Spatio-Temporal Constraint for Fast Face Tracking in Movies

Yue Wu[1], Yuan Dong[1], Peng Li[2], Kun Tao[2]
[1]Beijing University of Posts and Telecommunications
[2]Orange Labs International Center Beijing
wuyuebupt@gmail.com, yuandong@bupt.edu.cn,{peng.li, kun.tao}@orange.com

Abstract—**In this paper, a new unified framework of the face tracking is presented. It is based on a new tracking-by-detection tracker. Three different face detectors are utilized in the tracker. By exploiting the spatio-temporal constraint between the intra and inter frames, a Bayesian formulation is proposed to merge different detections from the three detectors and link the faces into tracks. In addition, a salient region in a frame is found by employing context prior knowledge. The tracking procedure is efficiently accelerated because that the three sliding-window based detectors scan in the smaller salient region instead of a whole frame. Our method is evaluated on the standard Hannah dataset, which contains a feature-length movie. The performance is demonstrated to match or exceed the sate-of-the-art. Furthermore, our system is much faster than previous methods.**

1. Introduction

Face-based analysis used for the video content description, indexing and retrieval has long been established. Face detection and face tracking are important components in face recognition systems, which automatically detect and locate the face region. After the process of face detection and face tracking, *face tracks* are extracted from the videos. These face tracks are the basic unit of the analysis, e.g., face clustering [1], face identification [2][3][4] and character naming [5][6][7] . However, this task is challenging. In the real-world videos, the environmental changes (e.g., illumination) may change the appearance of faces. Head pose changes caused by the motion also bring problems. In addition, the blur and occlusions are also common. Hence, the accurate face tracking in real-world videos is very difficult.

Face detection plays a crucial role in face-related vision application. Numerous techniques have been proposed for face detection. The real-time face detection scheme proposed by Viola and Jones [8][9] is arguably the most commonly employed front face detector, which consists of a cascade of classifiers trained by Adaboost employing Harr-wavelet feature. And many other approaches for the multi-view face detection have been proposed (e.g., [10][11]). Due to a number of reliable face detectors have been built, one frontal detector [2][7][12][13] or two and more detectors [5][6][14] for frontal or profile faces are used. And most approaches [2][5][14] split the extraction of face tracks into two steps. The two steps are detecting all faces and linking the detections into face tracks, respectively. Moreover, some metrics under spatial-temporal constraint [5][14] are defined for clustering and generating the tracks. These methods can only merge faces that are detected. Their performance is limited by the face detector. To address the problem, Zhao et al. [12] use an efficient and robust tracker [15] based on facial landmarks to track. They employ a classifier embedded in the tracking system to detect the tracking failures and if the facial feature tracker fails, a re-detection will be conducted to confirm the end of a track. But face appearance variations are always largely changed when the drift happened. The frontal face detector can not handle such big variations. While Kalal etc. [13] use a validator after the detector to decide if the face represents a specific instance selected for tracking. The validator stores all positive and negative samples that have been collected during tracking and learns a multi-view model from the samples. Their approach needs to specific the faces which need to be tracked as a priori and initialize the corresponding validator by some samples. In unconstrained videos, the priori is hardly to get and the specific faces limit the generation ability to more faces.

The goal of online object tracking is to estimate the states of the target in the subsequent frames given a initialized state of a target object in a frame of a video. For most applications, the object to be tracked is unknown and arbitrary. And tracking algorithms can be generally categorized as either generative or discriminative. Generative tracking algorithms typically learn a model to represent the target object. While discriminative algorithms treat the problem as a classification task. Tracking-by-detection is particularly popular recently, which treats the tracking problem as a detection task applied over time. Inspired by this, a tracking-by-detection tracker is integrated into our system.

We propose a framework for face tracking based on a hierarchical structure. Our goal is to automatically track all faces in a video stream. A frontal face detector [16] is run periodically on the time line of the video, and to achieve a low false positive rate, a conservative threshold on detection confidence is used. The detector is used to find new faces and validate the tracking module. And a tracking module based on tracking-by-detection for all faces is built. A unified formulation of grouping different detections and generating the track is proposed in the tracking module.

Our contributions in this work are the following: 1. A new tracking-by-detection tracker that is robust to face variations is introduced. 2. A unified formulation of face tracking is proposed. The formulation is based on a Bayesian framework. 3. A method that exploits context prior knowledge is proposed to accelerate the system efficiently.

The rest of this paper is organized as follows. The framework of the face tracking is introduced in Section 2. Section 3 describes each part of our system, especially the spatio-temporal constraint and how to accelerate the tracking procedure by exploiting context information. Section 4 gives the experiment results, and Section 5 gives the final conclusion of this paper and the future work.

2. System framework

The framework of our system is shown in Figure 2. Firstly, shot boundary detection(SBD) [17] is used to prevent tracking across different shots. When a shot ends, all trackers will be terminated. Next, the reliable periodical face detection is performed to detect new faces and update the existing trackers. Once a face has been detected in this step and the face is not under tracking, a new face tracker will be created. Otherwise, the face will reinitialize the related existing tracker. The tracking-by-detection tracker consists of three Viola-Jones [9] face detectors and a Kanade-Lucas-Tomasi(KLT) based tracker [13]. Different form the periodical face detector with a low false positive rate, these three detectors have a normal threshold. They are used to get an initial set of detections. Then the spatio-temporal constraint is applied to generate tracks from these detections. Context prior knowledge is used to find the salient region in the current frame based on the previous tracking results. All the three detectors will only scan this region instead of the whole frame to accelerate the tracking procedure. Finally, a tracker will be terminated when it fails and the related faces will make up a face track.

3. Method description

In this section, we will describe our method of generating the face tracks. It is based on the following main steps: (1) Shot boundaries detection, (2) Periodical Face detection, (3) Tracking-by-detection tracker. Below, we present these steps in details and discuss their improvement to existing methods.

3.1 Shot boundary detection

Classically, shots are identified through the detection of their boundaries. Many sound technologies can be used for this target. Our method is based on RASH image descriptors [17].

3.2 Periodical Face detection

A frontal face detector [16] with a low false positive rate is run periodically on every K frames of the video. The different values of K are evaluated on the experiments.

3.3 Tracking-by-detection tracker

Our tracking-by-detection tracker involves three main stages: building spatio-temporal constraint, detecting with multi-detectors and tracking based on KLT. This tracker will conduct bi-directional tracking.

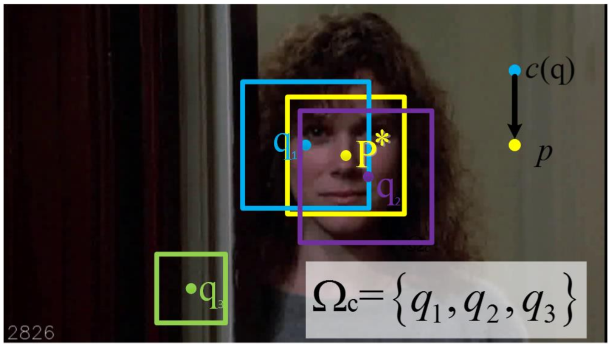

Fig. 1. The schematic diagram of spatial relationship between a face and its local context. The local context region Ω_c is the set of detection results by multi-detectors and it consists of $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{q}_3$,. The context feature at location $\mathbf{q}$ is denoted by $\mathbf{c}(\mathbf{q}) = \mathbf{q}$ including local information.

1) Building spatio-temporal constraint: A local context consists of a face region and its immediate surrounding background within a determined region. In our method, the local context of a face is defined as a set of the detection results of the different face detectors in the frame(see the rectangles in Figure 1). The tracking problem is formulated by computing a probability response which estimates the face location likelihood:

$$r(\mathbf{p}) = P(\mathbf{p}|o) \tag{1}$$

where $\mathbf{p} = (x, y, w, h)$ is a face bounding box and o is the face presenting in the scene. Figure 1 shows its schematic diagram.

In the current frame, we have the face bounding box $\mathbf{q}^*$ (i.e., coordinate of the tracked face center and the width and the height of the bounding box). The context feature is defined as $\mathbf{X}^c = \{\mathbf{c}(\mathbf{q}) = \mathbf{q}|\mathbf{q} \in \Omega_c)\}$ where $\mathbf{q}$ is a bounding box in the image and Ω_c is the set of detection results by multi-detectors. By marginalizing the joint probability $P(\mathbf{p}, \mathbf{c}(\mathbf{q})|o)$, the face location likelihood function in (1) can be computed by

$$\begin{aligned} r(\mathbf{p}) &= P(\mathbf{p}|o) \\ &= \sum_{\mathbf{c}(\mathbf{q})\in\mathbf{X}^c} P(\mathbf{p}, \mathbf{c}(\mathbf{q})|o) \\ &= \sum_{\mathbf{c}(\mathbf{q})\in\mathbf{X}^c} P(\mathbf{p}|\mathbf{c}(\mathbf{q}), o)P(\mathbf{c}(\mathbf{q}), o) \end{aligned} \tag{2}$$

where the conditional probability $P(\mathbf{p}|\mathbf{c}(\mathbf{q}), o)$ models the spatial relationship between the face bounding box and its context information which helps merge simultaneous detections of same faces(i.e., replace them by a single bounding box), and $P(\mathbf{c}(\mathbf{q}), o))$ is a context prior probability which models appearance of the local context.

A. Spatial Context Model

The conditional probability function $P(\mathbf{p}|\mathbf{c}(\mathbf{q}), o)$ in (2) is defined as

$$P(\mathbf{p}|\mathbf{c}(\mathbf{q}), o) = \left\| \frac{\mathbf{q} \bigcap \mathbf{p}}{\mathbf{q} \bigcup \mathbf{p}} \right\| \tag{3}$$

where $\mathbf{p}$ is a bounding box to be evaluated and $\mathbf{q}$ is the bounding box detected by a face detector.

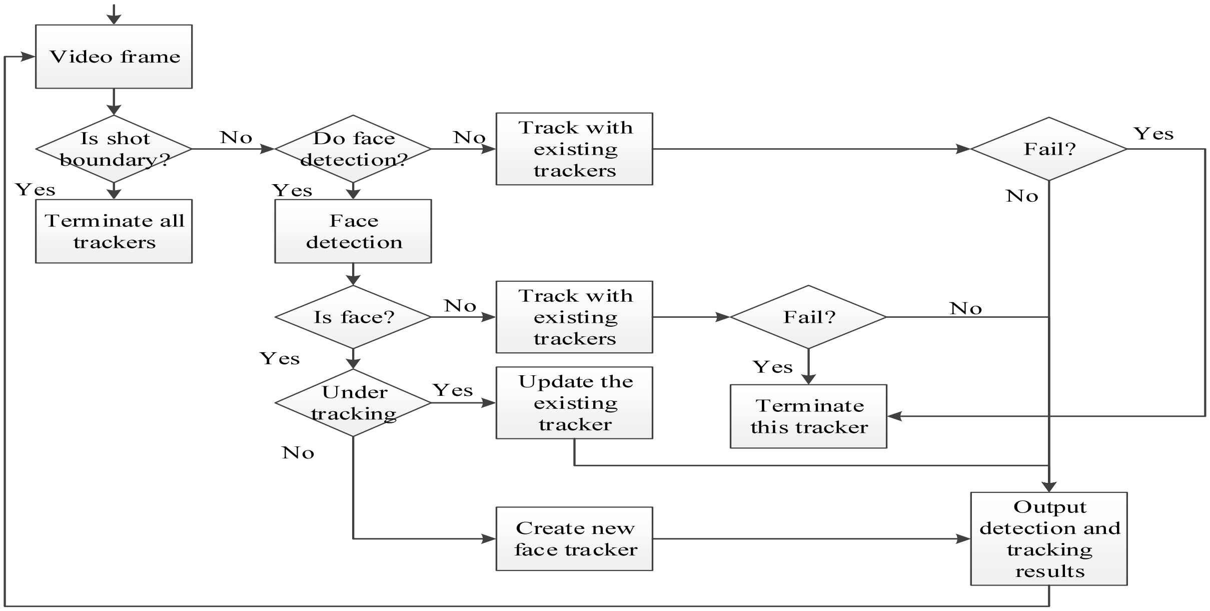

Fig. 2. The framework of our fast face tracking system

B. Context Prior Model

In (2), the context prior probability is simply modeled by

$$P(\mathbf{c}(\mathbf{q})|o) = e^{-\frac{\|\mathbf{q}-\mathbf{p}^*\|^2}{\sigma^2}} \tag{4}$$

where σ is a scale parameter and it is set to $a\frac{w_{q^*}+h_{q^*}}{2}$. The $w_{\mathbf{q}^*}, h_{\mathbf{q}^*}$ is the width and the height of $\mathbf{q}^*$. And a is set to 0.1 empirically.

In (4), it models the focus of attention that is motivated by the biological visual system which concentrated on certain image regions requiring detail analysis [18]. This spatially weighted function indicates the importance of a bounding box at different positions. The closer the new bounding box $\mathbf{q}$ is to he currently tracked face region $\mathbf{p}^*$, the more important it is to predict the face location in the coming frame, and a larger weight should be set.

2) Detecting with multi-detectors: three independent face detectors are used: one for approximately frontal faces based on MS-LTP [16], one for multi-view faces [11], and one Viola and Jones (VJ) detector [9] for profiles.

These three detectors are used to get an initial set of detections. As Figure 1 shows, these detections form the local context region Ω_c.

As is common with sliding-window detectors, the face detectors output multiple responses at nearby locations and scales. The faces between frontal views and profiles are especially easy to be detected by multiple detectors. By exploiting spatio-temporal constraint proposed before, the spatial context model is applied to remove multiple detections and merge ambiguous detections while the context prior model is used to connect faces in temporal and spatial continuity. And conventional non-maximum suppression is applied to (1) to get the final face location.

3) Tracking based on KLT: Though three different face detector are used, there still are many missed detections due to movement, pose variations and illumination changes. The Median Flow tracker [19] is added into our system to track the face when all detectors failed. This will not only interpolate missed detections, but also make the face tracks' boundary more reliable.

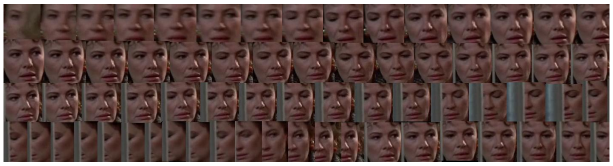

Fig. 3. Extracted faces is shown. Note that the tracker can handle frontal faces and profiles very well.

3.4 Fast Tracking using context prior knowledge

In (4), it is obvious that if $\|\mathbf{q}-\mathbf{p}^*\|^2$ is large, which means the new detection $\mathbf{q}$ is far from the previous tracked face location $\mathbf{p}^*$, the context prior probability $P(\mathbf{c}(\mathbf{q})|o)$ is low. When calculating (1) by (2), the $\mathbf{q}$ makes less contribution to the (1).

To simplify the problem, the width and height of the possible bounding box $\mathbf{q}$ is fixed as the same as $\mathbf{p}^*$. By setting the $P(\mathbf{c}(\mathbf{q})|o) < 0.01$, the envelope of the context region estimated by (4) is a circle centered at $\mathbf{p}^*$ with a radius of $2.1 * a * (\frac{w_{p^*}+h_{p^*}}{2})$. When $a = 0.1$, the radius of the circle is $0.21 * (\frac{w_{p^*}+h_{p^*}}{2})$ For the convenience of sliding-window based detectors, the circle is expanded to a

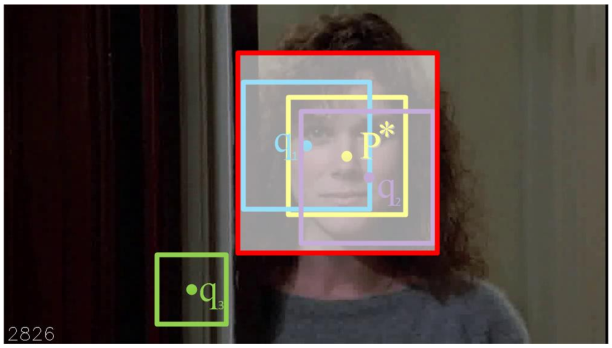

Fig. 4. The context region of a tracked face is inside the red rectangle which includes face region surrounding by yellow rectangle $\mathbf{p}^*$. $\mathbf{q}_1$ and $\mathbf{q}_2$ are two bounding boxes detected by detectors in the context region. The set Ω_c in this frame consists of $\mathbf{q}_1$ and $\mathbf{q}_2$. And $\mathbf{q}_3$ is in the irrelevant region which will not be detected.

rectangular region which centers at the center of $\mathbf{p}^*$, has a width of $w_{p^*} + 2 * 0.21 * (\frac{w_{p^*}+h_{p^*}}{2})$ and has a height of $h_{p^*} + 2 * 0.21 * (\frac{w_{p^*}+h_{p^*}}{2})$. By this way, the context region is formed.

The detection is applied in the context region instead of the whole frame, which greatly saves the time in detecting the irrelevant region for sliding windows based detector.

3.5 Difference with related work

In [3][6], merging different detections in the intra-frame uses the agglomerative clustering method. And in [6], the face tracks is created by extracting the Kanade-Lucas-Tomasi feature while in [3], linking the faces into the tracks by solving an optimization problem. We propose a spatial context model and a context prior model to represent these two steps and merges them into a union formulation based on a Bayesian framework. This idea is also inspired by [20], which exploits the context information to model the object. Figure 3 shows an example of tracked faces with signification subject motion and variations in pose of the face between frontal faces and profiles. In addition, the context prior knowledge is used to accelerate the tracking procedure easily and this is hard for the previous approaches.

4 Experiment

4.1 Data set

The Hannah dataset [14] is used to evaluate the performance of the proposed method. This dataset consists of a feature movie of 1 hour and 40 minutes with dense annotation. The ground-truth annotation contains a frame-by-frame description of all visible faces. Ignoring the "Crowd" boxes [14] , there are 198337 bounding boxes and 1983 face tracks.

The best result in [14] is our "baseline". Choosing different combination of the frequency of periodical frontal face detection and whether to use the context prior knowledge, we build 3 versions of the system:

Step by one frame with no context (**S1NC**): the step of the periodical frontal face detection is one frame and the context prior knowledge is not used.

Step by one frame with context (**S1C**): the step of the periodical frontal face detection is one frame and the context prior knowledge is used.

Step by ten frames with context (**S10C**): the step of the periodical frontal face detection is ten frames and the context prior knowledge is used.

The approaches are all implemented in C++ and tested on a single-core Intel Xeon E4570.

4.2 Experiment results

Using the methods proposed above, we measure the performance of the systems in terms of boxes(all face bounding boxes) , tracks(the number of all face tracks) , MTL(mean track length). And as mentioned in [21], $\overline{FP}$, $\overline{FN}$, $\overline{OP}$, $\overline{TP}$ and $Purity$ are also evaluated in the same condition.

TABLE I
Basic statistics on ground truth and all three systems outcomes

	boxes	tracks	MTL	time
Ground-Truth	198337	1983	100.019	-
baseline	136904	847	160.7	-
S1NC	139811	635	220.174	520m23s
S1C	138763	627	221.31	311m40s
S10C	102256	410	249.405	101m57s

TABLE II
Evaluation of three tracking system in terms of instantaneous detection

	$\overline{FP}(\%)$	$\overline{FN}(\%)$
baseline	17.4	39.2
S1NC	15.71	39.00
S1C	15.51	39.40
S10C	14.12	44.48

TABLE III
Evaluation of three tracking system in term of purity

	$\overline{OP}(\%)$	$\overline{TP}(\%)$	$Purity(\%)$
baseline	22.5	50.6	31.2
S1NC	20.87	58.84	30.81
S1C	20.31	58.90	30.20
S10C	13.03	61.91	21.54

As illustrated in table I, table II and table III. The best performance of our system is "**S1NC**". Compared with the "baseline", it extracts more boxes but less tracks and the $\overline{TP}$ is higher but the $\overline{OP}$ is lower. That's because our periodical frontal face detector has a low false positive rate, which will miss some not-well face and the false alarm will be lower at the same time. That explains that the MTL of our methods is longer than "baseline" and why "**S10C**" has less tracks than "**S1C**". And as expected, the $\overline{FP}$,$\overline{FN}$ of "**S1NC**" are both better than the "baseline".

Then, after employing the context prior knowledge to accelerate the system, "**S1C**" runs much faster than "**S1NC**" with less performance sacrificing.

Doing face detection in every frames is very time-consuming because of the huge number of the frames(e.g., 153,803 frames in the Hannah dataset). By enlarging the step of the periodical frontal face detection from 1 frame to 10 frames, "**S10C**" runs much faster than "**S1C**" with acceptable performance. To our knowledge this is the first time that face tracks are extracted from such a challenge video in real time.

All of the experiment results show that the proposed method has reached the state-of-the-art in an efficient way.

5. Conclusion

In this paper, a unified formulation of the face tracking is presented based on Bayesian framework. Meanwhile an efficient method to accelerate the tracking procedure is proposed. From the experimental results, we could conclude that our method is better and efficient for face tracking compared with the state-of-the-art method. And the fastest version can deal with the challenged video in real time with acceptable performance. In the future work, we plan to investigate incorporating other (non-facial) cues, such as clothing and hair and pay more attention to short tracks in videos.

Acknowledgment

This work is sponsored by collaborative Research Project (SEV01100474) between Beijing University of Posts and Telecommunications and France Telecom R&D Beijing, the National Natural Science Foundation of China (61372169), and the National High Technology Research and Development Program of China (863 Program)(No.2012AA012505).

References

[1] Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, and Qiang Ji, "Constrained clustering and its application to face clustering in videos," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3507–3514.

[2] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Unsupervised metric learning for face identification in tv video," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1559–1566.

[3] Martin Bauml, Makarand Tapaswi, and Rainer Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3602–3609.

[4] Makarand Tapaswi, M Bauml, and Rainer Stiefelhagen, ""knock! knock! who is it?" probabilistic person identification in tv-series," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2658–2665.

[5] Josef Sivic, Mark Everingham, and Andrew Zisserman, ""who are you?"-learning person specific classifiers from video," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1145–1152.

[6] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009.

[7] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Hello! my name is... buffy–automatic naming of characters in tv video," 2006.

[8] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001, vol. 1, pp. I–511.

[9] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[10] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[11] Shuisheng Liu, Yuan Dong, Wei Liu, and Jian Zhao, "Multi-view face detection based on cascade classifier and skin color," in *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on*. IEEE, 2012, vol. 1, pp. 56–60.

[12] Ming Zhao, Jay Yagnik, Hartwig Adam, and David Bau, "Large scale learning and recognition of faces in web videos," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–7.

[13] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Face-tld: Tracking-learning-detection applied to faces," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 3789–3792.

[14] Alexey Ozerov, Jean-Ronan Vigouroux, Louis Chevallier, Patrick Pérez, et al., "On evaluating face tracks in movies," in *IEEE International Conference on Image Processing (ICIP 2013)*, 2013.

[15] Thomas Maurer, Egor Valerievich Elagin, Luciano Pasquale Agostino Nocera, Johannes Bernhard Steffens, and Hartmut Neven, "Wavelet-based facial motion capture for avatar animation," Aug. 7 2001, US Patent 6,272,231.

[16] Zhe Wei, Yuan Dong, Feng Zhao, and Hongliang Bai, "Face detection based on multi-scale enhanced local texture feature sets," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 953–956.

[17] Cedric De Roover, Christophe De Vleeschouwer, Frédéric Lefèbvre, and Benoit Macq, "Robust video hashing based on radial projections of key frames," *Signal Processing, IEEE Transactions on*, vol. 53, no. 10, pp. 4020–4037, 2005.

[18] Antonio Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[19] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Forward-backward error: Automatic detection of tracking failures," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 2756–2759.

[20] Kaihua Zhang, Lei Zhang, Ming-Hsuan Yang, and David Zhang, "Fast tracking via spatio-temporal context learning," *CoRR*, vol. abs/1311.1939, 2013.

[21] Kevin C Smith, "Bayesian methods for visual multi-object tracking with applications to human activity recognition," 2007.