Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 24 (2006) 564–572
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Face detection and tracking in video sequences using the modified census transformation

Christian Küblbeck *, Andreas Ernst

Department of Electronic Imaging, Fraunhofer Institute for Integrated Circuits, Am Wolfsmantel 33, 91058 Erlangen, Germany

Received 7 December 2004; received in revised form 8 July 2005; accepted 23 August 2005

Abstract

We present the combination of an illumination invariant approach to face detection combined with a tracking mechanism used for improving speed and accuracy of the system. We introduce illumination invariant local structure features for object detection. For an efficient computation we propose a modified census transform, which enhances the original work of Zabih and Woodfill [19] [Ramin Zabih, John Woodfill. A non-parametric approach to visual correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996.]. The tracking is performed by means of continuous detection. We show that the advent of new rapid detection algorithms may change the need for traditional tracking. Furthermore the mentioned problems have a natural solution within the presented tracking by continuous detection approach. The only assumption on the object to track is its maximal speed in the image plane, which can be set very generously. From this assumption we derive three conditions for a valid state sequence in time. To estimate the optimal state of a tracked face from the detection results a Kalman filter is used. This leads to an instant smoothing of the face trajectory. It can be shown experimentally that smoothing the face trajectories leads to a significant reduction of false detections compared to the static detector without the presented tracking extension. We further show how to exploit the highly redundant information in a natural video sequence to speed-up the execution of the static detector by a temporal scanning procedure which we call 'slicing'. A demo program showing the outcomes of our work can be found in the internet under http://www.iis.fraunhofer.de/bv/biometrie/ for download.

Keywords: Face detection; Face tracking; Census transformation; Kalman filter; Illumination invariance

1. Overview

The paper arranged in three main parts. The first part deals with the face detection mechanism on each single frame. It deals with the problem of illumination invariant detection of faces. Often compensation with a simple lighting model followed by histogram equalization is performed like first proposed by Sung [16] in the context of face detection. Many other detection approaches adopt this strategy with minor modifications, see Refs. [13,14,18]. Schneiderman [15] and Viola [17] apply a simpler normalization to zero mean and unit variance on the analysis window. Illumination compensation demands more computational power than the classification of an image patch itself. For example Yang [18] presented the simple linear SNoW classifier for detection, which operates directly on pixel intensities. It would require only to accumulate the outcome of 400 lookup operations to classify an image patch of size 20×20 while alone the histogram equalization on the same patch takes a higher effort. Therefore, we advocate the use of inherently illumination invariant image features for object detection, which convey only structural object information. We propose the new feature set local structure features computed from a 3×3 pixel neighborhood using a modified version of the census transform [19]. We use this feature set with a four-stage classifier cascade. Each stage classifier is a linear classifier, which consists of a set of lookup-tables of feature weights. Detection is carried out by scanning all possible analysis windows of size 22×22 by the classifier. In order to find faces of various size the image is repeatedly downscaled with a scaling factor of $S_{\mathrm{mra}}=1.25$. This is done until the scaled image is smaller than the sub window size. For typical images of size 388×284 pixels and a sub window size of 22×22 we obtain 10 downscaled images, which constitute an image pyramid. To further speed up processing the pyramid is scanned using a grid search, which we introduced with our former work [8].

* Corresponding author. Tel.: +49 9131 776 549; fax: +49 9131 776 588.
E-mail address: christian.kueblbeck@iis.fraunhofer.de (C. Küblbeck).

0262-8856/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.08.005

The second part shows how to use tracking mechanisms to increase speed as well as accuracy of the face detector. In literature many methods are described that can be used to determine the new object location. One simple method is to take the pixel block covered by the object at the current location and perform block-matching to locate the block in a new image [1]. In face-tracking mostly more elaborate model-based approaches are used. Edwards [5] for example uses a 2D active appearance model for tracking, while DeCarlo [3] proposes a 3D approach to also handle out-of-plane rotation. Similar work was done by Dornaika [4] or La Cascia [2], but based on a different feature set.

Pure tracking approaches as the ones mentioned above have two major deficiencies which often limit their application in real-world scenarios, namely:

- Initialization problem, i.e. how does the tracker know when a new object enters the scene and thus a new track has to be started.
- 'Lost track' problem, i.e. can we decide if the tracker lost track because the object has left the scene or due to a tracking error.
- In Ref. [10] the problem is for example termed 'Entry and Exit' of interesting objects from the scene and its solution is identified as being one of the key issues in fully automated tracking systems. Many approaches therefore need manual segmentation of the object in the first frame [12] or use color segmentation or image difference energy based methods [10], which fail in case of a moving camera. If tracking breaks down due to occlusion or tracking errors the algorithms are widely unable to recover. Most algorithms also have no criterion to decide whether the tracked object exits the scene, or at least do not give any experimental evidence that they are able to detect such events. With the advent of a new class of rapid detection algorithms like those proposed by Viola [17] and Li [11] or that developed in this work, see also [7], the motivation for tracking may change fundamentally. Now it is possible to treat each frame of a video sequence as static image and perform a full search for desired objects. As those algorithms have good detection performance and low error rates the 'Entry and Exit'-problem is solved implicitly. In this work we suggest a tracking by continuous detection approach which means, that each new frame is processed by a real-time static face detector, which has state of the art detection properties and error rates [7]. The outcome of this detector is combined in an association step, using a prior motion model. The main difference to existing approaches is that the motion model is not used to restrict the search area of the detector but only to connect detections from different frames.

In the third part we show the outcome of our experiments on three videosequences. We compare the recognition rates between static detection and detection using the tracking feature.

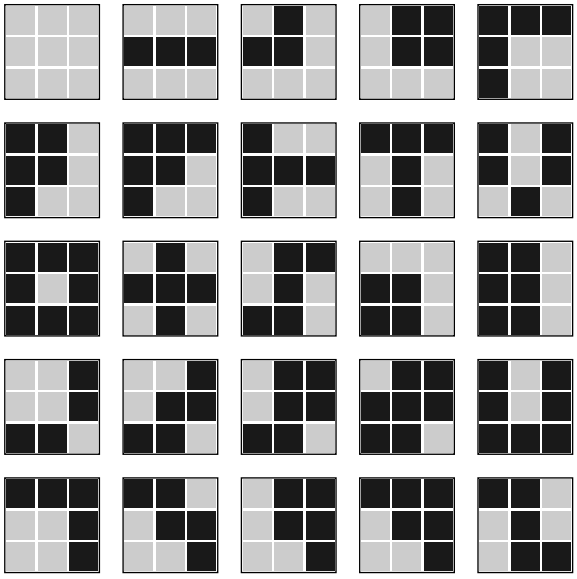

Fig. 1. A randomly chosen subset of 25 out of 2^9-1 possible local structure kernels in a 3×3 neighborhood.

2. Feature generation

2.1. Local structure features

The features used in this work are defined as structure kernels of size 3×3 which summarize the local spatial image structure. Within the kernel structure information is coded as binary information {0,1} and the resulting binary patterns can represent oriented edges, line segments, junctions, ridges, saddle points, etc. Fig. 1 shows some examples of this structure kernels. On a local 3×3 lattice there exist $2^9=512$ such kernels. Actually there are only 2^9-1 reasonable kernels out of the 2^9 possible because the kernel with all elements 0 and that with all 1 convey the same information (all pixels are equal) and so one is redundant and thus is excluded. At each location only the best matching kernel is used for description. The whole procedure can be thought of as non-linear filtering where the output image is assigned the best fitting kernel index at each location. In the next section we show the index of the best kernel can be obtained by a modified version of the census transform (Fig. 2).

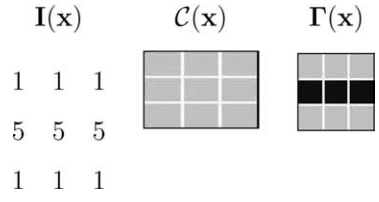

Fig. 2. Example for the difference between the original and the modified Census Transform: **I** shows an example (micro) image consisting of 3×3 pixels. $\mathcal{C}(\mathbf{x})$ shows the result of the original transform and $\mathbf{\Gamma}(\mathbf{x})$ shows the more accurate result of the modified transform.