Face Recognition and Head Tracking in Embedded Systems

Personalization and gaze-based user interfaces in vehicles and other devices

New, efficient algorithms make it possible to embed head tracking and face recognition with high frame rates in devices with simple, mobile CPUs. This provides an unintrusive basis for personalization of cars, TV sets, and many other devices. Head-tracking enables the creation of simple and intuitive user interfaces that are sensitive to the user’s gaze point.

Face-Recognition and head-tracking methods apply a sequence of processing steps of varying complexity to sequences of video frames. The first step localizes faces, the subsequent face feature detection step localizes a number of fiducial positions in the face. Based on these positions, the spatial orientation and direction of gaze can be inferred and a frontal view of the face can be reconstructed. In the face recognition step, a classifier uses reference images to determine the identity of the person on the image.

Use cases

In the automotive environment, face recognition creates benefit for a variety of use cases. Driver identification can be connected to personalization functionality such as the pre-selection of infotainment settings, and mirror and seat adjustment. Face recognition can serve as an anti-theft measure for delivery and utility vehicles that is robust against possible negligence of drivers.

Today, driver attention assist systems detect sluggishness in the driver’s response to minor deviations from the center of the lane by monitoring lateral vehicle acceleration and steering wheel manipulation. As advanced driver assistance systems let drivers take their hands off the steering wheel while cruising on a highway, attention assistance systems will have to resort to other input. The direction of gaze and the status of the eye lids captured by interior cameras are most directly linked to the driver’s attention.

Head tracking can contribute to higher road safety and lower insurance costs for
commercial fleets. Today, telematics units that supply drivers with logistic
information, can at the same time transmit speed and acceleration statistics to the
fleet operators. Fleet supervisors use this information to specifically encourage
employees to drive safely and economically.

A number of logistics companies has been able to leverage these data to
significantly reduce both their risk of accidents and their insurance costs.
Gesture-based user interfaces that are sensitive to the user’s direction of gaze are
an exciting outlook; and as the number of devices that can be controlled by
gestures will increase, head tracking will become a crucial means of identifying
which object a gesture has been directed at. A wiping gesture might open or close
the window that the driver is looking at; the same gesture might be used to
navigate through a menu when the user is looking at the infotainment screen or a
head-up display. A hushing gesture while looking at a radio or stereo set would
arguably be the most intuitive way of muting it.
TV sets can benefit particularly from face recognition: streaming services such as
Netflix ask the user to log on as a registered user using the remote control in order
to receive personalized recommendations. Having the TV set itself identity all of the
users who are jointly watching at any time would not only be more comfortable, but
it would also allow the TV set and connected streaming services to recommend
offerings that are suitable for this particular combination of family members.

Face detection

The first step of face detection localizes faces within the image. Face detection has
been embedded in digital cameras, smart phones, and many other devices. The
most prominent face recognition algorithm iterates over all sizes and positions of
image regions. For each region, a cascade of classifiers decides whether it shows a
face. On each level of the cascade, an ensemble of simple classifiers tests a small
set of image features; several types of texture features are popular, including Haar,
HOG, DSIFT, LBP, and binary pixel features. Based on the outcome of the test, the cascade terminates with a negative decision or moves on to the next level. Since for most image regions a negative decision can be made on one of the topmost levels of the cascade, face detection is relatively fast. By implementing face detection on specialized hardware such as FPGAs it can further be accelerated.

**Face feature tracking**

The following step, face feature detection, localizes between 5 and around 50 fiducial points in the face – for instance, points that mark the outlines of upper and lower lips, upper and lower eye lids, and nose ridge (Fig. 1). Once their position have been determined, these fiducial points are tracked in subsequent video frames without a new face detection.

Recently, the shape regression approach has proven to be more accurate and orders of magnitude faster than earlier methods. Shape regression uses images of faces in which the fiducial positions have been labeled manually as training data. The method solves an optimization problem in order to find the parameters of a regression model which maps initial fiducial positions and a vector of texture features at these positions to offsets that move the fiducial markers towards their correct positions. This regression model is applied repeatedly until the markers converge towards their final positions. A confidence model that is also trained from labeled data determines whether there really is a face at the final marker positions.

A new method that uses binary tree features reduces the execution time of shape regression by another order of magnitude. In a first stage, the method uses labeled training data to generate ensembles of decision trees that predict the offset of the fiducial markers based on local image features. Each nonterminal node of a decision tree implements a case distinction based on one texture feature extracted from the vicinity of the marker. The advantage of this approach is that in order to process each video frame, only the features that are tested in one single branch of each tree have to be calculated. In this respect, the approach alludes to the Viola-Jones face detection procedure. The trees are then combined in a global regression model that predicts the final offset vector for all markers.

We have extended face feature tracking by a model of partial occlusions – see Fig. 2. An occlusion model determines which fiducial points are visible in an image. Efficient shape regression algorithms are suitable for tracking facial features on simple, mobile CPUs in real time at a high frame rate. All computations can be implemented using integer arithmetic, so that they can be carried out efficiently even without a floatingpoint unit. On a single-core ARM A9 CPU architecture, efficient shape regression can easily exceed rates of 10 frames per second.
**Head tracking and gaze estimation**

Head tracking determines the position and orientation – the yaw, pitch, and roll angles – of the head. The three-dimensional orientation of an object can be inferred from a single, two-dimensional image if a three-dimensional model of the object is known and corresponding fiducial points are localized both in the image and in the model. The level of accuracy of this inference depends on the number of fiducial points, and the accuracy of the three-dimensional model. Here, head tracking and face recognition are interleaved and leverage each other: With identification, the three-dimensional model can be adapted to the shape and size of a specific user’s head over time.

Since humans move their head to change their direction of gaze substantially, the head’s spatial orientation allows to approximate the direction of gaze; eye movements primarily serve to scan the frontal field of vision with the foveal area that allows highly resolved vision. However, in order to determine – for instance – whether a user has focused on a particular pedestrian in a street scene or on a particular advertisement on the screen, it may be necessary to additionally determine the exact fixation points with high precision. To this end, remote eye tracker locate retinal and corneal reflections of infrared light using cameras that resolve the eye region in high resolution and with high frame rates, and infer the orientation of the eyes. Head tracking can be based on simple cameras with a wide viewing angle, and inference of the facial orientation is achieved by efficient shape regression using simple, mobile CPUs. Eye tracking – while adding a higher precision to the gaze estimation – requires hardware that is still too expensive to become ubiquitous in the near future.

Sufficiently accurate face feature tracking can trace the outline of upper and lower eye lids (Fig. 1). The area between the eye lids determines the degree to which the eyes are opened.

**Face recognition**

Using a three-dimensional model, it is possible to infer a frontal perspective of the face. From a semi-profile image, a little over half a frontal face can be reconstructed; the second half is facing away from the camera and can only be reconstructed under symmetry assumptions.

From a frontalized greyscale image of a face, filter banks infer vectors of typically tens of thousands of texture features. From these elementary features, more abstract features are inferred, often in several layers of sequential feature transformations. These higher-level features reflect differences between individuals rather than differences in lighting conditions, poses, and other factors. They can be
created by searching for transformations that specifically discriminate between certain pairs of individuals, or by training neural networks on millions of images of faces. Using the entirety of all generated features, a classifier decides whether an image shows a person that has been enrolled using reference images and determines a confidence score.

For a long time, research on face recognition has focused on fully frontal, biometric images with good lighting. For applications such as immigration and law enforcement, these assumptions may be realistic; in an automotive environment, they are not. Today, the accuracy of the best known face recognition algorithms reaches human recognition accuracy for images with good lighting conditions and no occlusions but some variations in head pose and expression. Robust face recognition for images that simultaneously vary in head pose, expression, lighting conditions and that may contain partial occlusions are still the subject of research.

Face recognition is generally fairly expensive in terms of computation time and memory. Filter banks that infer tens of thousands of features can easily reach many megabytes in size. Feature selection algorithms make it possible to reduce the model size at the cost of a lower recognition accuracy. For many use cases, user identification has to be completed in the order of one second. The computationally expensive application of filter banks imposes requirements on the CPU that are not met by all mobile CPUs.

We have developed a method that computes and evaluates only those texture features that are strictly necessary to make an identification decision for the image at hand. This method reduces the computational costs of face recognition by more than an order of magnitude. It makes it possible to execute the entire pipeline of face feature tracking, head tracking, and face recognition in real time, at a high frame rate on energy-efficient and inexpensive mobile CPUs and embed it on a wide range of devices. For instance, on a single-core A9 CPU it is possible to obtain rates of 10 frames per second for the entire pipeline. The implementation is fully self-contained, has its own memory management and does not require any specific operating system.

A highly efficient face recognition does not only open up a wider range of hardware platforms on which face recognition can be executed in real time. When a higher number of different frames can be analyzed in the time that is available for user identification, the recognition becomes more robust, false-positive and false-negative identifications become less likely.

**Conclusion**

New, efficient and robust algorithms enable the deployment not only of face detection, but also of real-time face feature tracking, head tracking, and face
recognition on embedded systems that use simple, energy-efficient and inexpensive CPU. Face recognition and head tracking can be embedded jointly since they share most of the processing pipeline and have the same low hardware and camera requirements. Head tracking at a high frame rate provides an approximation of the user’s direction of gaze that meets the accuracy requirements of many use cases; the spatial and temporal resolution can be augmented with eye tracking for others. Face-recognition-based user identification allows to personalize the interaction between users and devices. In vehicles, this allows to activate infotainment preferences, seat and mirror settings; for delivery trucks, driver identification can serve as a security feature. At the same time, head tracking and tracking of the eye lids can monitor drivers’ attention even when advanced assistance systems allow the drivers to take their hands off the steering wheel. In a future in which a multitude of items can be controlled by gestures, head tracking will be able to resolve which item a gesture was directed at.

**Autor(en)**

Lenka Ivantysynova, managing director and co-founder of Asaphus Vision GmbH, Berlin, Tobias Scheffer, professor for computer science at the University of Potsdam and co-founder of Asaphus Vision GmbH

**Kontaktieren**

*Asaphus Vision GmbH*
*Bismarckstraße 10 - 12*
*10625 Berlin*
*Telefon: +49 30 850 191 77*