Asian Journal of Applied Science and Technology (AJAST)
(Open Access Quarterly International Journal) Volume 2, Issue 2, Pages 453-458, April-June 2018

Utility Person Detection and Multi-View Video Tracking Annotation Model

Keerthika S[1] and Thiruvenkatasuresh M.P[2]

[1]*PG Student, Department of Computer Science and Engineering, Excel Engineering College, Komarapalayam, India.*
[2]*Associate Dean, Department of Computer Science and Engineering, Excel Engineering College, Komarapalayam, India.*

Article Received: 01 March 2018 *Article Accepted: 09 April 2018* *Article Published: 28 April 2018*

ABSTRACT

This paper proposes a generic methodology for the semi-automatic generation of reliable position annotations for evaluating multi-camera people-trackers on large video data sets. Most of the annotation data are automatically computed, by estimating a consensus tracking result from multiple existing trackers and people detectors and classifying it as either reliable or not. The proposed framework is generic and can handle additional trackers. We present results on a data set of ~6 h captured by 4 cameras, featuring a person in a holiday flat, performing activities such as walking, cooking, eating, cleaning, and watching TV. When aiming for a tracking accuracy of 60 cm, 80% of all video frames are automatically annotated. The annotations for the remaining 20% of the frames were added after human verification of an automatically selected subset of data. This involved ~2.4 h of manual labor. According to a sub-sequent comprehensive visual inspection to judge the annotation procedure, we found 99% of the automatically annotated frames to be correct. We provide an exploratory study for the multi-target case, applied on the existing and new benchmark video sequences.

Keywords: Multi-camera tracking, semi-automatic annotation, performance evaluation, people tracking.

1. INTRODUCTION

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Nowadays, image processing is among rapidly growing technologies. It forms core research area within engineering and computer science disciplines too. In this project analysis a few fundamental definitions such as image, digital image, and digital image processing. Different sources of digital images will be discussed and examples for each source will be provided. The continuum from image processing to computer vision will be covered in this project. The latest digital cameras combined with powerful computer software now offer image quality that is comparable with traditional silver halide film photography. Moreover, digital cameras are also easier to use and offer greater flexibility for image manipulation and storage.

TABLE I

COMMON MULTI-CAMERA DATASETS USED IN TRACKING EVALUATION (V=VIEWS, L=LENGTH (MIN) - NOTE THE SHORT LENGTH OF MOST SEQUENCES AND THEIR SPECIFIC CONTENT)

Dataset	V	L	fps	Content and/or application
ETISEO [13]	4	18	12.5	Surveillance in an airport apron
PETS2001 [14]	2	3	30	Surveillance of people and cars
PETS2009 [15]	4	2	7	Surveillance in an University Campus
APIDIS [16]	7	1	25	Basketball game (Sports analysis)
UvA-T [5]	3	7.1	20	Surveillance in train platform
UvA-H [5]	4	7.6	20	People walking (Surveillance)
EPFL-L [17]	4	2	25	People walking around in a lab-room
EPFL-C [17]	3	5.2	25	Surveillance inn a University Campus
EPFL-P [17]	4	1.7	25	Surveillance in a passageway
ISSIA [18]	6	2	25	Football game (Sports analysis)

2. BACKGROUND AND RELATED WORK

In this paper we demonstrate and evaluate the approach on a dataset of 6 hours, taken from 4 views, with furniture present, persons performing different activities, involving different poses and under different light conditions. The consensus tracking result is based on 6 multi-camera trackers, using publicly available software and our own implementations. We quantify the manual labour involved in the proposed methodology by the fraction of video frames that has to be visually inspected and the time that this operation takes.

2.1. Related Work

In this section we present related work on the evaluation of tracking results. For methods that use position references, we limit our discussion to those where annotations are generated directly on video sequences, either manually or assisted by computer algorithms. Although we are aware of existing methods that generate position references using external sensors and different modalities (e.g. infrared [19], MoCap [20], inertial [21]), we do not include them here as such sensors are not always available or their use is limited to controlled environments.

Recent work on the evaluation of visual trackers focuses on defining objective metrics for accuracy, failure and robustness [22]–[28] and using these methods to estimate the performance of existing trackers [22], [29]–[31].For the purpose of generating such reference data, several manual annotation tools have been developed for videos from single and multiple calibrated views [8], [10], [11], [32]–[35]. However, these methods do not scale to long video sequences, as manual input is required in every frame. In [36], the editors of a recent special issue on ground truth collection discuss the importance of building efficient semi-automatic tools to generate ground truth for multi-media applications and computer vision research in general. In the following we limit our discussion to tools for positional annotation at the level of image bounding boxes and point references in world coordinates. Thus, we will not consider tools for pixel-wise label propagation in video, like those presented in [37] and [38].

2.2. Semi-Automatic Tools for Single Camera Annotation

Editing tools can help to significantly reduce human effort in semi-automatic ground truth generation. One such annotation tool is proposed in [39] and made available as a web-based collaborative platform [40]. Generic contour detectors and a tracker are used to reduce human effort for the annotation of bounding boxes in single camera video. However, their contour detection method works only on targets with stable shape and with boundaries that are easily distinguishable from the background, while in more cluttered scenes segmentation errors may cause the tracker to drift. Thus, although a well segmented target may help the tracker, the human has to manually edit or re-draw wrong contours in complicated scenes, which is a tedious task. This method was evaluated on the problem of annotating fish in a fish tank.

An extensive quantitative evaluation of video annotation with crowd-sourced marketplaces and an annotation tool called VATIC, both validated on MTurk, are presented in [7]. In order to reduce human input, humans are asked to manually annotate targets with bounding boxes at periodically selected keyframes. These bounding boxes are then

enforced as positional constraints in a dynamic programming framework, which also takes into account target appearance and a simple motion model. The authors propose an annotation cost measure which takes into account both human effort and computational complexity. However, the analysis in the paper is restricted to variants of one tracker which are compared in a single example. In a related contribution by the same authors [41] active learning is proposed to reduce the annotation time of the VATIC tool: an estimate of annotation uncertainty is used as reference to iteratively propose new keyframes to annotate manually. Methods that fuse the results of multiple algorithms for the purpose of semi-automatic annotation have been proposed recently. A tool that speeds-up the generation of ground truth is described in [42]. The tool is applied in annotations of faces, and combines different face detectors and a tracker for the automatic part of the method, and the ViPER tool [8] for the human intervention. The fusion of the different algorithms relies on the score given by the external detectors. Compared to our methodology, statistics of the errors are not taken into account to fuse the results of the different algorithms. Furthermore, the tool is designed for single camera sequences and there is no estimation of the necessary effort in the manual part of the method.

2.3. Related Work on Fusing Tracking Results

The idea of fusing the results of several trackers to increase tracking performance appeared in [12], [43]–[45], and [46]. Specifically, Zhong et al. [12] approach single-view object tracking as a weakly supervised learning problem, in which the outputs of several trackers are treated as noisy labels. The probabilistic framework for optimal integration of labellers proposed in [47] is used to jointly infer the object position and the accuracy of each tracker. A heuristic that measures agreement among trackers is used for training data selection and to update the trackers models. By sampling not only the state of the target but also the trackers involved in its computation, the method proposed in [43] generates proposals for the best combination of tracker components. The target position is then estimated from the selected combination. In [44], a disagreement-based approach calculates the weights of a linear combination of probability maps obtained from several trackers. Such a disagreement-based approach is also presented in [45]. In order to find an optimal target bounding box, a mea-sure of attraction in a 4 dimensional space (position and size) between bounding boxes from tracking fusion candidates and the trackers is maximized. A heuristic for bad trackers removal is proposed to improve the estimate of the bounding box. Other researchers [46] have proposed a symbiotic tracker ensemble framework for learning an optimal combination of results from several trackers. This combination is based on estimates of temporary consistency of individual trackers and correlation of pair-wise trackers. In another form of consensus tracking [48], a factorial hidden Markov model learns both the unknown trajectory of the target and the reliability of each tracker. . In this case, the consensus position is computed recursively using a particle filter in order to handle multiple hypotheses. The likelihood of each observation (position estimate output by a tracker) is based on the distances between the bounding boxes and the actual target state, where the scale parameter of the chosen distributions is proposed as a reliability measure for each tracker.

The aforementioned methods focus on building a better tracker rather than on presenting a framework for minimal-effort reference data generation by fusing the results of several trackers. However, such consensus tracking improves automatic annotations as well, as we will discuss in Sect. V-B. Note that all of these papers deal

with single camera tracking only. Moreover, they measure consensus in terms of degree of overlap between bounding boxes, which is only an indirect measure of real-world positional accuracy as the overlap is measured in pixel coordinates which do not have a one-to-one relationship with real-world coordinates. Also, most of these papers report results on small, albeit challenging sequences only, except by the work in [45], which uses the benchmark presented in [29].

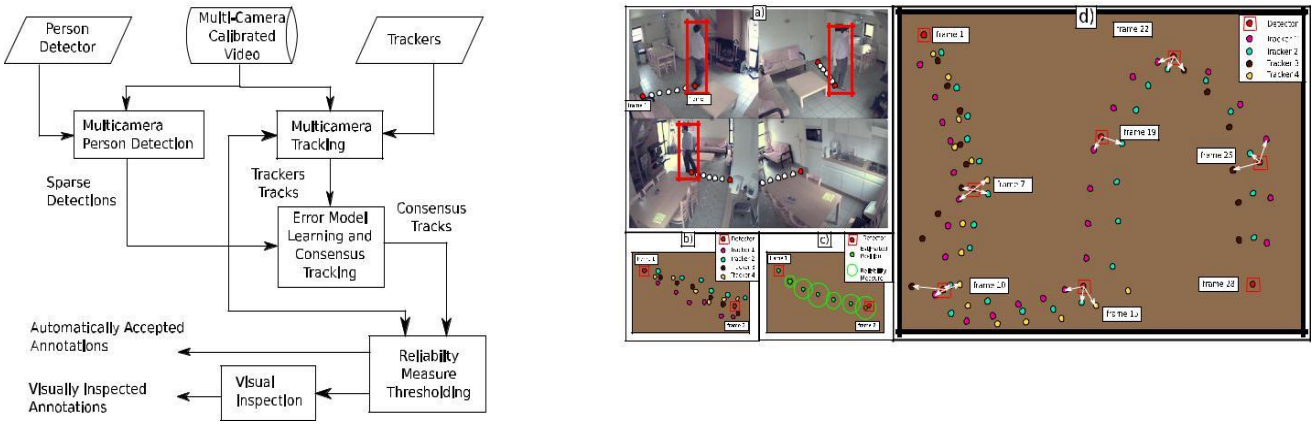

Fig. 1. Flow-chart of the method proposed for semi-automatic annotation.

2.4. Semi-Automatic Tools for Multi-Camera Annotation

A semi-automatic tool for annotating players in football video footage proposed in [18] generates proposal tracks using several algorithms: A foreground/background detector extracts blobs for target initialization; next a simple tracking strategy based on colour information and supported by heuristics, links the blobs over time. A graphical interface is used to edit the bounding boxes produced. Compared to our method, the main disadvantage of this tool is that the user has to browse the video frame-by-frame in order to check/correct the resulting annotations. The authors claim that the annotation tool can also be used for surveillance scenes, even though that type of video content is very different from football. While this is correct in principle, it requires adapting the heuristic rules to specific video content, which is not a trivial task.

Fig. 2. Illustration of the consensus tracker. (a) The white dots represent the real trajectory of the person. A highly reliable multi-camera person detector produces reference data for only some frames; two successful detections are shown (red dots at frames 1 and 7). (b) Several multi-camera trackers are initialised at frame 1 and track the person independently. (c) Consensus tracking result (green dots) and a reliability measure estimate (dotted green circles: the smaller the circle, the higher the consensus among trackers.).

Gathering multi-camera tracking statistics. The trackers are re-initialised at frame 28. In between initialisations, statistics on the position error (white arrows) are gathered by comparing to the output of the reliable person detector (red dot with red square). Note that although ideally a tracker would always output a position estimate, in practice trackers may lose the target (trackers 3 and 4 in this example). Additionally, some trackers may detect target loss and inhibit their output until new evidence is found (tracker 3). Thus, properties like robustness and accuracy are

often conflicting between each other and therefore it is desirable to collect tracking statistics of complementary trackers.

3. CONCLUSIONS

We proposed a novel method for semi-automated generation of positional annotations in calibrated multi-camera systems, targeted for long video sequences. A key contribution of this paper is a novel procedure to fuse the outcomes of multiple trackers that involves estimating the error statistics of the individual trackers (and their joint failure modes) by comparing their output to a reliable people detector. The consensus tracker based on these statistics results in more accurate tracking than averaging the trackers outputs, and the individual output of the trackers. Another novelty is that we proposed a method to estimate the time needed for the manual part of the proposed procedure. We also estimate the precision of the resulting semi-automatically annotated data set, under given values of desired accuracy.

We demonstrated the performance of our method with experiments on a multi-camera video dataset of about 6 hours duration, showing scalability in semi-automatic annotation for long multi-camera sequences for the first time. This leads us to the final contribution of this paper: the annotated data itself, which is available in two versions: the first version results from the proposed procedure; the second from the exhaustive inspection to create ground truth. Since annotated multi-camera data sets of such length are not yet reported in literature, we hope these annotated sequences will be useful to the research community.

REFERENCES

[1] V. Lepetit and P. Fua, “Monocular model-based 3D tracking of rigid objects: A survey,” Comput. Graph. Vis., vol. 1, no. 1, pp. 316–323, 2005. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” ACM Comput. Surv., vol. 38, no. 4, 2006, Art. no.13.

[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, “Recent advances and trends in visual tracking: A review,” Neurocomputing, vol. 74, no. 18, pp. 3823–3831, Nov. 2011.

[3] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. Van Den Hengel, “A survey of appearance models in visual object tracking,” ACM Trans. Intell. Syst. Technol., vol. 4, no. 4, Sep. 2013, Art. no. 58.

[4] C. Vondrick, D. Patterson, and D. Ramanan, “Efficiently scaling up crowdsourced video annotation,” Int. J. Comput. Vis., vol. 101, no. 1,184–204, Jan. 2013, doi: 10.1007/s11263-012-0564-1.

[5] D. Doermann and D. Mihalcik, “Tools and techniques for video performance evaluation,” in Proc. ICPR, vol. 4. Sep. 2000,167–170. Sorokin and D. Forsyth, “Utility data annotation with Amazon mechanical turk,” in Proc. CVPRW, Jun. 2008, pp. 1–8.

[6] Utasi and C. Benedek, “A multi-view annotation tool for people detection evaluation,” in Proc. VIGTA, 2012, pp. 1–6.

[7] G. Miller et al., “MediaDiver: Viewing and annotating multi-view video,” in Proc. 30th Conf. Human Factors Comput. Syst. Extended Abstracts, 2011, pp. 1141–1146.

[8] Zhong, H. Yao, S. Chen, R. Ji, T.-J. Chin, and H. Wang, "Visual tracking via weakly supervised learning from multiple imperfect oracles," Pattern Recognit., vol. 47, no. 3, pp. 1395–1410, Mar. 2014.T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin, "ETISEO, performance evaluation for video surveillance systems," in Proc. AVSS, Sep. 2007, pp. 476–481.

[9] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in Proc. PETS-Winter, Dec. 2009, pp. 1–6.

[10] De Vleeschouwer and D. Delannay. (2009). Basketball Dataset From the European Project Apidis. [Online]. Available: http://www.apidis.org/Dataset

[11] P. Fua, E. Tu, J. Berclaz, F. Fleuret, and E. Turetken, "Multiple object tracking using K-shortest paths optimization," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 9, pp. 1806–1819, Sep. 2011.