Unsupervised Learning of Face Detection Models from Unlabeled Image Streams

Thomas Walther
Universität Paderborn, Fakultät für Elektrotechnik, Informatik und Mathematik

Rolf P. Würtz
Ruhr-Universität Bochum, Institut für Neuroinformatik

Abstract: Modern artificial face detection shows impressive performance in a variety of application areas. This success comes at the cost of supervised training, using large-scale databases provided by human experts. In this paper, we propose a face detection system based on *Organic Computing* [vdM08] paradigms that acquires necessary domain knowledge autonomously and learns a conceptual model of the human face/head region. Performance of the novel approach is experimentally compared to state-of-the-art face detection, yielding competitive results in scenarios of moderate complexity.

1 Introduction

Humans with their social perception skills are quite adept in vision-based face identification and interpretation of other peoples' facial expressions. Although the biological foundations of these abilities are still hardly understood it can safely be assumed that these skills rely on fast and robust cerebral face detection (FD) mechanisms. Recent technical solutions mimicking human abilities in face detection show good performance and target a prospering market.

This success comes at a cost: artificial FD solutions are critically dependent on the availability of purposeful *face models*. These patterns can be provided in a variety of ways (cf. [YKA02]), nevertheless, but are always based on *domain knowledge* [Wal11] provided by human supervisors. This is a tedious, time-consuming and costly task, and the resulting databases often have to be tailored to the system's expected operating conditions. Worse, the vast majority of FD solutions are unable to extend such restricted databases autonomously; thus, failure in unforeseen scenarios is programmed.

The above problems seem to have no counterpart in the biological systems: the visual cortex acquires useful domain knowledge 'from the input itself' [Gor06] in a completely autonomous manner. Herein, sophisticated 'concept building' [Wal11], 'generalization', and 'nontrivial learning' [PB04] mechanisms are of prime importance. Endowing standard face detection with these central abilities will potentially reduce human effort and render the resulting systems more reliable in novel scenarios. This is studied within the *Organic Computing* [vdM08] (OC) domain; recently, OC principles have been used to enhance autonomy in articulated human body modeling [Wal11]. Linking to and extending this

work, the current paper modifies elements of the *meta model* proposed in [Wal11] in order to form an autonomously assembled, OC-inspired face detector based on 'Gabor wavelet' information.

2 Learning meta models of the upper human body

The face detection scheme proposed here relies on an autonomously learned, abstract representation of the human body, the so-called *meta model* [Wal11]. Assume an array of N_{M} unlabeled input video streams of $150 - 300$ frames, capturing a single human subject performing smooth, fronto-parallel upper-body motion in front of a static and moderately cluttered background. Using basic motion segmentation techniques in combination with *graph cut* [BJ01] allows to separate the moving foreground subject reliably. Afterwards, a set $\mathcal{F} = \{\mathbf{f}_0, ..., \mathbf{f}_{N_{\mathrm{F}}-1}\}$ of trackable features is distributed uniformly on the extracted foreground entity; *Kanade-Lucas-Tomasi* tracking [TK91] yields motion trajectories for each feature. A self-tuning[ZMP04] variant of spectral clustering [vL07] groups features according to the similarity of their trajectories, resulting in a set of *feature groups* $\mathcal{G}_i,\ i \in \{0...N_{\mathrm{G}} - 1\}$. Assuming body parts to move as coherent entities allows to correlate each feature group with an observed limb; kinematic constraints between the identified feature groups are introduced using a modified version of the skeleton assembly techniques found in [Kra03]. To complete limb extraction, the sparse feature groups have to be converted into compact limb templates; on that behalf, each foreground pixel $\mathbf{x}$ is assigned to limb template i via (cf. [Wal11])

$$i = \arg \min_{k \in \{0,...,N_{\mathrm{G}}-1\}} \min_{\mathbf{f}_j \in \mathcal{G}_k} \|\mathbf{f}_j - \mathbf{x}\| \ .$$

The derived limb templates and the skeleton structure are combined within a *sequence-specific* [Wal11] 'pictorial structure' [FH00] (PS) model of the observed upper human body. Repeating the above model extraction procedure for each input sequence yields a PS model array $\mathcal{M} = [\mathbf{M}_0, ..., \mathbf{M}_{N_{\mathrm{M}}-1}]$. Each single $\mathbf{M}_i$ can be expected powerless for matching purposes in generic scenarios (see [Wal11]). However, sophisticated learning strategies can be employed to combine all sequence-specific PS models to form a much 'more generic and powerful *meta model*' [Wal11]$\mathbf{M}_{\text{meta}}$. This process is described elsewhere [Wal11]. For this paper we describe the involved *Gabor limb prototypes* [Wal11].

3 Gabor prototypes

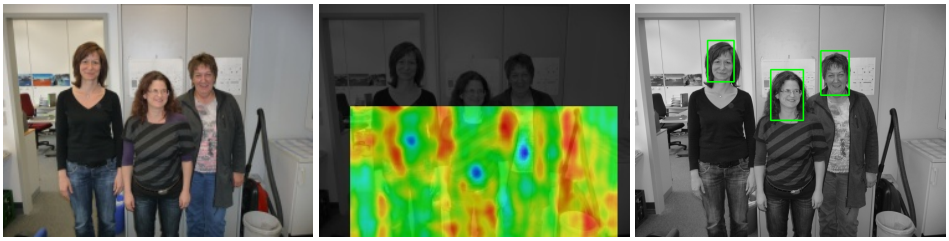
Gabor wavelets represent a widespread method to access spatial frequency information in gray scale images [LVB$^+$93, WFKvdM97]) and defined as

$$\begin{aligned}\Psi_{\mathbf{k}_{\nu\mu}}(\mathbf{x}) &= \frac{\|\mathbf{k}_{\nu\mu}\|^2}{\sigma_{\mathrm{G}}^2}\cdot\exp\left(-\frac{\|\mathbf{k}_{\nu\mu}\|^2\|\mathbf{x}\|^2}{2\sigma_{\mathrm{G}}^2}\right)\left(\exp\left(i\mathbf{k}_{\nu\mu}^T\mathbf{x}\right)-\exp\left(-\frac{\sigma_{\mathrm{G}}^2}{2}\right)\right),\\ \mathbf{k}_{\mu,\nu} &= \begin{pmatrix}k_\mu\cos\varphi_\nu\\ k_\mu\sin\varphi_\nu\end{pmatrix};\ k_\mu=2^{-\frac{2+\mu}{2}}\pi;\ \varphi_\nu=\frac{\nu\,\pi}{8}.\end{aligned} \tag{1}$$

Like in [WFKvdM97, Wal11] scales are sampled according to $\nu\in\{0,...,4\}$ and orientations are discretized by $\mu\in\{0,...,7\}$.

A Gabor *jet* [WFKvdM97] 'collects wavelet filter responses at a dedicated image position from all $\nu\cdot\mu$ subband images' [Wal11]. Image features based on such Gabor jets are widespread in 'biologically motivated' [KS07] face recognition based on *'elastic (bunch) graph matching'* [WFKvdM97]. Herein, single jets correlate to dedicated 'facial landmarks' [Wal11] and are attached to the nodes of a deformable 'face graph' [WFKvdM97]. The edges of this graph employ spring-like constraints to ensure spatial coherence of the face model (cf. [WFKvdM97]). Face recognition is eventually based on graph comparison; similarity of the connected jets is evaluated by a variety of 'comparison functions' [Wal11]. Within the current context, access to the jets' absolute values will be sufficient for jet comparison; nevertheless, more sophisticated comparison methodologies exist (cf. [GW09]).

For face *detection* purposes autonomously learned grid graphs of Gabor jets (the aforementioned Gabor limb prototypes) allow for reliable face detection in complex images. Note that the system proposed in [Wal11] originally learns Gabor prototypes for all observed N_{G} body parts. Leaving detailed discussion of this procedure to [Wal11], assume that the nodes of the generated Gabor prototypes $G_{\mathrm{G},i}, i\in\{0,...,N_{\mathrm{G}}-1\}$ carry texture information that remains stable across all input sequences. As Gabor wavelets are not rotationally invariant [Gün11] and cloth texture varies 'significantly between sequences' [Wal11], stable texture samples will evolve exclusively in the face region (due to negligible head rotation and stable face texture) of the meta model's torso element. Thus, all Gabor prototypes $G_{\mathrm{G},i}$ except the torso become depleted of nodes and are eventually pruned. The remaining $G_{\mathrm{G,torso}}$ 'turns into a generic *texture-based torso detector* that optimally responds to human torsi in upright position' [Wal11]. To actually perform torso detection in an input image $I(\mathbf{x})$, $G_{\mathrm{G,torso}}$ is swept (at scales $s\in\{0.7,...,1.0\}$, discretized in N_{dis} steps, with upright orientation) over a Gabor jet representation $\mathcal{G}_I(\mathbf{x})$ of some query image $I(\mathbf{x})$. Comparing Gabor jet information from the swept graph's nodes with Gabor jet information from $\mathcal{G}_I(\mathbf{x})$ yields a 'Gabor cue map' [Wal11] $G_s^{\mathrm{torso}}(\mathbf{x})$; minima of this map correlate to the barycenters of putative torso candidates in $I(\mathbf{x})$. Fig. 1b demonstrates application of $G_{\mathrm{G,torso}}$ to the *query image* in fig. 1a; the pronounced bluish minima indicate reliable detection of each observed torso, grayish regions correspond to areas where 'the projected Gabor graph has at least one node outside the image area' [Wal11] and thus could not reliably be evaluated. In [Wal11], use of the described torso detection scheme



(a) Query image (b) Gabor cue map (c) Face detection

Figure 1: The face detection cycle: Gabor information is in heat map style; red color indicates high values, while blue color corresponds to minima. Detected faces are highlighted by green rectangles.

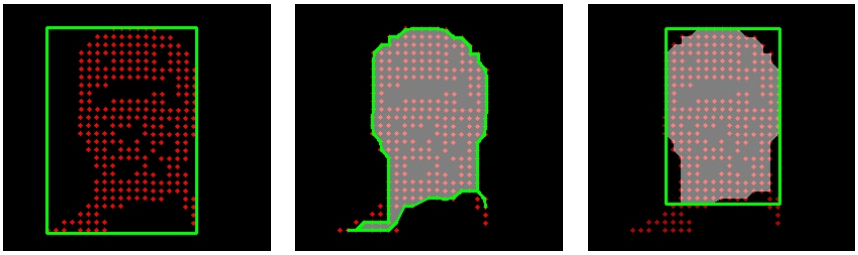
was deliberately restricted to search space reduction in articulated posture analysis. By the modifications described below, however, $G_{\text{G,torso}}$ demonstrates potential in multi-person face detection and will experimentally be shown to compete with state-of-the-art *Viola-Jones face detection* [VJ04] in moderately complex scenarios.

4 Gabor-based face detection

Turning $G_{\text{G,torso}}$ into a reliable face detector is quite straightforward, the powerful *OpenCV* [Bra00] library provides most of the necessary algorithms. In a first step, the face region of the meta limb's torso is localized: given ideal circumstances, this could be achieved by finding the bounding rectangle of all nodes in $G_{\text{G,torso}}$. However, stable nodes of the torso's grid graph might not evolve exclusively on the true face area, but also on the head/shoulder transition; such 'outlier' nodes would unnecessarily inflate the face rectangle, as indicated in fig. 2a.

The initial bounding rectangle can seed an 'active contour model' [KWT88]: evolving this 'snake' [KWT88] over < 1000 time steps, it eventually yields a tight perimeter (shown as green line in fig. 2b) that fits all graph nodes. During the evolution process, nodes close to the convex hull of $G_{\text{G,torso}}$ are charged with higher attraction weights, in order to prevent the snake from excessive shrinking. Morphological opening of the perimeter's inside area (sketched as white overlay in fig. 2b) eliminates outlier influences and gives a compact approximation of the true face region (white overlay in fig. 2c). This compact structure can well be represented by a single encompassing rectangle $\mathbf{R}_{\text{face}}$ (overlaid as green line in fig. 2c). Let $\mathbf{b}_{\text{face}}$ be the barycenter of this face rectangle in torso-centric coordinates. With that, the torso detection method of [Wal11] can directly be employed as a single-face detection scheme: by adding $\mathbf{b}_{\text{face}}$ to the position of the most pronounced Gabor cue map minimum (across all sweeping scales), $\mathbf{R}_{\text{face}}$ can be projected into the query image domain.

However, *multi-face* detection requires a more sophisticated approach: assume that minima in each $G_s^{\text{torso}}(\mathbf{x})$ cluster tightly around the true barycenter positions of all torsi ob-



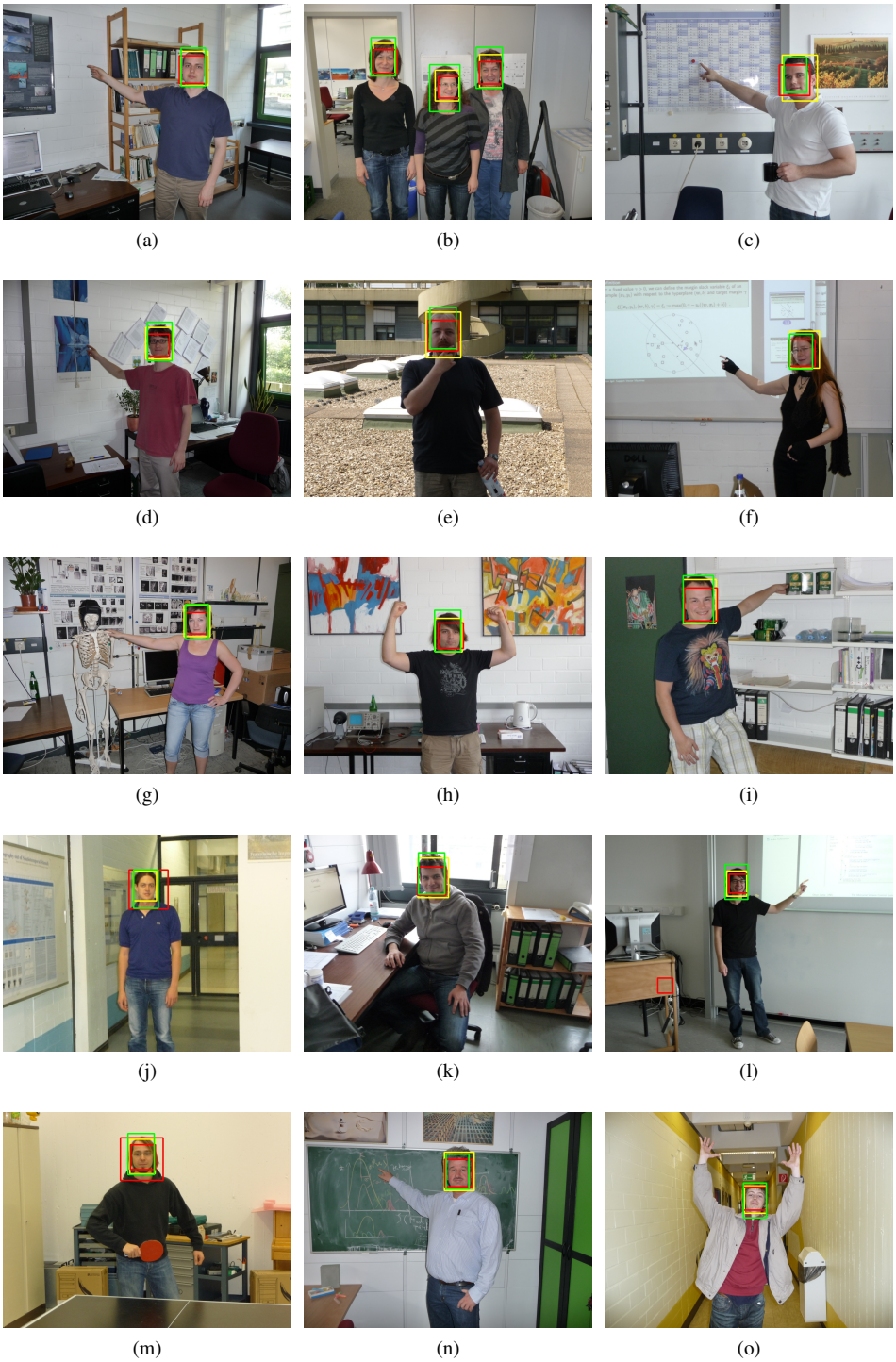
(a) Bounding rectangle (b) Evolved snake (c) Face rectangle

Figure 2: Face rectangle evolution for Gabor-based face detection: rectangle and snake perimeters are shown in green, Gabor graph nodes are indicated as red dots.

served in $I(\mathbf{x})$. Multiway spectral clustering (cf. [MX03],[vL07]) based on Euclidean distances reliably identifies the single clusters and provides, via the *eigengap* [vL07] criterion, a good estimate of the cluster number K. This allows to perform *K-means* clustering on the 'spectral embedding' [vL07] coordinates of all minima positions and thus identifies the final *torso clusters.* Each cluster's most pronounced minimum is then extracted and stored; this procedure is repeated on all scale levels. In a concluding spectral clustering step, the stored minima are re-grouped; clusters with less than N_{dis} members can safely be deemed instable across scales and are eliminated. The most pronounced minima of the remaining K_{r} clusters are considered true torso detections and are charged with $\mathbf{b}_{\text{face}}$ to arrive at the final face detections $\mathbf{d}^i_{\text{face}}$, where $i \in \{0, ..., K_{\text{r}} - 1\}$. Using $\mathbf{d}^i_{\text{face}}$ to project $\mathbf{R}_{\text{face}}$ into the query image plane is straightforward and demonstrated in fig. 1c; observe that the projected face rectangles (green overlays) trace the true face/head regions neatly. It remains to assess the quality of the proposed Gabor-based face detection (GBFD) scheme w. r. t. other state-of-the-art face detection approaches; competitive behavior of the former method will be demonstrated in the following experiments.

5 Experimental evaluation

Being popular in the computer vision community (cf. [ZZ10]), the *Viola-Jones face detection* (VJFD) scheme is a good contemporary candidate to compare Gabor-based face detection to. Coarsely speaking, VJFD finds multiple faces in a given query image by cascaded evaluation of 'Haar-like features' [VJ04] within a 'sliding window' [ME07] approach. Here, the VJFD method implemented in OpenCV is employed; threshold-based skin color detection (cf. [EMH09]) in the *Lab* color space biases this standard VJ scheme and counters detection of false positives. To the same end, the proposed GBFD method is enhanced with autonomously acquired color information from the meta model's *color prototypes*; for details on this technique, refer to. [Wal11]. Comparison of VJFD and GBFD is performed on the *INIPURE* (Institut für NeuroInformatik – PostURe Estimation)database [Wal11]; using this proprietary database instead of publicly available ones (cf., e. g., [SR03] or [HRBLM07] for a comprehensive overview) has three main reasons:



(a) (b) (c)

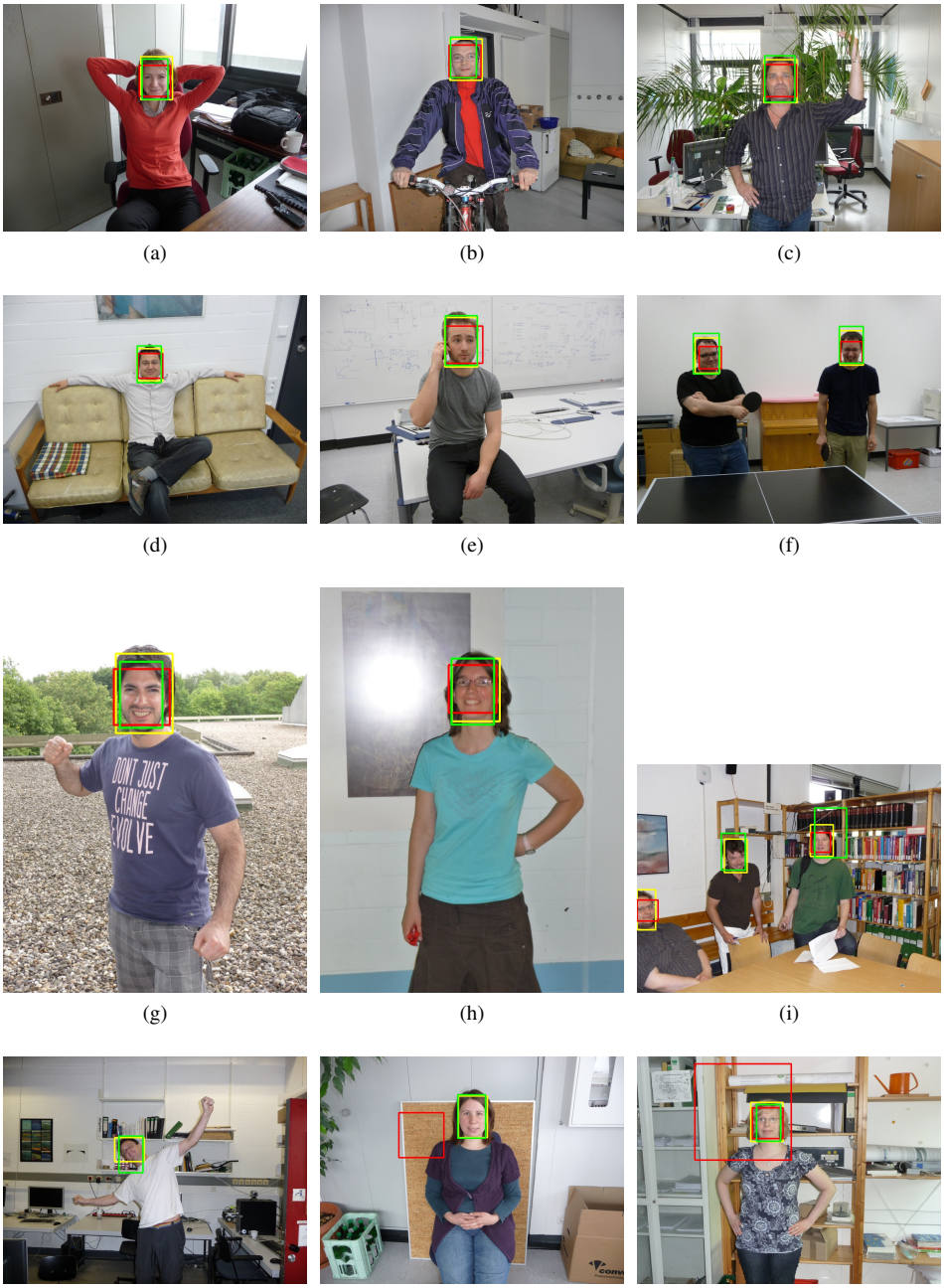
(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure 3: Examples where both methods detected all faces successfully.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 4: Some more successful examples (4a through 4h). In 4i GBFD failed to find the leftmost person and VJFD failed to find the person in the center. In figures 4j through 4l, VJFD misdetected the face, while GBFD found it correctly.

Face detection scheme	μ_{Q}	σ_{Q}
VJFD	0.44347	0.86635
GBFD	0.66675	0.19107

Table 1: Face detection results

first, complexity of most public datasets is either too low (close to plain mug shots) or unacceptably high (e. g., including strong horizontal and vertical face tilt or large scale variation). Second, the GBFD approach is, in its current form, too slow for processing the large numbers of images found in the aforementioned databases. Last not least, hardly tractable copyright conditions hamper use of many of the most interesting face detection benchmarks.

For the following experiments, a subset of $N_{\mathrm{I}} = 54$ images is picked from the INIPURE collection; each picture shows upper body shots of people of varying gender, 'physique and worn attire' [Wal11]. Shots either contain single individuals or groups of people; herein, 'background clutter and scene illumination are assumed unconstrained' [Wal11]. Running VJFD and GBFD on the INIPURE images yields the results shown in fig. 3 and fig. 4: faces detected with the color augmented Viola-Jones method are drawn as red rectangles, detections from the color augmented Gabor-based approach are sketched as green rectangles. The yellow rectangles indicate manually supplied ground-truth face regions.

Qualitatively, GBFD is shown to be at least on par with VJFD when run on the INIPURE test set, locking reliably on single or multiple face instances. To quantitatively underpin this impression, the ground-truth face rectangles in benchmark image $I^i(\mathbf{x})$ are used to set up a binary *ground-truth foreground map* M^i_{F}. Let N^i_{F} be the number of '1'-pixels in M^i_{F}. Further, use the inverse of M^i_{F} to construct a binary *ground-truth background map* M^i_{B}. In addition, let O^i_{F} be the overlap between all face rectangles (retrieved either by VJFD or GBFD) and M^i_{F}. Similarly, O^i_{B} identifies the overlap between all face rectangles and M^i_{B}. With that, the face detection *quality* Q^i for any benchmark image i can be defined as

$$Q^i = \frac{O^i_{\mathrm{F}} - O^i_{\mathrm{B}}}{N^i_{\mathrm{F}}} \tag{2}$$

Running i over all images in the chosen INIPURE subset allows to find the *detection quality mean* μ_{Q} and the corresponding *detection quality standard deviation* σ_Q. High values of μ_{Q} indicate a good average face detection performance, while a low σ_{Q} shows stable behavior of the selected face detection scheme. In ideal case, μ_{Q} should approach 1, while σ_{Q} should tend to 0. Table 1 shows that the proposed, color-augmented GBFD method is not only on par with, but seems to outperform color-augmented VJFD w. r. t. to both μ_Q and σ_Q on the INIPURE database. However, some care has to be taken in the interpretation of these figures: due to the strikingly different training methodologies used in VJFD and GBFD, the former tends to generate detections that tightly encompass the 'pure' face area (the region between forehead and chin), whereas the latter prefers detections that embrace the whole head area. As the manually provided ground-truth rectangles tend to extend beyond the pure face region, inherent positive bias is given to the GBFD method. Further,

spurious false positives in VJFD (which could have been eliminated using a more sophisticated color augmentation scheme or appropriate size restrictions) tear up the statistics via integration of O_{B}^{i} in eq. 2. With that background information, it is reasonable to assume that an appropriately tuned VJFD approach is going to leave behind GBFD on larger, more unconstrained datasets. Nevertheless, tbl. 1 shows that Gabor-based face detection principles can well compete with one of the most powerful face detection paradigms currently available, on benchmark images of moderate complexity.

6 Concluding remarks

By the promising results given above, Organic Computing ideas seem to have at least a threefold impact when applied to the face detection domain: first, Gabor-based face models can be learned in a completely autonomous manner using techniques that already proved useful in articulated body modeling [Wal11]. By that, human supervision is no longer required for face model construction, having GBFD the edge over most contemporary face detection solutions concerning system autonomy. Second, GBFD experimentally proved on eye level (using the INIPURE benchmark) with one of the most popular, state-of-the-art face detection schemes, namely the Viola-Jones algorithm. Accounting for the massive amount of manually provided training data fed into VJFD, this result shows the potential assistance that OC might provide in face detection and computer vision as such. Last not least, GBFD could be modified to enter *nontrivial learning loops*: reliable face detections from moderately complex input images were then used to update the Gabor-based face concept, thereby allowing for reliable FD in query images of high complexity. However, it should be made absolutely clear that the current paper is basically a proof of concept and the enumerated OC advantages do not come for free: the proposed GBFD approach is still way behind Viola-Jones methods (and most other modern face detection schemes) w. r. t. computational speed (finding the faces with the readily trained model takes about 1 minute per image). Thus, the employed INIPURE test set had to be restricted to small size and moderate complexity; other, more complex and extended databases will have to be tested in order to assess the 'real-world' behavior of Gabor-based face detection. Further, GBFD currently learns from a single individual; on the one hand, this renders the detection results above even more impressive w. r. t. generalization capabilities of the proposed system. On the other hand, the resulting face model does integrate information beyond the true face region (s. above). Left to future work, learning from a larger variety of individuals likely amends this issue: the resulting face models are expected to become more compact and to allow for precise detection of the true face area.

Acknowledgments

The authors gratefully acknowledge funding from the DFG in the priority program "Organic Computing" (MA 697/5-1, WU 314/5-2, WU 314/5-3). We thank our colleagues at the Institut für Neuroinformatik for posing for the testing data.

References

[BJ01] Y. Boykov and M.-P. Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In *Proc. ICCV*, volume 1, pages 105–112. IEEE Computer Society, 2001.

[Bra00] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–125, November 2000.

[EMH09] A. Elgammal, C. Muang, and D. Hu. Skin Detection. In S. Z. Li and A. K. Jain, editors, *Encyclopedia of Biometrics*, pages 1218–1224. Springer, 2009.

[FH00] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Matching of Pictorial Structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 66–73, Hilton Head, SC, USA, June 2000. IEEE Computer Society.

[Gor06] Pam Frost Gorder. Neural Networks Show New Promise for Machine Vision. *Computing in Sciene and Engineering*, 8(6):4–8, November 2006.

[Gün11] M. Günther. *Statistical Gabor Graph based techniques for the Detection, Recognition, Classification and Visualization of Human Faces*. PhD thesis, Institut für Neuroinformatik, Ruhr-Universität Bochum, 2011.

[GW09] M. Günther and R. P. Würtz. Face Detection and Recognition Using Maximum Likelihood Classifiers on Gabor Graphs. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(3):433–461, 2009.

[HRBLM07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Kra03] N. Krahnstoever. *Articulated Models from Video*. PhD thesis, Pennsylvania State University, 2003.

[KS07] B. Vinay Kumar and D. R. Sai Sharan. Pattern Recognition with Localized Gabor Wavelet Grids. In *Proc. Intl. Conf. Computational Intelligence and Multimedia Applications*, volume 2, pages 517–521, Washington, DC, USA, 2007. IEEE Computer Society.

[KWT88] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.

[LVB+93] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R. P. Würtz, and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[ME07] Tomasz Malisiewicz and Alexei A. Efros. Improving Spatial Support for Objects via Multiple Segmentations. In *Proc. BMVC*, September 2007.

[MX03] M. Meilă and L. Xu. Multiway cuts and spectral clustering. Technical report 442, University of Washington, 2003.

[PB04] T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature*, 431(7010):768–774, 2004.

[SR03] P. Sharma and R. B. Reilly. A Colour Face Image Database for Benchmarking of Automatic Face Detection Algorithms. In *Video/Image Processing and Multimedia Communications Conference*, Zagreb, Croatia, July 2003.

[TK91] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[vdM08] C. von der Malsburg. The Organic Future of Information Technology. In R. P. Würtz, editor, *Organic Computing (Understanding Complex Systems)*, pages 7–24. Springer, 2008.

[VJ04] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[vL07] Ulrike von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17:395–416, 2007.

[Wal11] T. Walther. *Human Motion Analysis Based on Organic Computing Principles*. PhD thesis, Fakultät für Elektrotechnik und Informationstechnik, Ruhr-Universität Bochum, 2011.

[WFKvdM97] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. PAMI*, 19(7):775–779, 1997.

[YKA02] M.-H. Yang, D. J. Kriegmann, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans. PAMI*, 24(1):44–58, 2002.

[ZMP04] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Proc. NIPS*, volume 17, 2004.

[ZZ10] C. Zhang and Z. Zhang. A Survey of Recent Advances in Face Detection. Technical report MSR-TR-2010-66, Microsoft Research, Microsoft Corporation, 2010.