Fantastic Faces: Where to Find Them and How to Understand Them

Yiming Lin
Intelligent Behaviour Understanding Group
Department of Computing, Imperial College London, UK

***Abstract*— In my past two FG submissions, I have been focusing on the problem of finding target faces. In 2019, I published the first dataset for mobile face tracking. Later in 2020, I have proposed a novel tracking framework that achieves state-of-the-art results on face tracking in both mobile settings and surveillance settings. In this manuscript, I briefly summarise these works. Once the location of the target face is available, the next step is understand it. I will also discuss my ongoing work on the face parsing with polar RoI Tanh-warping. The goal of the warping is to find a joint representation of the face and the context in the polar system rather than simply cropping out the face with fixed margins. Finally, I will discuss my future plan on using the RoI-warped representations to develop models for various face analysis tasks. The ultimate vision of my research would be to build a systematic pipeline for face analysis in the wild.**

I. FACE TRACKING IN VIDEOS

Knowing the location of the target face in an image is a key initial step in a face analysis system. As shooting videos with smartphones is becoming increasingly popular, mobile face tracking is gaining attention as a significant module in video face analysis systems. In the following, I briefly summarise my FG works on face tracking.

A. MobiFace: The First Dataset for Mobile Face Tracking

Compared to face detection, face tracking has received little attention, mainly due to the scarcity of dedicated face tracking benchmarks. In my FG2019 work [8], we introduced MobiFace, the first dataset for single face tracking in mobile situations. It consists of 80 live-streaming mobile videos captured by 70 different smartphone users in fully unconstrained environments. Over $95K$ bounding boxes are manually labelled. The videos are carefully selected to cover typical smartphone usage. Some exemplar sequences are shown in Fig. 1.

Although it might appear that the mobile face tracking problem can be readily solved using existing generic object tracking methods, extensive experiments on our dataset show this to be incorrect. Four key differences between the two problems illustrate why. First, the target faces can undergo large scale variations due to the mobility of smartphones, whereas the target's size and the aspect ratio rarely change in most object tracking videos. Second, due to the use of hand-held devices in the mobile footage, the motion of the target can be fast and sometimes unpredictable. Third, rarely does object tracking have similar objects in the same video, whereas in mobile face tracking the tracker can often encounter multiple faces. Finally, due to the smaller field of view of mobile cameras, targets can be easily occluded or out of view. Nevertheless, domain adaptation from generic object tracking to face tracking can still provide a promising starting point given a sufficient amount of data.

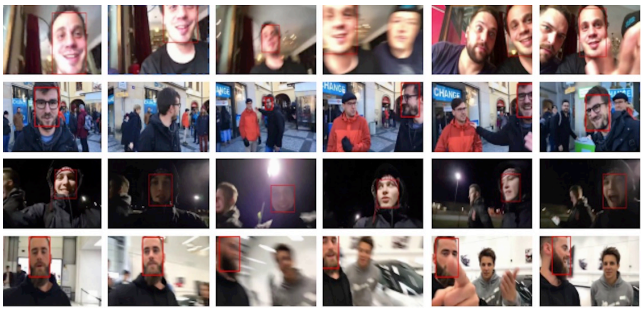

Fig. 1. Exemplar videos from our proposed MobiFace. Red rectangles indicate the ground truth bounding boxes.

To prove the usefulness of MobiFace, we benchmarked 36 state-of-the-art trackers, including facial landmark trackers, generic object trackers and trackers that we have fine-tuned or improved. Our contributions and key observations are summarised below:

1) We introduce MobiFace which consists of 80 unedited in-the-wild mobile video uploaded by 70 smartphone users. We provide bounding box annotations for all the $95,635$ frames. We also define 14 attributes for these mobile videos and provide annotations for each one.
2) We benchmark 36 state-of-the-art tracking methods and models, including 4 facial landmark trackers, 14 object trackers and 18 trackers that we improved or fine-tuned. Our results suggest that mobile face tracking is still a very challenging problem that cannot be fully solved by existing landmark or object trackers, neither by a simple concatenation of face detection and verification method.
3) We demonstrate that fine-tuning on MobiFace significantly boosts the performance of deep learning-based trackers. This suggests MobiFace captures the unique characteristics of mobile face tracking, demonstrating its use potential for the research community.
4) Long-term components in trackers can boost the performance. However, simply concatenating face detection and verification not only results in a slow tracker, but also the performance is unsatisfactory. This suggest an organic combination of the two fields can be a promising direction.

B. FT-RCNN: Face Tracking with Region-based CNN

Motivated by our observations from the MobiFace benchmark, in FG2020, we proposed an efficient face tracker called FT-RCNN, short for Face Tracking with Region-based CNN [7]. FT-RCNN extends the FasterRCNN [11] detection framework with a simple yet effective tracking branch to perform face detection and tracking jointly. The tracking pipeline at the inference stage is depicted in Fig. 2.

FasterRCNN [11] is an advanced object detection framework and has been successfully applied to face detection with promising results [3]. One important, if not the most important, factor of its success is the abundant imagery data with bounding box labels for training its complex components (*e.g.* Region Proposal Network). From the experience of annotating MobiFace, we understand the complexity of labelling face videos and it is infeasible for us to prepare large-scale training data comparable to face detection datasets [14].

To address the problem of insufficient training data for face tracking, we propose a novel pairwise training strategy that enables us to train face tracker by leveraging existing face detection datasets, thus eliminating the need to collect and annotate video data specifically for face tracking. Furthermore, we devise a novel loss function, termed Pair-hard Triplet Cosine Loss, that employs a pair-hard triplet mining strategy to increase the discriminative power of our face tracker.

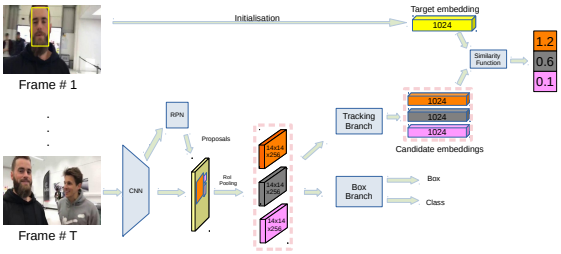

Fig. 2. FTRCNN

Our contributions can be summarised as follows:

1) We introduce a real-time face tracker, FT-RCNN, that is based on FasterRCNN with a novel tracking branch.
2) We propose a pairwise training strategy to allow face trackers to be trained on face detection image datasets.
3) We propose to use the pair-hard triplet cosine loss that adaptively mines triplets from the training pairs. The integration of a metric-learning trained embedding allows a face detector to perform tracking effectively.
4) FT-RCNN has achieved state-of-the-art results on three popular face video datasets, MobiFace, ChokePoint and Youtube Face, while the overall tracker runs at real-time speed. Some visualisation results can be found in Fig. 3 and Fig. 4.

II. FACE PARSING WITH POLAR ROI TANH-WARPING

In this section, I introduce one of our ongoing works that deals with face parsing. We propose the polar RoI (Region-

Fig. 3. Qualitative results of the proposed method on some challenging sequences in MobiFace dataset.

Fig. 4. Qualitative results of the proposed method on some challenging sequences in ChokePoint dataset.

of-Interest) Tanh-warping operator that transforms the whole image to the log-polar coordinate system with the non-linear Tanh function. The motivations, baselines and preliminary results are discussed below.

A. To Crop or Not to Crop

Once we have the location of the target face, the next question to answer is how to take the face out with the bounding box and input it to downstream tasks. In the literature, some tasks require the face to be cropped out tightly without any context information [10] while some [12] claim some cropping with a large margin can lead to better recognition results.

In the face parsing task [13], not only the rigid facial components needs to be segmented but also the hair. Cropping faces with bounding boxes is unable to handle the unpredictable area of the hair. Therefore, many works simply ignored the hair to avoid answering the question of how much

to crop the face out. Is cropping the only option? We try to answer that in the following subsections.

B. Baseline: RoI Tanh-Warping

Recently, Lin *et al.* [6] proposed a RoI Tanh-warping for face parsing. They were inspired by the peripheral vision [5] in the human vision system. Peripheral vision covers the non-focused area of the receptive field. Although peripheral vision is not as sharp as central vision, it makes the brain aware of the environment, helping us detect events even when we are not looking at the direction. They proposed RoI Tanh-warping operator which maps the whole image into a fixed-size with the non-linear Tanh function. It addresses the dilemma between fixed input size and the unpredictable area of hair while reserving the amplified resolution on important regions. The warping process is illustrated in Fig. 5.

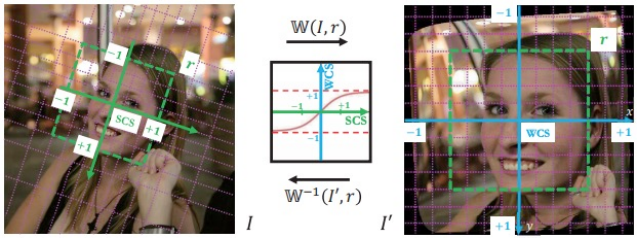

Fig. 5. Baseline: RoI Tanh Warping. SCS stands for source coordinate system and WCS stands for warped coordinate system. Based on the ROI, central face regions are warped to fixed locations while peripheral regions are also kept but "squished" by the Tanh function. Landmarks are required to warp the components to fixed locations. The warped coordinate system is not homogeneous.

Since the warped coordinate system are not homogeneous, face landmarks are required to warp the inner components to predefined locations. The central region and the peripheral region need to be handled separately using different models. The overall model requires complex component designs and training the model requires multiple stages and loss functions. This also makes it difficult to embed the warping operator into the model to facilitate end-to-end training.

C. Ongoing Work: Polar RoI Tanh-Warping for Face Parsing

The quest for equivariant representations is as old as the field of computer vision and pattern recognition itself. Whether in the classical SIFT [9] descriptor, or in the recent deep Convolutional Neural Networks (CNNs), equivariant representations are highly sought after as they encode both class and deformation information in a predictable way. CNNs achieve translational equivariance by its convolution kernels and invariance to local deformations by pooling layers [4]. However, they are not naturally equivariant to some other transformations such as changes in the scale or rotation of the image, which are especially common in faces.

We propose to improve the RoI Tanh-warping by warping the face to the log-polar coordinates. Fig. 6 illustrates the warping effects. The variations of rotation and scale are transformed to translation variations along the two axes. Therefore, the face alignment is not required and the inner (*i.e.* eyes/eyebrows/nose/mouth) and outer parts (*i.e.* hair) do not require to be treated separately as in [6]. The face and the context information are preserved even when the face detection box contains errors.

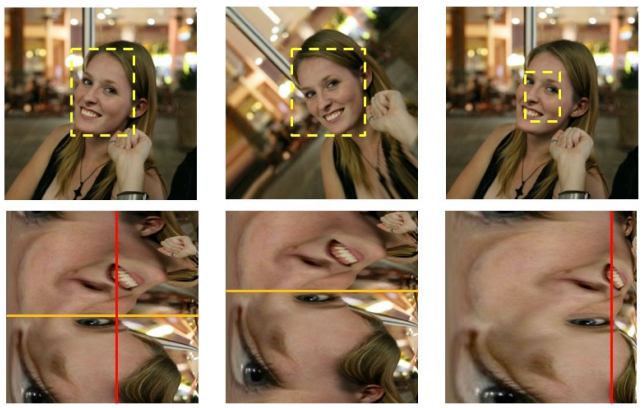

Fig. 6. Our Polar RoI Tanh-Warping. In the log-polar representation, rotations around RoI centre become vertical shifts, and scales of RoI become horizontal shifts. The position of the green/red lines corresponds to the rotation angle/scale factor. Top rows: original image with RoI, rotated image, and original image with scaled RoI. Bottom rows: the corresponding polar images.

With the proposed operator, we have implemented several face parsing models. Preliminary results have shown the operator bring consistent improvements in face parsing, even without major modifications to the baseline segmentation models. So far we have outperformed all state-of-the-art methods in parsing inner facial components, facial skin, and hair on the HELEN dataset [13]. Our model is simpler and the training is straightforward without complex tricks. Some qualitative results are shown in Fig. 7. The proposed operator is fully differentiable and can be easily plugged into CNN models. Thus, the gradients can back-propagate from the Cartesian system to the log-polar system, and vice versa.

Fig. 7. Qualitative results of our ongoing work. Our model can handle hairs with various lengths, and is simpler and more efficient thanks to the proposed warping operator.

As this is an ongoing work, we will publish the details and open-source the codes in the future.

III. FACE ANALYSIS WITH POLAR ROI TANH-WARPING

Upon finishing the previous work, it would be interesting to see if the polar RoI Tanh-warping operator can also contribute to other face analysis tasks. Since we have implemented the differentiable operator in PyTorch, we plan to incorporate it to large CNN models. Not only can it be applied to the input image, but can also be plugged into the intermediate convolutional layers. However, as discussed in Sec. I-B, training such differentiable operators in a CNN model may require large-scale annotated datasets, and face parsing datasets are limited in scale because of the intensive labelling labour for per-pixel annotations.

On the other hand, researchers in the community have contributed large-scale datasets on different face analysis tasks, such as face recognition [1], facial age estimation [12] and facial expression recognition [2]. With such large-scale datasets, the parameters in the warping operator, such as the polar origin or the angular offsets, may be learned and novel CNN modules may be found for different tasks. However, this requires more time and effort to validate our hypothesis, as well as more GPUs to explore the model choices. We will leave it for future work.

REFERENCES

[1] Q. Cao et al. "VGGFace2: A dataset for recognising faces across pose and age". In: *FG*. 2018.

[2] Shiyang Cheng et al. "4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications". In: *CVPR*. 2018.

[3] Huaizu Jiang and Erik Learned-Miller. "Face Detection with the Faster R-CNN". In: *FG*. 2017.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[5] Jerome Y. Lettvin. "On Seeing Sidelong". In: *The Sciences* 16.4 (1976), pp. 10–20.

[6] Jinpeng Lin et al. "Face Parsing With RoI Tanh-Warping". In: *CVPR*. 2019.

[7] Yiming Lin et al. "FT-RCNN: Real-time Visual Face Tracking with Region-based Convolutional Neural Networks". In: *FG*. 2020.

[8] Yiming Lin et al. "MobiFace: A Novel Dataset for Mobile Face Tracking in the Wild". In: *FG*. 2019.

[9] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[10] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition". In: (2015).

[11] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *NeurIPS*. 2015.

[12] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "Deep expectation of real and apparent age from a single image without facial landmarks". In: *International Journal of Computer Vision* 126.2-4 (2018), pp. 144–157.

[13] Brandon M Smith et al. "Exemplar-based face parsing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3484–3491.

[14] Shuo Yang et al. "WIDER FACE: A Face Detection Benchmark". In: *CVPR*. 2016.