Video Based Group Tracking and Management

Américo Pereira[1,2], Alexandra Familiar[1,2], Bruno Moreira[1,2], Teresa Terroso[1,4], Pedro Carvalho[1,3], and Luís Côrte-Real[1,2]

[1] INESC TEC, Portugal
[2] Faculty of Engineering of the University of Porto, Porto, Portugal
[3] School of Engineering, Polytechnic Institute of Porto, Porto, Portugal
[4] The School of Management and Industrial Studies, Polytechnic Institute of Porto, Vila do Conde, Portugal
americo.j.pereira@inesctec.pt

Abstract. Tracking objects in video is a very challenging research topic, particularly when people in groups are tracked, with partial and full occlusions and group dynamics being common difficulties. Hence, its necessary to deal with group tracking, formation and separation, while assuring the overall consistency of the individuals. This paper proposes enhancements to a group management and tracking algorithm that receives information of the persons in the scene, detects the existing groups and keeps track of the persons that belong to it. Since input information for group management algorithms is typically provided by a tracking algorithm and it is affected by noise, mechanisms for handling such noisy input tracking information were also successfully included. Performed experiments demonstrated that the described algorithm outperformed state-of-the-art approaches.

Keywords: Video, groups, tracking, management.

1 Introduction

Video object tracking has been an increasingly growing area of research, mainly in video-surveillance scenarios, but with applications in many other areas. In nearly all of these scenarios we can have groups of people. Due to the proximity of people in groups, its hard to understand the movement of each individual, and traditional detection and tracking algorithms tend to be less effective on these scenarios. Occlusions, unpredictable movements and merging/splitting of groups are just some associated problems. Group analysis and tracking can brings advantages, such as predicting the position of the persons in the group even under heavy occlusion. However, it also adds several challenges, including: the number of occlusions; temporal changes in the group structure; different individual trajectories within the group. An important challenge is the group definition itself. Correctly defining a group is a critical step for subsequent group handling.

This paper proposes enhancements to a group tracking and management algorithm with the main focus of increasing robustness. This translates into a

new algorithm that, as the base one, receives individual tracks as input and assists in the detection, creation and management of groups but enables increased performance, especially in the presence of tracking errors. Results show that it outperforms state-of-the-art proposals, as well as the original proposal, even when introducing common tracking errors on the input data.

The remaining of this paper is structured as follows. A brief literature review is presented in section 2. The proposed algorithm enhancements are presented in section 3 along with a description of experiments that were performed. The datasets and metrics used for testing and evaluation are described in section 4. Finally, the conclusions and observations, as well as future work is presented in section 5.

2 Group Concepts and Tracking

Detection is typically the basis of any tracking system since it is responsible for obtaining representations of objects of interest to be tracked. A survey of recent algorithms dedicated to person detection is present on [1, 2]. For a more in-depth study of the underlying principles, techniques and algorithms related to video object tracking, the reader is referred to some of the many existing surveys. Aggarwal and Ryoo [3] provided a recent update to their previous surveys describing a vast number of publications with a special focus on the interpretation of human motion. Another survey was presented by Smeulders et. al. [4], where a set of nineteen tracking algorithms were thoroughly evaluated and experimented.

The concept of group is viewed socially as a set of people who are in spacial proximity and interact with one another with a common goal [5]. While this is a good principle, it's not enough to identify a group. Other factors, such as size, duration, velocity and structure are also fundamental in defining and managing groups. When considering a group tracking scenario, three entities can be defined: person, group and crowd. An entity is considered to be a crowd when there is a set of people dense enough, that it becomes impossible to distinguish between individuals [6]. Some authors proposed treating a group as a set of individual entities [7] when individual segmentation is possible and see the group as a single entity [8], otherwise. Work related to group tracking is present in [9] in which the counting of pedestrians moving in groups is addressed. The estimation of the number of people present in a group is based on projection information, enabled by accurate camera calibration information. The approach presented in [10] creates a framework that includes both detection and tracking for individuals and groups with sharing of information between them. The authors used a Decentralized Particle Filter [11] to model individuals with a position and speed; for groups, a match was made between the groups and the individuals in it. In [12], the authors focused on group tracking and behaviour recognition in long sequences. The proposal started by segmenting the people in the scene, detecting the blobs, following the several objects, grouping them in more complex entities and using that information to detect events. A common problem when dealing with groups is the need to handle the exit and re-entering in the

scene of its members. This adds the difficulty of deciding whether it should be considered the same group or not. In [13], a re-acquisition process was proposed using a descriptor based in co-variance matrices to model the group and deal with these situations.

3 Group Management and Tracking

3.1 Base algorithm

The proposed solution is an evolution of the state-of-the-art algorithm described in [12] and was assessed under the same conditions. The inputs of the base algorithm are intended to be the results from a people tracking algorithm, which are first filtered in order to reduce errors. The algorithm uses a metric named Group (In)Coherence (GI) [12], which represents the probability of a set of people being a group. It is defined as the average of distance between individuals ($\bar{d}$), standard deviation of speed (σ_{speed}) and direction (σ_{dir}), each weighted differently (see Equation 1).

$$GI = w_1 \cdot \bar{d} + w_2 \cdot \sigma_{speed} + w_3 \cdot \sigma_{dir} \ . \tag{1}$$

These values are measured over a time window T; a common value of T is 20 frames, since its sufficient time for trajectories to be long enough without adding too much delay to the system. The weights w_1, w_2, w_3 were normalized.

The algorithm consists of 4 phases: creation, update, split/merge and termination. In the creation step, trajectories of objects are analysed through the T time window and a group is created if objects are close to each other and the associated GI is valid. The update step consists on validating the GI of a group through the time window. The split step is responsible for splitting a object/objects that consistently move away from their group, resulting in a two groups. As the name suggests, the merge step is responsible for merging two groups if they are linked. Finally, the termination step erases empty groups and groups without new physical objects for a long period of time.

Identification (ID) management is performed by considering that the group ID is given by the set of people in the group. Finally, a group termination module is employed to delete any objects that have not been present in the scene for an extended period of time, which include both people and groups.

3.2 Proposed enhancements

The base algorithm uses GI as the metric for group classification decision. The authors proposed the following weights: $w_1 = 0.7$, $w_2 = 0.15$ and $w_3 = 0.15$, which reflects the importance of the average distance between people. While using the GI criterion with a single threshold provides good results, it may cause unnecessary fragmentations. We propose several changes so that the robustness and performance of the base algorithm is augmented.

Hysteresis We argue that using hysteresis can reduce the number of fragmentations and propose two different threshold values for the splitting and merging events, as they should only happen when the algorithm is fairly confident, even if that delays the decisions. Merge events should only happen when $GI < t_{min}$ and, likewise, splitting events should only happen when $GI > t_{max}$, where t_{min} and t_{max} represent the hysteresis thresholds values. This avoids unwanted jitter in the decisions caused by excessive splitting or premature grouping.

Average Speed Instantaneous speed and direction represent the most immediate movement, but are noisy and often suffer from inconsistencies derived from tracking errors [12]. We propose the use of the average speed $s^*_{i,k}$ (Equation 2) in the previous t frames, for the calculation of the speed deviation.

$$s^*_{i,k} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{(P_{x,i,k} - P_{x,i,k-t})^2 + (P_{y,i,k} - P_{y,i,k-t})^2}\ , \tag{2}$$

where $P_{x,i,k}$ $P_{y,i,k}$ are the X and Y coordinates of person i in frame k. Experiments show that generally a value of 5 for T is sufficient to obtain good results.

Group Elements Distance When a group grows, individual distances will tend to increase; but the average distance, used in the base algorithm, can remain similar (see Figure 1a). To address this, we propose the use of distance d^*:

$$d^* = (0.75 \times \bar{d}_{center} + 0.25 \times d_{closest2})\ , \tag{3}$$

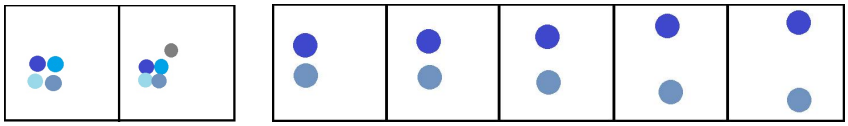
where $\bar{d}_{center}$ is the average distance of members to the group centroid, $d_{closest2}$ is the average distance to the closest two persons. The weights were empirically determined.

Angle and Direction of Movement The direction has been used to assess the type of movement, but in cases such as Figure 1b where the movement is only present in one direction, it may contribute to a wrong decision in the original GI formula as the direction displacement might not suffice. Even though the two represented persons are splitting, this only happens in a single axis. Equation 4 represents how the group is spreading from their movement angle. Equation 5 represents the smallest angle between the person and the group.

$$\theta_{deviation} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_{min}(i)\right)^2}\ , \tag{4}$$

$$\theta_{min} = min(|max(\theta_i, \bar{\theta}) - min(\theta_i, \bar{\theta}) + 180|, |\theta_i - \bar{\theta}|)\ , \tag{5}$$

where θ_i is the angle of movement of the person i and $\bar{\theta}$ is the mean angle of a group. We argue that the use of σ_{dir} to characterize the group movement is not



(a) (b)

Fig. 1: Example of scenarios. (a) Two situations that have similar average distances and the base algorithm may fail and merge all persons. (b) Simulated movement of people splitting in one axis.

enough and, as such, we propose the addition of a weighted contribution of the group spreading angle (see Equation 6).

$$\sigma^*_{dir} = 0.75 \times \sigma_{dir} + 0.25 \times \theta_{deviation} \ , \tag{6}$$

where σ_{dir} is the direction deviation used in GI. The weights were determined empirically.

Non-Linear Evaluation The previous proposed enhancements are alternative components for the calculation of GI (Equation 1). However, we argue that it is necessary to go further and change the way GI is used. We propose a non-linear formulation of GI, depending on the group motion (Equation 7).

$$GI = \begin{cases} 0.85 \times d^* + 0.10 \times \sigma^*_{speed} + 0.05 \times \sigma^*_{dir} \text{ if } \bar{s} < 0.2 \\ 0.60 \times d^* + 0.15 \times \sigma^*_{speed} + 0.25 \times \sigma^*_{dir} \text{ if } \bar{s} > 0.75 \\ 0.75 \times d^* + 0.15 \times \sigma^*_{speed} + 0.10 \times \sigma^*_{dir} \text{ otherwise} \ , \end{cases} \tag{7}$$

where $\bar{s}$ is the normalized average speed of the group and σ^*_{speed} is the standard deviation of speed between entities, calculated using the speed formula s^*_i. These presented weights were obtained empirically for each type of motion.

4 Results

Group tracking data can be obtained from well known datasets such as CAVIAR [14] or BIWI [15], they have annotated data for both individuals and groups. However, these datasets miss some specific group evolution situations. One dataset that contains prominent group social interactions and annotated data is the Friends Meet dataset [16], used in different proposals on the literature. As such, we performed the evaluation of both the base algorithm and the proposed changes using the subset of 13 sequences from the Friends Meet Dataset depicting real scenarios. These sequences contain interesting and difficult group situations and have associated reference information for objective assessment. The evaluation was two-fold: without and with noise. The following subsection (4.1) describes the noise addition process and its importance. Next the results are presented and discussed in subsection 4.2.

4.1 Addition of Noise

State-of-the-art group management algorithms are often assessed using noise free tracking data. While this has several challenges by itself, we also analyse the algorithms with noisy data resulting in a significantly challenging task. For the latter, we added typical tracking errors to the reference data, as characterized in [4]: localization errors (type I); false positives (type II); false negatives (type III). Three increasingly levels of noise were implemented, adding tracking errors of the three types. The levels represent the quantity of noise added, resulting in more difficult tracking data as the noise increases. To simulate errors of type I, we performed random perturbations on the localization and size of bounding boxes. By doing this we create a jitter in the localization of the detections and subsequently degrade the ability to track them. False positives were simulated by randomly adding bounding boxes with typical size and position and an associated identification across the videos. For the errors of type III, portions of existing tracks were randomly selected and cut from the data as a way to simulate miss detections. For all these perturbations, uniform distributions were used.

4.2 Proposal Evaluation

Traditionally, metrics such as precision and recall have been used in detection and tracking, but are not sufficient for evaluating groups; hence other metrics are needed. Track Fragmentation (TF) [17] is used to capture the number of discontinuities in the group trajectory when compared to the ground truth. Another useful metric is the Group Detection Success Rate (GDSR) [10], which represents the success hit rate of detecting groups. Therefore, the results of the algorithms were evaluated using precision, recall, GDSR and TF* ($TF^* = 1 - TF$). Figure 2a depicts the effect of noise in the initial algorithm. Before the addition of noise (first bar of the columns), the metrics are above 90% and the average fragmentation is 0.59. The fragmentation occurs because the GI value fluctuates around the threshold value. In the presence of tracking errors, the performance drastically degrades with the increase of the noise intensity. With this its noticeable that the base algorithm fails to handle severe errors. Figure 2b depicts the average results of the proposed algorithm with the increase of noise. The comparison clearly shows that our proposal achieves better performance. An additional experiment was performed, comparing the initial algorithm [12], the proposed algorithm and the state-of-the-art algorithm $DEEPER - JIGT$ [10]. Since for the latter only GDSR values were available, the comparison was made using this metric. Our proposal obtained the best results of 93%, while the initial algorithm and $DEEPER - JIGT$ obtained 81% and 88.46%, respectively. Its noticeable that the proposed algorithm modifications enable better performance, namely in the presence of noise inputs.

5 Conclusion

Object tracking in video is an unsolved problem, with proposals tending to focus on specific applications. In particular, tracking and management of groups has

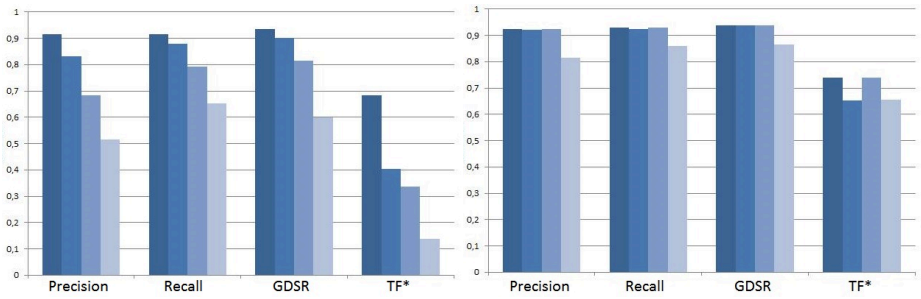


Fig. 2: Performance comparison of the initial algorithm in (a) and proposed algorithm in (b). The bars represent, from left to right, no noise and increasing levels of noise.

received significant less attention from the research community, and proposals still lack maturity and sufficient robustness.

This paper presents a group tracking and management algorithm that is an evolution of a state-of-the-art algorithm. Enhancements to the GI metric are proposed so that noisy tracking data and group dynamics have less impact in the decision criteria, resulting in a better group management. Evaluation was performed using well known sequences. The algorithms were tested in the same conditions and the addition of noise for the tracking data was also performed in order to simulate the effects of tracking in a real and uncontrolled scenario. The results show that the proposed method outperforms state-of-the-art algorithms, namely in the presence of noisy inputs.

The experiments reported show that our proposal enables a better and more robust performance and has the potential to be improved. Next steps will primarily include the preparation of additional realistic datasets and exhaustive experiments to improve the parametrization. Information about the scene, known a priori or automatically extracted, could also be used for an automatic adaptation of the parameters. For an even more in-depth and demanding assessment, the proposed algorithm should be integrated in a real people tracking algorithm and promote the exchange of information between them.

Acknowledgment

This work was partially funded by project "NORTE-01-0145-FEDER-000020" is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

1. Dollar, Piotr and Wojek, Christian and Schiele, Bernt and Perona, Pietro: "Pedestrian detection: An evaluation of the state of the art", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no.4, pp. 743–761, 2012.
2. Nguyen, Duc Thanh and Li, Wanqing and Ogunbona, Philip O: "Human detection from images and videos: A survey", *Pattern Recognition*, vol. 51, pp. 148–175, 2016.
3. J. R. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, pp. 16:1–16:43, April 2011.
4. A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1442–1468, 2014.
5. GAUQUELIN, Michel, and Françoise GAUQUELIN, "Dicionário de Psicologia: as idéias, as obras, os homens," *Paris: Centre d'Étude et de Promotion de la Lecture*, 1987.
6. S. J. Junior *et al.*, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.
7. D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 1187–1190.
8. S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE, 2007, pp. 1–6.
9. P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110, no. 1, pp. 43–59, April 2008.
10. L. Bazzani, M. Cristani, and V. Murino, "Decentralized particle filter for joint individual-group tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* IEEE, 2012, pp. 1886–1893.
11. T. Chen, T. B. Schon, H. Ohlsson, and L. Ljung, "Decentralized particle filter with arbitrary state decomposition," *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 465–478, 2011.
12. C. Gárate, S. Zaidenberg, J. Badie, F. Brémond *et al.*, "Group tracking and behavior recognition in long video surveillance sequences," *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on.* Vol.2. IEEE, 2014.
13. S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human reidentification by mean riemannian covariance grid," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on.* IEEE, 2011, pp. 179–184.
14. "Caviar dataset," `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`, accessed: 2015-02-08.
15. "Biwi dataset", `http://www.vision.ee.ethz.ch/datasets/index.en.html`, accessed: 2015-02-09.
16. "Friends meet dataset", `http://www.iit.it/en/datasets-and-code/datasets/fmdataset.html`, accessed: 2015-02-09.
17. F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Rio De Janeiro, Brazil*, 2007.