Proceedings of the 2nd World Congress on Electrical Engineering and Computer Systems and Science (EECSS'16)
Budapest, Hungary – August 16 – 17, 2016
Paper No. MVML 105
DOI: 10.11159/mvml16.105

Real-Time Face Tracking for Audience Engagement

Chengyuan Peng, Sari Järvinen, Johannes Peltola
VTT Technical Research Centre of Finland LTD
P.O. Box 1000, FI-02044, Espoo, Finland
Chengyuan.Peng@vtt.fi; Sari.Jarvinen@vtt.fi; Johannes.Peltola@vtt.fi

Abstract – Audience measurement is an application area where face tracking and analysis could automatically provide reliable answers about how consumers and users behave spontaneously in real environments. The solutions can detect age, gender, size of the audience, distance from a display, dwell times and more. By knowing audience emotional reactions and engagement towards the products, content and campaigns, one can offer audience an improved experience and customized content addressing their interests and behaviour. In this paper, we present a pose-invariant face tracking method to automatically monitor audience's attention time in real-time. The current system can handle 3D pose variation up to ±55 in yaw and ±30 in pitch angles.

Keywords**:** face tracking, face detection, 3D face anthropometry, pose estimation, camera calibration

1. Introduction

To communicate with people, one has to get their attention. Digital Signage which enables a two-way communication between digital displays and viewers is a natural choice for reaching large audiences. This application can gather real-time, anonymous information about the audiences while they behave spontaneously in different life environments. One approach to measuring and learning more about audiences in front of screens would be automatically detecting their facial expressions, age, gender, emotion as well as attention time and then mining the reported information to understand changes and trends in the demographics, interests and styles of the audiences.

Some related work can be found in [1] [2]. In [1], the authors concluded that face images are a powerful source of information for attention detection, but they may not always be robustly available. Even with high-resolution video cameras, such a technical improvement may still not capture cases where users' faces are simply not turned towards the system. They thus recommended to combining both depth-based features and face features. In [2], several features characterizing facial and head gestures were used as several aggregation methods over a short time window to capture the temporal dynamics of engagement. They show improved performance over baseline methods that mostly rely on head-pose orientation.

In this paper, we use pose estimation method to obtain audience attention time, therefore, face tracking is a critical prior step that localizes the region of the face, from which a relevant feature set can be extracted and subsequently served as input to the face analysis. However, human face is a dynamic object having high degree of variability in its appearance, accurate and robust face detection and tracking still remains one of the most challenging problems due to appearance variations of face poses, facial expressions, illumination, partial occlusion and motion blur in the real life applications [3].

When tracking a face of a person, in some situations, the face being tracked is not frontal, for example, when an audience turns its face to one side. A face tracking method based only on face detection from a frontal camera would lose the face at this point. These pose changes make the work more challenging. In this paper we focus on improving pose-variant face tracking accuracy. We will present a method which uses the accurate 2D facial landmark feature points extracted from face detection.

We introduce some prior work on pose-invariant face tracking. Affine face tracker based on the eye position mentioned in [4]. This tracker is firstly initialized by the face detector based on the Viola-Jones method, and then uses the template matching algorithm to find out the translation, rotation and scale factors of a frontal face very precisely. The only way to do so is tracking independently two or more points from the face, and according to how one of these moves with respect to the others it is possible to compute the affine motion of the face [5].

In [6], authors presented an automatic technique for handling pose variations for face recognition, which involves learning a linear mapping from the feature vector of a non-frontal face to the feature vector of the corresponding frontal face.

In [7], authors used a 3D morphable model to fit a non-frontal face image and then synthesize a frontal view of the face. In [8], authors learned pose-specific locally linear mappings from patches of non-frontal faces to patches of frontal faces. Their method only handles a discrete set of poses and requires some manual labelling of facial landmarks. In [10], authors used a single AAM to fit non-frontal faces but also require manual labelling. In [11], authors presented a set of prototype non-frontal face images that are in the same pose as the input non-frontal face. In [5], authors used locating facial feature points but use 2D affine warp and apparently rely on manual initialization.

In the following sections, we first introduce our proposed methods. Then in the following section we demonstrate our experimental results. Section 4 concludes the paper and highlights the future work.

2. Our Methods

This section we outline the steps of our methods. The first step is to detect faces. Haar-like feature-based Cascade Classifiers were used with pre-trained cascade classifiers for straight frontal face detection. However, faces detected by Viola and Jones might be not always detectable by the Viola-Jones algorithm, for example, when a subject turns his back and walks away from the camera, his/her face totally disappears. In this scenario face detection techniques cannot be applied, but it can still be tracked because we already know where some of face feature points are in the previous frame.

We designed face tracking techniques that, for the present frame and having the face detected in the previous one, should be able to find its location in the present one. That is, the idea was to find the motion between any two given frames. Lucas and Kanade proposed a method for estimating sparse optical flow within a window around a pixel and it is easy and fast to compute because there is no descriptor computation and no scale-space analysis is involved [6]. The Lucas-Kanade algorithm has become one of the most important sparse optical flow techniques, mainly because it can be easily applied to a subset of points in an input image. This method relies only on the local information that comes from the surrounding area of each point of interest. The optic flow is used to predict the new positions of the facial feature points.

Good initialization in Lucas and Kanade method is particularly important when there is large pose variation. To achieve the goal, our method uses a robust fitting method to find 68 facial landmark points [8]. Using these points as initialized input points to Lucas and Kanade model, the facial feature points that cannot be detected by face detection can be predicted using this method.

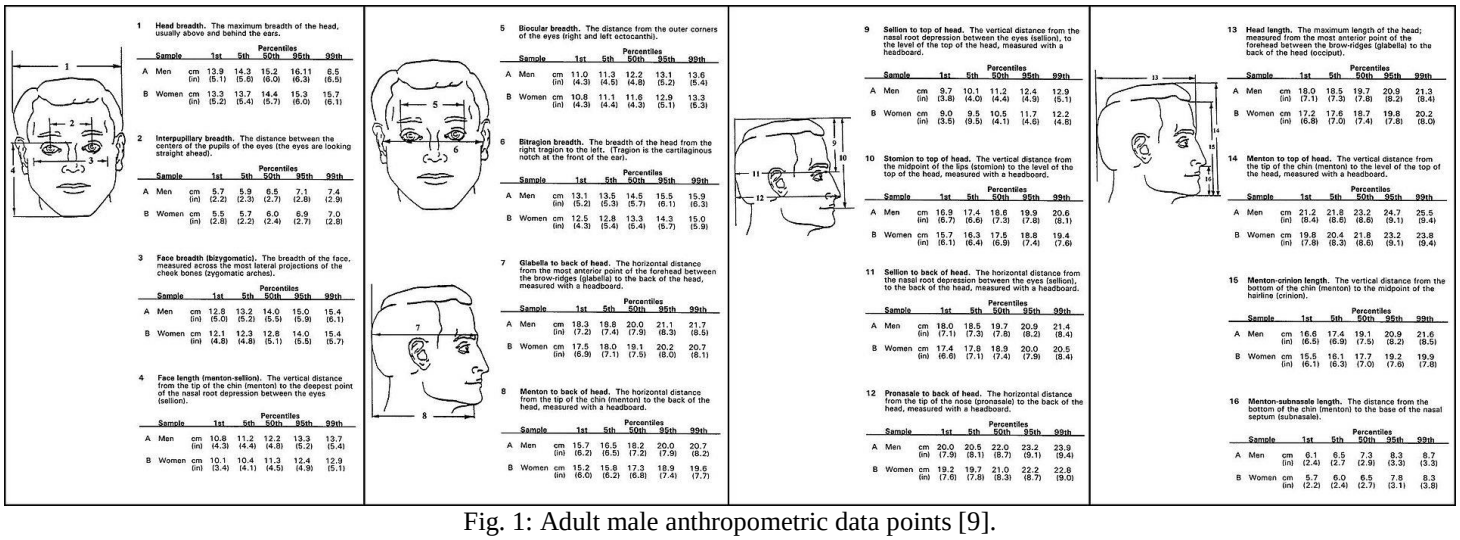

Fig. 1: Adult male anthropometric data points [9].

Instead of using manual marking of several facial feature points, an initial set of facial feature points are automatically extracted that ensure good initialization for optical flow. Facial feature point extraction is incorporating landmark-based features and uses adult male anthropometric data to match a real 3D head to the projected face on an image. 3D points on a human head are relative to sellion values (see Fig. 1, enlarge the figure to see clearly) [9]. The facial features take the form of 68 landmarks. These are points on the face such as the corners of the mouth, along the eyebrows, on the eyes, and so forth. Finally the 3D head points are projected onto 2D face image.

In real-time face tracking, large amount of time is wasted on searching faces in faceless frames. Typical face detection and tracking are conducted frame by frame and window by window. In terms of face detection, the time spent on filtering out a faceless frame is comparable to that on identifying a frame containing faces as every search window needs to be checked to ensure all possible faces are detected.

We define the audience attention time as the time that the person looked into a camera centre. Therefore, we estimate the yaw and pitch angles of the faces (see Fig. 3). In order to measure audience's attention time, we use automatic head pose estimation method. Estimating the 3D head pose from a single 2D image automatically is a key step in our approach to pose-invariant face tracking. Our method for estimating the face pose angles from 3D-2D point correspondences, that is given a set of face landmark points, their corresponding image projections, as well as the camera intrinsic parameters and the distortion coefficients, the pose can be estimated.

3. Implementation and Results

In this section, we present the proposed framework for face detection, tracking and attention estimation. It supports detection and tracking of multiple faces at the same time, and runs in real-time. It also supports face identification feature. Experiment results demonstrate the performance improvement brought by the proposed method. The Viola and Jones method based on Haar-like features and weak classifiers is currently the most robust. It was able to detect up to 95% of the frontal faces in images, with a very low false positive rate.

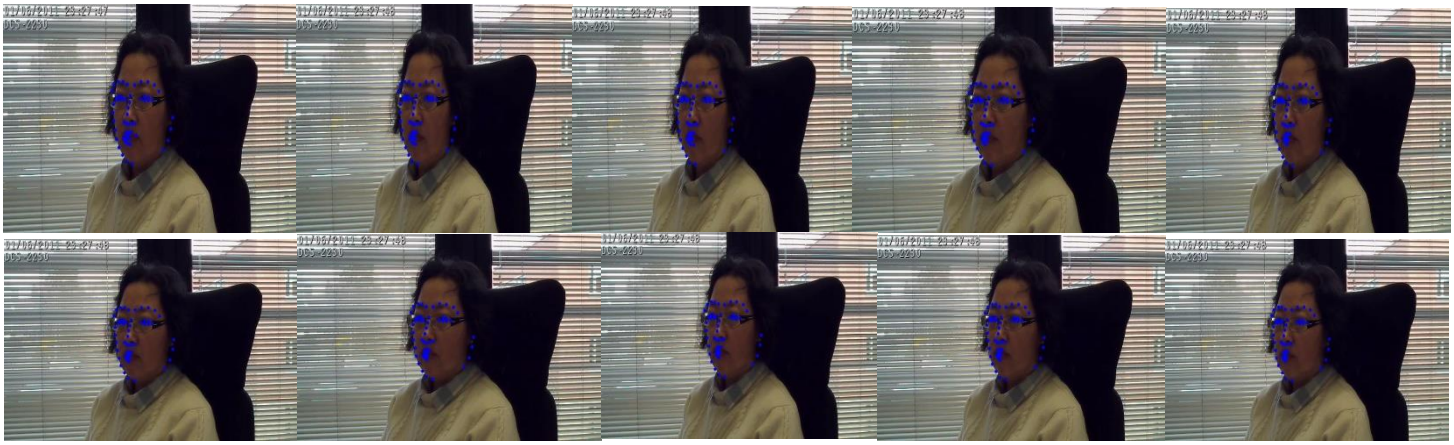

Fig. 2: Predicted facial landmark feature points.

Fig. 2 shows ten consecutive frames where the tracker is running in which the facial landmark feature points cannot be detected but can be predicted using the proposed optical flow method. The implementation demonstrates a significant improvement achieved in face tracking. For pose estimation, the camera must be calibrated to get camera's intrinsic and extrinsic parameters. The face and feature detectors we use are Viola-Jones type cascades of haar-like features trained using AdaBoost algorithms.

Our detector can handle faces with yaw angles roughly from −75 to +75 and pitch angles roughly from −50 to +50. Fig. 4 gives the attention time estimation obtained and analysed from yaw and pitch angles. This leads to a tracker that significantly improves robustness against abrupt appearance changes and occlusions. It is critical for the subsequent recognition phase. The resulting tracker is robust and provides accurately face tracking. The confidence of the tracking result is measured and it was 8:10.

In addition, our face recognition method is based on 3D non-rigid face representation and uses sparse local features as observation which includes both off-line trained major facial features and online tracked image features.

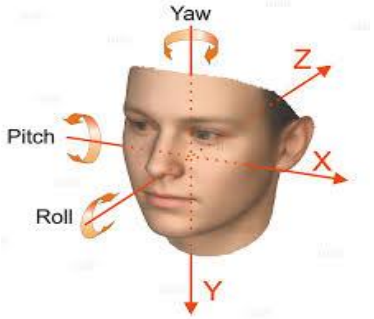

Fig. 3: Yaw, pitch, roll.

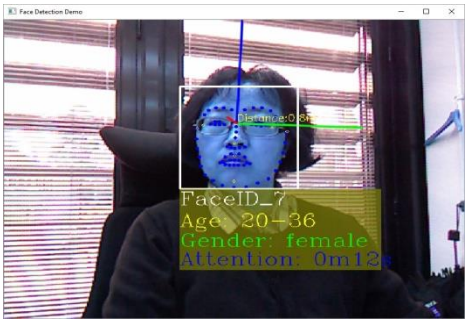

Fig. 4: Attention time estimation.

4. Conclusions

Measuring audience attention as well as expression, emotion, intention and age is an active application area. We demonstrated that our tracker improves robustness against abrupt appearance changes. With the accurate face localization together with pose estimation in the presence of illumination and pose variations, we can improve face tracking and recognition performance. It can be used as a basis of similar applications.

Furthermore, our method handles a continuous range of poses and is thus not restricted to a discrete set of predetermined pose angles. Our main contribution is a fully automatic system for audience measurement application. Other contributions include the use of pose-dependent correspondences between 2D landmark feature points and 3D head model, a method for 3D pose estimation as well as age, gender and face identification in real-time.

In the future, we plan to measure audience's emotion such as smile, sad, etc. in real-time applications.

References

[1] F. Alt, A. Bulling, L. Mecke, and D. Buschek, "Attention, please! Comparing Features for Measuring Audience Attention Towards Pervasive Displays," in *Proceedings of the ACM SIGCHI Conference on Designing Interactive Systems (DIS)*, Brisbane, QLD, Australia, 2016.

[2] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of TV viewers," in *Proceedings of Automatic Face and Gesture Recognition (FG), 10th IEEE International Conference and Workshops*, Shanghai, China, 2013, pp. 1-7.

[3] F. Ahmad, A. Najam, and Z. Ahmed, "Image-based Face Detection and Recognition: 'State of the Art'," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 6, pp. 169, 2012.

[4] D. Duan and J. Ma, "Sensor-Assisted Face Tracking," *International Journal of Distributed Sensor Networks*, vol. 2015, 2015.

[5] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," arXiv: Tech Report, pp. 1-27.

[6] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and Rohith MV, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *Proceedings of International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 939-944.

[7] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. Anchorage, AK, pp. 1-8.

[8] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, pp. 1867-1874.

[9] Wikipedia. [Online]. Anthropometry. Available: https://en.wikipedia.org/wiki/Human_head.

[10] J. Cheney, B. Klein, A. K. Jain, and B. F. Klare, "Unconstrained Face Detection: State of the Art Baseline and Challenges," in *Proceedings of the 8th IAPR International Conference on Biometrics (ICB)*, Phuket, Thailand, 2015.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 815-823.