Benchmarking face tracking

Mika Fischer Martin Bäuml Hazım K. Ekenel Rainer Stiefelhagen
Karlsruhe Institute of Technology
Adenauerring 2, 76227 Karlsruhe, Germany
{baeuml, mika.fischer, ekenel, rainer.stiefelhagen}@kit.edu

Abstract

Face tracking is an active area of computer vision research and an important building block for many applications. However, opposed to face detection, there is no common benchmark data set to evaluate a tracker's performance, making it hard to compare results between different approaches. In this challenge we propose a data set, annotation guidelines and a well defined evaluation protocol in order to facilitate the evaluation of face tracking systems in the future.

1. Introduction

Face tracking is a basic building block in many computer vision areas. Any face analysis in videos, such as face recognition, gender classification or expression analysis, usually relies on first reliably detecting and tracking a face in the scene. The literature on face tracking is vast [2, 3, 5, 6, 7, 8, 9, 12, 13]. However, no commonly agreed-on diverse data set has been established to measure and compare the performance of the different approaches (unlike for face detection, for which a considerable number of such data sets exist, for example [4, 10]).

The goal of the face tracking challenge is to provide a common data set and well-defined metrics in order to evaluate the performance of vision-based face tracking. The focus lies on single-view 2D face tracking since in many real-world scenarios there are only 2D image sequences available. We propose a data set consisting of videos from various sources and with a large range of challenges in image conditions. This allows to evaluate a tracker across all of those scenarios and image conditions, opposed to only one specialized setting.

There have been a few other face tracking evaluation efforts. To the best of our knowledge, the most prominent has been CLEAR [11], which however only focused on a smart room scenario. We build upon their effort, especially by using the MOT metric [1] which has become widely accepted in person and general object tracking. A few other researchers have kindly provided their data along with ground truth labels [5, 7, 8]. They too mostly focused on a single setting. The generalization properties of face trackers to real-world scenarios remain unclear if evaluated only on such data.

In the following, we will first briefly describe the proposed data set. The larger remainder of this challenge proposal will specify guidelines according to which the data is annotated, and define the evaluation protocol in detail.

2. Data set

We provide a video data set exhibiting a large range of challenging image conditions, such as differences in lighting, low image resolution, non-frontal head poses and occlusions. In order to cover different areas in which face tracking is employed in computer vision, we collected the videos from various different sources, for example:

- TV series and movies
- news casts
- webcam recordings
- youtube
- surveillance camera networks

The data set is split into three different (mostly subjective) difficulties. For example, sequences with from a stationary view point containing only a single, non-occluded face are considered easy. On the other hand, sequences with a large number of faces, in difficult image conditions such as low lighting or with many occlusions are considered hard (cf. Table 1).

We further split the data set into a development set and an evaluation set. The development set can be used in order to train and fine-tune a tracker (e.g. estimate parameters). The evaluation set is used to measure the tracker's performance. We will provide ground truth annotations for both the development set and the evaluation set, so that experiments can be performed continuously. However, we want to

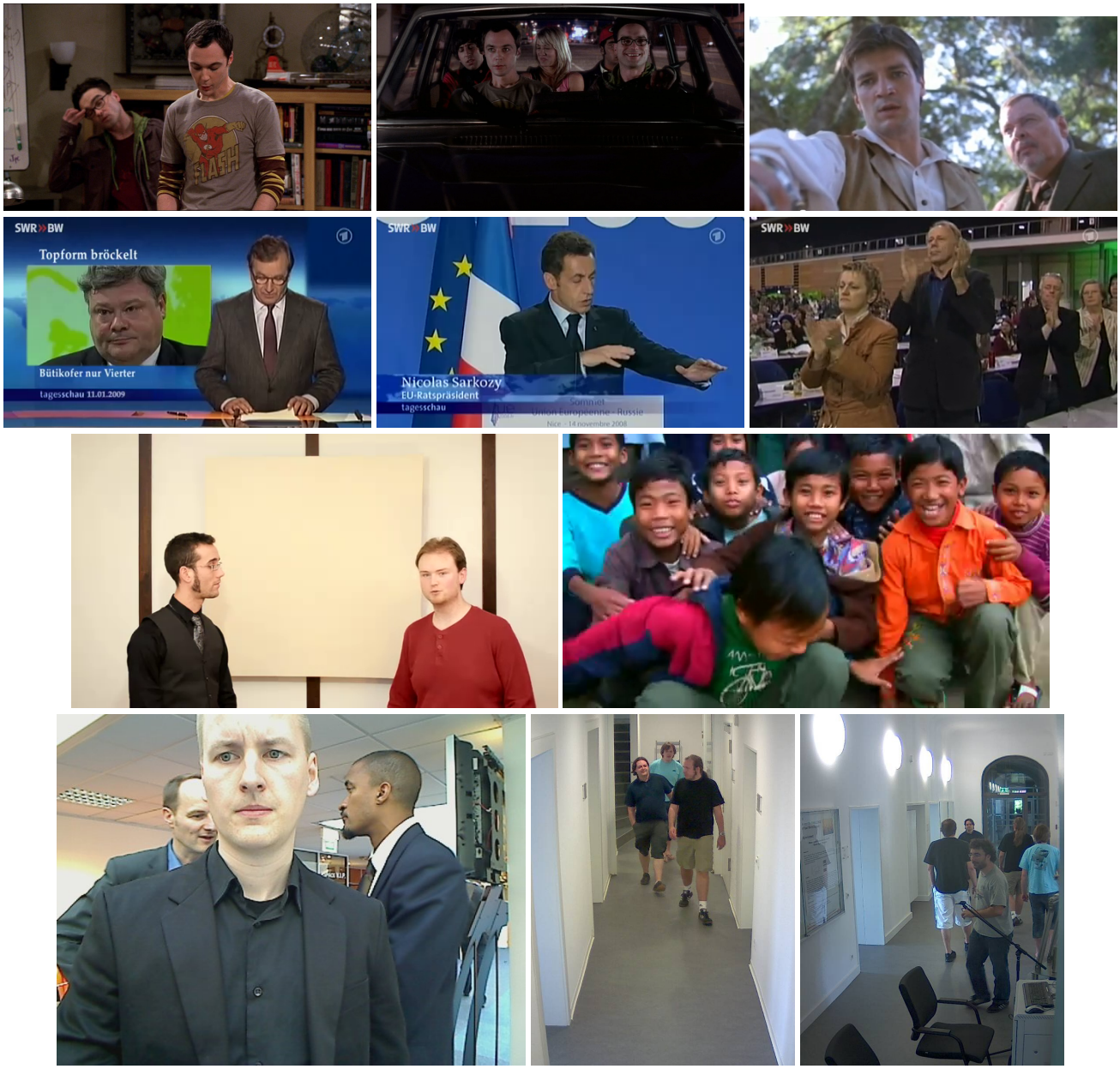

Figure 1: Samples from the data set (row 1: TV series, row 2: news cast, row 3: youtube, row 4: webcam and surveillance data).

stress that *it is imperative that the evaluation ground truth is only to be used for the final performance evaluation of a system and never used for training and/or tuning a tracker.*

We will provide neither intrinsic nor extrinsic camera calibration information for any of the sequences, since usually these are not available in many real-word scenarios (and we just do not have them for most of the sequences).

3. Annotations

The following manual annotations will be provided for each face in the annotated frames:

- Face bounding box: (x, y, width, height)
- Eye center positions: (x, y) (x, y)
- Mouth center position: (x, y)
- ID: 1 ... N

	ILL	RES	POS	MUL	MOT	COM
webcam	+		+			
surveillance	++	++	++	++		+
youtube	++	++	++	+	++	++
TV series	+		++	+	+	
news cast	+		+	+	+	

Table 1: Overview over the different challenges usually contained in videos from the different scenarios. ILL: difficult illumination, RES: low resolution, POS: non-frontal head pose, MUL: many faces, MOT: camera motion, COM: compression artifacts.

The annotations are performed according to the following guidelines.

3.1. Guidelines

The frames will be annotated in 0.2s intervals (i.e. every 3rd, 5th and 6th frame for videos with 15fps, 25fps and 30fps, respectively). Only those frames will be used for the evaluation. A system may output results on other frames, too. However, those will not be taken into account for the performance evaluation.

The face bounding box will be annotated tightly to the face, excluding the ears and the forehead. This means that the lower limit of the box should be the chin, the upper limit should be the end of the forehead and the limits on either side of the face should be the cheek excluding the ears. In this way the annotations should be largely independent of hairstyles, hats, etc. Faces with a bounding box smaller than 15 px in either direction will not be labeled. The bounding boxes are not necessarily quadratic.

Note that this does not mean that the trackers have to return bounding boxes exactly like this. For one, a perfect overlap is not required for a match (see Section 4 and Figure 3). Furthermore, the bounding boxes returned by the tracker can be resized if they are significantly larger or smaller than the annotated bounding boxes. Such scaling should however be determined on the development set exclusively.

The locations of eyes and mouth center will be labeled if they are visible. If they are not, they will be labeled as (-1, -1).

For some faces in the video, it is very difficult to track them, for instance because they are occluded by other objects or because their view angle is larger than 90 degrees. In order to neither penalize trackers that find such faces nor penalize those that don't, we consider the following faces in the ground truth as *Don't Care Objects* (DCOs):

- Faces, where either side of the bounding box is between 15 and 20 pixels

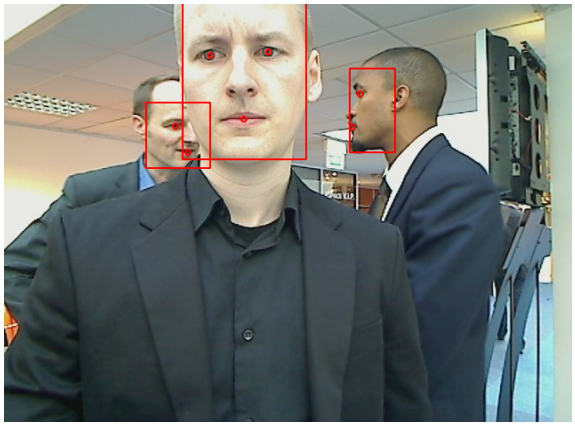

Figure 2: Face labels according to the guide lines. None of the above faces is considered as DCO since for each face at least two facial features are visible.

- Faces, where two of the three facial features (left eye, right eye, mouth) are not visible

These DCOs and the bounding box hypotheses that are associated with them are completely removed from the scoring process. For details, see Section 4.

The ID will be labeled as an integer, in such a way that a person has the same ID in all frames of the same video.

Note that it is only required that the tracker assigns the same ID for the same person in subsequent frames. If the person is invisible and returns later, a new ID can be assigned to the person without incurring a penalty (cf. Section 4).

Further note that since this is a face tracking evaluation, tracking a head where the face is completely invisible is considered an error. If it is desired to use a head tracker, the tracker must be able to determine whether the face is visible. The parts of the tracks where the face is invisible should then be removed for the evaluation.

3.2. Label Format

The labels will be provided in XML format. For each video in the data set, there will be one XML file with the following format:

```
<?xml version="1.0" encoding="UTF-8" ?>
<video filename="foo.avi">
  <frame number="0" timestamp="123.456">
    <face id="1"
          bbox_x="123"
          bbox_y="456"
          bbox_width="50"
          bbox_height="50"
          left_eye_x="-1"
          left_eye_y="-1"
```

```
            right_eye_x="160"
            right_eye_y="523"
            mouth_x="150"
            mouth_y="520" />
        ...
    </frame>
    <frame number="5" timestamp="123.656">
        ...
    </frame>
    ...
</video>
```

The `video` element has exactly one attribute:

1. `filename`: The name of video file for which this is the label file.

The `video` element contains one or more `frame` elements, one for each labeled frame. The `frame` element has exactly two attributes:

1. `number`: The frame number.
2. `timestamp`: The timestamp of this frame.

Each `frame` element can contain zero or more `face` elements. In the ground truth labels, each `face` element has exactly 10 attributes:

1. `bbox_x`: The x-coordinate of the bounding box containing the face.
2. `bbox_y`: The y-coordinate of the bounding box containing the face.
3. `bbox_width`: The width of the face bounding box.
4. `bbox_height`: The height of the face bounding box.
5. `left_eye_x`: The x-coordinate of the left eye.
6. `left_eye_y`: The y-coordinate of the left eye.
7. `right_eye_x`: The x-coordinate of the right eye.
8. `right_eye_y`: The y-coordinate of the right eye.
9. `mouth_x`: The x-coordinate of the mouth center.
10. `mouth_y`: The y-coordinate of the mouth center.

The eyes are labeled biologically correctly, i.e. the `left_eye_*` coordinates denote the location of the person's left eye, not as seen from the viewer's perspective.

The expected format for the submissions is the same, except that all attributes of the `face` element except the `bbox_*` attributes may be omitted, since they are not necessary for the evaluation.

An example for a valid result file:

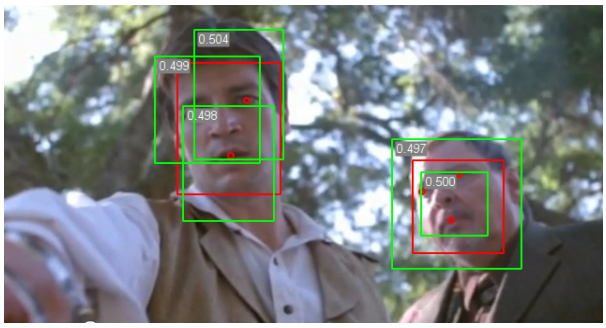

Figure 3: Allowed deviation from bounding box labels. The hypothesis bounding box can deviate quite a bit and still be considered a match (overlap distance to the red ground truth bounding box is given for each green bounding box hypothesis).

```
<?xml version="1.0" encoding="UTF-8" ?>
<video filename="foo.avi">
  <frame number="0" timestamp="123.456">
    <face id="1"
          bbox_x="124"
          bbox_y="454"
          bbox_width="49"
          bbox_height="51" />
      ...
  </frame>
  <frame number="1" timestamp="123.486">
      ...
  </frame>
  ...
</video>
```

4. Metrics

To measure the tracking performance, the well-known Multi-Object Tracking (MOT) metrics [1] will be used.

As a first step, a correspondence between the face location hypotheses returned by the tracker and the ground truth locations has to be established. For this, we follow the method described in [1]. We use a distance derived from the rectangle overlap:

$$d(r_1, r_2) = 1 - \frac{|r_1 \bigcap r_2|}{|r_1 \bigcup r_2|} \quad (1)$$

Only pairs of bounding boxes with a distance smaller than $T = 0.5$ are considered as potential correspondences. Figure 3 illustrates exemplary hypotheses that are still valid correspondences for a given label.

As described in Section 3.1, Don't Care Objects (DCOs) take part in the correspondence determination, but do not contribute to the misses, false positives, mismatch errors or the number of ground truth labels. A bounding box associ-

ated to a DCO, are completely discarded, as are the DCOs themselves.

Using the correspondences, the MOT Accuracy (MOTA) can be computed as follows:

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad (2)$$

where m_t is the number of misses, fp_t is the number of false positives and mme_t is the number of mismatch errors in the frame at timestamp t, as defined in [1].

In addition to the MOTA score, the individual components of the MOTA score should be reported as well:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t} \quad (3)$$

$$\overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t} \quad (4)$$

$$\overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t} \quad (5)$$

At this time we see no need to use the MOTP score in this evaluation. The reason is that it mainly scores the similarity of the bounding boxes returned by the tracker to the way the ground truth bounding boxes are annotated. Since this can differ significantly from tracker to tracker, we see no significant information in the score.

The MOTA score for each scenario is defined as the average of the MOTA scores for the videos in the scenario. The MOTA score for each difficulty is defined as the average of the MOTA scores for the videos of the difficulty. The total MOTA score of the evaluation is defined as the average of the MOTA scores for all scenarios.

5. Evaluation tool

An evaluation tool will be provided that computes the metrics described above, given the track hypotheses and the ground truth as input.

6. Submissions

Results can be submitted to the authors at any time. The authors will evaluate the performance using the evaluation tool and post the results on the challenge website. A link to an academic article explaining the system used to obtain the submitted results is strongly encouraged.

The format of the submission is a zip file including hypotheses for all videos in the evaluation data set in XML format (see Section 3.2). Additionally, the zip file should include a file `description.txt`, which should include a short description of the system as well as an indication of the runtime complexity of the system. In particular, the hardware used for the experiments and the frame rate that the system achieved should be given. For instance: "Our system runs with 10 fps on an i7-720 CPU, using all eight cores, but no GPU."

References

[1] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 1, 4, 5

[2] G. Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Workshop on Applications of Computer Vision, WACV*, pages 214–219. IEEE, 1998. 1

[3] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003. 1

[4] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings, 2010. 1

[5] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, Oct. 2003. 1

[6] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Conference on Computer Vision and Pattern Recognition*, number 1, pages 1–8. Ieee, June 2008. 1

[7] E. Maggio and A. Cavallaro. Hybrid particle filter and mean shift tracker with adaptive transition model. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, number 4. Citeseer, 2005. 1

[8] E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro. Particle PHD filtering for multi-target visual tracking. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 1. IEEE, 2007. 1

[9] D. a. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, 77(1-3):125–141, Aug. 2007. 1

[10] H. Rowley, S. Baluja, and T. Kanade. Rotation Invariant Neural Network-Based Face Detection. In *Conference on Computer Vision and Pattern Recognition*, pages 963–963. IEEE Comput. Soc, 1998. 1

[11] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 Evaluation. In *International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 3–34, Baltimore, USA, 2007. 1

[12] R. Verma, C. Schmid, and K. Mikolajczyk. Face detection and tracking in a video by propagating detection probabilities. *Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1215–1228, Oct. 2003. 1

[13] J. Yang and A. Waibel. A real-time face tracker. *Workshop on Applications of Computer Vision, WACV*, pages 142–147, 1996. 1