Statistical Modeling and Performance Characterization of a Real-Time Dual Camera Surveillance System

Michael Greiffenhagen*, Visvanathan Ramesh*, Dorin Comaniciu*, Heinrich Niemann‡

*Imaging & Visualization Department
Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540, USA
{michael,rameshv,comanici}@scr.siemens.com

‡Lehrstuhl für Mustererkennung
Universität Erlangen-Nürnberg
Martensstr. 3
91058 Erlangen, Germany
niemann@informatik.uni-erlangen.de

Abstract

The engineering of computer vision systems that meet application specific computational and accuracy requirements is crucial to the deployment of real-life computer vision systems. This paper illustrates how past work on a systematic engineering methodology for vision systems performance characterization can be used to develop a real-time people detection and zooming system to meet given application requirements. We illustrate that by judiciously choosing the system modules and performing a careful analysis of the influence of various tuning parameters on the system it is possible to: perform proper statistical inference, automatically set control parameters and quantify limits of a dual-camera real-time video surveillance system. The goal of the system is to continuously provide a high resolution zoomed-in image of a persons head at any location of the monitored area. An omni-directional camera video is processed to detect people and to precisely control a high resolution foveal camera, which has pan, tilt and zoom capabilities. The pan and tilt parameters of the foveal camera and its uncertainties are shown to be functions of the underlying geometry, lighting conditions, background color/contrast, relative position of the person with respect to both cameras as well as sensor noise and calibration errors. The uncertainty in the estimates is used to adaptively estimate the zoom parameter that guarantees with a user specified probability, α, that the detected person's face is contained and zoomed within the image [1].

1 Introduction

Rapid improvement in computing power, cheap sensing and more flexible algorithms are facilitating increased development of real-time video surveillance and monitoring systems [6]. The deployment of video understanding systems in certain critical applications in the real-world can be done only if performance guarantees can be provided for these systems. This paper illustrates the use of systematic engineering methodology outlined in [13] to design and validate a real-time system with given computational and accuracy constraints. We show that by judicious choice of the intermediate transforms (components of the system) along with a careful analysis of the influence of various parameters in the system, it is possible to perform proper statistical inference, automatically set the control parameters and quantify the limits of a dual-camera real-time video surveillance system.

The following section discusses a review of methodologies for analysis and synthesis of vision systems and outline our approach. Subsequent sections describe surveillance system example, statistical modeling and performance analysis, validation, and experimental results.

2 Statistical Methodologies for Vision System Design

Past works have addressed methodological issues and have demonstrated performance analysis of components and systems ([5], [4], [8], [16]). However, it is still an art to engineer systems that meet a given application requirement in terms of computational speed as well as accuracy. The trend in the community is to emphasize statistical learning methods, more appropriately Bayesian methods for solving computer vision problems (See for example [9]). However, there still exists the problem of choice of the right statistical likelihood model and right priors that suit an application. Even if this were possible, it is still computationally infeasible to satisfy real-time application needs.

Sequential decomposition of the total task into manageable sub-tasks (with reasonable computational complexity) and the introduction of pruning thresholds is the common way to tackle the problem. This introduces problems because of the difficulty in approximating the probability distributions of observables at the final step of the system so that Bayesian inference is plausible. This approach to perceptual Bayesian inference has been attempted, (see for example [13], [7]). [13]'s work places more emphasis on performance characterization of a system, while [7] attempted Bayesian inference (using Bayesian networks) for visual recognition. The idea of gradual pruning of candidate hypotheses to tame the computational complexity of the estimation/classification problem has been presented in [1]. Note that none of the works identify how the sub-tasks (e.g. feature extraction steps) can be chosen automatically given an application context. There has been prior

[1] Note, the higher the probability α the more conservative the zoom factor would be. We set α to 0.99 in our current system.

work to identify system or module configurations that best perform a given task using contextual models of algorithms.

Our approach involves the following key steps (based on [13]):

System Configuration choice: The first step is to choose the modules for the system. This is done by use of context (in other words: application specific prior distributions for object geometry, camera geometry and error models, illumination models). Real-time constraints are imposed by choosing pruning methods or indexing functions that restrict the search space for hypotheses. The choice of the pruning functions are derived from the application context and prior knowledge. The choice of the indexing function is not necessarily critical, except that the only criterion that needs to be used is that it be of the form that simplifies computation of the probability of false hypothesis or the probability of missing a true hypotheses as a function of the tuning constants.

Statistical modeling and Performance Characterization: The second step involves the derivation of statistical models for errors at various stages in the chosen vision system configuration, so that one can quantify the indexing step and to tune the parameters to achieve a given probability of miss-detection and false alarm rate. In addition, we perform a validation of theoretical models for correctness (through Monte-Carlo simulations) and closeness to reality (through real experiments). For more details on the methodology see [13].

Hypotheses verification and parameter estimation: Bayesian estimation is used to evaluate candidate hypotheses and estimate object parameters by using a likelihood model, $P(measurements|hypothesis)$, that takes into account the effects of the pre-processing steps and tuning parameters. In addition, the uncertainty of the estimate is derived in order to predict system performance.

The rest of the paper is organized as follows. Section 3 describes an overview of the surveillance system concept, the mathematics governing the geometry of the people detection and zooming problem, system modules and statistical characterization. Section 4, 5, and 6 describe the validation of the models. Real experiments point out the validity of the models in indoor settings. We conclude in section 7.

3 System Description

The task of the two camera surveillance system is to continuously provide zoomed-in high resolution images of the face of a person present in the room. These images represent the input to higher-level vision modules, e.g. face recognition, compaction and event-logging (not discussed in this paper).

3.1 Application Requiremen ts:

The application requirements are as follows: 1) real-time performance on a low-cost PC [2], 2) person miss-detection rate of x_m, 3) person false-alarm rate of x_f, 4) adaptive zooming of person irrespective of background scene structure (with maximal possible zoom based on uncertainty of person attributes estimated (e.g. location in 3D, height, etc), with performance of the result characterized by face resolution attainable in area of face pixel region (as a function of distance, contrast between background and object, and sensor noise variance and resolution) and bias in the centering of the face. In addition to these requirements, the following assumptions can be made about scene structure: (e.g) the scene illuminant consists of light sources with similar spectrum (e.g. identical light sources in an office area), the number of people to the detected and tracked is bounded, the probability of occlusion of persons (due to other persons) is small.

[2]Not all system resources in the PC are allocated for visual processing.

3.2 System Hardware and Software Configuration:

To continuously monitor the entire scene we use an omnidirectional sensor (OmniCam [11]) mounted below the ceiling. Omni-images are used to detect and estimate the precise location of a given person's foot in the room and this information is used to identify the pan, tilt and zoom settings for a high-resolution foveal camera. Figure 1 shows the system's interface: the overview image, and the high resolution zoomed image of the detected person's face.

Figure 1: Top: omnidirectional overview image. Red sector: region of interest. Radial lines (green and red) show detected persons. Crosses denote estimated foot/head position. Insert: foveal camera view.

Before we describe the details of how the application requirements translate to the design of individual modules, we discuss the prior distributions (of the 3D scene) reasonable for the given application and identify how these priors induce image priors. The choice of the various estimation steps in the system are motivated from these image priors and real-time requirements. The camera control parameters (pan and tilt) are selected based on the location estimate and its uncertainty (that is derived from statistical analysis of the estimation steps) so as to center the person's location in the image. The zoom parameter is set to maximum value possible so that the camera view still encloses the persons head within the image.

3.2.1 Priors Camera models, Illumination models

The general Bayesian formulation of the person detection and location estimation problem does not suit the real-time constraints imposed by the application. Our approach is to use this formulation only after a pruning step that rules out a majority of false alarms by

designing an indexing step motivated by the 2D image priors (region size, shape, intensity characteristics) induced by the prior distribution in the 3D scene. The prior distributions for person shape parameters: size, height, and his/her 3D location are reasonably simple. These priors on the person model parameters induce 2D spatially variant prior distributions in the projections (e.g. the region parameters for a given person in the image depends on the position in the image) whose form depends on the camera projection model and the 3D object shape.[3]. In addition to shape priors, the image intensity/color priors are of importance in our application. Typically we do not assume anything about the object intensity (e.g. homogeneity of object since people can wear variety of clothing and the color spectrum of the light source is not necessarily constrained). However, in the surveillance application, the background is typically assumed to be a static scene (or a slowly time varying scene) with known background statistics (Gaussian mixtures are typically used to approximate these densities). To handle shadowing and illumination changes, these distributions are computed after the calculation of an illumination invariant measure from a local region in an image. The prior distribution of the spectral components of the illuminants in our application are assumed to have same but unknown spectral distribution. Finally, the noise model for the CCD sensor noise is to be specified. This is typically chosen to be i.i.d. zero mean Gaussian noise in each color band.

3.2.2 System Software Configuration

The software is composed of four functional modules: calibration, illumination-invariant measure computation at each pixel, indexing functions to select sectors of interest for hypothesis generation, statistical estimation of person parameters (e.g. foot location estimation), and foveal camera control parameter estimation. Figure 2 illustrates the step by step transformations applied to the input. The input color image, $\hat{R}(x,y), \hat{G}(x,y), \hat{B}(x,y)$, is transformed ($T : R^3 \rightarrow R^2$) typically to compute an illumination invariant measure $\hat{r_c}(x,y), \hat{g_c}(x,y)$. The statistical model for the distribution of the invariant measure is influenced by the sensor noise model and the transformation $T(.)$. The invariant measure mean ($B_o(x,y) = (r_b(x,y), g_b(x,y))$) and covariance matrix $\Sigma_{B_o}(x,y)$, is computed at each pixel (x,y) from several samples of $R(x,y), G(x,y), B(x,y)$ for the reference image of the static scene. A change detection measure $\hat{d^2}(x,y)$ image is obtained by computing the Mahalanobis distance between the current image data values $\hat{r_c}(x,y), \hat{g_c}(x,y)$ and the reference image data $B_o(x,y)$. This distance image is used as input to two indexing functions $\psi_1()$ and $\psi_2()$. $\psi_1()$ discards the radial lines θ by choosing hysteresis thresholding parameters that satisfy a given combination of probability of false alarm and miss-detection values, while $\psi_2()$ discards segments along the radial lines in the same manner. The result is a set of regions with high probability of significant change. At this point we employ our full blown statistical estimation technique that uses the 3D model information, camera geometry information, priors on objects, shape, and 3D location to estimate the number of objects and their positions. The last step is to estimate the control parameters for the foveal camera based on the location estimates and uncertainties.

The following sub-sections describe the system, the statistical models, rationale for indexing steps chosen based on application priors, foot position estimation scheme and control parameter selection scheme. First, we outline the omni-directional/pan-tilt camera geometric relations.

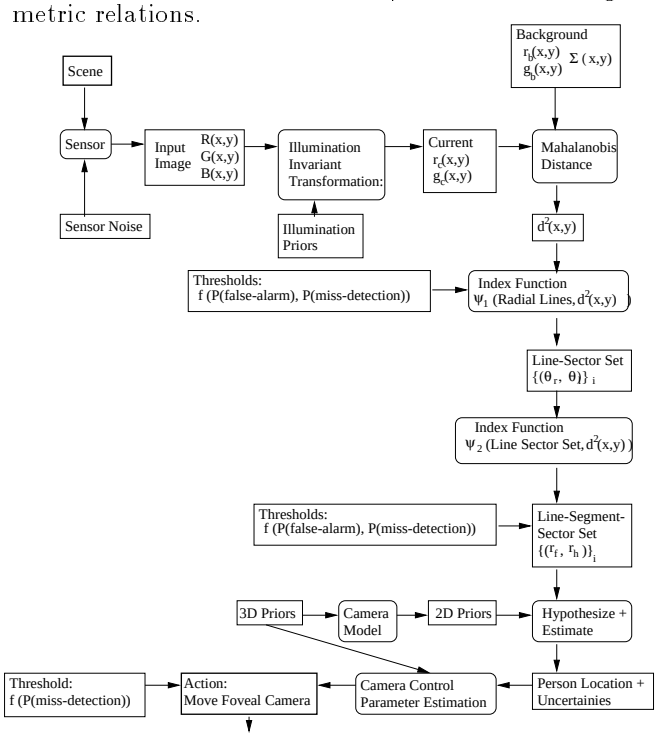

Figure 2: Block diagram: boxes with rounded corners represent transformations while boxes represent data objects.

3.2.3 Calibration and Geometry

The projection model for the two cameras are discussed in this section. We denote the geometric model parameters as follows (see Figure 3, 4):

- H_o height of OmniCam above floor (inches)
- H_f height of foveal camera above floor (inches)
- H_p person's height (inches)
- R_h person's head radius (inches)
- R_f person's foot position in world coordinates (inches)
- D_c on floor projected distance between cameras (inches)
- $p(x_c, y_c)$ position of OmniCam center, (in omni-image) (pixel coordinates)
- r_m radius of parabolic mirror (in omni-image) (pixels)
- r_h distance person's head – (in omni-image) (pixels)
- r_f distance person's foot – (in omni-image) (pixels)
- ϑ - angle between the person and the foveal camera relative to the OmniCam image center (Please see figure 4).

[3]For this application we found that modeling the person as an upright cylinder is a reasonable approximation.

- θ - angle between the radial line corresponding to the person and the zero reference line (please see figure 4).

where capital variables are variables in 3D, and small variables are given in image coordinates. During the calibration step (combination of real world and image measurements) H_o, H_f, D_c, r_m and $p(x_c, y_c)$ are initialized.

Using the geometric features of OmniCam (see [11]), and under the hypothesis that the person is standing upright, the relationship between r_f respectively r_h and R_p can be shown to be:

$$R_p = aH_o \quad \text{with} \quad a = 2\frac{r_m r_f}{r_m^2 - r_f^2} \tag{1}$$

$$R_p = b(H_o - H_p) \quad \text{with} \quad b = 2\frac{r_m r_h}{r_m^2 - r_h^2} \tag{2}$$

Let α, and β be the foveal camera control parameters

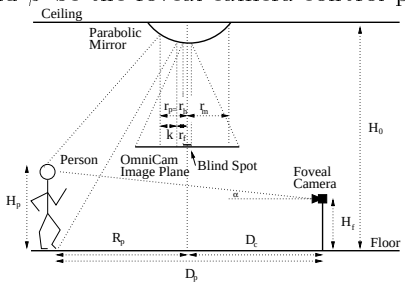

Figure 3: Geometry (viewed from the side) in real world and omni-image coordinates. OmniCam is looking into the parabolic mirror at the ceiling.

for the tilt respectively pan angle, and D_p the projected real world distance between the foveal camera and the person. Assuming, the person's head is approximately located over his/her feet, and using basic trigonometry in Figure 3 and 4, it can easily be seen that D_p, α, and β are equal to:

$$D_p = \sqrt{D_c^2 + R_p^2 - 2D_c R_p \cos(\vartheta)} \tag{3}$$

$$\tan(\alpha) = \frac{H_p - R_h - H_f}{D_p}; \qquad \sin(\beta) = \frac{R_p}{D_p}\sin(\vartheta) \tag{4}$$

where ϑ is the angle between the person and the foveal camera relative to the OmniCam position.

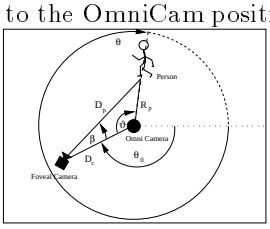

Figure 4: 3D geometry, viewed from atop.

3.2.4 Illumination Invariant Measure Estimation

This step is the module that takes in as input the current color image $(\hat{R}(x,y), \hat{G}(x,y), \hat{B}(x,y))$, normalizes it to obtain $(\hat{r_c}(x,y), \hat{g_c}(x,y))$ and compares it with the background statistical model $(B_o(x,y), \Sigma_{B_o}(x,y))$

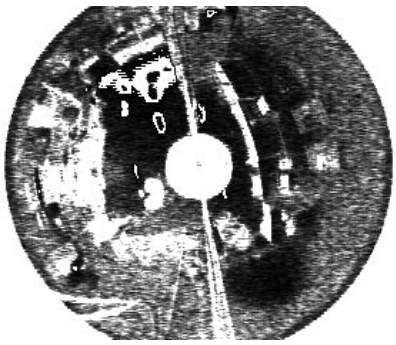

Figure 5: Covariance image for $\sigma_{\hat{r},\hat{r}}$. Bright regions in the normalized space denote high variance; these regions correspond to dark areas in the RGB image.

to produce an illumination invariant measure image $\hat{d}^2(x,y)$. This section illustrates the derivation of the distribution of $\hat{d}^2(x,y)$ given that the input image measurements $\hat{R}, \hat{G}$ and $\hat{B}$ are Gaussian with mean R, G, B, and identical standard deviation σ.

As indicated in section 3.2.1 the illumination prior assumption is that the scene contains multiple light sources with the same spectral distribution with no constraint on individual intensities. To compensate for shadows which are often present in the image we employ a shadow invariant representation of the color data ([17], pp.347). Let $S = R + G + B$. The illumination normalizing transform $T : R^3 \to R^2$ appropriate to our assumption is: $r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}$ It can be shown that, the uncertainties in the normalized estimates $\hat{r}$ and $\hat{g}$ are dependent not only on sensor noise variance, but also on the actual true unknown values of the underlying samples (due to the non-linearities in the transformation $T(.)$). Based on the assumption of a moderate signal to noise ratio (i.e. $\sigma \ll S$), we approximate $(\hat{r}, \hat{g})^T$ as having a normal distributed with pixel-dependent covariance matrix

$$\begin{pmatrix} \hat{r} \\ \hat{g} \end{pmatrix} \sim N\left(\begin{pmatrix} r \\ g \end{pmatrix}, \Sigma_{\hat{r},\hat{g}}\right)$$

$$\text{with} \quad \Sigma_{\hat{r},\hat{g}} = \frac{\sigma^2}{S^2}\begin{pmatrix} 1 - \frac{2R}{S} + 3\frac{R^2}{S^2} & -\frac{R+G}{S} + 3\frac{RG}{S^2} \\ -\frac{R+G}{S} + 3\frac{RG}{S^2} & 1 - \frac{2G}{S} + 3\frac{G^2}{S^2} \end{pmatrix} \tag{5}$$

The values of $\sigma^2_{\hat{r},\hat{r}}$ are shown in Figure 5 for an entire OmniCam frame. Note, that in the normalized space the covariance matrix for each pixel is different: Bright regions in the covariance image correspond to regions with high variance in the normalized image. These regions correspond to dark regions in RGB space.

Since the covariance matrices in the normalized space are pixel-dependent we calculate the test statistic, i.e. the Mahalanobis distance d^2, that provides a normalized distance measure of a current pixel being background. Let $\hat{\mu}_\mathbf{b}$ be the vector of mean r_b, and mean g_b at a certain background position (mean b_b is redundant, due to normalization), and $\hat{\mu}_\mathbf{c}$ be the corresponding vector of the current image pixel. Since

$$\begin{pmatrix} \hat{r}_c - \hat{r}_b \\ \hat{g}_c - \hat{g}_b \end{pmatrix} \sim N\left(\begin{pmatrix} r_c - r_b \\ g_c - g_b \end{pmatrix}, \Sigma_{\hat{r}_c,\hat{g}_c} + \Sigma_{\hat{r}_b,\hat{g}_b}\right) \tag{6}$$

we can define for each pixel a metric d^2 which corresponds to the probability, that $\hat{\mu}_\mathbf{c}$ is background pixel:

$$\hat{d}^2 = (\hat{\mu}_{\mathbf{b}} - \hat{\mu}_{\mathbf{c}})^T (2\Sigma_{\hat{r}_b,\hat{g}_b})^{-1} (\hat{\mu}_{\mathbf{b}} - \hat{\mu}_{\mathbf{c}}) \quad (7)$$

For background pixels, $\hat{d}^2$ is approximately χ^2 distributed with two degrees of freedom. For object pixels $\hat{d}^2$ happens to be non-central χ^2 distributed with two degrees of freedom, and non-centrality parameter c.

3.2.5 Indexing for Hypothesis generation

To address real-time computational requiremen ts of the application it is crucial to identify sectored segments in the image that potentially contains people of interest. To perform this indexing step in a computational efficient manner we define two index functions $\psi_1()$ and $\psi_2()$ that are applied sequentially as shown in the system block diagram (2). Essentially $\psi_1()$ and $\psi_2()$ are projection operations. For instance, define $\hat{d}^2(R,\theta)$ as the change detection measure image in polar coordinates with coordinate system origin at the omni-im age center $p(x_c, y_c)$. Then, $\psi_1()$ is chosen to be the projection along radial lines to obtain $\hat{M}_\theta$, the test statistic that can be used to identify changes along a given direction θ. This test statistic is justified by the fact that the object projection is approximated by a line-set (approximated as an ellipse) whose major axis passes through the omni-im age center with a given length distribution that is a function of the radial foot position coordinates of the person in the omni-image. This section derives the expressions for the probabilities of false alarm and misdetection at this step as a function of the input distributions for $\hat{d}^2(R,\theta)$, the prior distribution for the expected fraction of the pixels along a given radial line belonging to the object, and the noncentrality parameter of $\hat{d}^2(R,\theta)$ in object locations.

Let $L_\theta^{x_c,y_c}$ be a radial line trough $p(x_c, y_c)$, parameterized by angle θ, and $\hat{M}(\theta) = \sum_r d_\theta^2(r)$ denote the accumulative measure of d^2 values at image position $p(\theta, r)$ parameterized by angle θ and distance r in a polar coordinate system at $p(x_c, y_c)$. Applying Canny's hysteresis thresholding technique ([13]) on $\hat{M}(\theta)$, provides the sectors of significant change bounded by left and right angles θ_l respectively θ_r. Let r_m be the total number of pixels along a radial line $L_\theta^{x_c,y_c}$, and k be the expected number of object pixels along this line (The distribution of k can be derived from the projection model and the 3D prior models for person height, size, and position described previously, see [15]). The distribution of the cumulative measure is:

$$\text{Background} \qquad M_\theta \sim \chi^2_{2r_m}(0) \quad (8)$$

$$\text{Object} \quad M_\theta \sim (r_m - k)\chi^2_{2(r_m-k)}(0) + k\chi^2_{2k}(c) \quad (9)$$

with $c \in [0 \ldots \text{inf})$.

To guarantee a false-alarm rate for false sectors of equal or less than $x_f\%$ we can set the lower threshold T_l so that

$$\int_0^{T_l} \chi^2_{\hat{M}_\theta}(\xi) d\xi = 1 - x_f\% \quad (10)$$

To guarantee a misdetection rate of equal or less than $x_m\%$, theoretically, we can solve for an upper threshold T_u similarly by evaluating the distribution in eqn. (9). Note that k is a function of H_p, R_f, and c (see [15]).

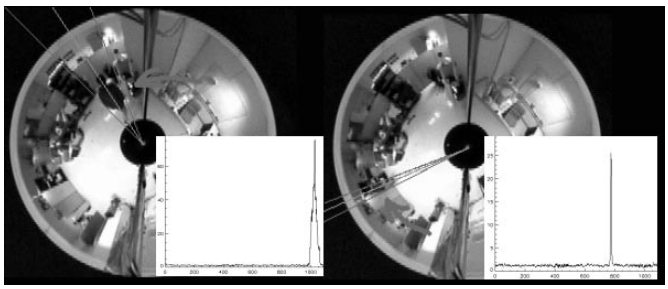

Figure 6: Area of significant change (Left and right lines correspond to θ_l and θ_r; Center line denotes the angular position $\hat{\theta}_f$.) Inserts show corresponding radial profile M_θ.

Therefore we would need to know the distributions of H_p, R_f, and c to solve for T_u. Unfortunately, we cannot make any assumptions about the distribution of non-central parameter c, so we have to resort to the use of a LUT $T_u(x_m)$ generated by simulations instead.

The second index function $\psi_2()$ essentially takes as input the domain corresponding to the radial lines of interest and performs an pruning along the radial lines R. This is done by the computation of $\hat{d}^2_{\theta_f}(r)$ the integration of the values $\hat{d}^2()$ along $\theta_f = \theta + \pi/2$ (within a finite window whose size is determined by the prior density of the minor axis of the ellipse projection), for each point r on the radial line θ. The derivation of the distribution of the test statistic and the choice of the thresholds are exactly similar to the above step.

3.2.6 Hypothesis and Estimation Step

We have derived the distributions of the $\hat{d}^2$ image measuremen ts, and have narrowed our hypotheses for people location and attributes. The next step is to perform the Bayes estimation of person locations and attributes. This step uses the likelihood models $L(\hat{d}^2|background)$ and $L(\hat{d}^2|object)$ along with 2D prior models for person attributes induced by 3D object priors $P(R_p), P(H), P(\theta)$ and $P(S)$. In our current application we make use of the fact that the probability of occlusion by persons is small to assert that the probability of a sector containing multiple people is rather small.[4] The center angle θ_f of a given sector would in this instance give us the estimate of the major axis of the ellipse corresponding to the person. It is then sufficient to estimate the foot location of person along the radial line corresponding to θ_f. The center angle θ_f of the sector defines the estimate for the **angular component** of the foot position, see Figure 6. We approximate $\hat{\theta}_f$ to be normal distributed with unknown θ_f and variance σ_{θ_f}. θ_f's are estimated as the center positions of the angular sectors given by $\psi_1()$. The standard deviation of a given estimate is currently obtained by assuming that the width of the angular sector gives the 99 percentile confidence interval.[5]

Given the line θ_f it is necessary to estimate the foot position of the person along this radial line. To find

[4] Estimation when occlusions occur in the OmniImage within a sector is a subject of future research.

[5] We observe that this estimate is an upper bound of the true standard deviation and our experiments have indicated that true variance is a small quantity for a wide range of person positions and background to person contrasts.

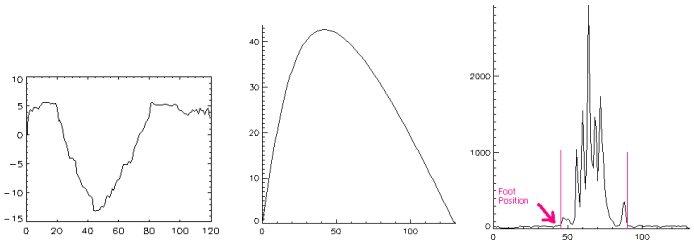

Figure 7: Top left: Bayes error as function of hypothesized foot position r_f', here: most probable foot position at position $r_f = 47$. Bottom left: Projected person length k as function of r_f. Note: $k(r_f' = 47) = 43$. Right: Profile $\bar{d}^2_{\theta_f}(r)$: by minimizing Bayes error, responses in interval [43...90] are classified as object responses.

this estimate and variance of the **radial foot position** r_f we choose the best hypothesis for the foot position that minimizes the Bayes error.[6] Let $P(h_i|m)$ denote the posterior probability to be maximized, where h_i denotes the ith out of multiple foot position hypotheses and m the measurements $(\bar{d}^2_{\theta_f}(r))$, that are statistically independent; hyper-script b or o denotes *background* respectively *object*:

$$\begin{aligned} & P(h_i|m) \\ &= P(h_i^b|m^b)P(h_i^o|m^o) = P(h_i^b|m^b)\left(1 - P(\bar{h_i^o}|m^o)\right) \\ &= \frac{p(m^b|h_i^b)P(h_i^b)}{p(m^b)} \frac{p(m^o) - p(m^o|\bar{h_i^o})P(\bar{h_i^o})}{p(m^o)} \end{aligned} \quad (11)$$

where p denotes the density function. $P(h_i|m)$ becomes maximal for maximal $p(m^b|h_i^b)$ and minimal $p(m^o|\bar{h_i^o})$, so that

$$r_f = \underset{r_f'}{\text{argmax}} \log\left(\frac{p(m^b|h_i^b)}{p(m^o|\bar{h_i^o})}\right) = \quad (12)$$

$$\underset{r_f'}{\text{argmax}} \left(\sum_{r=0}^{r_f'-1} \bar{d}^2_{\theta_f}(r) + \sum_{r=r_h(r_f')}^{r_m} \bar{d}^2_{\theta_f}(r) - \sum_{r=r_f}^{r_h(r_f')-1} \bar{d}^2_{\theta_f}(r) \right)$$

Finally, we estimate the uncertainty in the foot position r_f. We do not have a close form for that yet, though, our approach provides us with the pdf's up to the latest step in the algorithm. At this point it is affordable to simulate the distribution of r_f and generate $\sigma^2_{\hat{r}_f}$ via perturbation analysis, since only few estimates with known distributions are involved in few operations. Experiments show, that one can approximate $\hat{r}_f$ as Gaussian distributed with unknown mean r_f, and variance $\sigma^2_{\hat{r}_f}$.

[6] The prior distribution of person heights and sizes are assumed to be Gaussian. While we actually need to estimate the person height as well as width on the projection using the Bayesian formulation we use the assumption that the variance of the height and size is small and just fix H_p and S_p as constants. The geometric transformations are still taken into account to identify 2D projection lengths as a function of radial position along the radial line.

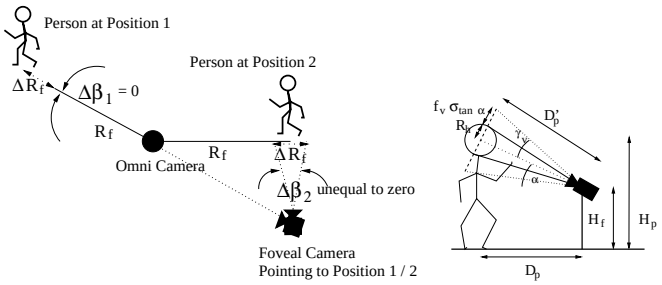

Figure 8: Left: local dependency - same uncertainty in R_f, different $\Delta\beta$. Right: geometric relations for vertical angle of view calculation (view from the side).

3.2.7 Foveal Camera Control Parameter Estimation

Once the foot position $p(\theta_f, r_f)$ is known, we can apply formula 1– 4 to estimate 3D distances R_p, D_p, and foveal camera control parameter tilt α, pan β and zoom factor z.

Figure 8 (left) illustrates how uncertainties in 3D radial distance R_p influence the foveal camera control parameters.

We have seen that the foot position estimate error can be approximated as a zero mean Gaussian random variate. For the following error propagation steps we will assume that $\hat{r}_m$, $\hat{r}_p$, $\hat{H}_o$, $\hat{H}_p$, $\hat{H}_f$, and $\hat{D}_c$ are Gaussian random variables with true unknown means r_m, r_p, H_o, H_p, R_h, H_f, and D_c, and variances $\sigma^2_{\hat{r}_m}, \sigma^2_{\hat{r}_p}, \sigma^2_{\hat{H}_o}, \sigma^2_{\hat{H}_p}, \sigma^2_{\hat{R}_h}, \sigma^2_{\hat{H}_f}$, and $\sigma^2_{\hat{D}_c}$ respectively (all estimated in the calibration phase). By applying linearization in the geometric transformations, and making independence assumptions on variables where applicable (see sections 4, 5) it is easy to show (see [15]) how the estimates and its uncertainties propagate through the geometric transformations outlined in section 3. For limited space reason we only print the final results for the uncertainties in tilt α, and pan β, which were used to calculate the zoom parameter z in section 3.2.7. (for more details, and derivations of $\sigma^2_{\hat{R}_p}, \sigma^2_{\hat{D}_p}$ see [15]):

$$\begin{aligned} \sigma^2_{\tan\hat{\alpha}} &= \frac{\sigma^2_{\hat{D}_p}}{D_p^4}\left((H_p - R_h - H_f)^2 + \sigma^2_{\hat{H}_p} + \sigma^2_{\hat{R}_h} + \sigma^2_{\hat{H}_f}\right) \\ &\quad + \frac{\sigma^2_{\hat{H}_p} + \sigma^2_{\hat{R}_h} + \sigma^2_{\hat{H}_f}}{D_p^2} \end{aligned} \quad (13)$$

$$\begin{aligned} \sigma^2_{\sin\hat{\beta}} &= \frac{{R_p}^2 {\sigma_{\hat{\vartheta}}}^2 \cos^2\vartheta}{{D_p}^2} + \left(\sin^2\vartheta + {\sigma_{\hat{\vartheta}}}^2\cos^2\vartheta\right) * \\ &\quad * \left(\frac{{R_p}^2{\sigma_{\hat{D}_p}}^2}{{D_p}^4} + \frac{{\sigma_{\hat{R}_p}}^2}{{D_p}^2} + \frac{{\sigma_{\hat{R}_p}}^2{\sigma_{\hat{D}_p}}^2}{{D_p}^4}\right) \end{aligned} \quad (14)$$

Given the uncertainties in the estimates, we can derive the horizontal and vertical angle of view for the foveal camera, γ_h respectively γ_v, which map directly to the zoom parameter z. Figure 8 (right) shows the geometric relationships for the vertical case. Following

equation provides the vertical angle of view.

$$\gamma_v = 2 \text{ atan} \left(\frac{\hat{R}_h + f_v \sigma_{\tan\hat{\alpha}} \hat{D}'_p}{\sqrt{\hat{R}_h^2 + \hat{D}_p'^2}} \right) \text{ with } \quad \hat{D}'_p = \frac{\hat{D}_p}{\cos\alpha} \quad (15)$$

where factor f_v solves for $\int_0^{\frac{f_v}{2}} N(0,1) d\xi = \frac{x_z}{2}\%$ given user specified confidence percentile x_z that the head is display in the foveal frame. Similar derivations apply for the horizontal case.

4 Validation of Assumptions

We verified the correctness of our theoretical expressions and approximations through extensive simulations. Due to lack of space, we only show plots validating expressions for illumination normalization (eqn. 5,5, Figure 9), and for foveal camera control parameters (eqn. 13, 14, Figure 10). This validation assumes correctness of the underlying statistical models. Validation of the models on real data is done in section 5. In the following we include plots showing theoretically predicted answers and the differences between these predictions and simulated results, based on 10000 samples of normal distributed parameters. For demonstration purpose, parameters that represent the worst-case system behavior were chosen (see [15] for details). The figures show results obtained by using the following settings for the standard deviations: $\sigma_{\hat{H}_o} = 1cm$, $\sigma_{\hat{H}_p} = 5cm$, $\sigma_{\hat{H}_f} = 1cm$, $\sigma_{\vartheta} = 1°$, $\sigma_{D_c} = 10cm$ $\sigma = 1$ graylevel. For validation of the distribution of the normalized color values, we used a value for $B = 50$ while varying R and G values in the range 0 through 255 (see Figure 9).

In reality uncertainties are calculated on-line from the current data and are functions of the object, background and location of the object as well as the sensor noise.

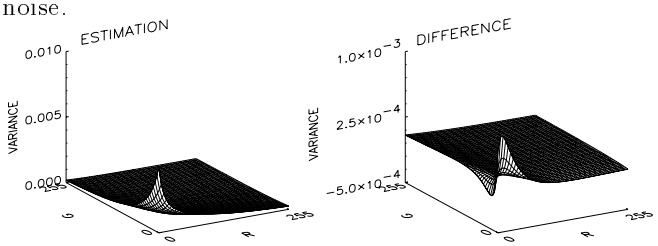

Figure 9: Color normalization: Variance $\sigma^2_{\hat{r}} + \sigma^2_{\hat{g}} + \sigma^2_{\hat{r}\hat{g}}$. Theoretical values (left) and difference between simulation and theory (right).

Plots show the correctness of the derivations and approximations, give insights of the system limitations depending on user defined tolerances, and show, where the assumptions hold. By examining parametric expressions for uncertainties (eqn. (13),(14), see also [15]) the differences between simulation, and derived prediction can be explained by the linearization error, where the assumption of low signal to noise ratio breaks.

5 Experimental Validation of Models

The correctness of our models is verified by comparing ground truth values against module estimates for mean and variance of the running system. First, we marked eight positions $P1 - P8$ of different radial distances and pan angles (Figure 11 upper left image).

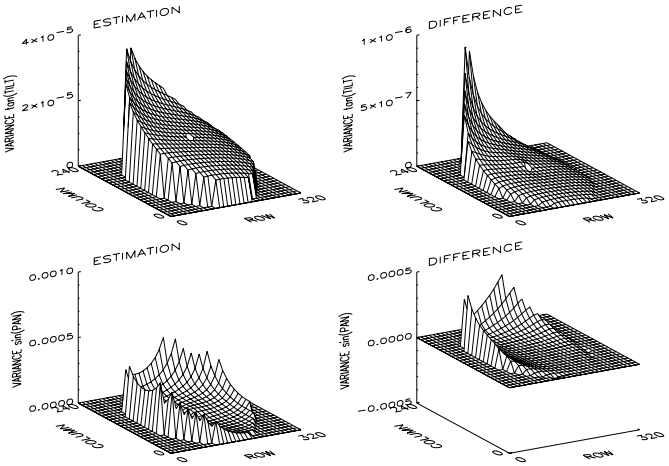

Figure 10: Variances of $\sin(\hat{\beta})$ and $\tan(\hat{\alpha})$ plotted as a function of person foot position in omni-image coordinates. Left: Theoretical. Right: Difference between simulation and theory.

Positions, and test persons were chosen to simulate different positions, illumination, and contrast. For limited space reasons we only show the table for the final foveal camera control parameters for one person. Ground truth values for the mean values were taken by measuring tilt angle α, and pan angle β by hand, and are compared against the corresponding mean of system measurements estimated from 100 trials per position and person. The variances calculated from the system estimates for pan and tilt angle are compared against the average of the corresponding variance-estimates calculated based on the analysis. The comparison between system output and ground truth demonstrates the correctness of the model assumptions in the statistical modeling process (see Table 1).

Table 1: Validation. First two lines show the predicted and experimental variances for the tilt angle, respectively. Next two lines correspond to pan angle.

$\times 10^{-5}$	P1	P2	P3	P4	P5	P6	P7	P8
$\hat{\sigma}^2_{\tan\hat{\alpha}}$	2.10	2.12	1.57	1.40	1.35	1.31	1.31	1.32
$\tilde{\sigma}^2_{\tan\hat{\alpha}}$	2.05	2.04	1.60	1.34	1.36	1.32	1.40	1.31
$\hat{\sigma}^2_{\sin\hat{\beta}}$	28.9	26.1	21.3	17.9	15.3	15.2	18.4	20.1
$\tilde{\sigma}^2_{\sin\hat{\beta}}$	25.9	24.1	19.5	15.1	14.9	15.0	18.1	19.3

6 System Performance

In this section we demonstrate the performance of the running system. Figure 11 demonstrates how the system can precisely track[7] a person's face and zoom, while guaranteeing that the face is in the frame. The output of the foveal camera proved sufficient as input for face recognition algorithms (not discussed in this paper.)

We now illustrate, how the statistical analysis is used to optimize the camera setup. The formulas 13, and 14 suggest that the configuration that minimizes these uncertainties is the one with large inter-camera distance

[7] The tracker uses a simple nearest neighbor prediction.

Figure 11: Top left: omni-image; tracked path, positions at which snapshots were taken. Left right, top down: corresponding foveal image.

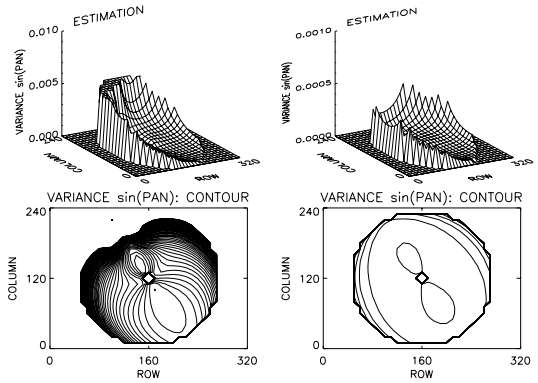

Figure 12: Influence of camera positioning on global and local performance. Top: variance $\sigma_{\sin\hat{\beta}}$. Bottom: corresponding contours. Left: close distance D_c between foveal camera and OmniCam. Right: larger distance, better performance.

D_c and foveal camera height H_f equal to the mean person eye-level height H_p. Figure 12 illustrates a comparison of the uncertainties in the pan angle for this setup (right plots), versus a camera position setup with lower distance D_c (left plots). Similar results are obtained for the tilt angle (omitted for limited space reasons).

The system proved reliable in terms of detection and zooming over longtime experiments within the operational limits denoted by the outer line of the upper right contour plot in Figure 11. Outside, the zooming became imprecise, and did not match the user defined tolerances in terms of zoom precision. This is anticipated, since the assumptions do not hold for these regions as one can see from the plots in section 4.

Figure 12 illustrates how the setup of the system (here placement of foveal camera) influences precision globally and locally. Note preferred directions of low uncertainties (top right to lower left in upper plots). This can be used to adapt the system to user defined accuracy constraints in certain areas of the room.

7 Conclusion and Future Work

This paper demonstrated how by careful statistical modeling it is possible to develop and quantify a system to perform a visual surveillance task. The essence of the message is that by careful decomposition of the global task into sub-pieces, statistical characterization of the system, and incorporation of application specific priors in various stages of the system, it is possible to build computationally efficient, but yet statistically well motivated systems. To present this essence it was necessary to make certain simplifying priors and illustrate a working system in a constrained environment[8]. Extensive amount of real and synthetic data experiments were used to validate the models derived. Although we mainly discussed person detection and location estimation alone, the actual video surveillance system has tracking algorithms implemented. We hope to address systematic characterization of the tracking algorithm in the next paper.

References

[1] Y. Amit and D. Geman, "A computational model for visual selection", Neural Computation, 1999.

[2] P.Allen and R. Bajcsy, "Two sensors are better than one: example of vision and touch", *Proceedings of 3rd International Symposium on Robotics Research*, pp. 48-55, Gouvieux, France, 1986.

[3] B. Efron, R. Tibishirani "An Introduction to the Bootstrap", Chapman & Hall, New York, 1993.

[4] Machine Vision & Applications, International Journal, Special Issue on Performance Evaluation; ed. by W. Forstner, Vol. 9, nos. 5/6, 1997.

[5] R. Haralick, "Overview: Computer Vision Performance Characterization", *Proceedings of the DARPA Image Understand Workshop*, Vol.1, pp.663-665, 1994.

[6] T. Kanade et al, "Advances in Cooperative Multi-Sensor Video Surveillance", *Proceedings of the DARPA Image Understanding Workshop*, Vol. 1, pp. 3-24, 1998.

[7] W. Mann and T. Binford, "Probabilities for Bayesian Networks in Vision,", Proceedings of the ARPA IU Workshop, 1994, Vol. 1, pp. 633-643.

[8] K. Cho, P. Meer, J. Cabrera: "Performance assessment through bootstrap". IEEE Trans. Pattern Anal. Machine Intell., 19, 1185-1198, 1997.

[9] D.Mumford, Pattern theory: a unifying perspective, in "Perception as Bayesian Inference", edited by D.Knill and W.Richards, Cambridge Univ. Press, 1996.

[10] S. Nayar and T. Boult, "Omnidirectional Vision systems: 1998 PI Report", *Proceedings of the DARPA Image Understanding Workshop*, Vol. 1, pp. 93-100, 1998.

[11] S. Nayar, "Omnidirectional Video Camera", *Proceedings of the DARPA Image Understanding Workshop*, Vol. 1, pp. 235-242, 1997.

[12] J. Nielsen "Characterization of Vision Algorithms: An experimental Approach", *ECVnet Workshop on Benchmarking*, 1995.

[13] V. Ramesh et al, "Computer Vision Performance Characterization," *RADIUS: Image Understanding for Imagery Intelligence*, edited by. O. Firschein and T. Strat, Morgan Kaufmann Publishers, San Francisco, 1997.

[14] D.Stoyan, W.S.Kendall, J.Mecke, *Stochastic Geometry and its Applications*, John Wiley and Sons, 1987.

[15] M.Greiffenhagen, V. Ramesh, "Auto-Camera-Man: Multi-Sensor Based Real-Time People Detection and Tracking System", Technical Report, Siemens Corporate Research, Princeton, NJ, USA, Nov. 1999.

[16] S. Wang and T. Binford, "Generic, Model-Based Estimation and Detection of Discontinuities in Image Surfaces,", Proceedings of the ARPA IU Workshop, 1994, Vol. 2, pp. 1443-1450.

[17] G.Wyszecki, W.S.Stiles "Color Science: Concepts and Methods, Quantitative Data and Formulae", John Wiley & Son, 1982.

[8] Although this is reasonable for the application chosen the challenge of accomplishing system characterization with more complicated priors remains.