acknowledgements

I would like to thank my supervisor Ing. Vítězslav Beran, Ph.D. for his help and support.

References

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *Computer Vision for Road Scene Understanding and Autonomous Driving*, September 2014.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, June 2005.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[4] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 32–39, 2009.

[5] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[6] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1409–1422, July 2012.

[7] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *CoRR*, 2014.

[8] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[9] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[10] Floris De Smedt and Toon Goedemé. Open framework for combined pedestrian detection. In *VISAPP (2)*, pages 551–558. SciTePress, 2015.

[11] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2011.