Robust Multi-Pose Face Tracking by Multi-Stage Tracklet Association

Markus Roth*, Martin Bäuml*, Ram Nevatia† and Rainer Stiefelhagen*

* *Institute for Anthropomatics*
Karlsruhe Institute of Technology
markus.roth@student.kit.edu
{baeuml, rainer.stiefelhagen}@kit.edu

† *Institute for Robotics and Intelligent Systems*
University of Southern California
Los Angeles, CA 90089
nevatia@usc.edu

Abstract

We propose an approach for multi-pose face tracking by association of face detection responses in two stages using multiple cues. The low-level stage uses a two-threshold strategy to merge detection responses based on location, size and pose, resulting in short but reliable tracklets. The high-level stage uses different cues for computing a joint similarity measure between tracklets. The facial cue compares facial features of the most frontal face detections in pairs of tracklets. The classifier cue learns a discriminative appearance model for each tracklet, using detection pairs within reliable tracklets and between overlapping tracklets as training data. The constraint cue observes the compatibility of motion of two tracklets. The association of tracklets is globally optimized with the Hungarian algorithm. We validate our approach on two challenging episodes of two TV series and report a Multiple Object Tracking Accuracy (MOTA) of 82% and 68.2%, respectively.

1 Introduction

Face tracking is important for many higher level tasks such as face recognition, gaze estimation, emotion recognition or interaction analysis. Despite recent advances, challenges remain for example due to camera ego-motion and when multiple persons are interacting with each other, leading to occlusions and possible track identity switches.

Well performing *person* detectors have led to association-based person tracking approaches, which associate corresponding detection responses or tracklets into longer tracks (*e.g.*, [7, 8, 9, 11]). Correspondence between two tracklets is found via appearance, motion and temporal affinities. A Hungarian algorithm is often used to find the global optimum (*e.g.*, [8, 9]). These approaches have been shown to perform very well on person tracking, and in this work we adapt them to the task of multi-pose face tracking.

Finding similarities between face detection responses needs additional features other than color histograms as used primarily for person tracking. Using additional information like face pose provides more robust affinities between tracklets. Our contributions are the following: (i) We adapt association-based tracking to multi-pose face tracking by associating multi-pose face detection responses (Sec. 2). (ii) We introduce face-specific affinities and similarity measures to associate tracklets in two stages (Sec. 2.2 and 2.3). (iii) We evaluate our approach on two episodes of two challenging TV series (~60 minutes in total) and consistently outperform other state-of-the-art methods in terms of tracking accuracy (Sec. 3). Fig. 1 shows a challenging scene in which our approach is able to successfully track three faces despite occlusions and missing detections.

1.1 Related work

Face tracking has a long history of research, but we will focus here only on detector-based approaches. Babenko *et al.* [1] proposed a general object tracking approach with online multiple instance learning (MIL) which they also apply to face tracking. While MIL helps to avoid a degradation of the detector from imprecise object localization, their approach still requires manual initialization by a user or another general detector. Sivic *et al.* [10] associate frontal and profile face detections over time if the number of tracked Kanade-Lucas-Tomasi features between two detections exceeds a threshold. In previous work [2], we used a particle filter to temporally associate detection responses from multi-pose face detectors. However, the particle filter only takes into account the track state in the previous frame in an online manner, which limits its capability to resolve ambiguities. Also, no appearance models were

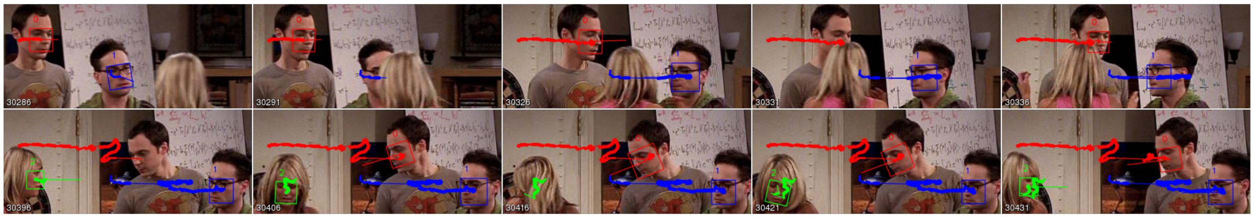

Figure 1. Face tracks as obtained by our approach in a challenging scene including camera motion, occlusions, pose changes and detector failures.

used to detect and avoid track switches or track over occlusions.

For *person* tracking, association-based tracking approaches have been shown to perform quite well (*e.g.*, [7, 8, 9, 11]). Huang *et al.* [5] associate person detections on three levels, with simple affinity measures on the lowest level, more complex affinity measures on the middle level, and estimating and taking into account scene structure and occluders on the highest level. Kuo *et al.* [7] introduce online-learned appearance models as one of the higher affinity measures which are learned on-the-fly during tracking.

2 Multi-stage tracklet association

Our proposed approach hierarchically associates oriented face detections to tracklets in two stages.

(i) The low-level stage constructs short, reliable tracklets from single detection responses in consecutive frames. The association is performed by using a two-threshold strategy based on size, location and pose affinities.

(ii) The high-level stage associates reliable low-level tracklets. Three different cues are used to compute a joint similarity measure between tracklets: The *face id* cue compares facial features of the most frontal face views. The *classifier* cue learns a discriminative appearance model per tracklet. The *constraint* cue encourages natural associations in terms of motion and pose compatibility between tracklets. The globally optimal assignment is obtained using the Hungarian algorithm.

2.1 Face detection

We obtain face detection responses by using Modified Census Transform (MCT) face detectors [6] trained for different pan/roll angle combinations. We train one detector for each $(pan, roll)$ combination with $pan \in \{0, 15, 30, 45, 60, 90\}$, $roll \in \{0, 22.5, 45\}$ and their respective mirrored versions. We also train one generic MCT-based eye detector which is used to localize the eyes within a detected face region.

An exhaustive search over all frames with all face detectors is done to obtain face detection responses r_i in the form of

$$r_i = (x_i, y_i, w_i, h_i, o_i, eyes_i, t_i)$$

where (x_i, y_i, w_i, h_i) denotes the face detection bounding box, o_i the pose of the face with $o_i = (pan_i, roll_i)$, $eyes_i$ the coordinates of the left and right eye, and t_i the time stamp of r_i.

Strongly overlapping raw face detections are clustered into a single oriented face detection response by using the confidence weighted mean.

2.2 Low-level association

Following [5], face detection responses of two consecutive frames r_i, r_j, $t_j = t_i + 1$ are linked by using a two-threshold strategy. The link probability between two detection responses $S_{ij} = P_{\text{link}}(r_i, r_j)$ is defined as the product of their affinities $A(r_i, r_j)$:

$$S_{ij} := A_{\text{location}}(r_i, r_j) A_{\text{size}}(r_i, r_j) A_{\text{pose}}(r_i, r_j)\ . \quad (1)$$

Detection responses r_i and r_j are associated, if their link probability is above a threshold θ_1 *and* exceeds the link probability of any conflicting pair by θ_2, *i.e.* if $S_{ij} > \theta_1$, and $S_{ij} > S_{ik} + \theta_2$ and $S_{ij} > S_{lj} + \theta_2$ for all $k \neq j$ and all $l \neq i$.

A set of associated detection responses forms a *reliable* tracklet. A tracklet T_k consists of its detection responses, which are sorted in ascending order by their timestamps:

$$T_k = \{r_i^{(k)} \mid t_i^{(k)} < t_{i+1}^{(k)}\} \quad (2)$$

By construction, each detection response belongs to exactly one tracklet. By setting θ_1 and θ_2 conservatively ($\theta_1 = 0.3$ and $\theta_2 = 0.03$ in our experiments), all detection responses within a reliable tracklet belong to the same target face with high probability. The low-level association stage results in a set of reliable tracklets $\mathcal{T}_L$, which form the input for the high-level stage.

2.3 High-level association

The high-level association stage merges tracklet pairs $T_i, T_j \in \mathcal{T}_L$ according to their *joint similarity* $\Sigma(T_i, T_j)$. Association between T_i and T_j is only considered if T_j follows T_i within a timeframe of Δt and they do not overlap, *i.e.* $\max(\{t_k | r_k \in T_i\}) < \min(\{t_k | r_k \in T_j\} < \max(\{t_k | r_k \in T_i\}) + \Delta t$.

The joint similarity $\Sigma(T_i, T_j)$ combines similarities from face id, classifier and constraint cues.

Face id cue. We compare the facial appearance of two tracklets T_i and T_j based on local Discrete Cosine Transform (DCT) features which have been shown to be robust against illumination changes and occlusions [2, 4]. In order to reduce pose influence as much as possible, the two most frontal face detection responses $r_i^0 \in T_i$ and $r_j^0 \in T_j$ are compared, and before feature extraction, the face is aligned to a canonical frontal pose of size 48×64 pixels via affine warping based on the detected eye locations $eyes_i^0$ and $eyes_j^0$. We divide the aligned face in 6×8 blocks, compute the DCT on each of the blocks and concatenate the top 5 coefficients of each block (excluding the 0th mean coefficient) as feature vector $\mathbf{f}$. The face id similarity $\sigma_{DCT}(T_i, T_j)$ is then computed as

$$\sigma_{DCT}(T_i, T_j) = sig(\omega_1 + \omega_2 ||\mathbf{f}_i - \mathbf{f}_j||_2) \qquad (3)$$

where $sig(z) = (1 + e^{-z})^{-1}$is the sigmoid function. We set $\omega_1 = 5$ and $\omega_2 = -0.2$ in our experiments.

Classifier cue. The classifier cue builds on Online Learned Discriminative Appearance Models [7] and learns discriminative appearance models $\mathcal{M}_i$ to distinguish each tracklet T_i from co-occurring ones. Training samples (pairs of detection responses) are collected on-the-fly based on spatio-temporal constraints. Positive training pairs are collected from within T_i, since it is a reliable tracklet. Negative training pairs are collected as pairs of detection responses between T_i and overlapping tracklets T_j, observing that two tracklets which overlap in time but are spatially separated belong to different targets.

As features we use 8-bin color histograms for each HSV channel, which we compute on 3×3 subregions for both the face detection bounding box and a clothing bounding box beneath the face. Using positive and negative training sample pairs, we train a strong classifier

$$H(r_i, r_j) = \sum_f \omega_f h_f(r_i, r_j) \qquad (4)$$

as a linear combination of weak classifiers h_f using AdaBoost as proposed in [7]. The weak classifiers stem from a feature pool $\mathcal{F}$ which consists of the Bhattacharyya histogram distances over all subregions and color channels. We apply the strong classifier to find the similarities between T_i and every following tracklet T_j. In contrast to [7], we use the maximum classifier output of detection response pairs from the tail of T_i and the head of T_j to obtain the classifier similarity $\sigma_{\text{c}}(T_i, T_j)$:

$$\sigma_{\text{c}}(T_i, T_j) := \max(\{H(r_i, r_j) \mid \begin{array}{l} r_i \in \text{tail}(T_i), \\ r_j \in \text{head}(T_j)\}) \end{array}$$

where head$(\cdot)$ and tail$(\cdot)$ denote the first and last N detection responses of a tracklet, respectively (we set $N = 4$ in our experiments).

Constraint cue. Both face id and classifier cue result in (possibly strong) similarities between tracklets despite their compatibility in location, pose and motion. The constraint cue defines tracklet similarities to model inertia in terms of location and pose changes and thus rewards more natural associations. The similarities used in the constraint cue also help for video sequences, where no classifier can be learned, *i.e.* when there are no overlapping tracklets.

The constraint distances d_* that underlie the similarities σ are computed between detection responses $r_l = \text{last}(T_i)$ and $r_f = \text{first}(T_j)$ as follows:

- $d_{\text{pose}}(T_i, T_j) = ||o_l - o_f||_2$
- $d_{\text{motion}}(T_i, T_j) = ||[x_{\text{pred}}, y_{\text{pred}}] - [x_f, y_f]||_2 \cdot w_l^{-1}$, where $[x_{\text{pred}}, y_{\text{pred}}]$ is the predicted T_j starting location based on the mean velocity of tail(T_i) and w_l is used to normalize the distance by the size of the faces.
- $d_{\text{time}}(T_i, T_j) = t_f - t_l$

Each distance d_* is transformed into a similarity σ_* by means of a sigmoid function similar to Eq. 3. Our implementation uses $\omega_1^{\text{pose}} = 5$, $\omega_2^{\text{pose}} = -0.2$, $\omega_1^{\text{motion}} = 3$, $\omega_2^{\text{motion}} = -1$, $\omega_1^{\text{time}} = 3$, $\omega_2^{\text{time}} = -0.7$.

Assembling the similarities. The similarities obtained from the three cues are incorporated into a joint similarity $\Sigma(T_i, T_j)$, if they are *promising* and otherwise ignored. A similarity $\sigma(T_i, T_j)$ is *only* considered promising, if its input data is sufficient for a trustworthy similarity. σ_{DCT} is defined promising, if the most frontal face detections of T_i and T_j are frontal enough ($pan < 15°$). σ_{pose} is defined promising, if the time gap is smaller than 0.3 seconds, since we assume that any arbitrary face pose change can be made in a period of time longer than that. σ_{motion} is considered promising, if there are at least four detections for estimating the predicted point. σ_{c} is promising for a positive classifier result. σ_{time} is always considered promising.

Table 1. Evaluation results

Method	The Big Bang Theory (Ep. 01-01)				Buffy the Vampire Slayer (Ep. 05-02)			
	MOTA	MR	FPR	IDS	MOTA	MR	FPR	IDS
PF (Bäuml *et al.* [2])	79.57%	12.08%	7.38%	103 (0.97%)	63.94%	21.37%	12.22%	211 (2.47%)
KLT (Sivic *et al.* [10])	-	-	-	-	53.23%	41.97%	2.98%	157 (1.83%)
Ours (w/o high-level)	77.77%	16.04%	3.92%	238 (2.27%)	61.83%	29.22%	3.88%	422 (5.07%)
Ours (w/ high-level)	81.95%	12.71%	5.02%	33 (0.31%)	68.19%	24.54%	6.09%	102 (1.18%)

We define the joint similarity as

$$\Sigma(T_i, T_j) := \prod_{\text{promising } \sigma} \sigma(T_i, T_j) \tag{5}$$

A global optimal association is obtained using the Hungarian algorithm on the joint similarities. If the Hungarian algorithm associates entries i and j, tracklets T_i and T_j are merged to form a longer tracklet. The high-level association can be repeated multiple times by taking the resulting set of tracklets $\mathcal{T}_L$ of the previous stage as input for the next round.

3 Experimental validation

We use the Multiple Object Tracker Accuracy (MOTA) [3] in order to evaluate the ability to consistently label faces over time. MOTA is defined as

$$\text{MOTA} = 1 - \frac{\sum_t \text{FP}_t + \text{MISS}_t + \text{IDS}_t}{\sum_t \text{GT}_t} \tag{6}$$

where FP_t, MISS_t and IDS_t are the false positives, misses, and identity switches at time t, respectively. We also report false positive rates (FPR), identity switches (IDS) and miss rates (MR) individually.

We evaluate our approach on two episodes of the TV series *The Big Bang Theory* (Ep. 01-01) and *Buffy the Vampire Slayer* (Ep. 05-02). The two episodes present different challenges to the face tracker. We manually annotated the test data with face bounding boxes and identities[1]. Annotations are given for every 5th frame of *The Big Bang Theory* and every 10th frame of *Buffy The Vampire Slayer*. All evaluations are perfomed using these annotations. Some statistics on the two datasets can be found in Table 2.

We compare our results to the particle filter approach from [2] using the same MCT-based face detectors. On the Buffy dataset we compare also to [10] for which we downloaded the tracking results from their website[2].

An overview over the results can be found in Table 1.

Table 2. Dataset statistics

	BBT 01-01	*Buffy* 05-02
frames	32,990	62,157
raw detections	3,685,587	5,518,699
clustered detections	63,713	135,087
face tracks low-level	906	1,337
face tracks high-level	589	803

The Big Bang Theory Ep. 01-01 *The Big Bang Theory* (BBT) is a sitcom with only few fast face movements, and most shots are well-lit. However, faces are rather small on average due to many full scene shots. We achieve a MOTA of 77.8% with the low-level stage of our approach, and 82.0% after the high-level stage. For comparison, the particle filter approach from [2] achieves a MOTA of 79.6%, and shows a larger false-positive rate and number of identity switches than our proposed approach. Our high-level stage is able to reduce both MR and IDS compared to the low-level stage, by connecting tracklets over missing detections. This however comes at the expense of a slight increase in FPR, which can be explained by connections (and interpolation) between tracklets over full occlusions of a face, for which we would expect the tracker to result in two tracks (as we defined in our ground truth). However, the reductions in MR and IDS outweigh the increase in FPR.

Buffy Ep. 05-02 This episode contains several dark and action scenes with fast face movements. Despite these additional challenges, our approach achieves a MOTA of 68.2%. The PF-based approach [2] and the KLT-based approach from [10] achieve a MOTA of 63.9% and 53.2%, respectively. In comparision, our approach achieves a relative improvement of 6.6% and 28.1%. Similar to the BBT dataset, MR and IDS are reduced with the high-level stage while FPR is increased.

The worse performance on the Buffy dataset compared to the BBT dataset can to some extend be attributed the challenging conditions in the Buffy dataset with which our detectors have some difficulties. Fig. 2 illustrates some false positives and misses on the Buffy dataset, which are both directly dependent on the detector performance. Consistent false positives over mul-

[1] The groundtruth can be downloaded from `http://cvhci.anthropomatik.kit.edu/projects/facetracking`

[2] `http://www.robots.ox.ac.uk/~vgg/data/nface/`

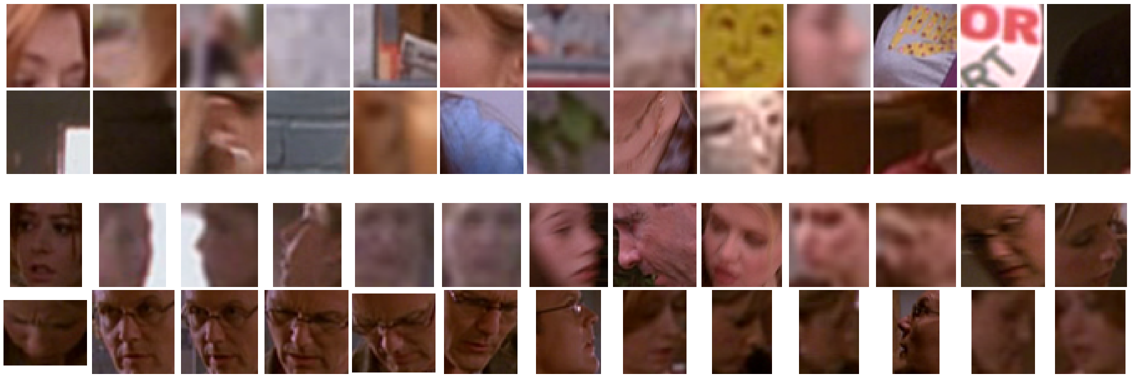

Figure 2. Excerpt of false positives (two top rows) and misses (two bottom rows) of our approach on the Buffy dataset.

tiple frames tend to be hard to prune in the tracking stage. Misses are either dark, blurry, highly non-frontal or tilted faces. The latter can be overcome by using additional face detectors trained for tilted angles.

4 Runtime notes

We implemented our approach in C++ and measure the runtime on an AMD Phenom^{TM}II X4 970 quad core processor. The different combinations of pan and roll and their respective mirrored versions result in 47 detectors. We measure an average detection time of 1.84 s per frame (1024×576 px), of which clustering and eye detection take 0.12 ms and 93 ms, respectively.

Low-level association and high-level association combined take on average 25.2 ms of CPU time per frame, of which low-level association runtime is negligible ($3.7\,\mu$s). Thus, the detection task takes as much as 98.6% of the total computation time.

5 Conclusion

We present an approach for obtaining face tracks from face detection responses. Tracklets are associated in two stages. In both the low-level and the high-level stage, we use face-specific features such as the face pose and facial features as affinities between tracklets. An online learned appearance model based on color histograms from the clothing of a person complements the facial features. We are able to report consistent reductions in false-positive rate and number of identity switches over the state-of-the-art, which will directly benefit higher level applications such as face recognition.

Acknowledgments We thank Bo Yang for many helpful discussions. Markus Roth gratefully acknowledges a scholarship from InterACT. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under contract no. 01ISO9052E. The views expressed herein are the authors' responsibility and do not necessarily reflect those of BMBF.

References

[1] B. Babenko, M. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009.

[2] M. Bäuml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *AVSS*, 2010.

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

[4] H. K. Ekenel and R. Stiefelhagen. Analysis of Local Appearance-Based Face Recognition: Effects of Feature Selection and Feature Normalization. *CVPR Biometrics Workshop*, 2006.

[5] C. Huang, B. Wu, and R. Nevatia. Robust Object Tracking by Hierarchical Association of Detection Responses. In *ECCV*, 2008.

[6] C. Küblbeck and A. Ernst. Face detection and tracking in video sequences using the modified census transformation. *IVC*, 24(6):564–572, June 2006.

[7] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-Target Tracking by On-Line Learned Discriminative Appearance Models. In *CVPR*, 2010.

[8] C.-H. Kuo and R. Nevatia. How does Person Identity Recognition Help Multi-Person Tracking ? In *CVPR*, 2011.

[9] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-Object Tracking Through Simultaneous Long Occlusions and Split-Merge Conditions. In *CVPR*, 2006.

[10] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – Learning person specific classifiers from video. In *CVPR*, 2009.

[11] B. Song, T.-Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury. A Stochastic Graph Evolution Framework for Robust Multi-target Tracking. In *ECCV*, 2010.