Fast Human Detection in Crowded Scenes by Contour Integration and Local Shape Estimation*

Csaba Beleznai
Austrian Research Centers GmbH - ARC
Vienna, Austria
csaba.beleznai@arcs.ac.at

Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology
bischofg@icg.tugraz.at

Abstract

The complexity of human detection increases significantly with a growing density of humans populating a scene. This paper presents a Bayesian detection framework using shape and motion cues to obtain a maximum a posteriori (MAP) solution for human configurations consisting of many, possibly occluded pedestrians viewed by a stationary camera. The paper contains two novel contributions for the human detection task: 1. computationally efficient detection based on shape templates using contour integration by means of integral images which are built by oriented string scans; (2) a non-parametric approach using an approximated version of the Shape Context descriptor which generates informative object parts and infers the presence of humans despite occlusions. The outputs of the two detectors are used to generate a spatial configuration of hypothesized human body locations. The configuration is iteratively optimized while taking into account the depth ordering and occlusion status of the hypotheses. The method achieves fast computation times even in complex scenarios with a high density of people. Its validity is demonstrated on a substantial amount of image data using the CAVIAR and our own datasets. Evaluation results and comparison with state of the art are presented.

1. Introduction

Reliable human detection is a key algorithmic component of many application-oriented computer vision systems, for instance in automated visual surveillance, automotive safety, human-computer interaction and multimedia processing. High detection rates and low false alarm rates are essential for achieving robustness in higher level vision tasks such as tracking or activity recognition. While many human detection methods perform quite well for spatially separated, unoccluded humans in more-or-less controlled environments, nevertheless, they undergo an ungraceful degradation of detection performance when facing a high density of humans (see Figure 1), clutter and varying illumination conditions. These problems have been recognized by the scientific community and substantial amount of research has been recently carried out to extend the operational domain of human detection frameworks beyond simple scenarios. Devising an adequate representation for humans seen in images still remains a challenging task since such a representation must meet requirements of specificity, generality and computational efficiency at the same time.

Figure 1. A sample input frame (left) and corresponding detection output generated by the proposed method (right).

Previous methods of human detection can be grouped according to the following criteria: *shape-based* approaches and *motion-based* methods. Approaches which employ a model-based representation can be further categorized as *monolithic* (full-body) and *part-based* detectors.

Shape-based monolithic detectors have been designed using hierarchically structured edge templates [7], learned edge-based models [6] or classifiers in combination with shape-encoding features [5, 8]. While these methods work well in cluttered scenes, their detection rates drop significantly in presence of occluded humans.

Motion-based approaches based on change detection and statistical modelling of the background [20, 16] have been popular due to their simplicity and computational efficiency, enabling systems applicable to simple scenarios with few

*This work was supported by the FFG COMET project ECV and the FFG project AUTOVISTA (813395) under the FIT-IT program.

persons. Nevertheless, foreground segmentation errors and detection errors become evident with increasing density of humans and clutter.

A recent shift of focus towards *part-based* representations has resulted in detection methods capable to detect parts of humans and perform occlusion reasoning based on the part-detection results. Zhao *et al*. [23] employ a multiple-part human model in conjunction with a global optimization step within the multi-object configuration space, but the outcome of the method strongly depends on the quality of motion-based foreground segmentation. Leibe *et al*. [9] present a generative approach where multiple-part assemblies are hypothesized from local features and a validation step based on global features selects optimum hypotheses along with a greedy optimization of occlusion states. Given the data-driven voting mechanism the detector generalizes well, but it generates many false alarms in presence of clutter, therefore validations steps are required. Rodriguez *et al*. [13] use Shape Context descriptors [2] to build a codebook of local shape distributions which vote for human locations in an image. Zhao *et al*. [22] use hierarchically organized contour templates coupled with color-based segmentation to delineate human hypotheses. Lin *et al*. [10] also propose a hierarchical contour template matching scheme combined with motion detection and human inter-occlusion analysis. Template-based search, despite of its hierarchical structure, represents a computationally intensive operation when performing a dense scan across the image. Wu *et al*. [21] introduce discriminatively learned edgelet-based part detectors which are used to infer presence of humans by analyzing individual detector responses and occlusion states. Similarly, Shet *et al*. [14] combine discriminatively learned part detectors with a logical occlusion reasoning approach. In these cases the learning process of part detectors for multiple poses and views typically becomes complex and the required high level of generalization necessitates a hierarchical multi-view detector.

In summary, despite the significant advances represented by the above approaches human detection in crowded situations with mutual occlusions still poses a challenge. Therefore we propose two novel generic concepts which complement existing detection approaches: (i) An integral image based concept for contour integration which enables computationally efficient matching when using sparse contour templates. The proposed matching scheme permits the computational evaluation of a vast number of shape hypotheses with varying translational, rotational and scaling parameters. Additionally, occlusion analysis can be performed in a simple and fast way given the individual contour segment probabilities. (ii) An approximated Shape Context descriptor, called aSC, which is applied to data obtained by background subtraction, and capable of hypothesizing object locations in presence of clutter and occlusions. The two computationally efficient approaches combine local and global shape cues in a similar manner as in [9] and [10].

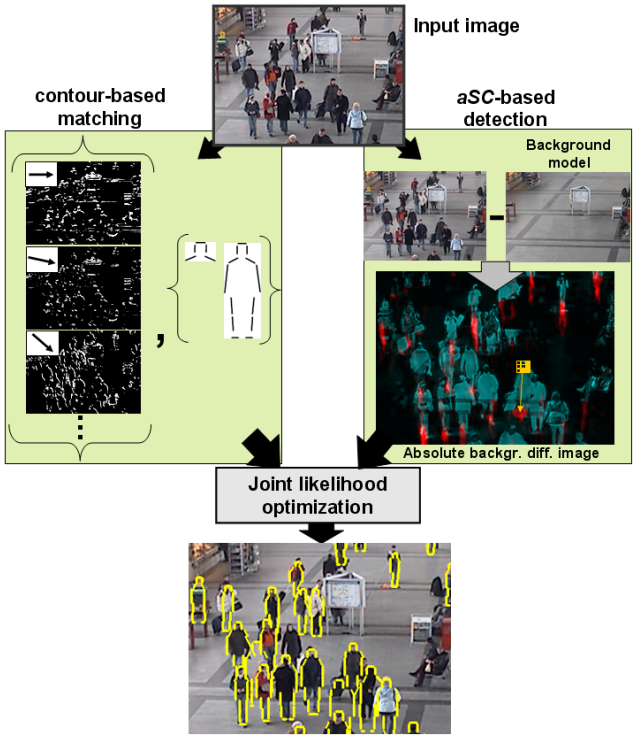

Figure 2. Outline of the proposed human detection method. Left: A set of anisotropically filtered images and a set of shape templates generate shape-based detection hypotheses. Right: A codebook of approximated Shape Context descriptors (aSC) is used to generate local shape-based detection hypotheses in a difference image (input image - background). The two sets of hypotheses are evaluated jointly in a final optimization step.

The paper is organized as follows: Section 2 describes the outline of the proposed human detection method. Section 3 presents the integral image based concept of fast contour integration. Section 4 demonstrates the concept of aSC descriptor in the context of motion-based detection, while section 5 describes the combination between the two detectors. Section 6 presents and discusses experimental results and their evaluation. Finally the paper is concluded in Section 7.

2. Outline of the detection method

Our approach relies on maximum a posteriori estimation (MAP) of the spatial configuration of humans ($\mathbf{c}^*$) best explaining the observed image features I:

$$c^* = \arg\max_{\mathbf{c}} \; P(\mathbf{c}|I), \tag{1}$$

We combine detection hypotheses generated by contour-based matching $\{\mathbf{h}^C\}$ (Section 3.2) and local shape estimation $\{\mathbf{h}^{LS}\}$ (Section 4): $\mathbf{c} = \{\{\mathbf{h}^C\}, \{\mathbf{h}^{LS}\}\}$. The process is illustrated in Figure 2. According to Bayes theo-

$\alpha_j = \arctan(b_j/a_j)$	$[a_j, b_j]$
0°	$[1, 0]$
26.57°	$[2, 1]$
45°	$[1, 1]$
63.43°	$[1, 2]$
90°	$[0, 1]$
116.57°	$[-1, 2]$
135°	$[-1, 1]$
153.43°	$[-2, 1]$

Table 1. The 8 orientations and corresponding offset components used in our experiments.

rem the posterior probability is proportional to:

$$P(\mathbf{c}|I) \propto P(I|\mathbf{c})P(\mathbf{c}), \tag{2}$$

where $P(\mathbf{c})$ denotes the prior probability of human models with respect to their parameters, which are described in Section 3.2. $P(I|\mathbf{c})$ is the joint likelihood of a configuration computed assuming independence between the information contained in individual cues:

$$P(I|\mathbf{c}) = P(I_c|\mathbf{c})\, P(I_m|\mathbf{c}), \tag{3}$$

where I_c denotes contour-based observation (shape matching using edge probabilities) and I_m is motion-based observation (shape estimation using background difference). The computations of $P(I_c|\mathbf{c})$ and $P(I_m|\mathbf{c})$ are described in Section 3.2 and Section 4, respectively.

3. Shape-based detection

Template-based matching is a versatile tool for various pattern matching problems, nevertheless, the measurement process - given the often existing uncertainties with respect to parameters defining translational, rotational, scaling, shape and other variations - imposes substantial computational requirements. Typical examples are chamfer matching [7] and edge-based detection approaches [21, 10], where usually hierarchical search strategies are used to minimize computational costs of the process locating the solution. Our proposed approach for computing line integrals along contour segments speeds up the measurement process and it still can be embedded into a hierarchical matching framework.

3.1. Contour integration by integral images

The integral image concept [4, 15] has been widely used to speed up the computation of region-based statistical measures, such as area sums [18], covariance [17] and co-occurrence [19]. We describe the construction of multiple integral images by oriented strings scans over the entire image (Figure 3) in order to efficiently compute integrals along oriented linear contour segments. Efficient integration permits fast evaluation of contour-based features.

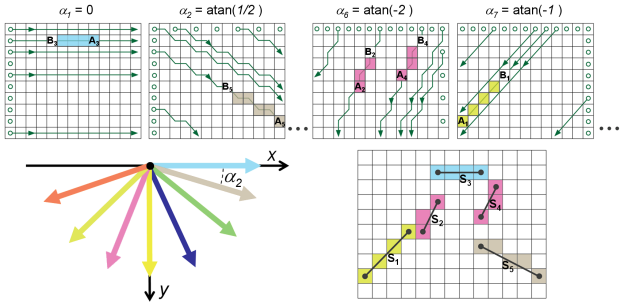

Figure 3. Illustration showing the construction of integral images by oriented string scans for different orientations. Dots represent starting locations of individual string scans. The bottom right image depicts an example contour template consisting of five line segments, where line integrals (sum of values at pixels color-coded according to orientation) can be efficiently computed based on the integral images.

Given the spatial discretization of digital images, we use discrete unit-integer orientations (see Table 1) as termed by Messom *et al.* [11]. A unit-integer orientation is an orientation $\alpha = \arctan(b/a)$ defined by horizontal and vertical offset components a and b, such that both components are integers and at least one of them is 1 or -1. An α_j oriented string scan refers to a spatial sequence of pixels starting at an image border, oriented according to the components $[a_j, b_j]$ and ending at another image border. Let I denote an image with a height of M and width of N pixels. Let $\{\alpha_j\}_{j=1..k}$ be a set of unit-integer orientations (Table 1). For each orientation α_j we partition I into a set of n scanlines $\{S_i\}_{i=1..n}$, such that their union equals I:

$$I = S_1 \cup S_2 \cup \ldots \cup S_n \ . \tag{4}$$

Each scanline S_i is uniquely defined by a set of parameters. For the set of orientations $O_1 = \{\{0° \leq \alpha_j \leq 45°\}, \{135° \leq \alpha_j \leq 180°\}\}$ the x-coordinates of a scanline are uniquely defined being $x \in \{1, \ldots, N\}$. For the complementary set of orientations $O_2 = \{45° < \alpha_j < 135°\}$ each pixel of a scanline has a uniquely assigned y-coordinate, $y \in \{1, \ldots, M\}$. Accordingly, each scanline is defined by slope-intercept forms (see Figure 3):

$$y^i = F_1(x^i, x_0^i, y_0^i, \alpha_j) \qquad \text{if } \alpha_j \in O_1 \tag{5}$$

$$x^i = F_2(y^i, x_0^i, y_0^i, \alpha_j) \qquad \text{if } \alpha_j \in O_2, \tag{6}$$

where the functions F_1 and F_2 are defined as:

$$F_1\left(x, x_0, y_0, a, b\right) = y_0 + \operatorname{sgn}\left(a\right)\left(x - x_0\right)\left(\frac{b}{a}\right), \tag{7}$$

$$F_2\left(x, x_0, y_0, a, b\right) = x_0 + \left(y - y_0\right)\left(\frac{a}{b}\right), \tag{8}$$

sgn is the sign function and $\{x_0^i, y_0^i\}$ denote the starting location of the i^{th} scanline. The set of all starting locations

is defined as:

$$\{x_0, y_0\} = \begin{cases} \{1, 1 \dots M\} & \text{if } \alpha_j = 0° \\ Q_1^j & \text{if } 0° < \alpha_j < 90° \\ \{1 \dots N, 1\} & \text{if } \alpha_j = 90° \\ Q_2^j & \text{if } 90° < \alpha_j < 180°, \end{cases} \tag{9}$$

where the pixel sets Q_1^j and Q_2^j are defined as:

$$Q_1^j = \{\{1, \text{mod}(y-1, b_j) = 0\}, \{\text{mod}(x-1, a_j) = 0, 1\}\} \tag{10}$$

$$Q_2^j = \{\{N, \text{mod}(y-1, b_j) = 0\}, \{\text{mod}(x-1, a_j) = 0, 1\}\}, \tag{11}$$

and mod denotes the integer modulo operator.
Integral images contain the cumulative sums computed along each oriented scanline for a given orientation:

$$ii(x, y, \alpha_j) = \begin{cases} \sum\limits_{x' \le x} I\left(x',\ F_1\left(x', x_0, y_0, \alpha_j\right)\right) & \text{if } \alpha_j \in O_1 \\ \sum\limits_{y' \le y} I\left(F_2\left(y', x_0, y_0, \alpha_j\right), y'\right) & \text{if } \alpha_j \in O_2 \end{cases} \tag{12}$$

Integral images ii can be built - similarly to integral images for area-based statistics - in a recursive manner:

$$ii(x, y, \alpha_j) = \begin{cases} ii_1(x, y, \alpha_j) & \text{if } \alpha_j \in O_1 \\ ii_2(x, y, \alpha_j) & \text{if } \alpha_j \in O_2 \end{cases}$$

where ii_1 and ii_2 are computed as:

$$\begin{aligned} ii_1(x, y, \alpha) =& ii_1\left(x-1, F_1\left(x-1, x_0, y_0, \alpha\right)\right) + \\ & I\left(x, F_1\left(x, x_0, y_0, \alpha\right)\right) \end{aligned} \tag{13}$$

$$\begin{aligned} ii_2(x, y, \alpha) =& ii_2\left(F_2\left(y-1, x_0, y_0, \alpha\right), y-1\right) + \\ & I\left(F_2\left(y, x_0, y_0, \alpha\right), y\right), \end{aligned} \tag{14}$$

and $ii_1(0, y, \alpha) = 0, ii_1(x, 0, \alpha) = 0$ and $ii_2(0, y, \alpha) = 0, ii_2(x, 0, \alpha) = 0$.

Figure 3 shows the rasterization of scanlines for some orientations. The bottom image shows an example for a contour-based template consisting of five line segments. Using the precomputed integral images $ii(x, y, \alpha)$, the sum of pixel values along the line segments can be computed as:

$$\begin{aligned} s =& \left(ii\left(A_1, \alpha_7\right) - ii\left(B_1, \alpha_7\right)\right) + \\ & \left(ii\left(A_2, \alpha_6\right) - ii\left(B_2, \alpha_6\right)\right) + \left(ii\left(A_3, \alpha_1\right) - ii\left(B_3, \alpha_1\right)\right) + \\ & \left(ii\left(A_4, \alpha_6\right) - ii\left(B_4, \alpha_6\right)\right) + \left(ii\left(A_5, \alpha_2\right) - ii\left(B_5, \alpha_2\right)\right) \end{aligned} \tag{15}$$

Thus, integration of values along one line segment requires a single arithmetic operation independent of location and scale.

The proposed contour integration scheme has an additional appealing property. A line segment with an orientation α_j can be rotated arbitrarily by an angle β to generate an approximated, unit-integer-oriented rasterized line segment:

$$\alpha_j \pm \beta \approx \alpha_l \in \{\alpha\}. \tag{16}$$

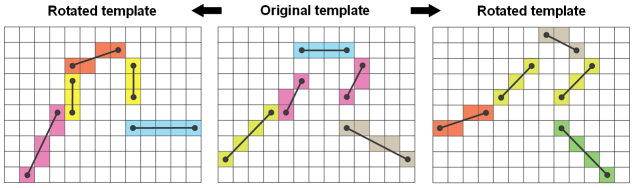

Figure 4. Illustration depicting a contour template (center image) and its rotated variants (left and right). Rotations of the contour template create a new set of edge segments, whose orientations are elements of the original set of discrete unit-integer orientations (see text for more details).

This property is visualized in Figure 4, where rotations of a contour template result in rasterized line segments, whose orientations are again unit-integer orientations for which integral images are available, as indicated by the color coding. This means that by using a set of integral images obtained by oriented string scans, contour templates and its rotated variants (although approximations of the original shape) can be used for template matching. The distinct advantage of the proposed contour based integration is that translated, rotated and scaled shape templates can be matched efficiently.

3.2. Human detection by sparse contour templates

We employ contour models incorporating typical shape variations by adopting a parametric shape model based on the *Point Distribution Model* [3]. For each image location multiple scaled contour templates are generated off-line with a scaling driven by an estimated height model prior.

Model of projected human height: We assume that pedestrians stand upright on a common ground plane. Similar to works [12] using stationary cameras, we perform an off-line calibration step estimating a model $H(y)$ of the projected 2D human height in the scene. The prior probability of human height at a given image location is computed by $P\left(H_i\right) = H_i\left(y\right) P\left(h_i\right)$, where $P\left(h_i\right)$ is a Gaussian distribution $N\left(\mu_h, \sigma_h^2\right)$ (μ_h=1.0, σ_h= 0.08).

Generating sparse contour templates: 120 pedestrian images of the INRIA dataset [5] were manually annotated by adjusting a prototype contour set consisting of 13 oriented line segments to the human shapes seen in the training images. Annotated shapes - obtained for frontal and side views - were registered into a common space using foot and head locations on a common vertical human axis. The dimensionality of the vector space - spanned by the segment end point coordinates - is reduced using PCA and 11 eigenvectors are retained explaining 95% of the total variance in the training set. By considering only the principal modes of variation, we generate k_T (k_T=30) shape samples $\{T_i\}_{i=1..k_T}$. The shape set is scaled for each y-position

of the image given $P(H_i)$ and line segment coordinates are approximated such that orientation of the line segments matches the nearest unit-integer orientation of Table 1.

Template matching: Using a filter bank of steerable Gaussian first derivative filters (σ=0.5), the input image is filtered along the unit-integer orientations of Table 1 and filter responses are thresholded to obtain edge probability maps, I_e. Contour-based likelihood at a given image location $\mathbf{x}$ is computed by matching head-shoulder (HS) and full-body (FB) templates in a dense scan:

$$P(I_c|\mathbf{x}) = w_1 P_{HS}(I_e|\mathbf{x}, T^*_{HS}(x)) + w_2 P_{FB}(I_e|\mathbf{x}, T^*_{FB}(\mathbf{x})), \quad (17)$$

where $T^*_{HS}(\mathbf{x})$ and $T^*_{FB}(\mathbf{x})$ denote the locally best matching head-shoulder and full-body templates, w_1 and w_2 are importance weights.

Computational complexity: The computational complexity in terms of number of arithmetic operations can be measured against the case of using conventional shape templates consisting of a chain of contour pixels. In our case computation of integral images requires approximately $8(M-1)(N-1)$ operations and integration for a single template needs $2n_t - 1$ operations, where n_t is the number of contour segments in the employed model. Straightforward integration along the contour of a given template requires $n_p - 1$ operations, where n_p is the number of contour pixels. Computational savings were quantified for our specific experimental settings (template scaling, sampling density) of Section 6. The proposed use of sparse contour templates requires 1500-2000 times less arithmetic operations than the use of conventional contour templates.

4.Detection using approximated Shape Context

We propose a method which uses a simple but informative local shape descriptor to infer human locations in images of absolute background difference obtained by motion detection [16]. Generating reliable shape cues by a data-driven process is inherently difficult given the high amount of ambiguity associated with extracted low-level motion features. Segmentation is one possible generative step leading to shapes, but obtaining a global high-quality segmentation is difficult.

Our approach follows a similar strategy as [13], however, we do not rely on segmented foreground and our descriptors can be easier turned into prototypical representations of local shape. Our approach is performed in two steps:

Training: In the training step we derive a pool of local shapes - in form of discretized binary foreground segments, which we call the approximated Shape Context (aSC) - along the boundary of humans. A small set (in our case 10 images) of manually segmented binary images of humans are used. The aSC descriptors are built as follows: A set of human contour locations are sampled and at each sampled

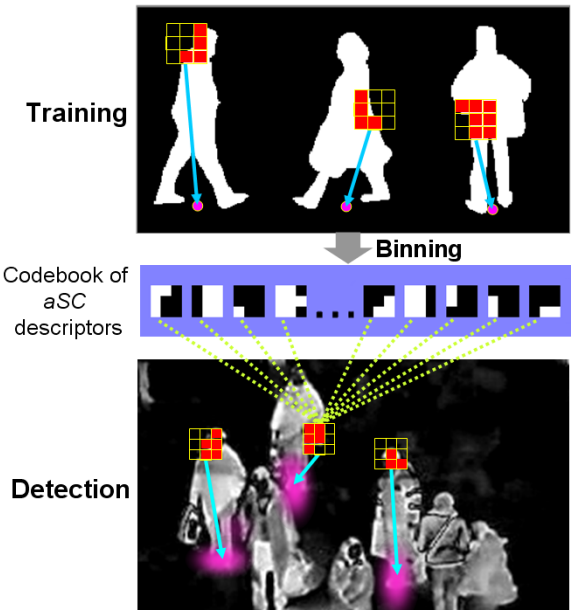

Figure 5. The training step (top) and the codebook-based detection step (bottom) hypothesizing human locations in an image of absolute differences.

location an $n_g \times n_g$ local grid centered on the location is defined (see Figure 5, top). The size of the grid D is defined to be proportional to the human height H: $D = zH$. Let $\{C_i\}_{i=1..(n_g)^2}$ denote the set of cells constituting the local grid. The number of cells are independent of grid scaling (in our case n_g=3). For each cell we compute the attributes $C_i = (s_i, l_i)$, where s_i denotes the number of foreground pixels relative to the total number of pixels in the i^{th} cell. l_i denotes a binary label indicating the status of the cell C_i ($0 = background, 1 = foreground$). l_i is computed simply as:

$$l_i = \begin{cases} 1 & \text{if } s_i > T \\ 0 & otherwise, \end{cases} \quad (18)$$

using a threshold T. The aSC is formed by the obtained vector of binary values $\boldsymbol{l}$ and encodes the local shape at a coarse level. After having sampled many locations, the obtained pool of aSC signatures can be simply clustered using binning and the spatial locations (foot position) of humans $p(x|\boldsymbol{l})$ - relative to the sampled location - are stored as a unit-normalized distribution in form of a codebook $\{\boldsymbol{l}, p(x|\boldsymbol{l})\}$ (see Figure 5, middle).

Detection: The learned codebook is used to find the best fitting codebook entry at sampled locations in an image I, where I is in our case an image of absolute differences obtained by background subtraction and normalized to the range $[0, 1]$.

First, we sample multiple locations with large-valued image gradients, locations which are assumed to be situated along potential object boundaries. The matching cost of a

given codebook entry evaluated locally is defined as

$$Ct\,(\boldsymbol{l}\,|\,I\,) = \frac{1}{A_F} \sum_{\{x,y \in C|\boldsymbol{l}=1\,\}} I(x,y) \;- \frac{1}{A_B} \sum_{\{x,y \in C|\boldsymbol{l}=0\,\}} I(x,y), \tag{19}$$

where A_F denotes the foreground area, A_B the background area within the local grid, defined by the cells belonging to foreground and background, respectively.

The best matching codebook entry maximizes the density within the hypothesized foreground region, while minimizing the density in the hypothesized background region:

$$\boldsymbol{l}^* = \arg\max_{\boldsymbol{l}} \; Ct\,(\boldsymbol{l}\,|I\,)\,. \tag{20}$$

The codebook entry meeting the above condition best explains the local structure of underlying distribution in the difference image.

Equation 19 can be evaluated and $\boldsymbol{l}^*$ can be determined very efficiently, since sums can be precomputed by integral image-based area sum computation for each grid cell only once and all foreground-background combinations (in our case 34) defined by the codebook entries can be efficiently formed using the precomputed cell sums. Thus all codebook entries at all sampled locations are evaluated and the best matching codebook entries vote - in a similar manner as in [9], [13] - for the hypothesized human locations yielding the motion-based likelihood of human presence:

$$P\,(I_m\,|x\,) = \sum_i p\,(x\,|\boldsymbol{l}_i\,)\,p\,(\boldsymbol{l}_i\,|I\,), \tag{21}$$

where $p\,(x\,|\boldsymbol{l}_i\,)$ denotes the learned spatial distribution of the best matching codebook entry and $p\,(\boldsymbol{l}_i\,|I\,) = Ct\,(\boldsymbol{l}_i\,|\,I\,)$ is its likelihood.

The presented concept of aSC-based shape estimation has several advantages: despite of ambiguous structures in the underlying distribution - the image of absolute differences is cluttered and textured due to textured foreground and background -, human locations hypothesized by the ensemble of local shape descriptors are associated with significantly less ambiguity. In addition, due to the non-parametric nature of the descriptor estimation step, the obtained aSC descriptors are to a great extent invariant with respect to linear intensity scaling of the underlying distribution, therefore local estimation is feasible at locations where otherwise thresholding would remove all information.

5. Detector combination, optimization

The two detector outputs are combined in a similar manner as in [10]. We select local maxima and perform non-maxima suppression on the computed likelihood maps $P\,(I_c\,|x\,)$ and $P\,(I_m\,|x\,)$, generating two sets of hypotheses $\{\mathbf{h}^C\}$ and $\{\mathbf{h}^{LS}\}$. The hypothesis sets are combined by merging spatially coinciding hypotheses. The spatial configuration of hypotheses is optimized in a greedy manner using the posterior probability and considering the occlusion status of individual hypotheses. Match scores for visible (unoccluded) contour segments are retrieved by table lookups and the final optimum configuration estimate is typically reached efficiently in a few iterations.

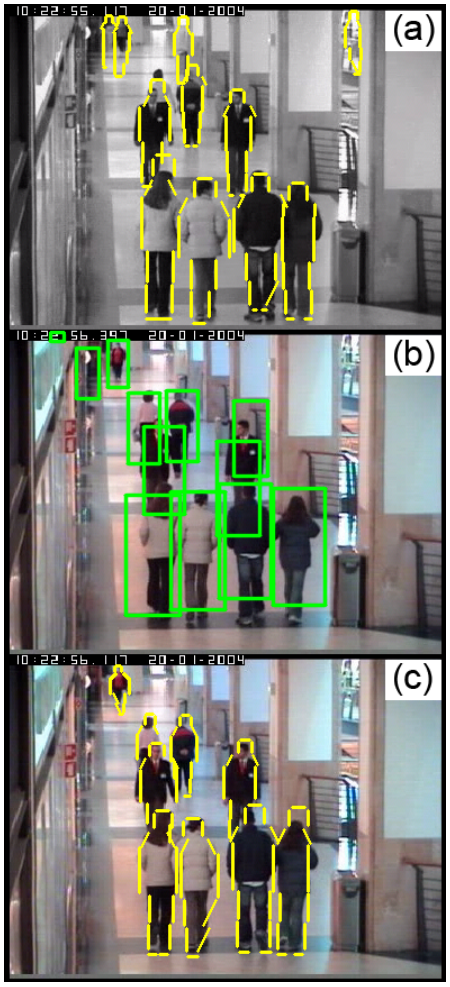

Figure 6. Sample detection results obtained for the CAVIAR sequence using (a) contour templates, (b) codebook of aSC descriptors and (c) combination of the two detectors.

6. Experiments and discussion

Detection experiments were performed on a sequence of the CAVIAR dataset [1] (*OneStopMoveEnter1cor*) and on two of our datasets (RailwayStation-A (RS-A) and RailwayStation-B (RS-B)). During evaluation following evaluation criteria have been taken into account: (i) a one-to-one match was enforced between all ground truth and detection instances with more than 50% overlap using a bounding box approximation of their spatial extent; (ii) per-

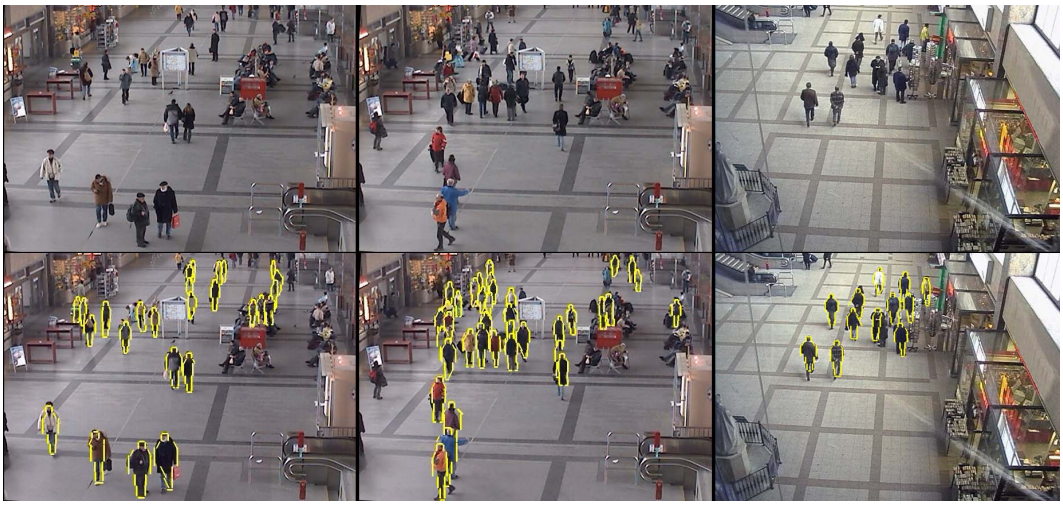

Figure 7. Sample detection results obtained for the Railway Station datasets (dataset A - left and center coloumns, dataset B - right coloumn) using the proposed combined algorithm. Top row shows original image content, the row below shows detection results by superimposed contour models.

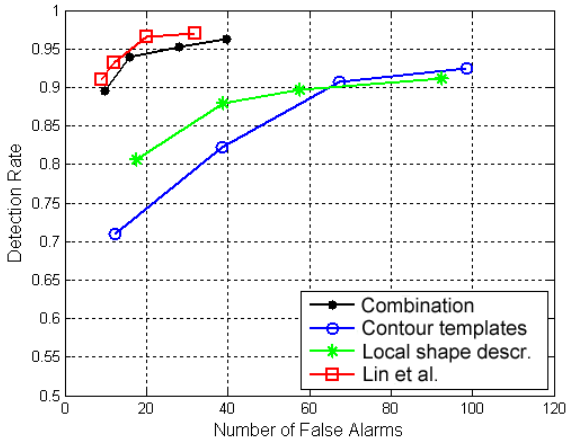

Figure 8. ROC curves obtained by evaluation on a subset of the CAVIAR dataset (200 images with 1800 humans) and compared to the results of Lin *et al*. [10].

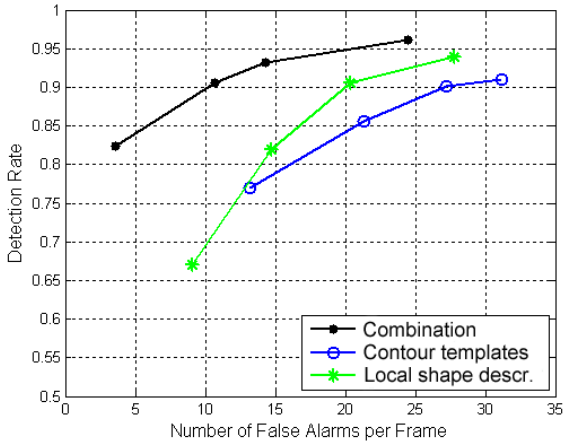

Figure 9. Detection performance of individual detectors and their combination evaluated on the Railway Station A dataset.

sons with less than 50% visibility at image boundaries and sitting persons of the RS-A dataset were not used in the evaluation process.

Detection results on the CAVIAR dataset: We evaluated detection results obtained for 200 frames (frames 800-1000, grayscale images of 384×288 pixels) of the dataset containing 1800 humans. Sample detection results for the proposed two algorithmic components (contour-based detection including occlusion reasoning when applied solely) and for their combination are shown in Figure 6, while a quantitative comparison of detection performance is given in Figure 8. The ROC-plot shows also results of Lin *et al*. [10] obtained on this dataset. As can be seen, contour-based detection and detection using the codebook of aSC descriptors - when applied independently - occasionally miss certain humans and produce false alarms in presence of clutter. Small-sized contour templates produce many false alarms - due to loss of specificity at small scales - given the clutter in the scene background. The combined detectors, however, achieve a detection performance comparable to the results obtained by Lin *et al*., but at a lower computational cost.

The combined detection approach achieves real-time performance. Our algorithm evaluates 30 shape templates in an exhaustive manner at every second pixel (along each image dimension). Contour-based detection is performed on graphics hardware, yielding approx. 22 fps on a low-end card (Nvidia 9800GT) and approx. 60 fps on a mid-range card (Nvidia GTX260) for a video resolution of 720×576 pixels. The codebook-based detection approach is implemented in a non-optimized form for the CPU (3.6 GHz). First tests show that framerates of at least 20 fps are possible for the complete detection system when processing

videos with the above resolution.

Detection results on the RS-A and RS-B datasets: Using an extensive annotation of the RS-A dataset (image resolution 640×480 pixels, 2982 frames containing 87373 annotated humans) we performed experiments using the individual algorithmic components and their combination, in the same manner as in the previous experiment. The individual ROC curves are shown in Figure 9. As can be seen, contour-based detection and codebook-based detection both produce good results for this difficult dataset containing frequent occlusions and substantial clutter. The combination of the two algorithms produces considerable improvement achieving detection rates around 90% with a small number of false alarms. A total of 6545 frames of the RS-A dataset were processed and some results are shown in Figure 7. In addition, 3500 frames of the RS-B dataset (resolution 720×576 pixels, interlaced) were processed and a sample result is shown in Figure 7.

7. Conclusions

We have presented two approaches and their combination for fast human detection in crowded scenarios. First, we have introduced an integral image based contour integration concept which (i) significantly speeds up sparse contour-based template matching and subsequent occlusion analysis and (ii) generates a segmentation of detected humans. A second simple detection concept based on an approximated form of a shape context is presented. The shape descriptor is estimated non-parametrically and generates reliable human hypotheses in presence of occlusions. The joint use of the two algorithms combines local and global shape cues and demonstrates highly accurate detection performance in complex scenes without using any temporal continuity information. C-implementation of the algorithms demonstrates real-time performance for scene complexities comparable to those of the presented datasets.

References

[1] Caviar dataset: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] S. Belongie and J. Malik. Matching with shape contexts. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 20–26, 2000.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[4] F. Crow. Summed-area tables for texture mapping. In *Proceedings of SIGGRAPH*, volume 18, pages 207–212, 1984.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[6] P. F. Felzenszwalb. Learning models for object recognition. In *CVPR*, volume 1, pages 1056–1062, 2001.

[7] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *ICCV*, pages 87–93, 1999.

[8] I. Laptev. Improvements of object detection using boosted histograms. In *BMVC*, volume 3, pages 949–958, 2006.

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, volume 1, pages 878–885, 2005.

[10] Z. Lin, L. S. Davis, and D. Doermann. Hierarchical part-template matching for human detection and segmentation. In *ICCV*, pages 1–8, 2007.

[11] C. Messom and A. Barczak. Fast and efficient rotated haar-like features using rotated integral images. In *Australian Conference on Robotics and Automation*, pages 1–6, 2006.

[12] J. R. Renno, J. Orwell, and G. A. Jones. Learning surveillance tracking models for the self-calibrated ground plane. In *BMVC*, pages 607–616, 2002.

[13] M. D. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 353–356, 2007.

[14] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. Bilattice-based logical reasoning for human detection. In *CVPR*, 2007.

[15] P. Simard, L. Bottou, P. Haffner, and Y. L. Cun. *Boxlets: a fast convolution algorithm for signal processing and neural networks*, volume 11, pages 571–577. Advances in Neural Information, Eds. M. Kearns, S. Solla, and D. Cohn, MIT Press, 1999.

[16] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, volume 2, page 2246, 1999.

[17] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, pages 589–600, 2006.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518, 2001.

[19] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007.

[20] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785, 1997.

[21] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int. J. Comput. Vision*, 75(2):247–266, 2007.

[22] L. Zhao and L. S. Davis. Closely coupled object detection and segmentation. In *ICCV*, volume 1, pages 454–461, 2005.

[23] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *CVPR*, volume 2, pages 459–466, 2003.