ICIC Express Letters
Part B: Applications
Volume **11**, Number **1**, January **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 85–91

FACE TRACKING AND DETECTION OF CHILDREN FACES IN OMNIDIRECTIONAL VIDEO

Emilien Briant and Keiichi Horio

Graduate School of Life Science and Systems Engineering
Kyushu Institute of Technology
2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan
briant.emilien282@mail.kyutech.jp; horio@brain.kyutech.ac.jp

Received July 2019; accepted October 2019

Abstract. *This work presents a long-term face tracking method based on existing tracking and face detection methods and introduces a tracking correction mechanism for an accurate long-term face and identities tracking in an omnidirectional video of children conversation. We introduce a three-step tracking correction mechanism, using face detection results to first apply a Jaccard coefficient check to identifying reliable tracking result, then a region of interest association to correct minor errors of tracking. Finally, we exploit the angular order of identities in the target videos to correct or recover the tracking of the remaining faces. The results of our method, validated on two tests videos, show that the proposed method achieves a much higher efficiency than uncorrected tracking and a good capacity to resist to challenging occlusions events.*

Keywords: Omnidirectional video, Face detection, Face tracking, Tracking correction

1. **Introduction.** Face detection and face tracking are essential for a number of applications and a first step for a subsequent analysis like face alignment, face recognition or face verification. In this work we focus our effort on face tracking in a video stream taken from an omnidirectional point of view in an unconstrained environment, depicting children discussing around a table. Long-term face tracking is a difficult problem in computer vision due to large appearance changes, and prolonged occlusions. Several studies are dealing with the problem of long-term tracking [1-3] to get the position of an object or a face when they appear on frame. The desired output for our work is the position of every children's face in the video, or in the case of occlusion, the best approximation of the children's heads position. Our scenario presents a particularly challenging setting with long facial occlusions, challenging poses with unconstrained roll, pitch and yaw angles, non-linear deformations due to the omnidirectional angle, and occasionally fast scale variations. Examples of normal situation compared with these kinds of challenges can be found in Figure 1. One of the main challenges is the capacity for our proposed method to resist to difficult events like long term occlusion of most of the frame by a face or a hand, as shown in Figure 1(e).

In our videos, the children are sitting around a table for a discussion, thus, considering the 360 degree angle of the camera, they will not leave the frame – they can however be occluded. They also stay seated in the same place throughout the discussion, which means they will not exchange position with one another, keeping the same order around the table, allowing us to use this order permanence in the proposed correction mechanism. Our work comes in the more general purpose of analysis of children behaviors through conversation videos and introduces a first step of accurate face and identities tracking. Our contributions in this paper can be summarized as follows.

DOI: 10.24507/icicelb.11.01.85

1) We propose a long term tracking method using off-the-shelf tracker on which we implement a tracking correction mechanism based on existing face detection methods for challenging face tracking in the context of omnidirectional videos of children in conversation.
2) We evaluate the accuracy of this method compared to non corrected tracking, and face detection algorithms.

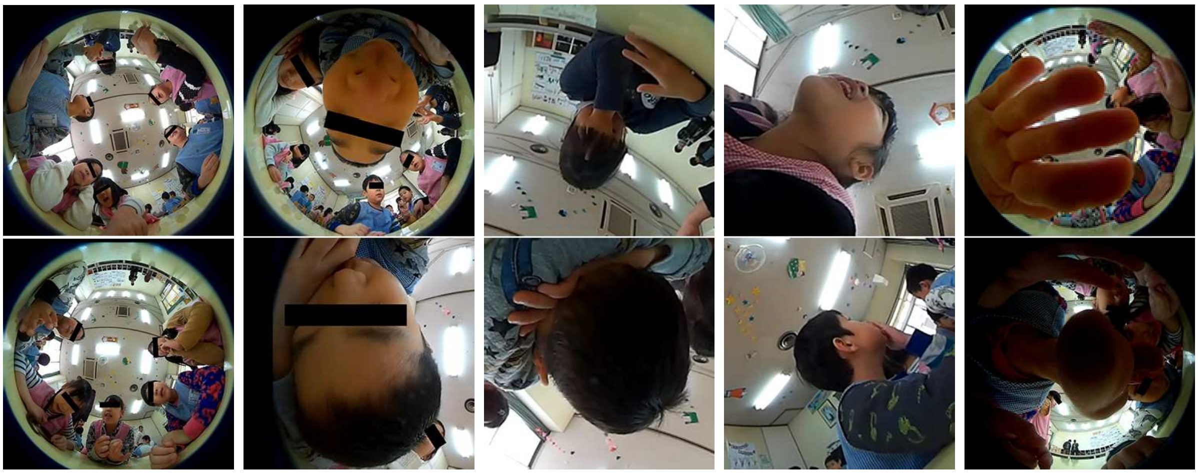

(a) (b) (c) (d) (e)

Figure 1. Examples of challenging situations that have an important impact on the performance of face detection and tracking: (a) normal situations, (b) deformations, (c) occlusions, (d) poses, (e) challenging events

The rest of the paper is organized as follows. Section 2 gives a brief introduction on related works on face detection and tracking, Section 3 details the mechanism we propose for correcting face tracking, Section 4 presents the results of our method on two test videos and evaluate the performances of such method compared to uncorrected tracking and face detection, and finally Section 5 concludes the paper.

2. Related Work.

2.1. Face detection.

The importance of face detection in a lot of different applications has made it an significant subject of research for the last few decades, starting with the seminal work of Viola and Jones [4], using rigid templates and boosting to detect faces in real-time. Since then, face detection has known an increase in efficiency, especially due to the emergence of Convolutional Neural Networks (CNN), that, when trained on large scale datasets, is able to learn more discriminating feature than hand-crafted filters. This growth is also due to the availability of large public datasets providing increasing challenges, mainly more and more unconstrained situations referred as “in-the-wild” face detection. Some of the most popular datasets for this kind of face detection are FDDB [5], WIDER FACE [6], UFDD [7], WIDER FACE being to our knowledge the largest dataset publicly available for face detection.

While most datasets present a large variety of challenges like illumination, make-up or pose, many state-of-the-art methods are focused on adjusting to scale problems, thus increasing the precision-recall performances. FAN [8], S^{3}FD [9] and PyramidBox [10] use feature pyramids, combining and using both high resolution but semantically poor feature layers and low resolution but semantically strong ones to detect tiny faces. HR [11], on the other hand, uses image pyramid to adjust to multi-scale detection. Face detection also benefits from methods inspired by other computer vision fields like general object detection, for example [12] inspired by [13] that adopts a two-stage object detection strategy, consisting of region proposal followed by a region classification. Other methods like [14] focus on the specific challenges of face detection though various poses; [14] uses three CNN in a coarse-to-fine cascade, to jointly detect faces and determine the roll angle of each face. The interested reader can refer to [15] for a comprehensive survey for a more in depth understanding of the field. This adaptability of the face detection methods in recent years makes it an interesting asset for our study with challenging poses and 360° range of roll angle. Our work takes advantage of face detection to improve the long-term performance of tracking methods.

2.2. **Tracking.** General object tracking has also benefited greatly from the rise of convolutional neural network, with methods that can be divided into two distinct strategies, online tracking and template matching. While tracking-by-detection like [16,17], fine-tune a convolutional model at every frame to search for the tracked object; template matching like [18,19] store the object representation. The interested reader can refer to [20] for a more thorough survey on the matter. Our method is based on the use of a short-term tracker, to which we add a correction mechanism in order to improve the performance in the long term.

Associations of different techniques of computer vision have been proposed by [1] for single target, long-term tracking with occlusion using a combination of tracking, face detection and face recognition. Other works like [21] take advantage of several state-of-the-art trackers, and fuse their result for better performance.

3. **Proposed Method.** This paper proposes a corrected tracking method for simultaneous tracking of several children in an omnidirectional children discussion video. Our method uses off-the-shelf tracker and face detection and introduces a correction pipeline for efficient long-term-tracking of faces. The details of the work flow are described as follows and represented in Figure 2.

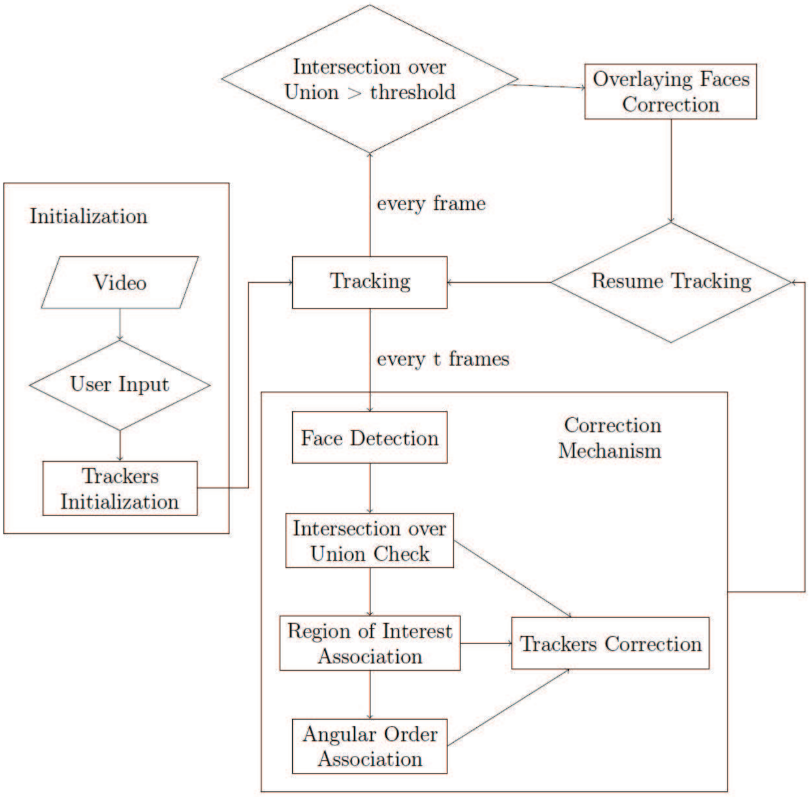


Figure 2. Workflow of our method

3.1. **Tracking with MemTrack.** We choose to use MemTrack [19] as the main tracker for our method. MemTrack is a template matching method, that stores template in external memory, with a Long Short-Term Memory (LSTM) to control memory reading and writing. It uses a combination of residual template and initial template, which in our case is a face detected on the first frame, to predict the target's bounding box position. We choose to use this tracker for its good performances in short and medium term, and its ability to resist to short occlusions, and to retrieve the tracked face after it.

3.2. **Face detection with PyramidBox.** For face detection, we need a reliable method to detect faces in a 360° range. We choose PyramidBox [10], which is one of the top performing methods on the WIDER FACE dataset, trained on augmented data for a detection of faces with any roll angle. While popular datasets have some challenging roll angles of faces, the distribution of such angles are centered around rather than being evenly distributed on the circle; therefore, it is crucial for our case to use methods trained on an augmented dataset, the omnidirectional point of view producing an even distribution of roll angles. Furthermore, in the face detection process, we exclude faces under a fixed size, in order to avoid detecting a face in the background, that could interfere with the tracking process.

3.3. **Jaccard and RoI association.** The first step of our method is a simple check of the tracking, for that we first perform an Intersection over Union (IoU) check: if a detected face has an IoU coefficient over a set threshold with one tracking result, the face is associated with the tracked identity. For this step, we accept faces detected with a relatively low threshold of confidence, considering that even a low confidence face can verify a tracking. Experiment also shows that taking a lower confidence rate for this step yields better results.

In a second time, for the remaining identities, we search if a face is in the Region of Interest (RoI) of the tracking result, defined as the tracking bounding box scaled up. This correction aims to correct slight drifting of the trackers from the faces. If a face is in several RoIs at the same time, it is associated with the closest identity. For this association, we use a higher face confidence than in the previous step.

3.4. **Angular order association.** On the first frame of detection, we use the faces chosen by the user to establish the identities' angular order, simply by computing the angle of their center from the center of the image, which is roughly the center of the omnidirectional video. This order will be maintained for the rest of the video.

In our correction process, the first two associations (IoU and RoI) are important steps for our method, as they allow us to confirm the positions of some faces that we consider as reliable, so they can be used to treat the other detected faces. The final correction we apply to the tracking is angular order association. The general idea is to associate the remaining detected faces to the unconfirmed identities by using the established order. For the association, we distinguish 3 separate cases.

- The first possibility is that a face is detected between two consecutive verified faces, then this face is simply ignored.
- The second case is a single face detected between two confirmed identities separated by a single unverified one in the angular order, then the detected face is simply associated to the missing identity.
- The last case, if a face is detected and several identities can correspond to it, the face is assigned to the closest tracking result by angular distance. The likelihood of this case is low and thus rarely used.

3.5. **System framework.** The goal of our system is to track the positions of the faces of all the children in an omnidirectional video, and to resist to long term occlusion events. The system is initialized with a user input to specify the initial positions and the identities of the children that have to be tracked. Trackers are then initialized for each subject, and their angular order is computed and saved in memory. After initialization, for the main process, tracking is performed for a fixed number of frames, with an extra step of limiting overlapping between the trackers, that mainly happens if two faces are close enough and one of them becomes occluded. To prevent this, we compute the Jaccard overlap coefficient of every tracking bounding box with each other, and if the Jaccard of two boxes exceeds a threshold, we re-position the trackers to their last positions. We perform face detection every t frame and apply the correction mechanism detailed in Sections 3.3 and 3.4 based on the detected faces. The association of the three steps allows us to either confirm the position of the tracked faces, or to recover faces after a tracking error. Once the correction is done, the tracking resumes and the process continues.

4. **Experiments and Results.** We evaluate the efficiency of the proposed method on several videos of children we own. We labeled the face position and identity of every child in these videos by a bounding box on each face every 10 frames, when it is possible. The complexity of the faces appearance and the large occlusions that can occur in the videos can make labeling impossible in certain frames, thus we were only able to label 93% of the faces. We evaluate the precision of our method by computing the intersection over union coefficient between the label and the tracking bounding box. A result is considered correct if the intersection over union coefficient is over 0.5.

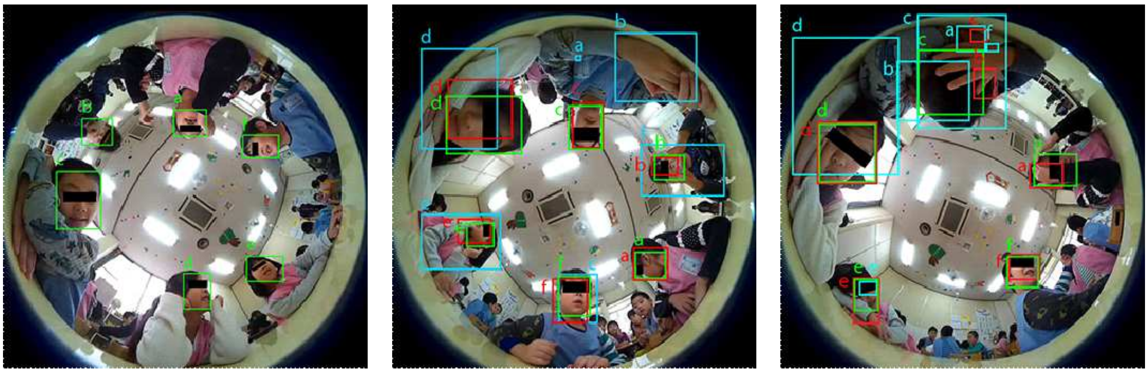
We present evaluation of our method over two videos, for 10,000 frames each. One of the videos contains six children to track, the other five. We choose to focus on these two videos, as the first one is exhibiting several large scale occlusion by a head getting too close to the camera, and the second presents frequent occlusion by one or several hands close to the camera lens. For this evaluation, we perform correction of tracking every 30 frames.

We want to evaluate both the ability to accurately track the faces, but also to associate it with the correct identity. The first metric we present corresponds to the accuracy of both face position and association to the correct identity. The second one is the accuracy of face detection alone, which corresponds to the ability to track the faces in the videos, but without considering the identities, meaning a face correctly detected but associated to the wrong identity would still be considered positive. Table 1 provides a comparison of our method with uncorrected tracking, and with face detection alone.

Table 1. Performance evaluation

Method	Position + Identity precision		Position precision	
	Video 1	Video 2	Video 1	Video 2
Our method	0.701	0.765	0.722	0.795
MemTrack [19]	0.188	0.282	0.333	0.290
PyramidBox [10]	–	–	0.770	0.875

This result shows our system greatly outperforms uncorrected tracking, both in the position and identity precision and in the position precision alone. Also note that our method has similar position precision with PyramidBox used alone; however, the comparison with face detection is just an indicator of how our method performs on the position of the faces, since face detection does not track the identities and does not treat the long occlusion phases. For a more visual representation of our results, Figure 3 provides a few images of the tracking process through the first test video.



(a) Frame 0 (b) Frame 5,000 (c) Frame 10,000

🟩 Labels, 🟥 Our Tracking, 🟦 MemTrack

Figure 3. (color online) Example of tracking over 10,000 frames

As can be seen in this example, our method correctly tracks the faces over the 10,000 frames, despite events of occlusion, deformation, and the camera itself being rotated by the children, which explains the change in their position after 5,000 frames. An error for one of the trackers (children c) in the end of the test video is due to a prolonged self-occlusion of one of the children's face. This kind of drifting is an example of the challenges encountered in this problem, and can be recovered by our correction mechanism, as soon as the child face is visible again. Meanwhile, an uncorrected tracker will quickly drift and loose the identities it is supposed to track, with no means to recover it, which will quickly lead to unusable results in the long-term, as seen in Figure 3(c).

5. **Conclusion.** In this paper we introduced a new tracking correction mechanism, comparing the result of tracking and face detection in three steps: intersection over union check, region of interest association and angular order association. We use this method to greatly improve the performance of long-term tracking on omnidirectional children discussion videos and we experimentally validate our method on two videos for 10,000 frames each. This work is a first achievement towards a task of tracking the gaze of children. This work is also part of an effort to analyse behaviours of children in conversation.

REFERENCES

[1] K. Zhang, E. Rashedi, E. Barati and X.-W. Chen, Long-term face tracking in the wild using deep learning, *CoRR*, abs/1805.07646, 2018.

[2] Z. Kalal, K. Mikolajczyk and J. Matas, Face-TLD: Tracking-learning-detection applied to faces, *2010 IEEE International Conference on Image Processing*, pp.3789-3792, 2010.

[3] J. Supancic and D. Ramanan, Self-paced learning for long-term tracking, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2379-2386, 2013.

[4] P. A. Viola and M. J. Jones, Rapid object detection using a boosted cascade of simple features, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.511-518, 2001.

[5] V. Jain and E. Learned-Miller, *FDDB: A Benchmark for Face Detection in Unconstrained Settings*, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[6] S. Yang, P. Luo, C. C. Loy and X. Tang, Wider face: A face detection benchmark, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] H. Nada, V. Sindagi, H. Zhang and V. M. Patel, Pushing the limits of unconstrained face detection: A challenge dataset and baseline results, *arXiv Preprint*, arXiv:1804.10275, 2018.

[8] J. Zhang, X. Wu, J. Zhu and S. C. H. Hoi, Feature agglomeration networks for single stage face detection, *CoRR*, abs/1712.00721, 2017.

[9] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang and S. Z. Li, S^{3}FD: Single shot scale-invariant face detector, *CoRR*, abs/1708.05237, 2017.

[10] X. Tang, D. K. Du, Z. He and J. Liu, Pyramidbox: A context-assisted single shot face detector, *CoRR*, abs/1803.07737, 2018.

[11] P. Hu and D. Ramanan, Finding tiny faces, *CoRR*, abs/1612.04402, 2016.

[12] H. Jiang and E. G. Learned-Miller, Face detection with the faster R-CNN, *CoRR*, abs/1606.03473, 2016.

[13] S. Ren, K. He, R. B. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *CoRR*, abs/1506.01497, 2015.

[14] X. Shi, S. Shan, M. Kan, S. Wu and X. Chen, Real-time rotation-invariant face detection with progressive calibration networks, *CoRR*, abs/1804.06039, 2018.

[15] S. Zafeiriou, C. Zhang and Z. Zhang, A survey on face detection in the wild: Past, present and future, *Computer Vision and Image Understanding*, vol.138, pp.1-24, 2015.

[16] J. Gao, T. Zhang, X. Yang and C. Xu, P2T: Part-to-target tracking via deep regression learning, *IEEE Trans. Image Processing*, vol.27, no.6, pp.3074-3086, 2018.

[17] M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg, ECO: Efficient convolution operators for tracking, *CoRR*, abs/1611.09224, 2016.

[18] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. S. Torr, Fully-convolutional siamese networks for object tracking, *CoRR*, abs/1606.09549, 2016.

[19] T. Yang and A. B. Chan, Learning dynamic memory networks for object tracking, *ECCV*, 2018.

[20] Y. Wu, J. Lim and M.-H. Yang, Object tracking benchmark, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.37, 2015.

[21] I. Leang, S. Herbin, B. Girard and J. Droulez, On-line fusion of trackers for single-object tracking, *Pattern Recognition*, vol.74, pp.459-473, 2017.