Using Local Context To Improve Face Detection

Hannes Kruppa, Bernt Schiele
Perceptual Computing and Computer Vision Group
ETH Zurich, Switzerland
{kruppa,schiele}@inf.ethz.ch
http://www.vision.ethz.ch/pccv/

Abstract

Most face detection algorithms locate faces by classifying the content of a detection window iterating over all positions and scales of the input image. Recent developments have accelerated this process up to real-time performance at high levels of accuracy. However, even the best of today's computational systems are far from being able to compete with the detection capabilities of the human visual system. Psychophysical experiments have shown the importance of *local context* in the face detection process. In this paper we investigate the role of local context for face detection algorithms. In experiments on two large data sets we find that using local context can significantly increase the number of correct detections, particularly in low resolution cases, uncommon poses or individual appearances as well as occlusions.

1 Introduction and Related work

Today, face detection is widely considered a mature technology with applications in surveillance, human computer interfaces and digital image enhancement to name a few. A recent survey of the field cites more than 150 references [12]. Researchers have built detectors that can detect both frontal and profile views [8], [2] as well as in-plane rotated faces [7]. The performance of more recent algorithms is typically evaluated on standardized test sets such as the MIT+CMU database [7] where current systems detect over 90% of the faces, have few false alarms and process within less than a second per image [5].

However, it is surprising to see how easily face detectors can be fooled by situations where humans have absolutely no problem to reliably detect faces. Such cases have been systematically studied in psychophysical experiments by Sinha and Torralba [10, 1]. One of the common findings is that the human visual system can robustly discriminate real faces from face-like patterns at very low resolutions. Computational systems on the other hand not only require a much larger amount of facial detail for detecting faces in real scenes, but also yield false alarms that are correctly rejected by human observers.

Torralba's experiments indicate that as the level of detail decreases humans make use of the *local context*, i.e. a local area surrounding the face. This contrasts the assumption behind the predominant object-centered approach that the only image features that are relevant for the detection of an object at one spatial location are the features that potentially belong to the object and not to the background. Following the formulation of [11] this

BMVC 2003 doi:10.5244/C.17.5

Figure 1: Examples of different faces (inner rectangle), their local context as proposed in this paper (outer rectangle) and their global context (outside the outer rectangle). For illustration purposes only one face per image is examined here.

paradigm can be formalized using Bayes Rule as

$$P(O \mid \mathbf{v}) \simeq P(O \mid \mathbf{v}_L) = \frac{P(\mathbf{v}_L \mid O)}{P(\mathbf{v}_L)} P(O) \tag{1}$$

where the image measurements $\mathbf{v}$ are in this case local measurements, that is $\mathbf{v} = \mathbf{v_L}$. The object-centered object likelihood is denoted by $P(\mathbf{v}_L \mid O)$ and $P(O)$ is the object specific prior. However, in order to capture dependencies between an object and its context the measurement vector can be extended to include features outside the target object

$$\mathbf{v} = \{\mathbf{v}_L, \mathbf{v}_C\} \tag{2}$$

where $\mathbf{v}_C$ are measurements of the object's context. Applying Bayes Rule now leads to an expression where all probabilities are conditioned on contextual information

$$P(O \mid \mathbf{v}) = \frac{P(O, \mathbf{v})}{P(\mathbf{v})} = \frac{P(\mathbf{v}_L \mid O, \mathbf{v}_C)}{P(\mathbf{v}_L \mid \mathbf{v}_C)} P(O, \mathbf{v}_C) \tag{3}$$

This idea has been implemented in a system to prime the search for various types of objects based on the *global context* (scene content) [11], and a specific representation for this has been proposed.

Contrastingly, this paper focuses on the role of *local* context as a predictive cue for computational face detection. The promise is to gain robustness that goes beyond the capabilities of the object-centered paradigm. For illustration figure 1 shows some examples of faces within their local and global context.

A modified version of the Schneiderman-Kanade detector is used to implement a *local context detector* which is described in the next section. The detection capabilities of this detector are then compared with a state-of-the-art object-centered approach on two

Figure 2: Examples of training instances used in the object-centered approach (top row) versus the proposed approach based on local context (bottom row). The resolution is 48x56 and 56x48 pixels, respectively.

large image databases in section 3. The proposed approach is also compared to a skin patch detector since a number of systems in the literature employ skin as an auxiliary cue for face detection. Particular attention is devoted to the role of image resolution versus the behavior of the different cues (section 4). The last section draws conclusions and discusses future work.

2 Detecting the Local Context of Faces

To test the feasibility of the local context idea we train an object detector with instances that contain a person's entire head, neck and part of the upper body. Intuitively this choice promises to contain important cues for the presence of faces. Other choices are still under investigation. The resulting training data is quite different from object-centered approaches where only the faces themselves are considered. Figure 2 shows training examples of both paradigms for comparison.

The employed appearance-based detector is a modified version of the Schneiderman-Kanade detector. The features of this detector capture local arrangements of quantized wavelet coefficients. To illustrate the differences of what is being modeled in the local context case and in the object-centered case an example instance is depicted in figure 3 together with a visualization of the corresponding wavelet decompositions.

In the case of local context (right side in the figure), the wavelet decomposition shows most parts of the upper body's contours, as well as the collar of the shirt and the boundary between forehead and hair. The shoulders, for example, appear in the HL subband (upper right subband) while the upper arms and the head contour are clearly visible in the LH subband (lower left subband). Also note that facial parts such as eyes and mouth are hardly discernible in the right wavelet transform and therefore *do not* contribute to the modeling of local context. Hence, this is quite different from data presented to traditional face detectors, where facial parts are typically the most informative cues.

Figure 3: This figure illustrates the differences in the data (wavelet coefficients) that are being modeled in the object-centered case (left box) and in the local context case (right). In the left wavelet decomposition the facial features such as eyes and mouth are clearly visible. These can hardly be discerned in the wavelet decomposition on the right side. However, other features such as the collar of the shirt, shoulders, head and body contours become visible here.

2.1 Implementation Details

At the core of the detector is a Naive Bayes classifier:

$$\prod_{k=1}^{n} \prod_{x,y \in region} \frac{p_k\left(pattern_k(x,y), i(x), j(y) \mid object\right)}{p_k\left(pattern_k(x,y), i(x), j(y) \mid nonobject\right)} > \theta \tag{4}$$

Here, the two likelihoods for the positive and the negative class $p_k\left(pattern_k(x,y), i(x), j(y) \mid object\right)$ and $p_k\left(pattern_k(x,y), i(x), j(y) \mid nonobject\right)$ are decomposed over n different feature types and the locations x,y within the detection window (overlap sampling). This decomposition is based on the assumption of independence, i.e. the Naive Bayes assumption. From the above equation follows that the likelihood functions themselves depend on $i(x)$ and $j(y)$ which are coarse quantizations of of the feature position within the detection window. This spatial dependency allows to capture the global geometric layout: within the detection windows certain features might be likely to occur at one position but unlikely to occur at another.

There are $n = 17$ different types of features involved. For lack of space we describe only their basic concept and refer the reader to [8]. Figure 4 shows one particular feature type $pattern_{\hat{k}}$ for illustration. An input instance is first decomposed by a two-dimensional 3-level wavelet transform using a biorthogonal 5/3 filter bank. Features are then extracted by examining local arrangements of a small number of wavelet coefficients (in this example eight coefficients). An arrangement can combine coefficients from one single subband or span over different subbands. In the example, four coefficients from the HL subband and four from the LH subband (both from the first level of the transform) are combined to form one feature value. This way of combining information from different subbands allows to capture dependencies across space (arrangements with different extends), frequency (different wavelet levels) and orientation (subbands with different filtering directions, i.e. LH and HL).

The individual class-conditional likelihoods $p_k\left(pattern_k(x,y), i(x), j(y) \mid object\right)$ and $p_k\left(pattern_k(x,y), i(x), j(y) \mid nonobject\right)$ are represented by high dimensional histograms.

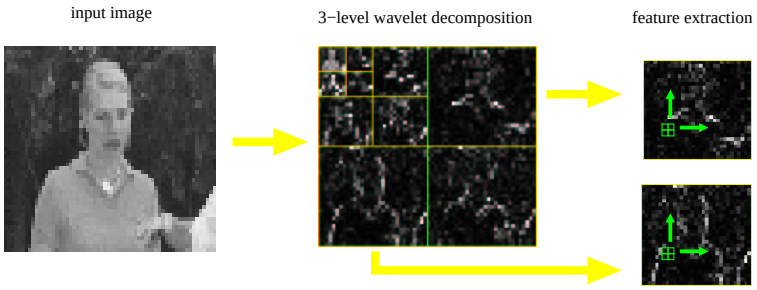

Figure 4: Feature extraction: local arrangements of quantized wavelet coefficients are combined to form one single feature value. In this example four coefficients from the HL subband (upper right subband in the wavelet decomposition) and four from the LH subband (lower left subband) capture the dependency between horizontal and vertical orientations. Such arrangements are examined over all locations within the involved subbands (oversampling). See text for details.

Depending on the feature type these can have up to 1.1 million bins. 2000 local context examples were collected for training the likelihood $p_k\left(pattern_k\left(x,y\right),i(x),j(y) \mid object\right)$. These instances were gathered from the world wide web and from private foto collections. In order to limit the amount of variation and hence increase discriminance only frontal views have been used. Further, taking the shoulders as landmark points the instances have been roughly aligned and vertically mirrored leading to an effective training set size of 4000 instances.

It is very important to collect good examples for the *nonobject* class. Training with random samples will most likely result in an inaccurately modeled decision boundary. Ideally, the likelihood should accurately model nonobjects which are very similar to objects, i.e. they can accurately represent the decision boundary. For collecting nonobject instances we apply an algorithm that is sometimes referred to as "bootstrapping": an initial version of the detector is run over images not containing the target object. Any false alarms are collected as negative training examples. These are then used to retrain $p_k\left(pattern_k\left(x,y\right),i(x),j(y) \mid nonobject\right)$ and the algorithms is iterated. We use a validation set to determine when to stop this iteration.

Larger instances are detected by succesive downscaling of the input image. The current implementation uses a scaling factor of $2^{\frac{1}{4}}$, i.e downscaling four times corresponds to reducing the original resolution by a factor of 2. Since the local wavelet transform is not shift-invariant an overcomplete transform of the entire input image is computed. This allows to apply the classifier at the original image resolution (except the right and bottom image border which is induced by the size of the detection window). For example, given a single 180 by 240 image, the detector searches seven scales, applying the classifier 74144 times altogether.

After all positions and scales have been classified independently, the final detection results are extracted by an arbitration scheme: the most confident detection is determined and nearby confidence values (both in space and scale) are tagged to belong to this same detection. This continues as long as the confidence exceeds the current acceptance threshold θ.

3 Quantative Analysis: Can Local Context Contribute To Face Detection?

In order to understand the relevance of local context several experiments have been carried out on two data sets, the MIT+CMU frontal face data set and a much larger data set collected by the authors. Both data sets are disjoint from the data used for training.

The MIT+CMU data set consists of 125 grayscale images with 448 frontal faces[1]. The second data set consists of 535 color images with 901 faces, 388 of which are profiles. Images in both data sets are upright. Note that the image resolution of the MIT+CMU data set is correlated with the image content: portraits have lower resolution than group pictures which effectively guarantees a certain minimum resolution for the contained faces. There is no such correlation for the second data set where all images are 180 by 240 pixels.

The left plot in Figure 5 shows the face detection performance on the MIT+CMU data set in terms of the ROC curve which has become common practice [2, 5]. The percentage of retrieved faces is shown on the vertical axis and the absolute number of false detections on the horizontal axis with a logarithmic scale. Both the performance of the object-centered and the local context detector are shown. For the object-centered detector the authors' original implementation was used which can be accessed through an on-line web interface [2]. Both frontals and profiles can be detected.

The threshold range in this implementation has a lower limit because the detector is cascaded: only for candidates that make it to the final classifier stage a full confidence score is computed. This means the ROC curve can be plotted only up to a certain number of false positives. Note that the local context detector has to infer the actual location of a face indirectly. We do this by assuming a fixed head position within the detected local context.

As can be seen in the left of figure 5 the object-centered approach detects 90% of the faces with very few false positive detections. The curve flattens from thereon not yielding any additional faces. On the other hand the local context detector produces many false alarms and at 10^3 reaches a detection rate of 40%.

However, the true potential of local context becomes apparent as the resolution of the test images is decreased. At half of the original resolution both detectors individually retrieve about 55% of the faces. Interestingly, their joint detection rate reaches more than 80%. This means that the joint detection rate surpasses the rates of each individual classifier by more than 25%.

This behaviour indicates that both cues are actually *complementary*. A look at the images where a face was found by one cue but not by the other revealed two cases: in portrait images the local context is often not contained. The image is cropped to contain the head and the neck only but not the full silhouette of the upper body. The object-centered approach is obviously well suited for this case while the local context cue is not. On the other hand the decrease in resolution deteriorates some of the facial features which may lead to failure of the object-centered detector. This case is typically well suited for the local context cue.

A similar analysis was performed on a larger image set consisting of 535 color images.

[1] As in [8] five images containing drawings of faces have been removed from the original set of 130 images

[2] http://www.vasc.ri.cmu.edu/facedemo/

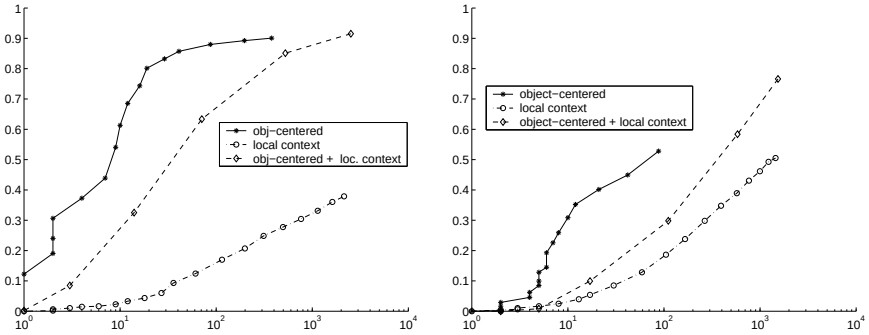

Figure 5: ROC curve showing the percentage of detected faces (vertical) vs. the absolute number of false positives (horizontal, logarithmic scale). At the original image resolution (left plot) the object-centered approach is sufficient to detect 90% of all faces. When the image resolution is decreased by 50% (right plot) the object-centered approach by itself retrieves only 55% of the faces. However, the combination with detections from the local context cue can yield an additional 25% of correct detections.

These are consumer digital photographs from one camera covering a wide range of real-world situations both indoor and outdoor (parties, ski-trips, beach scenes etc.).

A number of face detectors in the literature make use of skin color as an auxiliary cue – provided that color is available [6]. Many other face detectors assume surveillance cameras where color often is not available. Skin color is particularly popular in face tracking because color features can be computed very quickly [9].

To shed light on the relationship between skin color and local context the following experiments also evaluate the performance of a skin detector (the local context detector and the object-centered approach still use only grayscale information). The employed skin detector builds on a non-parametric skin color model trained from a vast collection of web images [3]. This color model captures the RGB statistics of the various skin complexions under different illumination conditions. Combining color information with an elliptical shape model the detector extracts compact skin regions. The full algorithm is described in [4].

The left plot in figure 6 shows the individual ROC curves of all three cues. The object-centered detector has the steepest slope and reaches a detection rate of 30% which proves the difficulty of the dataset. The skin detector reaches 40% and the local context detector 50%. On the right side in the figure the joint performance of all possible pairs as well as all three cues taken together is evaluated. The joint detection rate of object-centered and skin detector reaches 60%. Either combination with the local context detector, however, retrieves even 70% of the faces in the database. Combining all three cues does not further increase the detection rate (the curve lies on top of the local context + skin curve and is therefore hard to distinguish). This indicates that the 10% improvement between the object-centered+skin pair and the other two pairs is due to detections from the local context cue. For this database this means about 90 additionally retrieved faces that otherwise would have been overlooked.

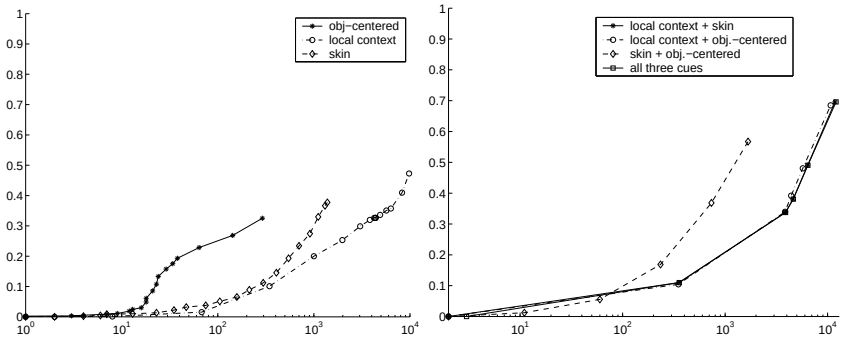

Figure 6: Performance of the individual cues and their combinations. The 10% improvement between the object-centered+skin pair and the other two pairs is due to detections from the local context cue (right plot). For this database this means about 90 additionally retrieved faces that otherwise would have been overlooked.

4 Qualitative Analysis of Novel Face Detections

As suggested by the above quantative results the local context cue yields novel correct detections. This section gives a qualitative analysis of the specific images where novel detections occured. A representative subset of these images is shown in figure 7. Novel detections are visualized as rectangles. The images in the top row show successful detections of very small faces the smallest being 15 by 20 pixels in size. The object-centered approach requires a higher resolution that is not available here. This supports the hypothesis that local context is indeed useful in low resolution cases. A typical example are group pictures where faces naturally occur at lower resolution than say in portraits. On the other hand group pictures provide rich contextual information such as the heads or shoulders of other people standing nearby. The local context for an individual face is thus affected.

While in these previous examples all the facial details are still visible they are partially concealed in the next few examples shown in row 2. This often happens because of a special facial pose, occlusion or low light. As for facial pose, an appearance-based object-centered approach could theoretically be trained accordingly, but in practice acquiring enough of this type of training data is very difficult. Occlusion occurs, for example, because people often lean onto their hands, or wear sunglasses or beards. The object-centered approach can sometimes overcome this problem at higher resolutions but fails as resolution decreases.

A particularly interesting low light situation is shown in the first image of row 5. Due to lack of contrast this face could not be detected by the object-centered approach even though the resolution would be well sufficient. Likewise skin color could not be detected given such difficult illumination. The local context could still be detected. The inferred face position is not very accurate because a fixed position within the detected local context is assumed. Alternatively one could issue a local search for head contours, e.g. within a the upper half of the detected local context.

5 Conclusion and Future Work

This paper has explored the idea of using local context for face detection. As a proof of concept it has been shown by quantitative and qualitative analyses that the detection of the local context of faces in grayscale images is feasible. This is in contrast to the traditional object-centered approach to face detection where the role of local context has so far been neglected. However, experimental results indicate that using local context yields correct detections that are beyond the scope of the classical object-centered approach and even beyond those of less specific cues such as skin detectors. This holds not only for low resolution cases but also for difficult poses, occlusion and difficult lighting conditions.

We are currently investigating mechanisms to integrate the local context detector as an auxiliary cue for the object-centered approach. Also we have started to systematically explore the possible shapes and ranges of local context that are optimal for face detection.

References

[1] Pawan Sinha Antonio Torralba. Detecting faces in impoverished images. In *AI Memo 2001-028, CBCL Memo 208*, 2001.

[2] S.Z. Li et al. Statistical learning of multi-view face detection. In *ECCV*, 2002.

[3] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. In *CVPR*, pages 274–280, 1999.

[4] Hannes Kruppa, Martin Bauer, and Bernt Schiele. Finding skin in real-world images. In *DAGM*, 2002.

[5] Michael Jones Paul Viola. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[6] Anil K. Jain Rein-Lien Hsu, Mohamed Abdel-Mottaleb. Face detection in color images. *PAMI*, pages 696–706, January 2002.

[7] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *PAMI*, 20(1):23–38, 1998.

[8] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, June 2000.

[9] Leonid Sigal and Stan Sclaroff. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *CVPR*, 2000.

[10] Pawan Sinha. Qualitative representations for recognition. In H.H. Buelthoff et al., editor, *Biologically Motivated Computer Vision (BMCV)*, pages 249–262, 2002.

[11] Antonio Torralba. Contextual modulation of target saliency. In *Advances in Neural Information Processing Systems*, 2001.

[12] Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *PAMI*, 24(1):34–58, January 2002.

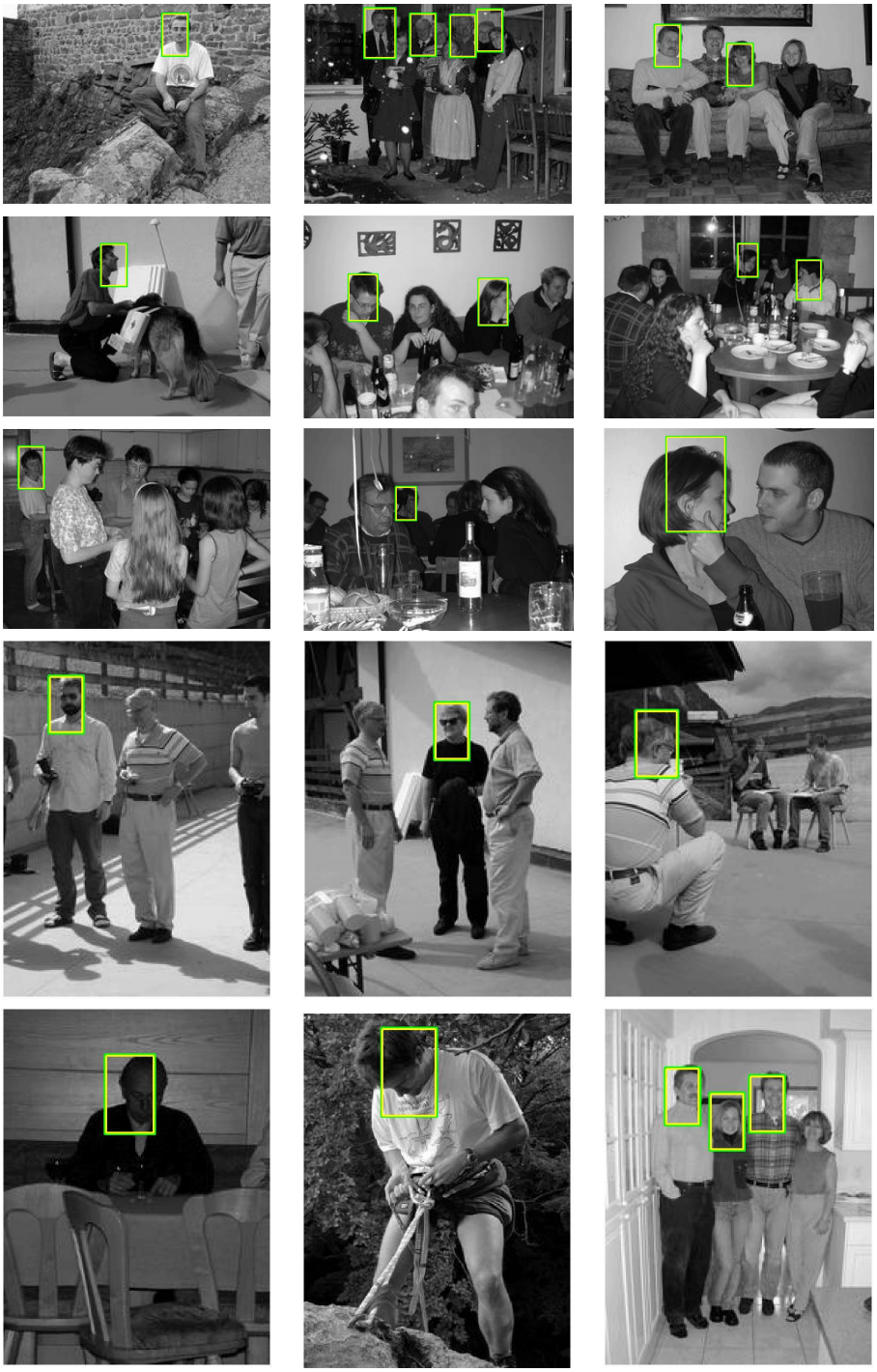

Figure 7: Novel face detections indirectly inferred from the local context cue