International Journal of Computer Vision 72(1), 47–66, 2007
© 2007 Springer Science + Business Media, LLC. Manufactured in The United States.
DOI: 10.1007/s11263-006-8892-7

Pre-Attentive and Attentive Detection of Humans in Wide-Field Scenes

J. H. ELDER, S. J. D. PRINCE, Y. HOU, M. SIZINTSEV AND E. OLEVSKIY
Centre for Vision Research, York University, Toronto, Ontario, M3J 1P3

Received December 15, 2004; Revised November 11, 2005; Accepted March 14, 2006

First online version published in July, 2006

Abstract. We address the problem of localizing and obtaining high-resolution footage of the people present in a scene. We propose a biologically-inspired solution combining pre-attentive, low-resolution sensing for detection with shiftable, high-resolution, attentive sensing for confirmation and further analysis.

The detection problem is made difficult by the unconstrained nature of realistic environments and human behaviour, and the low resolution of pre-attentive sensing. Analysis of human peripheral vision suggests a solution based on integration of relatively simple but complementary cues. We develop a Bayesian approach involving layered probabilistic modeling and spatial integration using a flexible norm that maximizes the statistical power of both dense and sparse cues. We compare the statistical power of several cues and demonstrate the advantage of cue integration. We evaluate the Bayesian cue integration method for human detection on a labelled surveillance database and find that it outperforms several competing methods based on conjunctive combinations of classifiers (e.g., Adaboost). We have developed a real-time version of our pre-attentive human activity sensor that generates saccadic targets for an attentive foveated vision system. Output from high-resolution attentive detection algorithms and gaze state parameters are fed back as statistical priors and combined with pre-attentive cues to determine saccadic behaviour. The result is a closed-loop system that fixates faces over a 130 deg field of view, allowing high-resolution capture of facial video over a large dynamic scene.

1. Introduction

A reasonable first goal for a wide-field visual surveillance system is to localize and obtain high-resolution footage of the people present in the scene. Research in the field has, however, focused on more narrowly-defined problems such as frontal or profile face detection, face- or hand-tracking at close range, detection and tracking of pedestrians. Solving these specific problems does not necessarily lead to a solution for the more basic problem of finding the people in the scene. This is the problem we address here.

Surveillance of a large, open environment demands a wide field of view. Unfortunately, wide field of view comes at the expense of image resolution. For example, a human face at a distance of 5m will subtend only about 4×6 pixels on a 640×480 sensor with 130 deg field of view. This is insufficent resolution for most biometric and security tasks.

This problem may be addressed by combining a fixed, pre-attentive, low-resolution wide-field camera with a shiftable, attentive, high-resolution narrow-field camera (Fig. 1, Elder et al., 2005). In this paper we focus on the problem of rapidly detecting and localizing human heads in the low-resolution pre-attentive stream. These locations then form saccadic targets for the attentive sensor component so that high-resolution facial information can be recorded and analyzed.

Unfortunately, the low resolution provided by a wide-field sensor and the unconstrained nature of realistic environments and human behaviour make form cues unreliable. We therefore propose an approach

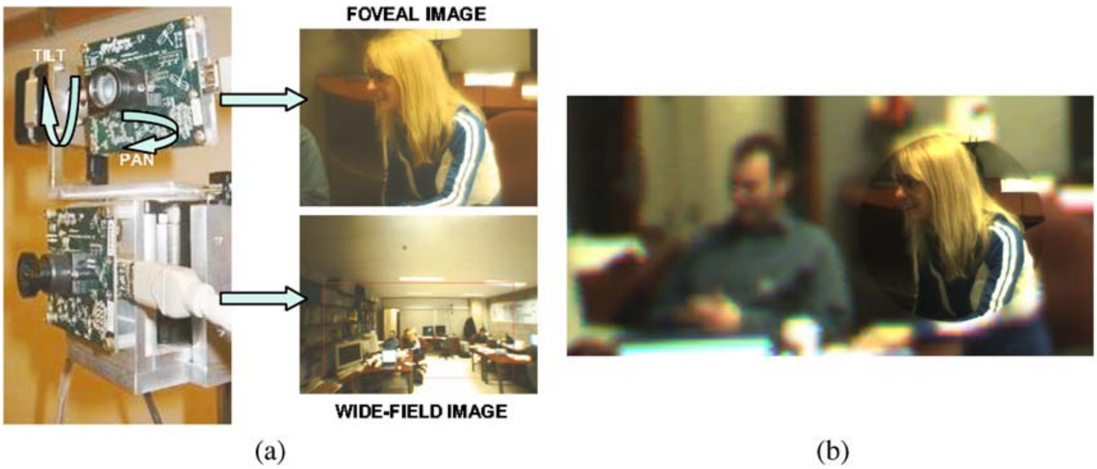


Figure 1. *(a)* Attentive wide-field sensor. *(b)* Fused output from the two sensors.

based upon the Bayesian combination of multiple weak, complementary cues that do not depend upon detailed spatial analysis.

The main contributions of this paper are:

1. We articulate what we hope is a clear first goal for visual surveillance: to localize and obtain high-resolution footage of the people present in a wide-field scene. To be relevant to realistic scenarios, it is important that observation be naturalistic, i.e., that the people in the scene are behaving naturally, not executing planned behaviours for the purposes of the project. We establish training and test databases that contain a balance of people who are sitting and standing, moving and stationary, typical of many work and public environments.
2. We lay out in detail the close relationship between this problem and the problem of peripheral detection in the human visual system, which provides inspiration for our approach.
3. We articulate the problem created by spatial correlations in the data in generating point estimates of human location. We introduce a novel solution to the problem involving two layers of supervised probabilistic modeling that bridges a pixel representation (e.g., Hayman and Eklundh, 2002) to a whole-body representation. We demonstrate the utility of a flexible norm for spatial integration that can be tailored to optimize the statistical power obtained from each modality.
4. We measure and report the relative importance of three distinct cues for person detection in naturalistic environments.
5. We perform a rigorous comparison between our proposed Bayesian cue integration technique and four state-of-the-art methods from the literature (Viola and Jones, 2001; Viola et al., 2003; Abramson and Freund, 2005; Xiong and Jaynes, 2003).
6. We implement a real-time version of our system, and integrate it within a pre-attentive/attentive sensor combination platform. We describe closed-loop methods for forming priors for Bayesian person detection based upon feedback from our attentive, high-resolution sensor. The result is a complete real-time system for localizing and obtaining high-resolution footage of the people in a wide-field scene.

2. Prior Work

2.1. *Pre-Attentive and Attentive Sensing*

There have been a number of prior efforts to combine narrow-field and wide-field sensors for human activity tracking. Scassellati (1998) studied the problem of localizing human faces and eyes using a binocular head in which each ‘eye’ consisted of separate narrow-field and wide-field sensors. Greiffenhagen et al. (2000) used a ceiling-mounted panoramic camera to provide wide-field plan-view sensing and a narrow-field pan/tilt/zoom camera at head height to provide high-resolution facial imagery. Marchesotti et al. (2003) combined a fixed wide-field sensor with a pan/tilt/zoom narrow-field sensor for outdoor surveillance.

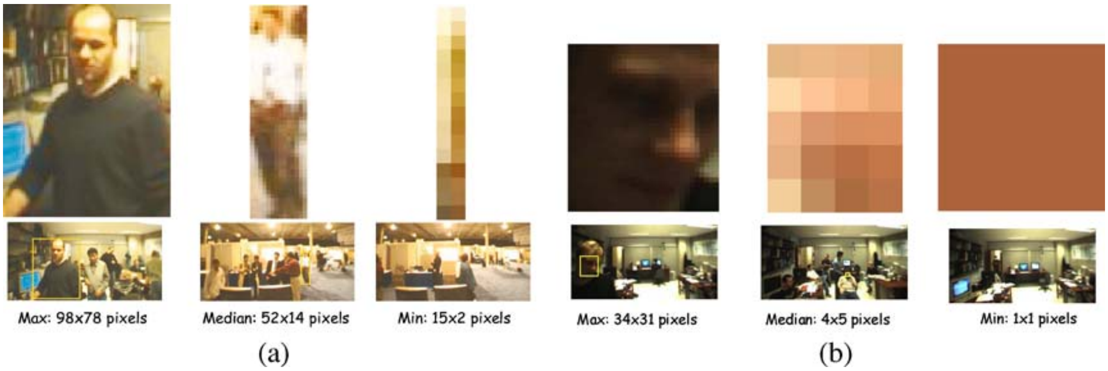


Figure 2. Range of body *(a)* and face *(b)* sizes in training database.

Functionality of prior systems has been limited in various ways. Scassellati's system was designed to detect frontal human faces at very close range (3–4 feet). Greiffenhagen et al. (2000) relied on extensive modeling of the scene and assumed people were standing upright. Marchesotti et al demonstrated only limited functionality of their system. A major goal of the present work is to go further than previous methods in demonstrating how this kind of sensor architecture can be useful for realistic surveillance problems.

2.2. *Detecting People*

Static face detection algorithms use combinations of local greyscale features to detect faces from images or single frames of a video sequence. Several of the most promising approaches use some version of AdaBoost to build strong classifiers from collections of weak classifiers in a supervised fashion (Viola and Jones, 2001; Schneiderman and Kanade, 2004; Schneiderman, 2004; Kruppa et al., 2003). More recently, systems trained to detect profile as well as frontal faces have been developed. Performance for frontal detection as high as 90% detection with 1 false positive for every 22 images has been reported. Results for profile detection are lower but still impressive: 87% detection with roughly 1 false positive for every 2 images (Schneiderman and Kanade, 2004).

A number of factors limit the utility of these methods for far-field or wide-field face detection (Bose and Grimson, 2004). First, static detection methods generally assume a minimum scale of around 24 × 24 pixels. By using local context (head and shoulder areas), reasonable performance has been achieved down to 12×16 pixel faces, in a highly constrained dataset (Kruppa et al., 2003). But to achieve surveillance over as large an area as possible, face detection must work reliably down to much lower resolutions. In our relatively unconstrained indoor testing environments, the median face size is 4 × 5 pixels, and many faces subtend less than a pixel (Fig. 2).

A second problem is that face pose in most environments is not restricted to being frontal or profile. For example, in a work environment people often look down as well as straight ahead, and may tilt their head as they are walking or talking (Fig. 3). We would also like to be able to detect a human head even when it is facing away from the camera, since the head may be turned toward the camera on subsequent frames.

Methods that make use of temporal information and coherence over frames have the potential to overcome some of these problems. Viola et al. (2003) extended their AdaBoost approach to the spatiotemporal domain to detect pedestrians in far-field scenes. They found that augmenting static greyscale features with 2-frame temporal features boosted detection performance by a factor of 10 for some scenes. They were able to achieve 80% detection with roughly 1 false positive for every 2 image frames.

In contrast to this integrated spatiotemporal approach, most dynamic approaches to human detection use foreground extraction or motion detection as a prefilter to select a subset of image locations for further analysis. Xiong and Jaynes (2003) use a foreground extraction prefilter followed by skin classification to select face region hypotheses. Haritaoglu et al. (2000) group foreground pixels into regions and then model the shapes of these regions to localize human bodies and body parts. Zhao et al. (2004) use a similar

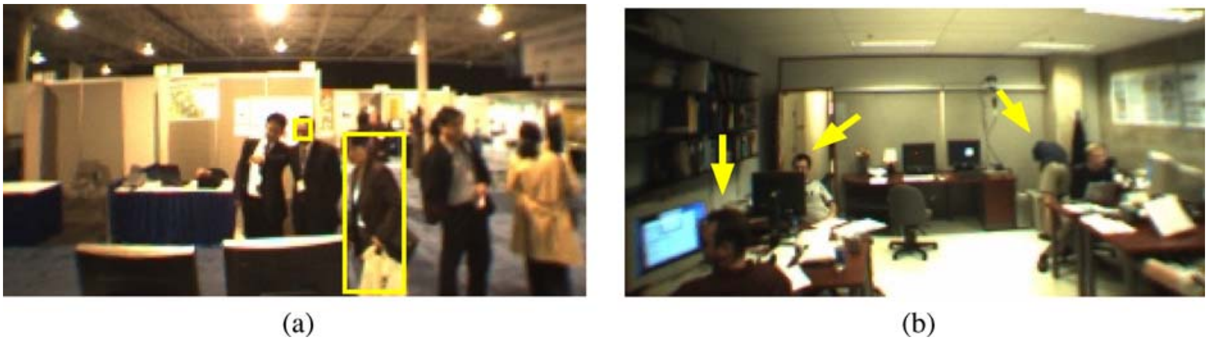


Figure 3. *(a)* Example training image from Environment 1 (conference hall). *(b)* Example image from Environment 2 (laboratory), demonstrating the variety of poses, scales and occlusions that occur in a typical working environment.

approach in conjunction with explicit 3D modeling to filter out pedestrian shadows.

Most recently, progress has been made toward adaptive methods that may generalize more readily over scenes or scene conditions. Bose and Grimson (2004) use foreground extraction as a prefilter to select moving regions in an outdoor scene. These regions are then classified based upon a set of spatial and temporal features, derived using semi-supervised methods. Nair and Clark (2004) also use foreground extraction to generate label hypotheses for unsupervised learning of static greyscale cues.

These dynamic approaches are all designed to detect walking (or running) pedestrians, translating along a ground plane, and will generally fail if pedestrians come to a standstill or become partially occluded. In contrast, we are interested in wide field detection of people engaged in a broader class of activities, e.g. talking, browsing, working. We would like to be able to detect people whether they are walking, turning, or remaining relatively still, whether they are standing or sitting, and whether or not they are partially occluded by a desk or other object. Successful detection over this broad range of conditions could enable a variety of surveillance applications in office, retail, airport and other types of environment.

The breadth of this objective suggests against many of the techniques that have been used with success for pedestrian detection. First, explicit three-dimensional modeling of scene parameters and body pose (Zhao and Nevatia, 2004) becomes difficult, since the number and range of parameters is much greater and the quality of the visual data poorer. Second, premature commitment (prefiltering) based upon a single modality is inappropriate. Motion detection will fail for people who are not walking or running, and adaptive methods for foreground extraction may fail if people sit or stand long enough to be incorporated into the background model, or become temporarily occluded. Skin or face detection methods will also fail if people are too distant or turned away from the camera. For such broad conditions no single modality will be able to filter out a significant number of candidate locations without generating numerous misses. Instead we suggest a feature integration approach in which each modality is treated as an independent cue. A decision based upon any one will generate errors, but a decision based upon a quantitative combination of all cues using reasonable probabilistic models has the potential to generate reliable behaviour.

2.3. Cue Integration

There has been considerable recent interest in cue integration approaches for related problems in video processing. Triesch & von der Malsburg (2001a,b) studied cue integration methods for the problems of face tracking and hand gesture recognition. For face tracking, they used motion, colour, shape and contrast cues, as well as a predictive cue from the results of processing previous frames. For hand gestures, the cues included Gabor jet responses, which represent local intensity and form cues, local colour, and colour Gabor jets. In both cases, cues were combined using a weighted sum. Spengler and Schiele (2001) have used a similar approach to integrate motion, skin colour, head shape and contrast cues for tracking moving people. Bayesian cue integration methods for high-resolution face-tracking, employing colour, motion, background subtraction, and ellipse-fitting cues have also been explored (Toyama and Horvitz, 2000; Sherrah and Gong, 2001).

Hayman and Eklundh (2002) studied methods for integrating motion, colour and contrast cues for dynamic scene segmentation. They also incorporated a predictive cue derived from segmentation decisions made in the previous frame. They focused on pixel-level segmentation, and employed a Bayesian method for cue combination at the pixel level.

The problems addressed by these algorithms are somewhat different from the problem we address in this paper. In many cases the foreground objects (e.g. the hands detected in Triesch and von der Malsburg (2001), the faces tracked in Toyama and Horvitz (2000) and Sherrah and Gong (2001), the people tracked in Hayman and Eklundh (2002)) are seen at relatively high resolution, and account for a large part of the image. In other cases, the people to be tracked are translating smoothly across the field of view, generating a strong motion signal (Triesch and von der Malsburg, 2001; Spengler and Schiele, 2001). Generally these methods are not real-time and are not compared with competing approaches.

3. Philosophy of Approach

A Bayesian framework is well suited to the problem of wide-field person detection for a number of reasons. The fact that none of the available cues are entirely dependable means that foreground and background likelihood distributions for each cue must be accurately modeled in order to extract what information is available from each. Moreover, the statistical approach prescribes a method for combination that maximizes the information gained from each complementary cue. Finally, a Bayesian framework defines a rigorous means for incorporating environmental and behavioural priors and specific system objectives to optimize performance.

The Bayesian approach has been successfully applied to many computer vision problems, including object recognition (Miller et al., 1997), perceptual organization (Elder et al., 2003), scene surveillance (Cox and Leonard, 1994; Buxton and Gong, 1995), dynamic scene segmentation (Hayman and Eklundh 2002), tracking (Isard and Blake, 1998; Sullivan et al., 2001; Toyama and Horvitz, 2000; Sherrah and Gong, 2001), and human appearance modeling (Sidenbladh and Black, 2003). The statistical framework we adopt here is very much in the spirit of this prior work. However, the particular problem of detecting people in wide- or far-field video, at low resolutions and in unconstrained poses and configurations, presents unique challenges.

Since this is a human-like task, we may be able to design better solutions by understanding and incorporating elements of the human system. The ability to detect other people in our environment is a fundamental human visual capacity, and due to the reduced spatial acuity of the human peripheral visual system, human observers face a similar resolution constraint.

Eye-tracking studies have shown that simple cues such as motion, flicker, colour, luminance and orientation can predict human saccadic behaviour (Parkhurst et al., 2002; Itti, 2005). Since in these experiments most saccades are less than 10 deg in amplitude, the effective cues likely lie in the parafovea and near periphery. However, psychophysical studies suggest that these simple detection mechanisms are largely preserved in the far periphery. Rovamo and Iivanainen, (1991) found that when stimuli are scaled according to the decline in cone density, sensitivity to hue is independent of eccentricity. Similarly, (Johnston and Wright, 1985) measured motion thresholds as a function of eccentricity, finding motion sensitivity to be constant across the visual field once stimulus size and speed are scaled by the cortical magnification factor, so that cortical size and speed are constant. These results suggest that the decline in motion and colour sensitivity in the periphery is completely explained by sampling: the basic mechanisms on which colour and motion discrimination is based remain intact.

This is not true for all visual capacities. Ikeda et al. (2005) have found that our perception of biological motion degrades qualitatively in the periphery, beyond what would be predicted by the decline in acuity. Similarly, (Hess and Dakin, 1997) found that contour integration mechanisms degrade markedly in the periphery, beyond what would be predicted by a decline in acuity or contrast sensitivity alone. Their results suggest that mechanisms for detecting curvilinear contours may exist only in the fovea. This does not mean that form vision is entirely absent in the periphery, but that it may be very primitive, involving only detection of dominant orientations.

In our own work (Velisavljevic and Elder, 2002) we find that colour is the most powerful amongst several cues for rapid encoding and recall of local scene information. We find that the strength of this cue remains relatively constant over the range of eccentricities tested (up to 15 deg; Velisavljevic and Elder, 2003). This can

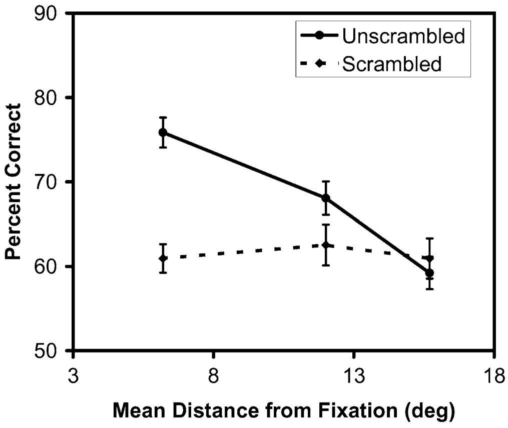


Figure 4. Recognition rates for local image patches as a function of visual eccentricity. The effects of cues related to image coherence vanish in the periphery. Velisavljevic and Elder (2003).

be contrasted with form cues, which we test by comparing local recognition and recall for normal and spatially scrambled images. While recognition rates are much higher for coherent images than for scrambled images when the region of the image to be recognized (the "target") is near the fovea, sensitivity to spatial coherence completely disappears for targets in the periphery (Fig. 4).

To summarize, the weight of evidence from human vision research suggests that mechanisms involved in peripheral object detection are likely to be based on simple motion and colour cues. Form cues, if used, are likely to be primitive, based on orientation detection. Of course, in a normally-sighted observer, human activity detected in the periphery often generates coordinated head/eye movements that bring the activity into the fovea, at which time detailed form processing can act to confirm or identify. This split between pre-attentive and attentive processing is reflected in our design approach to building attentive surveillance systems.

4. Algorithm Overview

A schematic flow diagram of our pre-attentive algorithm is shown in Fig. 5. The algorithm produces a map of the posterior probability that a human head is present at each point, derived from a vector $\overrightarrow{D}$ of complementary cues. Letting $\mathcal{H}_h$ denote the hypothesis that a head is present at a given pixel, and $\mathcal{H}_{\bar{h}}$ denote the hypothesis that a head is not present, the posterior probability of a head being present at a pixel is given by:

$$p(\mathcal{H}_h|\overrightarrow{D}) = \frac{p(\overrightarrow{D}|\mathcal{H}_h)p(\mathcal{H}_h)}{p(\overrightarrow{D}|\mathcal{H}_h)p(\mathcal{H}_h) + p(\overrightarrow{D}|\mathcal{H}_{\bar{h}})p(\mathcal{H}_{\bar{h}})} \tag{1}$$

Each individual cue in $\overrightarrow{D}$ is defined by a *modality*, *scale* and *offset*. In this paper we employ three different modalities: 2-frame motion differencing, foreground extraction and skin colour. Each of these is initially derived independently for each pixel in the image: see Section 6 for details.

The scale of each cue determines the size of the rectangular region over which these pixel cues are integrated. The offset of the cue determines the two-dimensional image displacement of the centre of the integration region relative to the hypothesized head location (Section 7). Treating each cue as conditionally independent[1], the joint likelihoods are formed by taking the product over all n cues:

$$p(\overrightarrow{D}|\mathcal{H}_h) = \prod_{i=1}^{n} p\,(D_i|\mathcal{H}_h) \tag{2}$$

$$p(\overrightarrow{D}|\mathcal{H}_{\bar{h}}) = \prod_{i=1}^{n} p\,\left(D_i|\mathcal{H}_{\bar{h}}\right) \tag{3}$$

Example likelihood distributions are shown in Fig. 8.

The spatial priors $p(\mathcal{H}_h)$ and $p(\mathcal{H}_{\bar{h}})$ incorporate weak knowledge about the height in the image at which heads are likely to occur, behavioural priors and novelty objectives (Section 11). Peaks in the resulting posterior map then form saccadic targets for our attentive sensor.

Likelihood distributions, scales, and spatial priors are learned in a supervised fashion from a hand-labelled training database. For each image, the centre of each head and boxes bounding the facial skin region and body were drawn (Fig. 3). These boxes were used to estimate the range of typical scales and offsets (Section 7).

We will first describe our training and test databases, and then how the three pointwise cues we employ are computed, and how scales and offsets are selected.

5. Training and Test Databases

Training and test databases were combined from two environments: (1) a university laboratory and (2) a demonstration session in a large conference hall. For Environment 1, training and test data were collected

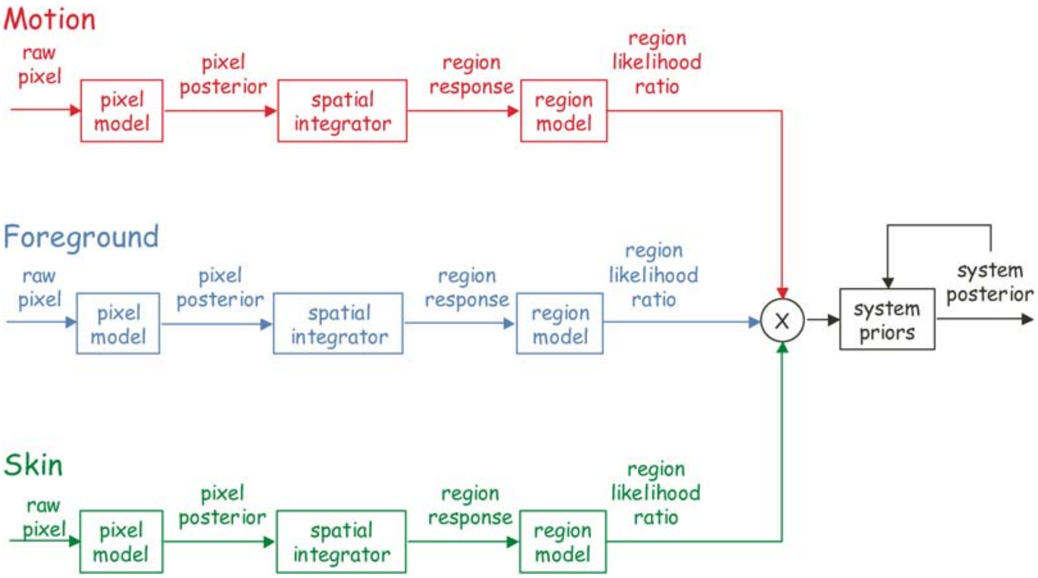


Figure 5. Pre-Attentive system design.

on different days. For Environment 2, training data were collected in the morning and test data were collected in the afternoon. The original resolution was 640×480 pixels, but processing was restricted to a central 512×256 region of interest. This was subsampled by a factor of 2 to yield a resolution of 256×128 pixels.

The frame rate for both databases was roughly 6 frames/second. Our unified training database consisted of 30 minute samples from each of the environments. Our unified test database was drawn from a 30-minute sample from Environment 1 and a 2 hour sample from Environment 2.

Every 10th frame of the training database was selected for manual labeling. This resulted in a labeled training dataset of 1,000 frames containing 5,035 heads from Environment 1 and 1,095 frames containing 6,170 heads from Environment 2. An example labeled training frame is shown in Fig. 3(a).

The test database consisted of five sets of 200 consecutive frames sampled uniformly from Environment 1, and five sets of 200 consecutive frames sampled uniformly from Environment 2. Head locations were labelled by hand: 3570 from Environment 1 and 4602 from Environment 2.

Likelihood distributions $p(\overrightarrow{D}|\mathcal{H}_h)$ for the head present condition $\mathcal{H}_h$ are based on cues observed at the 11,205 head locations in our training database. Likelihood distributions $p(\overrightarrow{D}|\mathcal{H}_{\overline{h}})$ for the head absent condition $\mathcal{H}_{\overline{h}}$ are based on cues observed at 11,205 random locations from randomly selected training frames, subject to the condition that each location lie at least 31 pixels horizontally and 57 pixels vertically from the nearest labeled body box. This ensured that cue computations at head absent locations did not overlap human bodies identified in the database.

Additional example frames from our databases are shown in Figs. 2, 6 and 12.

6. Modalities

We use motion, foreground and skin cues to detect human heads. In each case, the pointwise cue Y is the posterior probability of the hypothesis $\mathcal{H}$ that the pixel projects from a head or body, given the pixel data vector $\overrightarrow{X}$, i.e.,

$$Y = p(\mathcal{H}|\overrightarrow{X}) = \frac{p(\overrightarrow{X}|\mathcal{H})p(\mathcal{H})}{p(\overrightarrow{X}|\mathcal{H})p(\mathcal{H}) + p(\overrightarrow{X}|\overline{\mathcal{H}})p(\overline{\mathcal{H}})} \tag{4}$$

6.1. *Motion Differencing*

The data vector used for motion differencing at each pixel is based on an L_1 norm of pixel intensity differences between current and previous frames:

$$\begin{aligned}\Delta\overrightarrow{X}(t) &= |\overrightarrow{X}(t) - \overrightarrow{X}(t-1)|_1 \\ &= |\Delta r| + |\Delta b| + |\Delta g|\end{aligned} \tag{5}$$

The motion cue Y_m is computed using Eqn. 4, where $p(\mathcal{H})$ and $p(\overline{\mathcal{H}})$ are the prior probabilities of the pixel being generated by a human body or non-body region, calculated from training data. The likelihoods, $p(\overrightarrow{X}|\mathcal{H})$ and $p(\overrightarrow{X}|\overline{\mathcal{H}})$ are calculated from non-parametric representations of the pixel change distributions. We computed two-frame pixel differences for all the pixels within the body boxes in our labeled training dataset (Eqn. 5). We binned this distribution into 256 bins. Normalizing the result provides a direct representation of the likelihood distribution in the positive condition. A similar process was followed to determine the non-head likelihood, using the negative examples randomly sampled from the background regions of the training frames. An example motion posterior pixel map is shown in Fig. 6(c).

6.2. *Foreground Extraction*

We employ a pointwise adaptive algorithm for foreground extraction. We have found that operating directly in (r, g, b) space leads to many false positives due to shadows and other illumination effects, and have developed a subspace method to reduce these problems. We first computed a random set of colour changes $\Delta\overrightarrow{I}_i$ in individual background pixels from our training dataset, at random times t_1 and t_2: $\Delta\overrightarrow{I}_i = \overrightarrow{I}_i(t_1) - \overrightarrow{I}_i(t_2)$. Principal component analysis of the resultant pixel differences yields a new coordinate system in which the first component points roughly in the brightness direction (equal (r, g, b) weights). To reduce illumination effects, we discard this channel and model each pixel colour $\overrightarrow{X}$ as a mixture of two Gaussians in the residual 2D colour space.

The foreground cue Y_f is computed using Eqn. 4, where $p(\mathcal{H})$ and $p(\overline{\mathcal{H}})$ are the prior probabilities that the pixel was generated by the foreground and background process respectively.

The parameters of the foreground and background processes are estimated on-line in an unsupervised manner, using an approximation of the EM algorithm similar to (Friedman and Russel, 1997). The distribution with the greatest (learned) prior probability is assumed to represent the background process. The time constant (inverse adaptation rate) for the foreground and background process weights and for the foreground process parameters was fixed at 2 hours, to allow adaptation to slow changes in the environment. The adaptation rate for the background parameters was then optimized to maximize the evidence $p(\overrightarrow{X})$ over the training data: we found a relatively fast adaptation rate of 10 seconds to be optimal (Fig. 7(a)). The likelihood terms, $p(\overrightarrow{X}|\mathcal{H})$ and $p(\overrightarrow{X}|\overline{\mathcal{H}})$ are calculated from this mixture model. An example foreground posterior pixel map is shown in Fig. 6(e).

6.3. *Skin Detection*

The skin colour data vector is modelled in HSV space: $\overrightarrow{X} = (H, S, V)$, and the skin colour cue Y_s is computed using Eqn. 4, where $p(\mathcal{H})$ and $p(\overline{\mathcal{H}})$ denote the prior probabilities that the pixel was generated by skin and by non-skin, respectively. The likelihoods, $p(\overrightarrow{X}|\mathcal{H})$ and $p(\overrightarrow{X}|\overline{\mathcal{H}})$ are calculated from non-parametric representations of the color distributions. It is known from previous work that these distributions are not well represented by simple parametric models (Jones and Rehg, 1999)

Color distributions in HSV space were learned from hand-labelled training data: 19,000 skin pixels and 1.6 million non-skin pixels. The hue component was quantized into 60 equal-sized bins, and saturation and intensity were quantized into 20 bins each. An example skin posterior pixel map is shown in Fig. 6(g).

7. Spatial Integration

For reliable detection, cues must be spatially integrated over regions of the image that may project from the bodies and faces of humans in the scene. Due to strong spatial correlations in the data, inferring accurate probabilities for region responses directly from probabilities of pixel responses is difficult. Our approach is thus to remodel the conditional probability distributions of cues after spatial integration.

A danger with this approach is that collapsing a regional pattern of cues into a single scalar may eliminate important degrees of freedom and hence reduce statistical power. We have taken two measures to mitigate this risk. First, since the initial pixel cues have been converted into posterior probabilities prior to spatial integration, the posterior probability of a human presence at a location in the scene is likely to be monotonic with the region cue produced by integrating these local cues. Second, we employ a flexible L_γ norm for spatial integration that we optimize individually for each cue. This provides some degree of sensitivity to the nature of the spatial distribution of the cue within the integration region, in particular allowing sparse and dense cues to be handled differently.

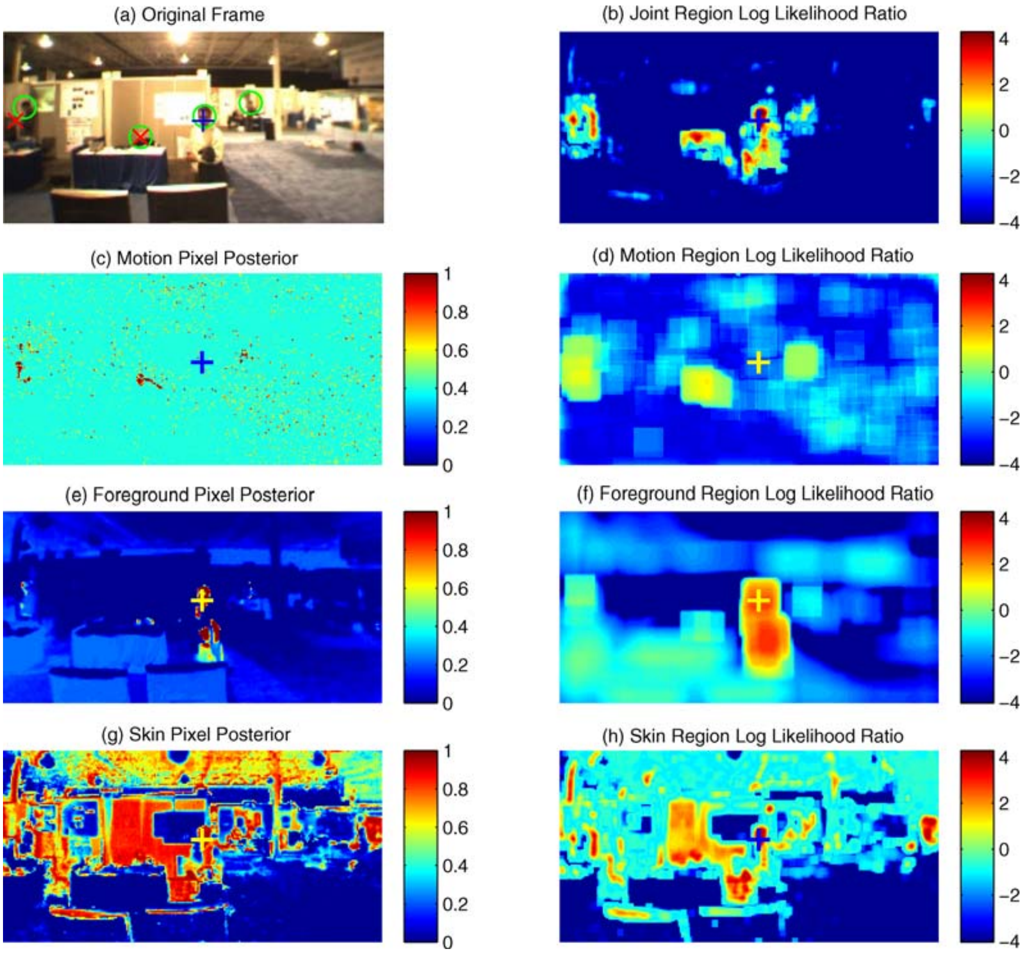


Figure 6. Likelihood computation for example frame from test database. 'O' symbols indicate hand-labeled locations of human heads. '+' and '×' symbols indicate the global and local maxima of the posterior probability, respectively. Log-likelihood ratios are clipped at −4 for display purposes only.

We use an integral image technique to rapidly integrate our pointwise cues over rectangular regions corresponding to faces and bodies. The integral image method allows a summation over any rectangular region to be computed in constant time (Viola and Jones, 2001). To determine a set of suitable scales for motion and foreground extraction, we examine the distribution of body boxes hand-labelled in our training dataset (Fig. 7(b)). We model the distribution of scales as a two-dimensional Gaussian (in height and width) and sample evenly within a 2σ confidence interval in log scalespace. A similar process is used to select candidate face scales for skin cue integration. The offset of each face and body box from the hypothesized head centre is modelled deterministically from the offsets witnessed in the training dataset using quadratic interpolation over log scalespace.

Pointwise cues are integrated over a selected N-pixel rectangular region Ω using L_γ normalization, i.e.,

$$D_i = \left(\frac{1}{N}\sum_{(x,y)\in\Omega} Y_i(x,y)^\gamma\right)^{1/\gamma} \qquad (6)$$

The likelihood distributions for the spatially integrated cues D_i are modelled by a mixture of 3 Gaussians, estimated using the EM algorithm.[2] Example fits are shown in Fig. 8.

Each resulting detector outputs a scalar value. To evaluate the sensitivity of each detector, these scalar outputs are thresholded to generate a binary decision

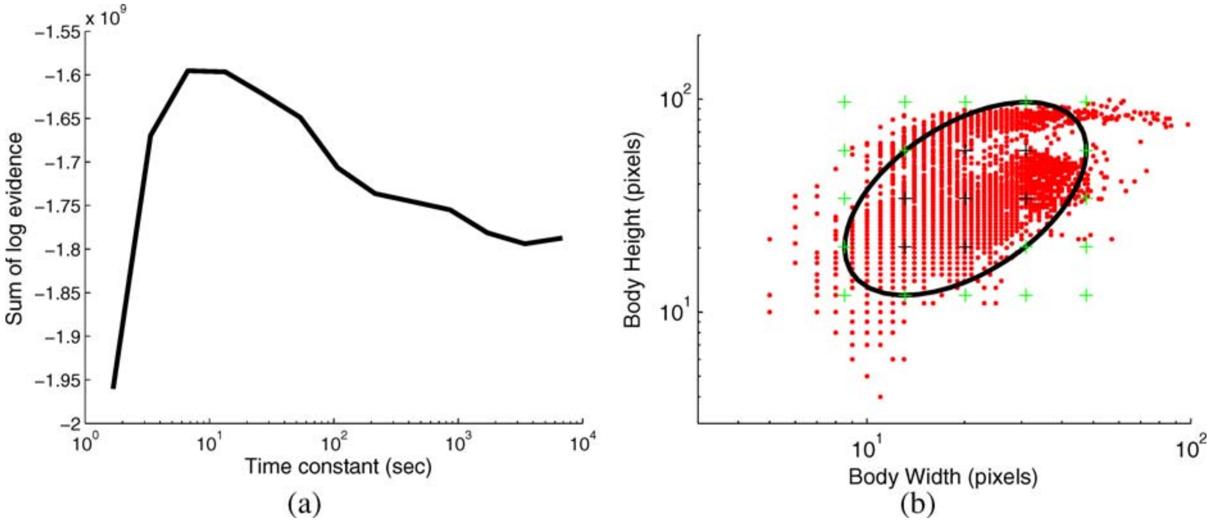


Figure 7. *(a)* Optimization of time constant for adaptation of background process parameters. The evidence over the training data is maximized for a time constant of about 10 seconds. *(b)* Multi-scale analysis. Scatterplot of body scales in the training database. Ellipse indicates the 2σ boundary of a fitted Gaussian model. Scale space is regularly sampled within this region.

for each input. Varying the threshold trades off the hit rate (proportion of true heads correctly identified) against the false alarm rate (number of background points erroneously identified as heads). The resulting curve is known as a receiver operating characteristic (ROC) and characterizes the sensitivity of the detector (Fig. 9) (Green and Swets, 1966).

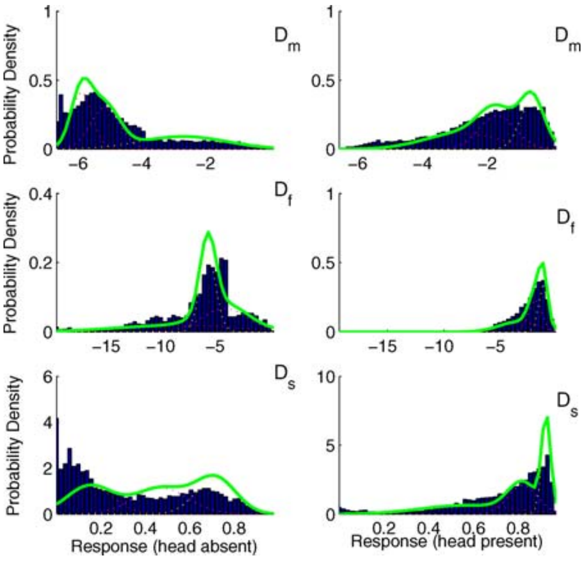


Figure 8. Example likelihood distributions for the three modalities for head-absent (left column) and head-present (right column) conditions, at representative scales and γ values. Distributions are fit with a mixture of 3 Gaussians (solid line).

The γ exponent in Eqn. 6 provides an important degree of freedom in how cues are spatially integrated. A small value ($\gamma < 1$) computes a robust statistic, in the limit counting pixels above a particular threshold. A large value ($\gamma > 1$) emphasizes large local changes, in the limit acting as a local winner-take-all mechanism.

Separate γ exponents were chosen for each modality by computing ROC curves for head/non-head classification based on each modality taken individually, over a range of γ values (Fig. 9(a)). The γ value yielding the largest area under the ROC curve was selected for each modality.

The results of this optimization were quite interesting. We found high γ values to be optimal for both skin and motion modalities. Motion detection is essentially a boundary cue, the weight of the cue being generated near sharp changes in colour, e.g., at the boundary of the foreground object. Further, human activity often does not involve large body motions, but rather small, isolated motions of the head or hands. These factors combine to make motion a sparse cue for human presence, leading to a high γ value that generates a strong response from a small number of highly active pixels.

We believe that a large γ value for skin detection arises for similar reasons. Because of the diversity in face pose, occlusion, and distance, the number of skin pixels actually visible within a face box is highly variable and often small. A large γ value allows a few strongly skin-like pixels to indicate strong evidence for human presence.

We found a more moderate value ($\gamma = 2.7$) to be optimal for foreground extraction. This makes sense, since foreground extraction is essentially a region cue, depending on the difference in colour between the foreground and background objects, and thus is likely to be less sparse than the other cues.

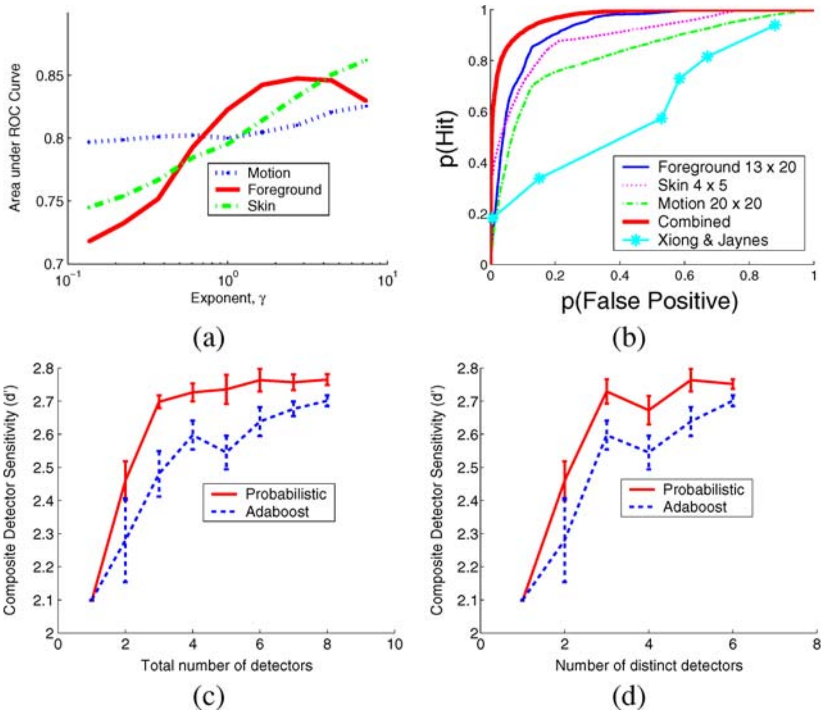
8. Detector Selection

Sampling 7 scales for for each modality leads to a total of 21 local detectors. Due to overlapping spatial support, many of these detectors are redundant. In order to select a small subset that provide complementary information, we employ a simple greedy selection strategy. We begin with the detector that yields the best solo performance, as measured by the area under its ROC curve for human/non-human classification on the training dataset (11,205 head-present and 11,205 head-absent examples). We then test the effect of adding each remaining detector and select the detector that yields the highest increase in performance. This process is repeated until performance asymptotes or declines.[3]

Table 1. List of detectors combined and resulting performance of composite system on training database, for probabilistic and Adaboost combination methods

Probabilistic			Adaboost		
Cue	Scale	d'	Cue	Scale	d'
Foreground	20 × 20	2.10	Foreground	20 × 20	2.10
Skin	4 × 5	2.46	Motion	13 × 20	2.28
Motion	20 × 20	2.70	Skin	4 × 5	2.48
Foreground	20 × 57	2.73	Skin	4 × 5	2.60
Skin	4 × 5	2.74	Motion	31 × 57	2.55
Foreground	13 × 20	2.76	Foreground	13 × 20	2.64
Motion	13 × 20	2.76	Skin	4 × 5	2.68

The result of this selection process is shown in Table 1 and Figs. 9(b-d). The power of multimodal statistical integration is confirmed: performance is increased most effectively by first integrating one detector from each of the three different modalities. Foreground extraction was found to be the most powerful individual cue, followed by skin and then



(a)

(b)

(c)

(d)

Figure 9. *(a)* Error rate for each module is plotted as a function of data exponent, γ. Data are averaged over all scales tested. *(b)* Performance on training database of individual and composite detectors, compared to representative logical combination method (Xiong and Jaynes, 2003). *(c–d)* Performance on training database of probabilistic and Adaboost methods for combining detectors.

motion. Adding additional detectors at various scales was found to improve performance by only 0.2%. We thus employ the 3-detector system for the remaining experiments in this paper.

Fig. 9(b) also shows the result of integrating foreground and skin cues using the conjunctive approach of (Xiong and Jaynes, 2003). The fact that performance lies well below the performance of our probabilistic approach based on only foreground or skin cues indicates that it is not just using multiple cues that matters: accurate probabilistic models and methods for integrating cues are also important. We discuss and evaluate this alternative conjunctive approach more thoroughly in Section 10.

Figs. 6(d, f and h) show log likelihood ratio maps for these 3 detectors following spatial integration, and Fig. 6(b) shows the joint log likelihood ratio obtained by summing these maps (under the assumption of conditional independence). Note that all four people in the scene generate local maxima in the joint likelihood ratio, but for different reasons. The central figure is quite still and generates very little motion energy, but shows up well in the foreground and skin modalities, while the others obtain support from all three sources.

For comparison, we have also implemented an Adaboost method for selecting from these 21 detectors and integrating their outputs.[4] In this case, the basis for selection is minimum weighted error rate, rather than area under the ROC curve (see, e.g., Viola and Jones, 2001).

Since the individual detectors in the Adaboost framework have binary output, the number of data points on the ROC curve for the Adaboost selection method is 2^{n-1}, where n is the number of detectors. As a result, the ROC curve is poorly defined for low-order systems. We therefore compare the two techniques using d', a measure of detector sensitivity based upon a normal noise model (Green and Swets, 1966). d' for a specific detector with a fixed threshold is defined as the difference between the z-score associated with the detector's hit rate and the z-score associated with the detector's false positive rate. To compare the different methods for detector selection, we compute and average d' values for each point on the ROC curve for each composite detector (Fig. 9(c)). The set of detectors selected is similar to those selected using our greedy method, but the order of selection is slightly different (Table 1).

Since repeated inclusion of the same detector is permitted in both combination methods we have also plotted our results against the number of *distinct* detectors employed (Fig. 9(d)). Since the computational cost of the algorithm is determined in part by how many distinct detectors are employed, this provides a more meaningful basis for comparison. The results show that probabilistic combination yields a more sensitive composite detector than Adaboost in all cases.

Our Bayesian integration method is based on approximating each cue as conditionally independent. How valid is this approximation? Correlation analysis[5] of the three selected detectors reveals r^2 values less than .03 in all cases, with one exception: we obtain an r-squared value of .15 relating the foreground and motion cues, conditioned on the presence of a head. In other words, variation in the foreground cue accounts for 15% of the variance in the motion cue, and vice-versa.

To put these values in perspective, they can be considered against the r^2 values found between scales within a modality, conditioned on the presence or absence of a head, which we find to range from 0.63-0.91. These strong correlations help to explain why we see limited benefit from incorporating multiple scales within modalities.

It is perhaps not surprising that our highest cross-modal dependence is found for foreground and motion cues, since both cues are based on deviations of pixels in the current frame from a model based upon observations from prior frames. In the case of the motion cue, this model is simply the values observed in the previous frame, while for the foreground cue the model is derived in a more sophisticated fashion from multiple prior frames.

The time constant for adaptation of background process parameters was based upon isolated analysis of the foreground cue (Fig. 7(a)). It is possible that joint analysis of foreground and motion cues would lead to a larger time constant, resulting in reduced correlation between the two cues. However, given the still modest r^2 value of 0.15, the potential benefit from such an analysis is limited.

9. Static Spatial Prior

Our pre-attentive detector generates a map indicating the likelihood ratio of human heads over the pre-attentive field of view (Fig. 6(b)). This likelihood ratio is updated at frame rate, which for our system is roughly 6 fps.

In most environments, heads do not occur with uniform probability over the entire scene. For example, heads rarely occur near the floor or ceiling of the scene.

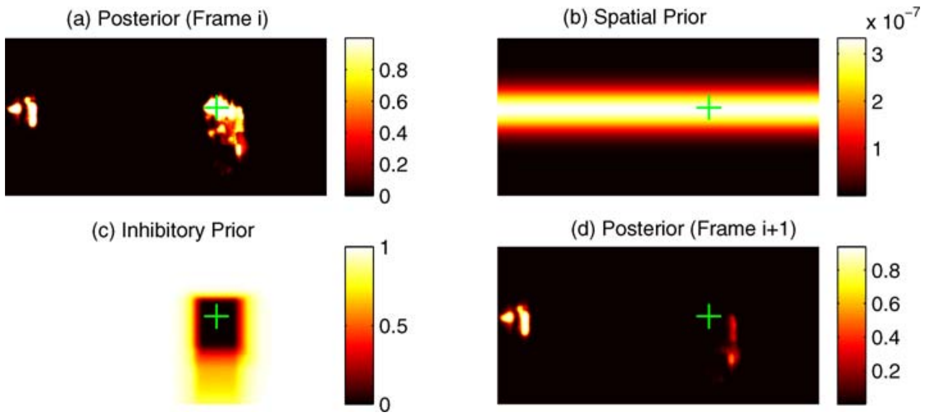


Figure 10. Priors and Posteriors. '+' symbol indicates the global maximum of the posterior in Frame *i*.

This information can be represented as a weak spatial prior that can be incorporated in a natural way with the likelihood ratios generated by the pre-attentive cues. Fig. 10(b) shows the spatial prior used for our system, modelled as Gaussian in the vertical direction, with parameters estimated from the locations of heads in our training database. Combination with the likelihood ratio map yields a map of posterior probabilities for heads in the scene (Fig. 10(a)).

10. Evaluation

Standard databases of frontal and profile faces have recently permitted quantitative analyses between competing approaches to static face detection. At the time of writing, however, there is still little comparative analysis of methods for solving the more general problem of detecting people in wide-field video. A major goal of the present work is thus to rigorously evaluate our Bayesian cue integration approach for wide-field human detection, and compare it with four other state-of-the-art approaches. The results are compared on our test database using an ROC analysis. All five approaches are run over each entire frame of the test dataset.

In our approach, local maxima of the posterior that exceed a threshold form the estimated head hypotheses. If two maxima above threshold lie within 12 pixels of each other, the lesser is suppressed. The requirement for a 'hit' is that the estimated head location lie within 12 pixels of the actual head location. This ensures that the centre of the head will lie in the field of view of our attentive sensor after a saccade, barring large movements of the subject between frames. Multiple estimates within a 12-pixel radius of a true head location are counted as a single hit. Varying the threshold for detection sweeps out an ROC curve for the system.

We now summarize each of the competing approaches we evaluate.

10.1. Static Adaboost Detection

The popularization of Adaboost methods in the computer vision community began with a paper by Viola and Jones (2001). The method employs a large number of simple, linear Haar-like filters operating on static greyscale frames. Each filter is used as a weak classifier, and these are combined to form strong classifiers using Adaboost. Cascading of strong classifiers allows successive pruning of potential face locations, resulting in fast system performance. The method was demonstrated on the problem of frontal face detection. We evaluate the OpenCV implementation of the Viola & Jones detector, which includes in-plane rotations of the front-line filters, potentially allowing greater variation in face pose (Lienhart and Maydt, 2002). The output of the system is a set of boxes bounding detected faces. We consider each box a hit if its centre lies within 12 pixels of a labelled head.

Since many of the faces in our database are below the minimum face size assumed by the Viola & Jones detector (20 × 20 pixels), we evaluate the system for images upsampled by factors of 1 to 8 to sweep out an ROC curve. Otherwise, default parameters were used. We used the pre-trained OpenCV face detection system: we did not retrain it on our database.

10.2. Semi-Automatic Visual Learning (SEVILLE)

One possible limiting factor of the Viola & Jones detector for wide-field human detection is that it uses only facial information. We therefore also evaluate a recent system called SEVILLE (Semi-Automatic Visual Learning) (Abramson and Freund, 2005) that employs features from the entire body of the subject.

As for the Viola & Jones approach, SEVILLE is based on simple boosted features derived from static greysycale frames. It differs in that these weak classifiers are based on a simple ordinal relationship between intensity values at a cluster of points. Specifically, binary classifer output is determined by whether intensities at one constellation of points are *all* greater than the intensities at another constellation of points. Selection from the large population of possible classifiers is based upon a genetic programming technique. We tested an interactive version of the system downloaded from www.yotam.net.

Positive and negative examples were isotropically scaled to 15 × 52 pixels, the median size of our training population. Since the training examples varied in aspect ratio, scaling required small adjustments in width or height, always made so that the human figure remained completely encompassed by the box. From our labelled training database we computed the mean displacement of the head centre relative to the top of the scaled body boxes to be 31% of the box height: this served as our estimate of head location.

Training was initially based on 3000 negative examples and 1875 positive examples selected randomly from the training database. A detector with 300 weak classifiers was trained and this was run on the training data. False positives from this run were used to bootstrap the negative training set to give a final set of 8170 negative examples. We then retrained with 1200 detectors.

The system provides a threshold parameter that can be used to adjust the balance of hits and false positives. We varied this parameter to sweep out an ROC curve for the system.

10.3. Dynamic Adaboost Detection

Viola et al. (2003) extended the work of Viola and Jones to the problem of detecting pedestrians, using both intensity and motion cues. Like the Viola and Jones approach, the method is based upon a cascade of boosted classifiers, but now a subset of these classifiers are built on difference images computed from successive frames with fixed spatial offsets. This provides a basis for features sensitive to the amount and direction of motion. These are combined with static features computed as in Viola and Jones (2001).

We implemented a version of their system using 45,000 detectors, trained on 3,000 positive and negative examples from our training database. Examples were isotropically scaled to 15 × 20 pixels, as in Viola et al. (2003). Since the training examples varied in aspect ratio, scaling required small adjustments in width or height, always made so that the human figure remained completely encompassed by the box. In other respects our implementation followed the design outlined in Viola et al. (2003).

The system outputs bounding boxes for regions detected as people in the scene. From these we need to extract an estimate of head location. From our labelled training database we computed the mean displacement of the head centre relative to the top of the scaled body boxes to be 13% of the box height: this served as our estimate of head location. We consider each estimate a hit if it lies within 12 pixels of a labelled head. Adjusting thresholds at each layer of the cascade sweeps out an ROC curve for the system, as in Viola et al. (2003).

10.4. Conjunctive Cue Integration

The potential benefit of cue combination for detection and tracking problems has been recognized for some time (e.g., (Triesch and Malsburg, 2001; Toyama and Horvitz, 2000)). It is less clear, however, whether the probabilistic approach to spatial integration and cue combination is critical, or whether it could be replaced with something simpler. To address this question, we have implemented a version of a head detection algorithm due to Xiong and Jaynes (2003). Their method also involves cue integration, but in their approach cues are integrated conjunctively.

To focus on the specific issue of how detectors are spatially integrated and combined, we use the same pixel posterior maps for both systems. The Xiong & Jaynes method uses these maps to classify each pixel as skin/non-skin and foreground/background. The subset of pixels classified as both skin and foreground are then selected and a morphological erosion process is applied to eliminate small regions. Bounding boxes are computed for the remaining regions, and these are considered face hypotheses. We consider each box a hit if its centre lies within 12 pixels of a labelled head.

The Xiong & Jaynes method has a number of free parameters: the thresholds for skin and foreground classification, the size of the erosion kernel, and a lower bound on the size of regions to be considered significant. We sample this parameter space over a range of reasonable values to sweep out an ROC curve for the system.

We emphasize that the conjunctive cue integration method we have evaluated is based on the approach described by Xiong and Jaynes (2003), however since it uses our own pixel features, it is not identical to their system, so definitive conclusions about the performance of their system cannot be drawn from this evaluation.

10.5. Results

Fig. 11 shows the results for all five systems. The Bayesian cue integration approach performs substantially better than the other approaches we tested. The advantage of probabilistic over conjunctive combination (Xiong and Jaynes, 2003) is confirmed. The relatively poor performance of the static Adaboost face detection system (Viola and Jones, 2001) is likely due to the low resolution and extreme pose variation that characterizes our database. The dynamic Adaboost detector (Viola et al., 2003) and the SEVILLE detector (Abramson and Freund, 2005), both designed to detect whole bodies of pedestrians, fair better.

Fig. 12 shows examples which provide some insight into these results. Detections for all systems except for the Static Adaboost face detector (Viola and Jones, 2001) appear to be correlated with human activity. The SEVILLE system (Abramson and Freund, 2005) is fooled by inanimate objects that appear vaguely human-like, e.g., the coat hanging in the background and the table legs in the foreground of Frame 4. The dynamic Adaboost approach succeeds in detecting a number of people who appear to be in motion, but does less well for people at rest (e.g., Frame 2). Some false alarms appear to be caused by poor localization of human activity (e.g., Frame 3), while others seem to be caused by mistaken static cues (Frame 4). The Bayesian cue integration approach appears to detect people over a broader range of conditions, including people walking (Frames 1 and 3), sitting (Frame 2), standing (Frame 4) and even people quite distant from the camera (Frame 5). A number of the people detected are partially occluded (e.g., Frames 1, 4, 5), although occlusion may be responsible for some misses as well (e.g., Frames 1, 2, 3, 5). False alarms are generally close to actual heads, but outside the criterion distance of 12 pixels. These ‘near misses’ are likely caused by poorly localized human activity: this suggests that performance might be improved with pre-attentive algorithms for refining location estimates once human activity is detected.

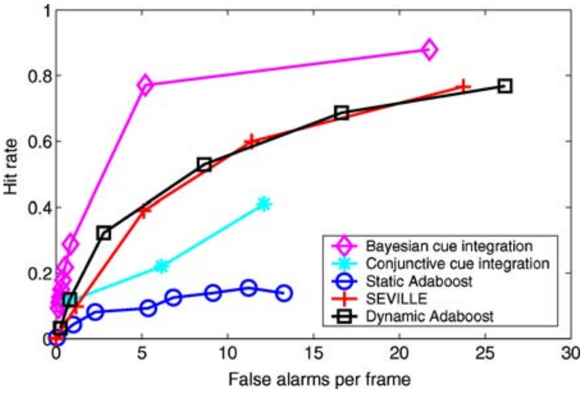


Figure 11. System performance comparison.

All systems, including the Bayesian cue integration method, generate numerous misses. In these examples, our misses occur most frequently for sitting subjects (Frames 1-3). This suggests that there might be some benefit in training separate detectors for upright and sitting subjects.

Although our method surpasses other state-of-the-art approaches, there is room for improvement: local maxima of the posterior generate a hit rate of 0.38 at a rate of 1.0 false positives per frame. For the purposes of attentive sensing, however, it is only the *global* maximum of the posterior that will form a saccadic target, and we find that 75% of saccadic targets defined by global maxima of the posterior correctly identify a human head. This level of performance is certainly sufficient to be useful, and can be further improved by incorporating attentional feedback (Section 11).

Fig. 13 provides information on the size distribution of detected heads. Labelled faces in the test database are distributed over a range of less than 1 to 16 pixels, with a median face width of 4 pixels. Assuming a normative face width of 17cm, this corresponds to a range of distances from roughly 60cm to more than 10m, and a median distance of roughly 2.4m. The distributions of detected faces (local maxima of the posterior above threshold) and targeted faces (global maxima above threshold) match the ground truth distribution quite closely; the median width of both detected and targeted faces is also 4 pixels.

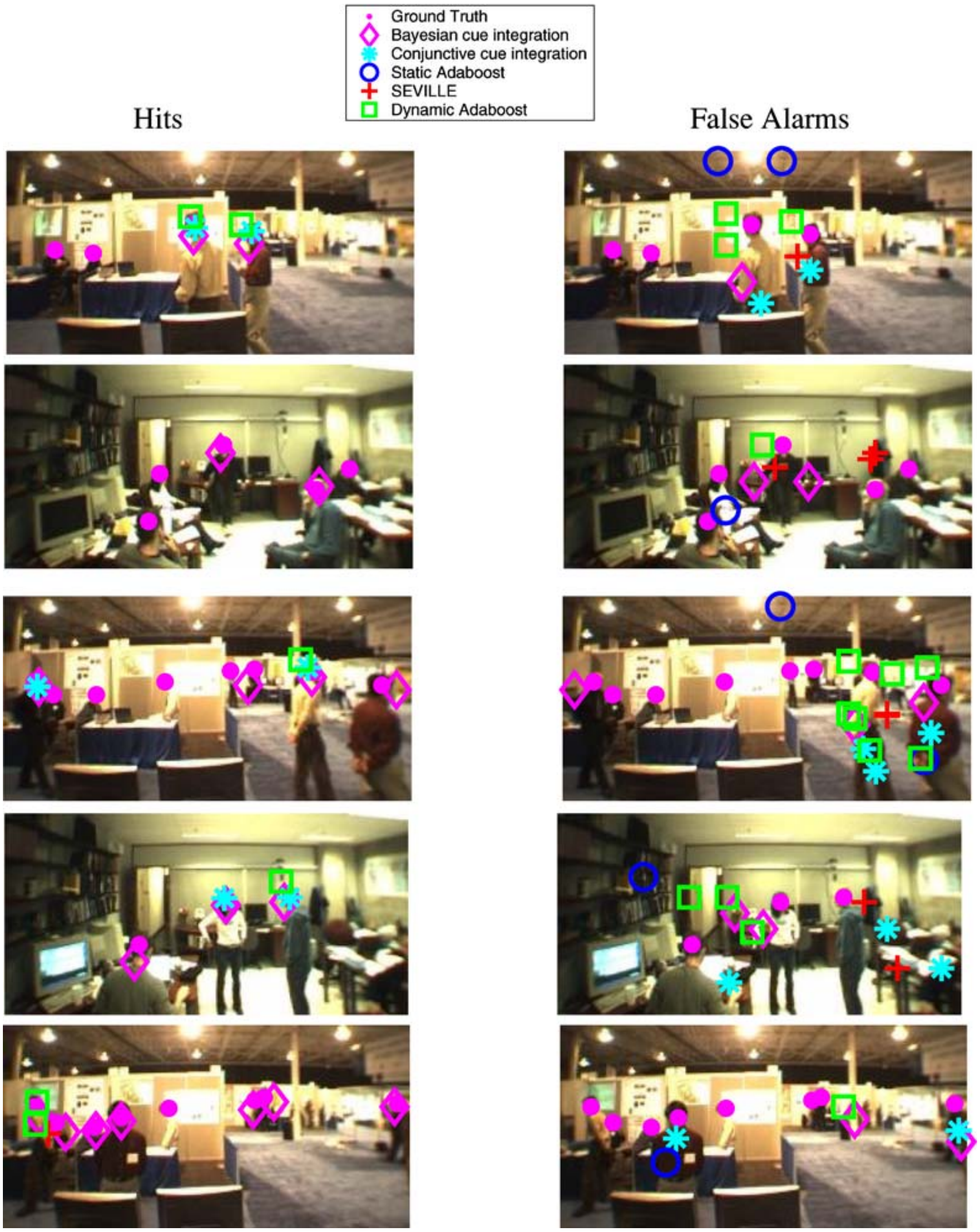


Figure 12. Example results.

11. Incorporating Attentive Feedback as Saccadic Priors

We have integrated a real-time version of our pre-attentive detector within an attentive sensing platform. The sensor consists of two 30Hz RGB Point Grey Dragonfly cameras. The wide field camera is fixed in position and has a 2.1 mm lens subtending a 130 deg horizontal field of view. The foveal camera is mounted on a pan/tilt platform and has a 24 mm lens with a 13 deg horizontal field of view. The pan and tilt motors are Lin Engineering step motors, with a step size of 0.1 degrees. The system runs on a Dual Processor 2.0 GHz PC at approximately 6 frames per second. The

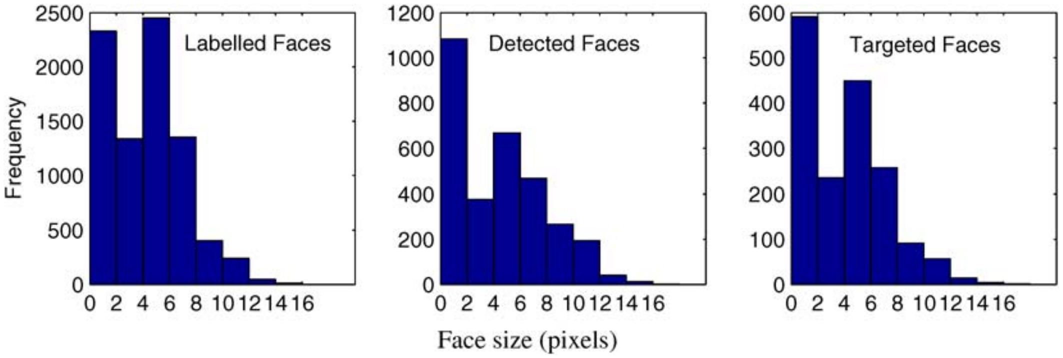


Figure 13. Distribution of detections over face size.

pan and tilt motors and video capture run on separate threads. Our algorithm detects human activity in the wide-field, low resolution image in order to orient the attentive camera.

Integration with an attentive device introduces the possibility of feeding back attentive sensing information as priors for the pre-attentive detection device. This information can include information about heads that have been confirmed attentively at high resolution, as well as information about the current fixation state.

Fig. 14 shows how these priors are used to shape the saccadic behaviour of our system. The likelihood ratio map determined from the current low-resolution (pre-attentive) frame is combined with a prior map to yield a posterior map that determines the gaze command sent to the attentive sensor. This gaze command is updated at frame rate (6 fps). Note that scenes containing many people may generate multiple extrema in the posterior map, which without further conditioning can cause rapid slewing of the attentive sensor back and forth between individuals, never providing sufficient time for attentive confirmation or identification.

To avoid this, the default state of the system is to maintain fixation until positive confirmation is obtained. This is effected by feeding the gaze command back as a delta function in pre-attentive space. In order

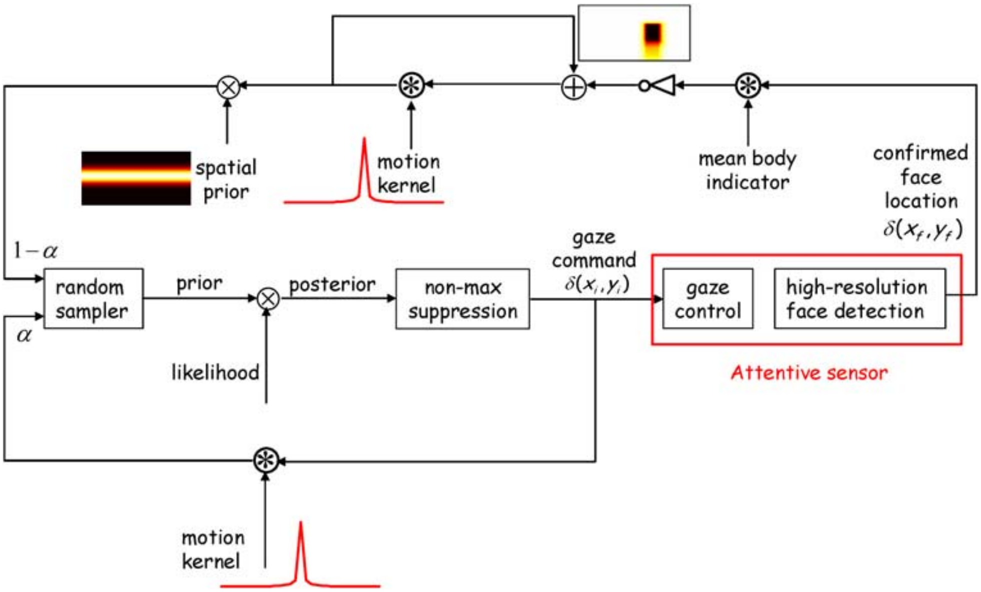


Figure 14. Fixation and Inhibitory Priors.

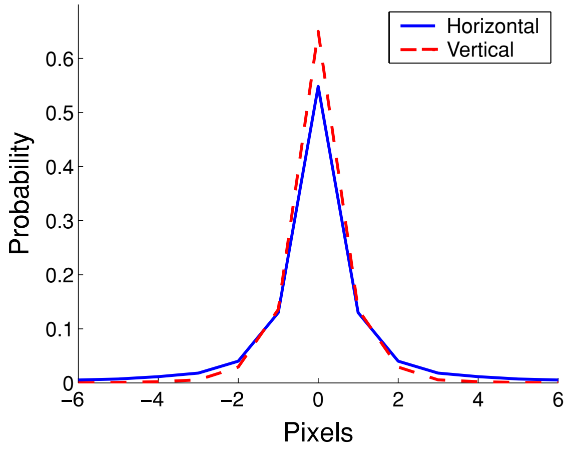


Figure 15. Motion kernel: probability density of head displacement between consecutive frames.

to account for possible movement of the individual between frames, this delta function is convolved with a 2D motion kernel representing the probability of two-frame displacement, learned from our labeled training data (Fig. 15). We call this the *fixation prior*, and its effect is to cause the attentive sensor to dwell on an individual, tracking them as they move around the scene.

In our current system, we use the OpenCV implementation of the Viola & Jones face detector (Viola and Jones, 2001; Lienhart and Maydt, 2002) to confirm faces at high resolution; in principle any method for high-resolution face detection could be used here. Some example faces confirmed at high resolution are shown in Fig. 16. When a face is confirmed, a feedback map is generated which inhibits the pre-attentive location of the detected face. This inhibition takes the form of a spatial map indicating the complement of the probability that each pre-attentive pixel projects from part of the body of the detected person: the goal here

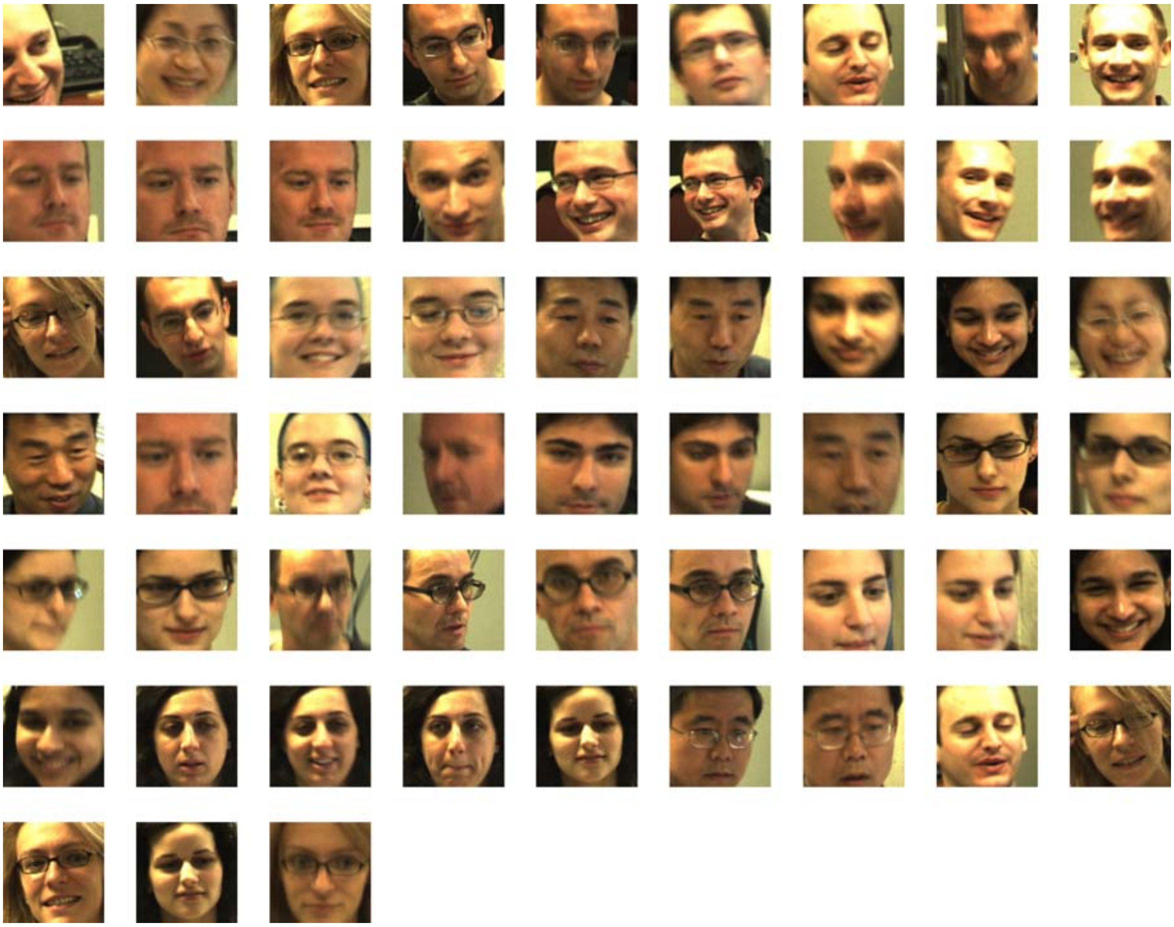


Figure 16. Faces confirmed at high resolution by attentive sensor.

is to null out all of the pre-attentive data generated by the person already detected. Once combined with the spatial prior, this inhibitory prior replaces the fixation prior on the next frame, typically generating a large gaze shift to a secondary maximum in the pre-attentive map. The fixation prior then takes over again, until the next face is confirmed. Note that the inhibitory prior maintains a memory of all locations at which faces have been confirmed, but this memory dissipates through repeated convolution of the inhibitory prior with the learned motion kernel on each frame, reflecting the increase in uncertainty about the location of the detected individual over time.

In our experiments, we find that the global maximum of the posterior corresponds to an actual head location on roughly 75% of frames. Unfortunately, this means that 25% of saccades land on non-head locations. This presents an obvious problem for our design: if the saccadic target is a false alarm, the attentive sensor may fixate indefinitely.

One possible solution is to also generate an inhibitory command when the attentive sensor *fails* to detect a face. The problem with this approach is that failure is often due to motion of the subject, or to a face that is outside the range of poses handled by the attentive detector. In these situations it is common that a face will be confirmed on subsequent attentive frames, when the subject turns more toward the camera, or slows down.

Our present solution is therefore to incorporate a random sampler that will maintain fixation with probability α, so that fixation durations follow an exponential distribution. In our experiments, we have set $\alpha = 0.95$, which yields an average fixation duration of 20 frames (roughly 3 seconds), in the absence of inhibitory commands.

Example videos of the system in operation can be found under *Current Research Projects* at elderlab.yorku.ca.

12. Conclusion

Detection of the people present in a wide-field scene is complicated by many factors, including low resolution, diversity in pose and occlusions. In this paper we have shown that a Bayesian cue-integration approach involving layered probabilistic modeling is capable of detecting humans in a variety of poses (sitting, standing, walking), over a large range of distances, and in various occlusion relationships. The Bayesian cue integration approach outperforms several competing methods based on conjunctive combinations of classifiers. Analysis suggests that both the multiplicity of complementary cues and the probabilistic method for integration are important for performance. We have developed a real-time version of our pre-attentive human activity sensor that generates saccadic targets for an attentive foveated vision system, and have shown how information from the attentive sensor can be fed back as fixation and inhibitory priors to shape saccadic behaviour.

Acknowledgments

We thank three anonymous reviewers for their comments and suggestions. We also thank Yotam Abramson for providing the SEVILLE software and guidance in its use. This work was funded by grants from IRIS and CRESTech.

Notes

1. We explore the validity of this approximation in Section 8.
2. We also tried using nonparametric representations of the likelihoods, but found a mixture of Gaussians model produced superior results. This may be due to the difficult problem of selecting optimal bin sizes (Izenman, 1991).
3. We have also implemented an exhaustive approach that tests all possible combinations of detectors and selects the best possible combination. We have tested this approach up to systems with $n = 4$ detectors, and found the greedy approach to be optimal for our dataset.
4. Here we are only testing Adaboost as a method for selecting and combining our probabilistic detectors. In Section 10 we will evaluate and compare complete systems for human detection based upon Adaboost techniques
5. Analyzing correlations is reasonable here, since the region cues are specifically designed to be monotonic with the probability of a human head.

References

Abramson, Y. and Freund, Y. 2005. Semi-automatic visual learning (Seville): a tutorial on active learning for visual object recognition, http://caor.ensmp.fr/∼abramson/sevilleCVPR/.

Bose, B. and Grimson, E. 2004. Improving object classification in far-field video. In *Proc. CVPR*, 2:181–188.

Buxton, H. and Gong, S.G. 1995. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1–2):431–459.

Cox, I.J. and Leonard, J.J. 1994. Modeling a dynamic environment using a bayesian multiple hypothesis approach, *Artificial Intelligence*, 66(2):311–344.

Elder, J.H., Dornaika, F., Hou, Y. and Goldstein, R. 2005. Attentive wide-field sensing for visual telepresence and surveillance. In *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, (Eds.) Academic Press/Elsevier, San Diego, CA.

Elder, J.H., Krupnik, A. and Johnston, L.A. 2003. Contour grouping with prior models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(25):661–674.

Friedman, N. and Russel, S. 1997. Image segmentation in video sequences: a probabilistic approach. In *Proc. UAI*, 175–181.

Green, D.M. and Swets, J.A. 1966. *Signal detection theory and psychophysics*. Wiley, New York.

Greiffenhagen, M., Ramesh, V., Comaniciu, D. and Niemann, H. 2000. Statistical modeling and performance characterization of a real-time dual camera surveillance system. In *Proc. CVPR*, 335–342.

Haritaoglu, I., HArwood, D. and Davis, L.S. 2000. W^4: Real-time surveillance of people and their activities, *IEEE PAMI*, 22(8):809–830.

Hayman, E. and Eklundh, J.O. 2002. Probabilistic and voting approaches to cue integration for figure-ground segmentation. In *European Conference on Computer Vision*, of *Lecture Notes in Computer Science*, 2352:469–486.

Hess, R.F. and Dakin, S.C. 1997. Absence of contour linking in peripheral vision. *Nature*, 390:602–604. Letters to Nature.

Ikeda, H., Blake, R. and Watanabe, K. 2005. Eccentric perception of biological motion is unscalably poor. *Vision Research*, 45:1935–1943.

Isard, M. and Blake, A. 1998. Condensation: conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28.

Itti, L. 2005. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes, *Visual Cognition*, 12(6):1093–1123.

Izenman, A.J. 1991. Recent developments in nonparametric density estimation, *Journal of the American Statistical Association*, 86(413):205–224.

Johnston, A. and Wright, M.J. 1985. Lower thresholds of motion for gratings as a function of eccentricity and contrast. *Vision Research*, 25(2):179–185.

Jones, M.J. and Rehg, J.M. 1999. Statistical color models with application to skin detection. In *Proc. CVPR*, 274–280.

Kruppa, H., Santana, M.C. and Schiele, B. 2003. Fast and robust face finding via local context. In *Proc. VS-PETS*, 157–164.

Lienhart, R. and Maydt, J. 2002. An extended set of Haar-like features for rapid object detection. In *IEEE International Conference on Image Processing*, 900–903.

Marchesotti, L., Marcenaro, L. and Regazzoni, C. 2003. Dual camera system for face detection in unconstrained environments. In *Proc. ICIP*, 1:681–684.

Miller, M.I., Grenander, U., O'Sullivan, J.A. and Synder, D.L. 1997. Automatic target recognition organized via jump-diffusion algorithms. *IEEE Transactions on Image Processing*, 6(1):157–174.

Nair, V. and Clark, J.J. 2004. An unsupervised, online learning framework for moving object detection. In *Proc. CVPR*, 2:317–324.

Parkhurst, D., Law, K. and Niebur, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123.

Rovamo, J. and Iivanainen, A. 1991. Detection of chromatic deviations from white across the human visual field. *Vision Research*, 31(12):2227–2234.

Scassellati, B. 1998. Eye finding via face detection for a foveated active vision system. In *AAAI/IAAI*, 969–976.

Schneiderman, H. 2004. Feature-centric evaluation for efficient cascaded object detection. In *Proc. CVPR*, 2:29–36.

Schneiderman, H. and Kanade, T. 2004. Object detection using the statistic of parts. *International Journal of Computer Vision*, 56(3):151–177.

Sherrah, J. and Gong, S. 2001, Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects. In *Proceedings of the International Conference on Computer Vision*, II:42–49.

Sidenbladh, H. and Black, M.J. 2003. Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54(1/2/3):183–209.

Spengler, M. and Schiele, B. 2001. Towards robust multi-cue integration for visual tracking. In *International Conference on Vision Systems*, Berlin, 2001, vol. 2095 of *Lecture Notes in Computer Science*, pp. 93–106, Springer-Verlag.

Sullivan, J., Blake, A., Isard, M. and MacCormick, J. 2001. Bayesian object localisation in images. *International Journal of Computer Vision*, 44(2):111–135.

Triesch, J. and von der Malsburg, C. 2001. Democratic integration: self-organized integration of adaptive cues. *Neural Computation*, 13:2049–2074.

Triesch, J. and von der Malsburg, C. 2001. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1449–1453.

Toyama, K. and Horvitz, E. 2000. Bayesian modality fusion: probabilistic integration of multiple vision algorithms for head tracking. In *Fourth Asian Conference on Computer Vision*.

Velisavljevic, L. and Elder, J.H. 2002. What do we see in a glance? [abstract]. *Journal of Vision*, 2(7):493.

Velisavljevic, L. and Elder, J.H. 2003. Eccentricity effects in the rapid visual encoding of natural images [abstract], *Journal of Vision*, 3(9):647a.

Viola, P. and Jones, M.J. 2001. Rapid object detection using a boosted cascade of simple features, In *Proc. CVPR*, 1:511–518.

Viola, P., Jones, M.J. and Snow, D. 2003. Detecting pedestrians using patterns of motion and appearance. In *Proc. ICCV*, 2:734–741.

Xiong, Q. and Jaynes, C.O. 2003. Mugshot database acquisition in video surveillance networks using incremental auto-clustering quality measures. In *Proc. AVSS*, Los Alamos, CA, IEEE, Computer Society, 191–198.

Zhao, T. and Nevatia, R. 2004. Tracking multiple humans in complex situations. *IEEE PAMI*, 26(9):1208–1221.