

Towards a Real-time and Distributed System for Face Detection, Pose Estimation and Face-related Features

J. Nesvadba¹, A. Hanjalic², P. M. Fonseca¹, B. Kroon^{1/2}, H. Celik^{1/2}, E. Hendriks²

¹ Philips Research, Eindhoven, The Netherlands

² Delft University of Technology, Delft, The Netherlands

Abstract

The evolution of storage capacity, computation power and connectivity in *Consumer-Electronics*(CE)-, in-vehicle-, medical-IT- and on-chip-networks allow the easy implementation of grid-computing-based real-time and distributed face-related analysis systems. A combination of facial-related analysis components - *Service Units* (SUs) – such as face detection, pose estimation, face tracking and facial feature localization provide a necessary set of basic visual descriptors required for advanced facial- and human-related feature analysis SUs, such as face recognition and facial-based mood interpretation. Smart reuse of the available computational resources across individual CE devices or across in-vehicle- or medical-IT-networks in combination with descriptor databases facilitate the establishment of a powerful analytical system applicable for various domains and applications.

Keywords

Face detection, pose estimation, face tracking, content management.

1 Introduction

Through the fast evolution of processing power, storage capacity and connectivity [1] in CE-, in-vehicle- and medical-IT-networks, generic *Multimedia-Content-Analysis*- (MCA-) and computer-vision-based analysis solutions start to reach human brain's semantic levels. Powered by smart usage of scattered processing power, storage and bandwidth available across those networks, realization of real-time high-level semantic analysis systems do not belong to the realm of fiction any more. Multiple cross-domain and cross-organizational collaborations [2], combinations of state-of-the-art network and grid-computing solutions, and usage of recently standardized interfaces facilitated the set-up of an advanced analytical system, further referenced to as *CASSANDRA Framework* (CF) [3]. This prototyping framework enables distributed computing scenario simulations for e.g. *Distributed Content Analysis* (DCA) across CE In-Home networks, but also the rapid development and assessment of complex multi-MCA-algorithm-based applications and system solutions. Furthermore, the modular nature of the framework - logical MCA and computer vision components are wrapped into so-called *Service Units* (SU) - eases the split between system-architecture- and algorithmic-related work and additionally facilitate reusability, extensibility and upgradeability of those SUs. Additionally, the modularization allows smart network management systems to balance the processing load across the available resources in applicable networks (e.g., CE In-Home networks). Such an elaborated DCA system can be seen as

basis for *Ambient Intelligence* (AmI) applicable in various domains, such as CE, medical IT, car infotainment and personal healthcare.

In many of these application domains, one of the most important elements is the human face. Therefore, indication of its location, its identity and even its expression provide useful semantic information. For this reason, one of the most prominent AmI-related problems is the availability of a reliable real-time face-analysis system. Consequently, various face-related SUs have been or are being jointly researched [2], implemented and integrated into the CF, further described in this paper. These comprise SUs such as omni-directional face detection, face tracking, face recognition, face online learning, facial features- and facial points-analysis. In combination, these SUs provide the basic visual descriptors for advanced facial- and human-related feature analysis and applications.

2 Distributed Face Analysis System

The realization of a real-time distributed face analysis system requires modularization of face analysis algorithms and standardization of face-related descriptors, which is the basic concept of the CF. In [1], the first attempt of such a modularization is described for the specific case of a face recognition system; this system includes the required underlying SUs *Face Detection* (SU FD) and *Face Tracking* (SU FT). CF-based evaluations highlighted the limiting capabilities of the implemented face detectors [1] in providing the necessary information for reliable face recognition. Consequently, new face detection algorithms are currently being researched that shall be able not only to localize faces regardless of their spatial orientation but also to achieve higher overall detection performances. Furthermore, these new algorithms will allow the implementation of mid-level SUs such as *SU Pose Estimation* (SU PE) (see *Figure 1*) - providing indication of the spatial orientation information of localized faces; additionally, *SU Facial Features* (SU FF) will determine position of ears, nose, eyes, etc. All collected facial data is thereafter used as input for SUs *Face Recognition* (SU FR), *Online Face Clustering* (SU OFC), *Facial Feature Points* (SU FFP) and *Facial Expression* (SU FE; emotion/mood interpretation) analysis, which are currently also under investigation. *Figure 1* illustrates the relation between such SUs.

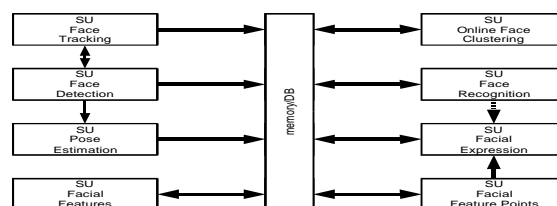


Figure 1 – Face-analysis-related SUs.

2.1 Existing Face Detection Algorithms

To extract face-related features like pose, gaze direction, identity, facial expression and mood, face detection is an essential step. With this in mind, face detection has been and still is extensively researched. One of the various face detection algorithms, a low complexity color-based method, performs detection in the compressed domain. This method is unequaled in computational efficiency but is not capable of handling monochrome video and due to its extreme low-complexity, it only performs satisfactorily under controlled conditions. To overcome these disadvantages, another algorithm was developed based on the Viola Jones-based learning method. However, this method has the shortcoming of only being able to detect upright frontal faces. Both algorithms are briefly described in the next sections.

2.1.1 Compressed Domain Face Detection

The compressed domain face detection algorithm [4] uses a feature-based approach to determine the presence and location of multiple frontal faces using only DCT coefficients extracted from compressed content (images). Face detection is accomplished by first performing skin color segmentation based on a model built from the statistical color properties of a large set of manually segmented faces. After applying binary morphological operators on the segmented image, specific subsets of the input AC coefficients are used, along with the brightness properties of the input image to determine in SU FF the location of specific facial features (eyes, eyebrows and mouth). Finally, using a model of typical frontal faces, face candidates are generated based on the location of these facial features. Face candidates are then ranked according to their size, their percentage of skin color pixels and the intensity of their facial features. Finally, the most relevant face candidate is chosen for each individual skin color region.

As illustrated in *Figure 2*, even though the face detector is intended for detection of frontal faces, it is also able to correctly determine the location of faces that are rotated and tilted up to a certain limit.

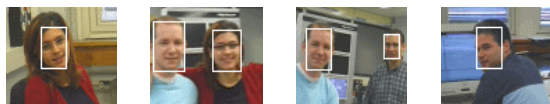


Figure 2 – Examples of correctly detected faces.

2.1.2 Viola-Jones-based Face Detection

Besides the compressed-domain-based face detector described in the previous section, a Viola-Jones based face detection algorithm [5][6] was implemented for evaluation purposes. This image-based detection algorithm works on uncompressed images and has proven to be robust under various lighting conditions. The method is based on a cascade of boosted classifiers of simple Haar-wavelet like features on different scales and positions. The features are brightness- and contrast-invariant and consist of two or more rectangular region pixel-sums that can be efficiently calculated by the Canny integral image. The feature set is overcomplete and an adaptation of the AdaBoost learning algorithm is proposed to select and combine features into a linear classifier. To speed up detection a cascade of classifiers is used such that every classifier can reject an image. All classifiers are trained to reject part of the candidates such that on average only a low amount of features are used per position and scale.

After all possible face candidates are obtained, a grouping algorithm reduces groups of face candidates into single positive detections.

This detection method has been mapped to a smart camera [7][8]. The smart camera detects multiple frontal faces of different sizes in images and allows small rotations ($\pm 10^\circ$). The face detection application is running at a rate of 4 frames per second.

2.2 Current Face Detection Research

As explained, the methods discussed in paragraph 2.1 are sensitive to color conditions and face pose. Current research addresses these limitations in an attempt to develop algorithms that allow the extraction of face-related features in uncontrolled scenarios regardless of pose and illumination conditions. The main difficulty in developing an omnidirectional face detector is related to the fact that the 2-D visual appearance of an object depends on its pose. To distinguish a face from other objects regardless of its pose, either a set of pose-dependent detectors operating in parallel, a complex “brute force” learning method, or a 3-D model fitting technique is required. For the first kind of detectors (parallel pose-dependent detectors), the in-plane and out-of-plane pose range (i.e.: rotation axis perpendicular or parallel to the image viewing plane respectively) is partitioned into a number of areas for which an independent detector is designed - this kind of omnidirectional face detector is called a multiview detector.

In the following sections, examples of face detectors that use the first two of these techniques are analyzed and their applicability for robust and real-time omnidirectional face detection in video content is discussed.

2.2.1 The Schneiderman-Kanade Method

In [9], Schneiderman and Kanade describe an object detection method applied to face detection. The proposed algorithm was one of the first efficient face detectors in literature that could determine the location of non-upright frontal faces. Besides being able to attain multiview face detection, it copes with variations in pose by using two specific classifiers trained separately: one for detection of frontal faces and another for detection of profile faces. The profile detector is trained for right profile view points and applying it on the vertical mirrored image allows for left profile face detection. As a result, faces with in-plane rotation between -15° and $+15^\circ$ and full-profile faces (-90° to $+90^\circ$ rotation out of plane) can be detected. For each view-point (profile, right-frontal and left-profile), the corresponding detector scans the original image and its downscaled versions at several locations. Images are analyzed with windows of size 48×56 for the frontal detector and 64×64 for the profile detector. The decision is based on a Bayesian classifier on joint values and positions of visual *attributes*. An *attribute* is here defined as a group of quantized wavelet coefficients in given sub-bands. In total, 17 different attributes are involved, a detailed description of which can be found in [9]. Attributes are sampled at regular intervals over the detection window (coarse resolution).

2.2.2 Viola-Jones-based Methods

In Section 2.1.2, a Viola-Jones frontal face detector was presented. In this section extension of that method for omnidirectional detection is discussed.

Omnidirectional face detection could be achieved simply by training a Viola-Jones detector with face images of all

poses. However, this would imply that a huge number of selected features would be needed in order to incorporate all different face appearances. Naturally, the complexity of the algorithm would become unbearable, especially for real-time implementations. In order to avoid this problem, a multiview Viola-Jones detector – i.e., in which a single detector is designed for each pose range – may be developed. It may be achieved according to one of the two following strategies: all detectors could perform classification in parallel or a single selector could be used for detection using the information of a pre-processing pose estimator, i.e. SU PE. Both approaches are described in existing literature.

In [10], Viola and Jones propose to train a C4.5 decision tree on 12 poses, 10 levels deep without pruning. The paper covers both in-plane and out-of-plane rotation, but does not present a complete solution. It is argued by the authors that a pose estimator would have approximately the complexity of one detector, which renders the method only twice as intensive as a frontal detector. The pose estimator/single classifier approach should thus be faster than the parallel classifiers approach. For this reason, a potentially robust real-time multiview Viola-Jones-based classifier system employing different kinds of base classifiers is envisioned. The classifiers in this system can be divided into two groups:

1. A pose estimator can quantize poses in order to reduce the classification problem for other classifiers. The pose estimator can be used on all image positions and scales prior to detection such that for non-face areas the pose will be arbitrary.
2. A pose-specific face detector classifies between face and non-face; detectors can be cascaded and of multiple types; simple detectors are used to quickly reduce false alarms without sacrificing recall, while more complex (and slower) detectors may be used to increase precision by validating remaining face candidates.

It may be observed that the original Viola-Jones detector is actually a cascade of classifiers; thus, an omnidirectional face detector may be actually built from a large tree structure of simple classifiers. Current research work may thus be regarded as an attempt to identify and design the optimal structure of such a system.

2.2.3 The Convolutional Face Finder

The third face detector, a *Convolutional Face Finder* (CFF) [11], is based on a multi-layer *Convolutional Neural Network* (CNN). CNNs were originally intended and designed for handwritten digit recognition.

It is designed for faces rotated between -20° and $+20^\circ$ in-plane, and between -60° and $+60^\circ$ out-of-plane and relies, unlike previous methods, only on a single detector.

The CFF consists of six successive neural layers. The first four layers extract characteristic features, and the last two perform the actual classification (face/non-face). The CFF is applied on several resized instances of the original image at several positions. The input of the system is a 32×36 window extracted from each rescaled image. The first step consists of convolving this input with 5×5 kernels and adding a bias; 4 kernel variants are applied, resulting in 4 different *feature maps*. The produced feature maps are then down-sampled by a factor of two, multiplied by a weight, and corrected by a bias before a sigmoid activation function is applied. Subsequently, this convolution/sub-sampling scheme is repeated with 3×3 masks resulting in 14 new feature maps which consist on the characteristic

features extracted for classification. The last two layers, comprised of traditional neural processing units decide on the presence of a face.

This face detector is an example of a monolithic “brute force” approach for the problem of omnidirectional face detection.

2.2.4 Comparative Analysis of the Methods

It is important to note that the abovementioned methods achieve omnidirectional face detection only for a limited range of in-plane and out-of-plane rotations. Upside-down oriented faces, for instance will likely not be detected. To achieve true omnidirectionality, multiple detectors have to be combined.

As explained earlier, the objective of this research work is twofold: while the aim is to efficiently detect faces regardless of their pose, this should be achieved on video content in real-time with a reasonable amount of processing power.

The Schneiderman-Kanade detector achieves high detection rates (above 90% on the CMU frontal set); it performs especially well on difficult profile face images (similar rates on the CMU profile test set) when compared to other multiview systems. The drawback of this approach lies on its computational cost, unacceptable for the purpose at hand, even if the heuristics described in [9] are included. Based on experiments conducted during current research, it was found that processing of each image takes several seconds.

The CFF is able to detect frontal and difficult semi-profile faces with a high detection rate and a very low false alarm rate, without using a specific detector for a given viewpoint or without running a pose estimator. Garcia and Delakis [11] report detection rates on the CMU Frontal set of around 90%, with an execution speed of approximately 4 frames per second for 384×288 images on a 1.6GHz P4 processor. Consequently, it appears suitable for our scope in terms of processing speed but this detector does not perform well on full-profile faces, which is a considerable disadvantage.

Finally, the combination of the omni-directional Viola-Jones pose estimator (SU PE) / pose-specific face detector (SU FD), as described in 2.2.2, proved to be the fastest of the methods analyzed. A frontal Viola-Jones FD runs at approximately 15 frames per second on a 3.2GHz P4 processor in images with 720×576 resolution, so the combination of pose estimator followed by a face detector is estimated to run roughly at 7 frames per second; current experiments point towards this assumption. *Figure 3* compares the detectors on a qualitative speed vs. detection performance plot.

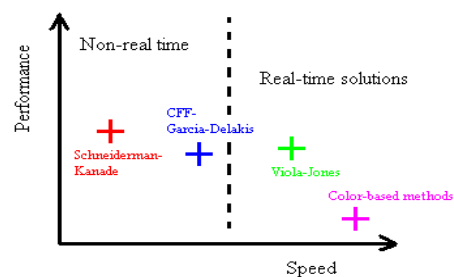


Figure 3 – Qualitative comparison of face detectors.

Viola-Jones-based detectors (frontal, described in Section 2.2, or omni-directional, described in Section 2.3.2) exhibit the best trade-off between speed and performance. Skin-color based methods, like the compressed-domain method described in Section 2.1, are extremely fast, but

have not proven to be sufficiently robust. On the other extreme, the Schneiderman-Kanade method shows a good detection performance but with a relatively low speed performance.

The Schneiderman-Kanade detector achieves the best performance on full profile face images. The drawback with this approach is that two different detectors trained for different views are used. The image is then scanned three times (once for each profile and one for frontal view), which further slows down the process. An original approach would be to apply this method after a Viola-Jones detector with pose estimation. The use of heuristics such as skin color filtering could also significantly improve the speed performance on color video or image content.

Concerning the CFF, it appears to be very robust, while covering a wide range of views (especially for semi-profiles in the range -60° to $+60^\circ$). Garcia and Delakis [11] evoke a more complex version with additive feature maps, in order to detect full profiles. Using two CFFs trained for frontal face and full profile could be a sound approach both in terms of execution time and detection performance. Another efficient procedure for Convolutional Neural Networks could be the combination of simultaneous pose estimator and face detector, which would also yield in a real-time system.

Finally, based on our experimentations and results reported in the literature, the conclusion is that an omnidirectional face detector should incorporate a pose estimator and a face detector, instead of consisting in several detectors applied separately on the image, if the objective is to achieve detection and speed performances suitable for the applications the algorithms are intended for.

2.3 SU Pose Estimation

As described in the previous section, pose estimation can be used as a valuable pre-processing step to face detection, being also very useful on its own. The pose of a face can be defined as one in-plane and two out-of-plane angles with a known low tolerance. The description of a face pose may provide useful semantic information. It may be used, for instance, to determine if people are facing one specific direction or if two persons are facing (possibly talking to) each other. This information can also facilitate the determination of facial points since it allows 3-D model fitting with the faces in the images. Pose estimation can thus aid in the determination of facial points of profile and non-upright faces; which in turn can help identifying and analyzing the expression of these faces.

2.4 SU Face Tracking

The previous section discussed several methods to detect faces in still images. However, to view a video as a collection of still images is a considerable naïve approach. Using the temporal dimension of video for object detection may lead to improvement in both localization and speed performances for two reasons, both related to the trivial observation that adjacent video frames are likely to share similar content:

1. False object detections and recognitions and wrong pose estimates may occur in single frames; by combining information from multiple frames, part of the false alarms can be removed and parameter accuracy can be increased without actually increasing computational complexity.

2. In frames that belong to the same shot, faces are unlikely to suddenly appear or disappear and objects do not change dramatically their position or size from frame to frame; this observation allows for a substantial reduction of the search window used for subsequent frames after initial detections have taken place.

Temporal localization of a face may also provide helpful cues for face identification.

3 Conclusions

In this paper, the potential of the Cassandra Framework's modular [3] approach – using SUs for individual services – in combination with face-related content analysis algorithms has been described. The framework provides an easy-to-use prototyping environment enabling the real-time execution of efficient and heterogeneous face-related algorithms, such as omnidirectional face detection, pose estimation and face tracking in a distributed environment. The high modularity of this real-time distributed system will trivially allow the addition of new face-based solutions, such as individual identification, facial expression recognition, or mood estimation. Current research on face detection was also discussed and some conclusions were drawn regarding the direction in which current work will proceed towards a robust efficient omnidirectional face detector.

References

- [1] J. Nesvadba, P. M. Fonseca, et al., *Face Related Features in Consumer Electronic (CE) device environments*, Proc. Int'l Conf. on Systems, Man, and Cybernetics, pp 641-648, The Hague - Netherlands, October 2004.
- [2] MultimediaN: www.multimedien.nl/, Cassandra: www.research.philips.com/technologies/storage/cassandra/, Candela: www.hitech-projects.com/euprojects/candela/
- [3] J. Nesvadba, P. Fonseca, et al., *Real-Time and Distributed AV Content Analysis system for Consumer Electronics Networks*, Proc. IEEE Int'l Conf. for Multimedia and Expo, Amsterdam - The Netherlands, June 2005.
- [4] P. Fonseca, J. Nesvadba, *Face Detection in the Compressed Domain*, Proc. IEEE Int'l Conf. on Image Processing 2004, pp. 2015-2018, Singapore – Singapore, October 2004.
- [5] P. Viola, M. Jones, *Rapid Object Detection using a Boosted Cascade of Simple Features*, Proc. IEEE Computer Vision and Pattern Recognition, 2001.
- [6] R. Lienhart, J. Maydt, *An Extended Set of Haarlike Features for Rapid Object Detection*, Proc. IEEE Int'l Conf. on Image Processing, Vol 1, pp. 900-903, 2002.
- [7] Philips Centre for Industrial Technology, *Inca 311: Smart Firewire Camera with Rolling Shutter Sensor*, <http://www.cft.philips.com/industrialvision>, 2004.
- [8] R. Kleihorst et al., *An SIMD Smart Camera Architecture for Real-time Face Recognition*, Abstracts of the SAFE & ProRISC/IEEE Workshops on Semiconductors, Circuits and Systems and Signal Processing, Veldhoven - The Netherlands, 2003.
- [9] H. Schneiderman, T. Kanade, *A statistical method for 3D object detection applied to faces and cars*. International Conference on Computer Vision, 2000.
- [10] M. Jones, P. Viola, *Fast Multi-View Face Detection*, MERL, TR2003-96, July 2003.
- [11] C. Garcia, M. Delakis, *Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, November 2004.