Ground-Truth Uncertainty Model of Visual Depth Perception for Humanoid Robots

D. Gonzalez-Aguirre, M. Vollert, T. Asfour and R. Dillmann
Karlsruhe Institute of Technology, Adenauerring 2, Karlsruhe-Germany.
{*david.gonzalez, michael.vollert, asfour, dillmann*}@kit.edu

***Abstract*—The visual perception of a humanoid robot bridges the physical world with the internal world representation through visual skills such as self-localization, object recognition, detection, classification and tracking. Unfortunately, these skills are affected by internal and external sources of uncertainty. These uncertainties are present at various levels ranging from noisy signals and calibration deviations of the embodiment up to mathematical approximations and limited granularity of the perception-planning-action cycle. This aggregated uncertainty deteriorates and limits the precision and efficiency of the humanoid robot visual perception.**

In order to overcome these limitations, the depth perception uncertainty should be modeled in the skills of the humanoid robots. Due to the complexity of the aggregated uncertainty in humanoid systems, the visual depth uncertainty can be hardly modeled analytically. However, the uncertainty distribution can be conveniently attained by supervised learning. The role of the supervisor is to provide ground-truth spatial measurements corresponding to the humanoid uncertain visual depth perception. In this article[1], a supervised learning method for inferring a novel model of the visual depth uncertainty is presented. The acquisition of the model is autonomously attained by the humanoid robot ARMAR-IIIB, see Fig.1.

I. Introduction

Intelligent humanoid robots should perceive the world in order to recognize objects, plan actions and interact with the environment. On the one hand, this perception should *quantitatively* determine the essential measurable properties of the world such as size, location and orientation. On the other hand, perception should *qualitatively* solve complex tasks such as recognition, classification and interpretation.

Specifically, in the *visual perception for model-coupling*[2], the quantitative perception provides the essential cues (length and depth) for the complex perception task of object recognition with 6D-pose estimation. This visual quantitative perception relies on different cues depending on the sensing (monocular or stereo) approach. For humanoid robots the most coherent and natural approach is the stereo vision to acquire quantitative information. This capability to obtain euclidean metric from images allows humanoid robots to recognize [1], grasp [2], manipulate [3], act upon environmental elements [4], self-localize [5] and categorize objects [6], [7]. Taking into account the perceptual uncertainty is of crucial importance to robustly realize complex actions. The integration of a visual depth uncertainty model in the task planning and execution in humanoid robots enables better perceptual scalability, ambiguities or error handling and consist sensor fusion.

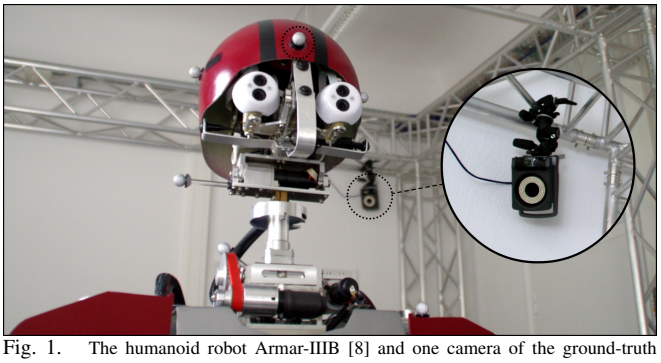

Fig. 1. The humanoid robot Armar-IIIB [8] and one camera of the ground-truth system [9] which provides the reference measurements for inferring the visual depth uncertainty model without reductionist assumptions.

The uncertainty analysis and modeling in stereo vision has been an active research field in the last three decades. Consequently, considerable results have been achieved and several models with different methodologies were proposed, see approaches from the pioneers [10], [11], [12] up to recent contributions in [13], [14], [15]. Despite their contributions, these approaches rely on at least one critical and fundamental assumption: i) They take for granted an underlying parametric distribution of the perceptual deviations and usually model it by arbitrary fitting parametric distributions. This tendency occurs either explicitly or quite subtle and implicitly. On the one hand, it occurs explicitly by modeling the uncertainty as the (weighted sum of) normal distribution(s) without rigorous validation of the uncertainty distribution profile, namely a methodical analysis of the normal-distribution plausibility, see [10]. This rigorous modeling should include the analysis of the *systematic-errors* (the mean behavior curve) and the *stochastic-spreading* (the standard deviation behavior curve) as functions of the depth. They also lack of the validation through rms-deviation analysis between their estimated normal distribution and one accurate ground-truth-based non-parametric model which tightly reflects the uncertainty nature of the visual depth perception. On the other hand, these assumptions are implicitly stated by *a priori* covariance propagation. This takes place assuming simplifications through analytical models satisfying strong mathematical constraints, usually linearization by truncated Taylor series, for instance [15]. ii) Previous approaches also ignore the actuator effects (see detailed discussion on this issue [16]) within the active visual perception and considering only one single combination of image content and feature extractor, see [14]. iii) Finally, these methods lack of independent and accurate source of measurements to widely (in dept range and amount of samples), strictly and trustworthy validate their models.

[1]**Acknowledgment**: The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no. 270273 (Xperience) and from the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) under the SFB 588

[2]To properly and efficiently answer the fundamental perceptual questions: *what?* and *where?*, the humanoid robot should establish a bidirectional link between external stimuli and internal symbols. This sensor-model-coupling comes into existence by means of sensor-transformation, recognition and state estimation algorithms.

II. Overview

In this paper, a novel uncertainty model of the visual depth perception is proposed. In contrast to the discretized approaches [12], the proposed method is based on supervised uncertainty learning in continuous space. In these terms, the uncertainty model of stereoscopic depth perception is the inferring function Ψ which maps the visual depth (distance between the camera and the target object) δ to its ground-truth depth γ in terms of the probability density function ζ

$$\underbrace{\Psi}_{\text{Ground-Truth Model}} : \underbrace{\{\delta, \gamma \in \mathbb{R}\}}_{\text{Perceptual and Truth Depths}} \mapsto \underbrace{\{\zeta(\delta, \gamma) : \mathbb{R}^2 \mapsto \mathbb{R}\}}_{\text{Ground-Truth PDF}}.$$

The realization of this innovative and non-reductionist approach takes four compounded and interconnected elements, see Fig.2. i) The ground-truth 6D-pose of the humanoid robot cameras and target objects to be visually recognized. This external reference is *partially*[3] attained with a marker-based system [9]. ii) The visual perception consisting of stereo camera calibration, feature extraction and depth estimation methods to be analyzed. iii) An autonomous process for collecting learning samples. This process generates a scanning plan, controls its execution on the humanoid robot 3 DoFs platform, 3 DoFs neck and cameras while simultaneously coordinates the network interfaces with the ground-truth system in order to gather sufficient learning samples necessary for soundly inferring the uncertainty model. iv) The inferring method analyzes the acquired learning samples to obtain the uncertainty model function Ψ. Afterwards, the detailed analysis of the attained model unveils the systematic-errors and stochastic-spreading of the visual perception. This produces an outstanding uncertainty model representation in terms of usability and minimal computational complexity.

In the next sections, a detailed description of the approach is provided. In Sec.III, particular aspects of the humanoid robot's depth perception are discussed. The Sec.IV describes the modeling setup. Afterwards in Sec.V, the visual recognition and 6D-pose estimation of two types of reference-rigs are introduced. Afterwards, the Sec.VI describes the ground-truth acquisition. Next, in Sec.VII, the integration of the visual recognition and ground-truth into the uncertainty model Ψ is presented. The Sec.VIII discusses the proposed model. Finally, the conclusions are stated in Sec.IX.

III. Visual Depth Perception Uncertainty

There are various uncertainty sources in the depth perception by stereo vision in humanoid robots. The embodiment imposes severe physical and computational restrictions. This restrains the characteristics of the cameras and objectives. In addition to the inherent sensor noise, the simultaneous operation of sensors and actuators generate mechanical, electrical and magnetic perturbations deteriorating the image quality, see [16]. The embodiment restrains the camera size

[3]The camera's kinematic frame C cannot be obtained by placing markers on the humanoid robot head nor by attaching markers close to the humanoid robot eyeballs. To obtain this frame, an automatic registration should be conducted by linking the visual recognition to the marker system using a reference-rig, see Sec.VI.

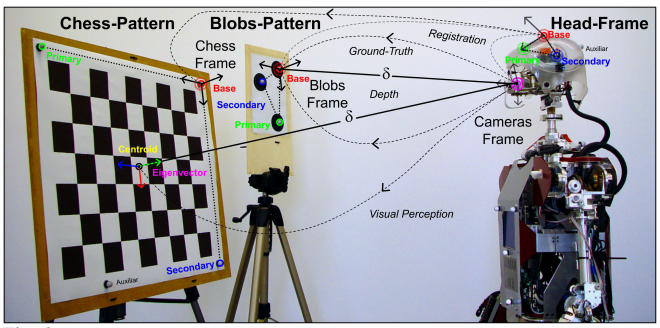

Fig. 2. The supervised learning of the visual depth uncertainty model includes four compounded elements: i) Three markers on the humanoid robot's head are labeled as *base*, *primary* and *secondary*. ii) The 2D blob-pattern calibration-rig with three frame-aligned spherical markers. iii) The chess calibration-rig with frame-aligned spherical markers. iv) The marker system with active-cameras. The exploitation of these elements partially provides the 6D-pose ground-truth for the humanoid robot's camera frame and the 6D-pose of the reference-rig. The aim of use two patterns is to analyze the image content and extraction methods in terms of differences of their uncertainty models.

making it unfeasible to integrate separation prism cameras [17] or vertically stacked photodiode sensors [18]. Thus, the commonly used bayer pattern cameras [19] generate notable aliasing in the color planes [20]. Additionally, the ubiquitous high-contrast image-content produces local over- and under-exposure resulting in severe rate-distortion quantization effects which substantially diminish the image quality, see [21]. Furthermore, most of the vision algorithms use undistorted images produced by interpolated unwrapping. This undistortion process usually employs fast but deficient interpolation methods with error-prone radial and tangential coefficients. More uncertainty is introduced by the estimated principal point, pixel-aspect ratio and focal length. These effects are noticeable at locations far from the principal point [11]. Moreover, the estimation of the 3D position by triangulation is subject to inaccuracies due to the lack of accurate subpixel-calculations [22] or because of the flaws in the extrinsic calibration, see [11].

The depth perception uncertainty is an intricate heterogeneous composition. Many attempts to partially model these effects have been done [10]-[15]. However, non of them has simultaneously regarded all the facts that a complete analytical integration of the uncertainty sources has to consider such as the particular image-content, the specific feature extraction method and the effects of the actuators during the sensing.

In real application humanoid robots are exposed to a wide variety of materials, lighting and operation conditions. Additionally, there are divers feature extraction methods which were successfully and complementary applied in complex humanoid robots application, see [23], [24]. Thus, a complete analytical formulation is unfeasible. Still, the humanoid robot depth perception most incorporate an effective uncertainty model which directly and consistently reflects the nature of the visual perception. A plausible model of such a complex composition is to represent the depth uncertainty Ψ in terms of the depth density distribution ζ, namely the PDF associating perceptual depth δ and ground-truth depth γ.

This holistic approach integrates all the uncertainty sources into a learnable compounded process providing a convenient description of the uncertainty distribution. In contrast to

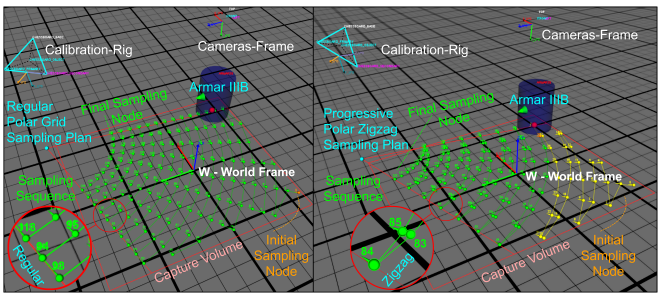

Fig. 3. The autonomous uncertainty sampling. The red rectangle on the floor is the boundary of the marker system. The sampling plan is a distributed set of linked 3D poses (2D for location and 1D for orientation of the robot platform) called sampling nodes. A sampling node includes a set of neck configurations. To determine these configurations, it is necessary to ensure the calibration-rig is within in the robot 3D field of view. On the left, the *regular polar grid* plan is shown. This direct and intuitive but naive and faulty plan produces irregularities and sampling artifacts. This occurs because the sampling nodes of each arc are located at the same distance from the calibration-rig. This produces isolated sample clusters preventing the proper generalization at other depths, see video. On the right, the *progressive polar zigzag* plan enables the proper acquisition of the uncertainty observations because the sampling nodes are located in regular depth progression while the angular distribution is adjusted, see Fig.5.

analytical approaches, the key concept is to learn the non-parametric uncertainty model Ψ as a probability density function by kernel density estimation. The resulting non-parametric PDF Ψ is further analyzed by means of nonlinear regression in order to compare and verify its shape and behavior against the existing parametric models.

IV. Modeling Setup

A learning observation $S_t \in \mathbb{R}^2$ of the depth uncertainty probability function consists of the visually estimated depth $\delta \in \mathbb{R}$ by means of stereo vision and the ground-truth depth $\gamma \in \mathbb{R}$ attained by the marker system. In order to collect these observations, the setup in Fig.2 is proposed:

Head markers: On the humanoid robot head, three spherical markers were placed in a non-collinear arrangement. They are used to estimate the head kinematic frame.

Reference-rigs: First, the *blob-rig* is made of two overlapped patterns: i) The 2D pattern consisting of three black asymmetrically distributed circles with rather large diameter ⌀60mm. ii) The 3D pattern consisting of three spherical markers respectively placed at the centers of the circles. The alignment between both patterns is done by a translation along the 2D pattern's normal $N_{\text{Blobs}} \in \mathbb{R}^3$. Hence, the reference-rig has two kinematic frames, the χ-printed pattern frame and the ς-spheres marker frame. Second, the *chess-rig* is a 800x600 mm standard calibration pattern with three reference markers located at the corners. This reference-rig also involves the two similarly defined kinematic frames χ and ς with an alignment vector $N_{\text{Chess}} \in \mathbb{R}^3$, the translation vector from the chess-pattern *center point* to the *base* sphere marker. The dimensions of the patterns are determined for the recognition at wide-depths further than ~3,500mm using 6mm lenses when using images of VGA resolution.

Labeled marker positions: The active-cameras and their multiple view fusion (see Fig.1) is done by the marker system [9]. All marker positions are calculated relatively to the world coordinate system $W \in SE^3$ established at the initial calibration of the marker system, see Fig.3.

Autonomous uncertainty sampling: In order to obtain highly representative depth observation samples, the regularly distributed locations and orientations of the humanoid head were planned and controlled using a scanning plan, see two labeled screenshots in Fig.3. This was done by sequentially transversing a previously computed path of sampling nodes. Within each sampling node, various configurations of the humanoid robot neck (pitch, roll and yaw) were planned and executed, in this manner the actuator uncertainty effects were introduced in the active sampling distribution. Furthermore, at each neck configuration several recognition trials were performed in order to obtain highly representative uncertainty sampling of the configuration-region.

This coordinated acquisition ensures the sound data association between the robot's visual perception and the marker reference system in terms of temporal consistency, spatial uniformity and active sensing representativeness, namely the wide inclusion of the uncertainty effects of the head actors. The systematic generation and execution of scanning plan has other advantages. For instance, it enables the *comparison* of different extraction methods. Since the process is totally autonomous, it was possible to extensively collect large amount of learning samples (more than 25,000) allowing the generation of a high quality uncertainty model, see video.

V. Humanoid Visual Perception

The humanoid robot visually estimates the 6D-pose of the blob's reference-rig as follows: To segment the pattern circles, the input color image $I_{rgb}(\mathbf{x} \in \mathbb{N}^2) \in \mathbb{N}^3$ is transformed to a normalized saliency image $I_S(\mathbf{x}, \tau) = 1 - \frac{1}{2^{24\tau}} \left[I_r(\mathbf{x}) \cdot I_g(\mathbf{x}) \cdot I_b(\mathbf{x}) \right]^\tau \in \mathbb{R}$, where $\tau \in \mathbb{N}^+$ improves the contrast between circles and background, see Fig.4. After, the binary active pixel image $I_A(\mathbf{x}) \in \{0, 1\}$ was obtained [25], see Fig.4-b. Next, a region growing algorithm extracts active blobs B_i. The blob's centroid $\bar{\mathbf{x}}_i$ is estimated by integrating its radial weighted saliency. The blobs in the left B_i^L and right image B_k^R were matched using epipolar geometry. Matches with low confidence or outside the depth interval (δ_0=500 $\leq |X_i| \leq$ δ_1=3,500mm) were removed. Based on these 3D blob's positions X_i^v (the superindex v denotes vision estimated), the marker's correspondence was performed using the blob's center distances. The matched blob centers X_B^v, X_E^v and $X_S^v \in \mathbb{R}^3$ described the kinematic transformation ${}^{v}\mathcal{T}_C^{\chi}$ from the camera frame C to the χ-printed pattern as an homogeneous matrix, namely

$$\begin{aligned} M^v &= X_E^v - X_B^v, \quad N^v = M^v \times (X_S^v - X_B^v) \\ P^v &= M^v \times N^v, \quad {}^{v}\mathcal{T}_C^{\chi} = \begin{bmatrix} \widehat{M}^v & \widehat{N}^v & \widehat{P}^v & X_B^v \\ 0 & 0 & 0 & 1 \end{bmatrix}, \end{aligned} \tag{1}$$

where $\widehat{\cdot}$ denotes an unitary vector. Finally, the kinematic transformation ${}^{v}\mathcal{T}_C^{\varsigma}$ from the camera frame C to the spheres marker frame ς was done by the alignment offset N_{Blobs} as

$${}^{v}\mathcal{T}_C^{\varsigma} = \begin{bmatrix} \widehat{M}^v & \widehat{N}^v & \widehat{P}^v & (X_B^v + N_{Blobs}) \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

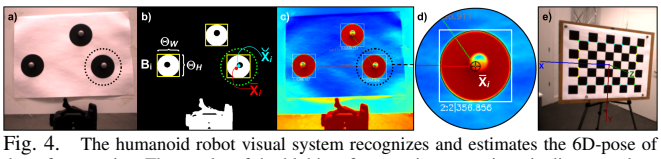

Fig. 4. The humanoid robot visual system recognizes and estimates the 6D-pose of the reference-rigs. The results of the blob's reference-rig processing pipeline at a close distance (~356.8mm). a) The left input color image $I_{RGB}(\mathbf{x})$. b) The active image $I_A(\mathbf{x})$. c) The saliency image $I_S(\mathbf{x}, \tau)$ with the recognition labeling and the estimated depth. d) The zoom circle shows the recognized blob with its centroid and bounding box, the lower text shows the matching identifier and the visual depth measurements. e) The chess reference-rig frame is robustly attained by analysis of the corner points.

This frame is the transformation from the camera's frame C to the reference-rig and is the key to bidirectionally relate the ground-truth system to the visual depth perception.

The extraction of corner points from the chess calibration pattern is a well studied problem and it has been properly solved, see [26]. The extracted noisy 3D corner points P_i^v were used to determine their centroid $\overline{P^v}$ corresponding to the *base* of the χ frame of the chess calibration-rig. Finally, from the matrix $G^v = \frac{1}{n}\sum_{i=1}^{n}(P_i^v - \overline{P^v})(P_i^v - \overline{P^v})^t$ using SVD decomposition, the Eigenvector with smallest associated Eigenvalue is the normal of the plane. The other axes are determine using the geometry of the pattern. The representation of this frame is analogous to Eq.2, see Fig.4-e.

VI. Ground-truth Measurements

In order to sample the uncertainty distribution of the visual depth, all kinematic frames should be linked in a kinematic tree, see Fig.2. This enables a globally unified temporal association of the measurements obtained from both systems.
World to head: The labels and position of the markers are obtained by a network interface. These measurements have a submillimeter accuracy [9]. The kinematic frame on the top of the humanoid robot head $\mathcal{T}_W^{H^m}$ is calculated as in Eq.1, here the superindex m denotes from marker measurements.
World to reference-rig: The transformation $\mathcal{T}_W^{\varsigma^m}$ is computed as in Eq.1 using the marker labels and positions.
Reference-rig to cameras: The transformation from the kinematic frame on the reference-rig to the camera kinematic frame is the inverse of Eq.2, namely $\mathcal{T}_{\varsigma^v}^C = [\mathcal{T}_C^{\varsigma^v}]^{-1}$.
World to camera: This transformation results from the forward kinematic chain of previous two transformations as $\mathcal{T}_W^C = \mathcal{T}_{\varsigma^v}^C \mathcal{T}_W^{\varsigma^m}$. It is the coupling of the visual perception ς^v to the marker ground-truth ς^m, the connection from the χ-printed kinematic frame to the sphere's kinematic frame ς.
Head to camera: Results from the kinematic chain coupling the inverse world to head transformation and the transform from world to camera, expressed as $\mathcal{T}_{H^m}^C = \mathcal{T}_W^C[\mathcal{T}_W^{H^m}]^{-1}$. This is the *camera registration* relative to the head's frame. To accurately achieve this registration, the humanoid robot should be close to the reference-rig while performing this procedure, see Fig.4a-d. This transformation is fixed and stored for sampling process. Thus, if the humanoid robot moves the transformation from the world's kinematic frame W to the camera's kinematic frame C is determined as $\mathcal{T}_W^C(t) = \mathcal{T}_{H^m}^C \mathcal{T}_W^{H^m}(t)$, where the t is the time stamp. This dynamic transformation unifies the visual perception and the ground-truth measurements.

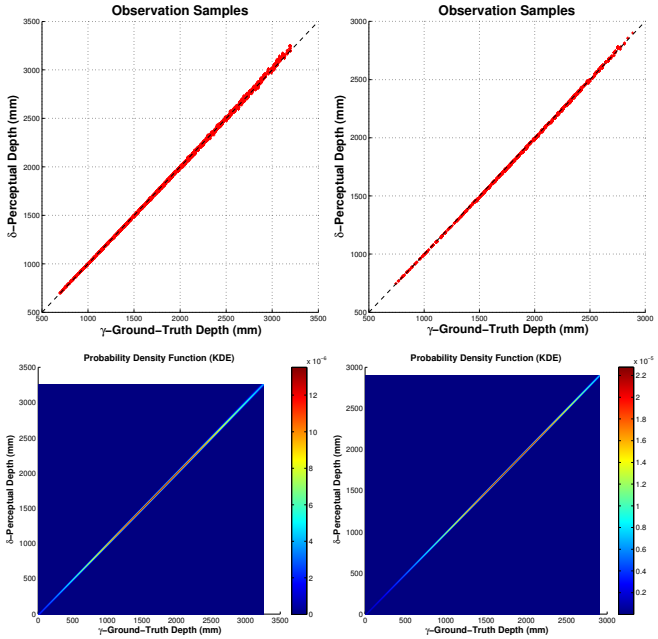

Fig. 5. The upper plots shows the learning observations S_t from Eq.3 represented by 2D points relating the visual depth perception δ to the ground-truth depth γ. The upper left and plots contains the blob-pattern and chess-pattern samples respectively. The lower plots shows the learned model by means of kernel density estimation using adaptive band-width Eq.5. The lower left and right plot show the resulting uncertainty models using blob-pattern $\zeta_{Blobs}(\delta, \gamma)$ and chess-pattern $\zeta_{Chess}(\delta, \gamma)$ respectively.

VII. Uncertainty Distribution Learning

A learning observation of the visual depth uncertainty is

$$S_t := \Big[\underbrace{\delta := \Phi\Big(\mathcal{T}_C^{\varsigma^v}(t)\Big)}_{\delta\text{-Visual Depth, Eq.2}}, \quad \underbrace{\gamma := \Phi\Big(\mathcal{T}_C^{\varsigma^m}(t)\Big)}_{\gamma\text{-Ground-Truth Depth, Eq.4}} \Big]^T \in \mathbb{R}^2, \tag{3}$$

where the depth function $\Phi : SE^3 \mapsto \mathbb{R}$ extracts the displacement length. The observation S_t (see Fig.5) integrate the vision $\mathcal{T}_C^{\varsigma^v}(t)$ from Eq.2 and marker information

$$\mathcal{T}_C^{\varsigma^m}(t) = \mathcal{T}_W^{\varsigma^m}(t)[\mathcal{T}_W^C(t)]^{-1}. \tag{4}$$

A. Model Learning

The uncertainty PDF of the depth perception Ψ is a random variable function sampled by a collection of learning observations using Eq.3. The inference based on the sampling set $L := \{S_t\}_{t=1}^m$ is done by kernel density estimation [27]. The continuous model implies that for a perception depth δ there is a corresponding uncertainty distribution ζ such as $\forall\ (\delta_0 \leq \delta \leq \delta_1)\ \exists \Psi(\delta) \Rightarrow\ \zeta(\delta, 0 \leq \gamma < \infty)$, consequently $\int_0^\infty \zeta(\delta, \gamma)\, d\gamma = 1$. The distribution is inferred as (see Fig.5)

$$\zeta(\delta, \gamma) = \sum_{t=1}^{m} K_{\lambda(\delta,\gamma)}\Big([\delta, \gamma]^T - S_t\Big), \tag{5}$$

where the locally adaptive Gaussian kernel $K_{\lambda(\delta,\gamma)}$ and its bandwidth were determined using the generalization of the Scott's rule [28] as $\lambda(\delta, \gamma) = m^{-\frac{1}{6}} \widehat{\Sigma}^{\frac{1}{2}}(\delta, \gamma)$, where $\widehat{\Sigma}(\delta, \gamma)$ is the sampling covariance matrix centered at $[\delta, \gamma]^T$.

The learned $\zeta(\delta, \gamma)$ is a non-parametric continuous distribution function sampled with fine discretization (κ=1mm) into a lookup table $\tilde{\zeta}(\delta, \gamma)$ with $\left\lceil \left(\frac{\delta_1 - \delta_0}{\kappa}\right) \right\rceil^2 = (3460)^2 = 11971600$ support points, see Fig.5.

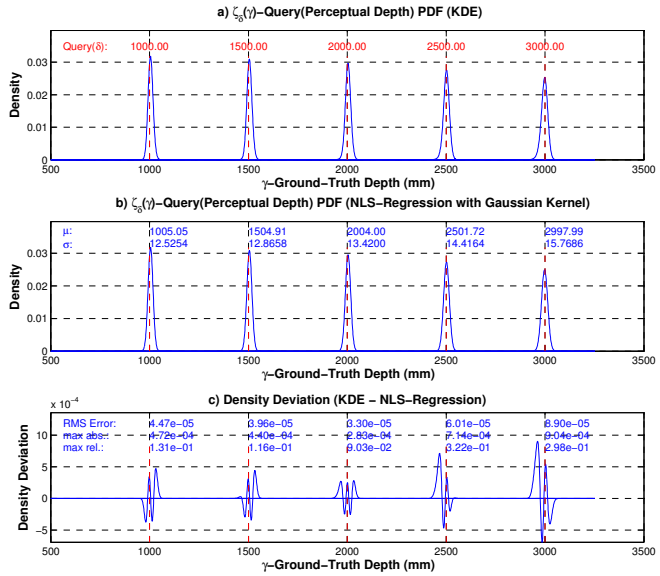

Fig. 6. Queries of the ground-truth uncertainty model of depth perception Ψ using blobs-pattern. a) In order to illustrate the results and usage of the model, five different queries plotted the inferred functions $\zeta(\delta, \gamma)$ from Eq.5. These plots are obtained by fixing the ground-truth depth (γ horizontal axis location in the lower left plot of Fig.5) and then conducting a vertically sampling of the density values δ. These curves are the learned uncertainty profiles at these exemplary depths. b) These are the resulting curves by applying NLS (nonlinear regression) with normal distribution to the previous profile curves from the upper plot. c) The rms-deviation analysis of the difference between the non-parametric model and the normal regression model. The maximal absolute and relative error are also displayed.

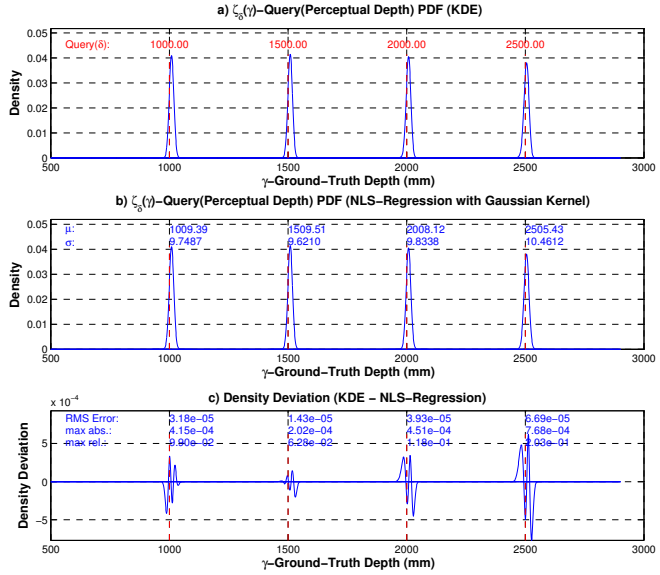

Fig. 7. Queries of the ground-truth uncertainty model of depth perception Ψ using chess-pattern, see the analogous caption in Fig.6.

When a point in space X_i has been visually estimated by the humanoid robot, it is possible to query its depth uncertainty distribution function $\Psi(|X_i|) \mapsto \zeta(|X_i|, \gamma)$, which efficiently, compactly and non-parametrically describes the visual depth PDF along the ground-truth depth γ. This inference $\zeta(|X_i|, \gamma)$ is the ground-truth learned model without any assumption or reduction of the visual perception process.

B. Model Analysis

A study of the uncertainty queries were conducted by applying nonlinear least squares regression at every depth interval κ=1mm as in Fig.6 and Fig.7 for each calibration-rig respectively. During the NLS-regression, the normal-distribution profile was used as the underlying regression shape in order to compare the parametric with the non-parametric approach. This was done by obtaining the systematic-error curve (mean *vs* depth), the stochastic-spreading curve (standard deviation *vs* depth) and their rms-deviation curve (normal-plausibility *vs* depth) to compare them against non-parametric queries, see the results in Fig.8 and Fig.9 for each calibration-rig respectively. Furthermore, these mean and standard deviation curves as functions of the depth can be coherently modeled (again by regression) as functions of the depth in a high order polynomial expression see figures 8-e and g, 9-e and g.

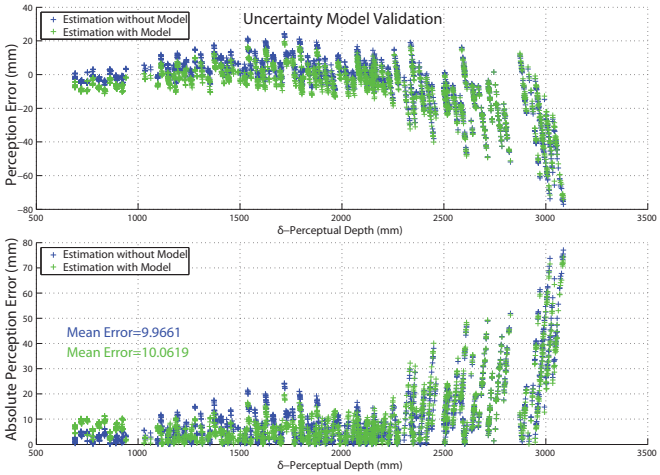

Fig. 10. The validation of the model was realized by comparing the predicted depths from the uncertainty model and the real depth attained by the marker system.

VIII. Discussion

The realization of the proposed uncertainty model is an experimental process. The successfully validation of the models was performed by collecting new samples S_t and compare the prediction of provided by both representation of the model the lookup table $\tilde{\zeta}(\delta, \gamma)$ and the polynomial functions of the depth in Fig.8 and Fig.9, see this validation in Fig.10. The proposed model Ψ can be used to better reflect the perceptual uncertainty of the landmarks by: i) *estimating a correction of the systematic-error of the perception* and *determining an accurate stochastic-spreading by a high oder polynomial function*. In this article, the focus is placed on the depth deviation directly formulated by the depth function Φ in Eq.3. This function maps the full learning space from a 6D-pose to a 1D visual depth subspace. Since the uncertainty of the 6D-pose was extensively and properly sampled, then it can be wider exploited in a similar manner using the S_t observations and a more general 6D-mapping Φ'. Further research in this direction is ongoing work.

IX. Conclusions

The contribution of this article is a novel ground-truth based uncertainty model of the depth perception for humanoid robots. The proposed method conveniently overcomes the analytical limitations of previous models while

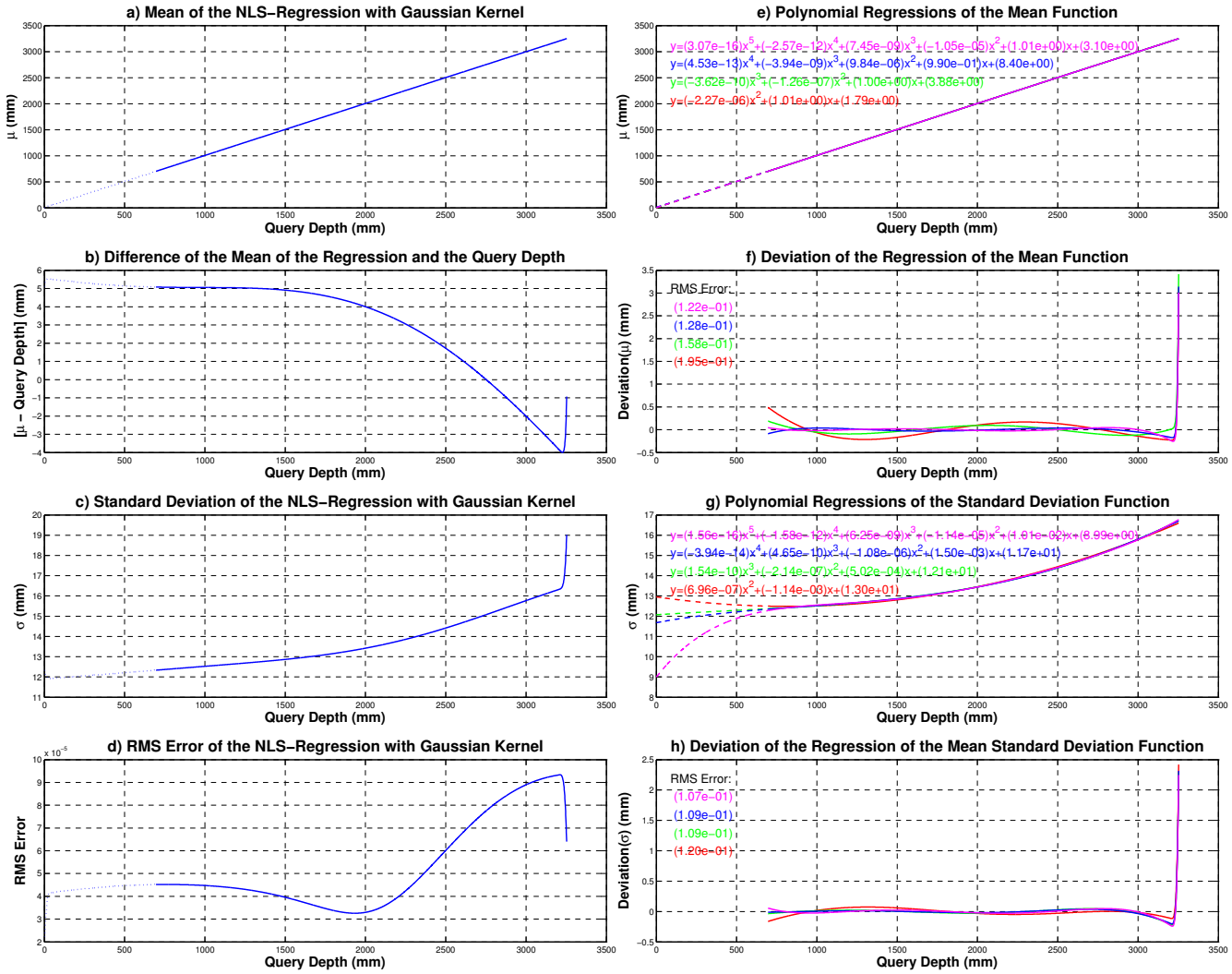

Fig. 8. NLS-regression of the ground-truth uncertainty model of depth perception Ψ using blobs-pattern. The plots in left columns were attained at every 1mm in the depth. The plots in the right side were attained by polynomial regression of those on the left respectively in each row. Notice the various polynomial orders and their deviations.

integrating all the uncertainty sources. The proposed setup and the developed recognition method exploit the accurate ground-truth attained by the marker system. The systematically collection of perceptual samples allows to obtain a high quality model in a fully automatic fashion. The attained uncertainty models presented in Fig.6 and Fig.7 corroborate the approach's motivation because the dispersion of the uncertainty behavior is tightly dependent on the feature extraction mechanism. These results support the need to learn the uncertainty model of the applied recognition mechanism. Furthermore, the representation and exploitation of the model into both lookup table and polynomial functions is a promising technique for a wide range of real-time visual perception applications. A unique and remarkable property of the proposed uncertainty model is its capability to correct the systematic depth errors in the visual depth perception. This novel property and a coherent standard deviation at each depth for the parametrization of the density were the successfully achieved objectives of this work. Additionally, the time varying kinematic tree formulation of the humanoid robot and the elements in the environment can also be used in many other context and applications (robot-machine interaction, motion graphs, imitation and kinematic learning) for robots were the marker positions and robot configurations are unified in a spatio-temporal world reference frame.

References

[1] P. Azad, T. Asfour, and R. Dillmann, "Accurate shape-based 6-dof pose estimation of single-colored objects," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009, pp. 2690 –2695.

[2] N. Vahrenkamp, S. Wieland, P. Azad, D. Gonzalez, T. Asfour, and R. Dillmann, "Visual servoing for humanoid grasping and manipulation tasks," in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, 2008, pp. 406 –412.

[3] N. Vahrenkamp, P. Kaiser, T. Asfour, and R. Dillmann, "Rdt+: A parameter-free algorithm for exact motion planning," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 715 –722.

[4] D. Gonzalez-Aguirre, S. Wieland, T. Asfour, and R. Dillmann, "On environmental model-based visual perception for humanoids," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, ser. Lecture Notes in Computer Science, E. Bayro-Corrochano and J.-O. Eklundh, Eds. Springer Berlin / Heidelberg, 2009, vol. 5856, pp. 901–909.

[5] D. Gonzalez-Aguirre, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Model-based visual self-localization using gaussian spheres," in *Geometric Algebra Computing*, E. Bayro-Corrochano and G. Scheuermann, Eds. Springer London, 2010, pp. 299–324.

[6] J. Bohg, C. Barck-Holst, K. Huebner, M. Ralph, B. Rasolzadeh, D. Song, and D. Kragic, "Towards grasp-oriented visual perception for humanoid robots," *International Journal of Humanoid Robotics (IJHR)*, vol. 6, no. 3, pp. 387–434, Sep 2009.

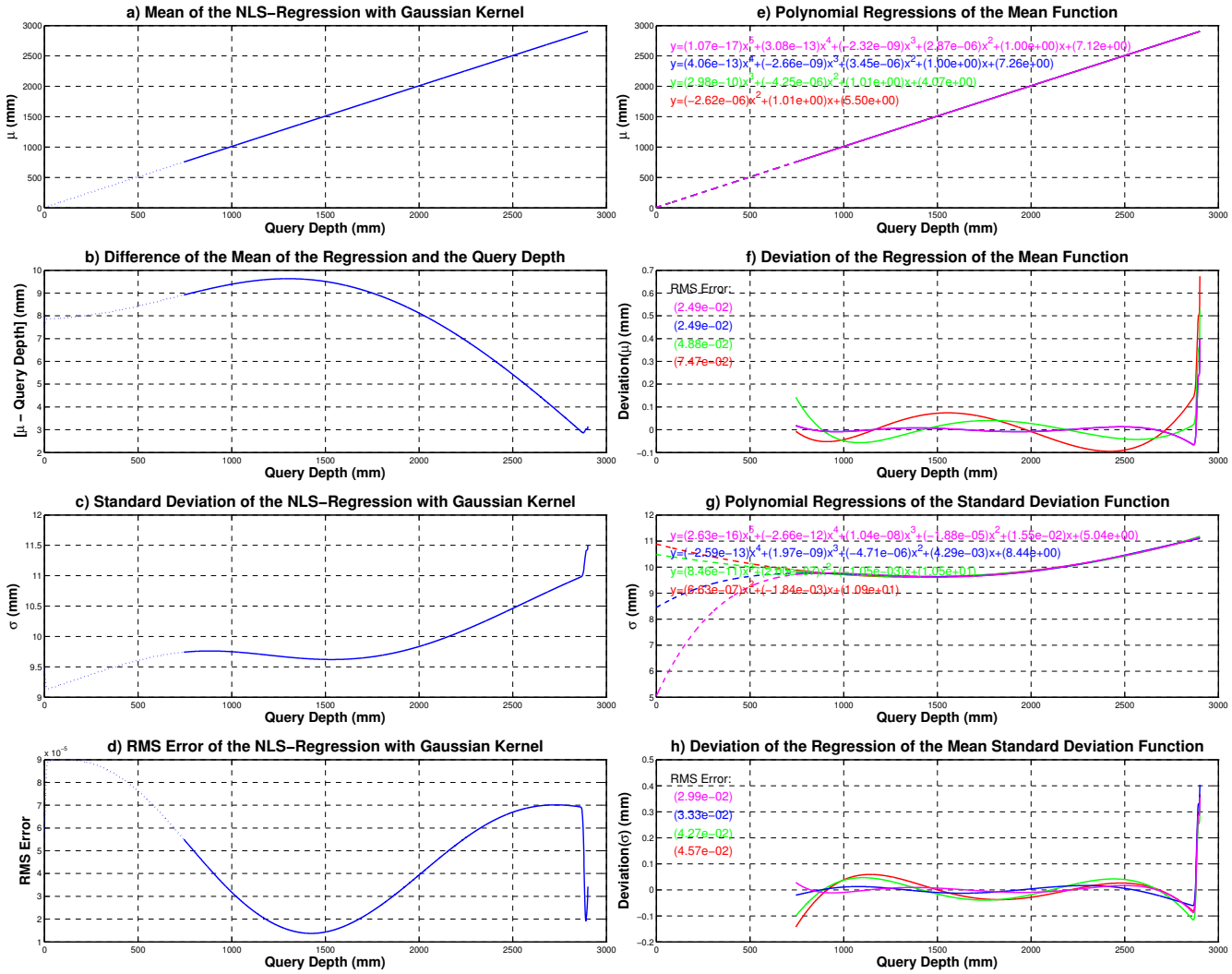

Fig. 9. NLS-regression of the ground-truth uncertainty model of depth perception Ψ using chess-pattern, see the analogous caption in Fig.8.

[7] D. Gonzalez-Aguirre, J. Hoch, S. Rohl, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Towards shape-based visual object categorization for humanoid robots," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 5226 –5232.

[8] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "Armar-iii: An integrated humanoid platform for sensory-motor control," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, 2006, pp. 169 –175.

[9] VICON, "Vicon Motion Systems and Peak Performance Inc," http://www.vicon.com, 2011, [Online; Accessed 1-July-2012].

[10] J. Miura and Y. Shirai, "An uncertainty model of stereo vision and its application to vision-motion planning of robot," in *In Proceedings of the 13th Int. Joint Conf. on Artificial Intelligence*, 1993, pp. 1618–1623.

[11] P. Swapna, N. Krouglicof, and R. Gosine, "The question of accuracy with geometric camera calibration," in *Electrical and Computer Engineering, 2009. CCECE '09. Canadian Conference on*, 2009, pp. 541 –546.

[12] M. Perrollaz, A. Spalanzani, and D. Aubert, "Probabilistic representation of the uncertainty of stereo-vision and application to obstacle detection," in *Intelligent Vehicles Symposium, 2010 IEEE*, 2010, pp. 313 –318.

[13] N. Alvertos, "Resolution limitations and error analysis for stereo camera models," in *Southeastcon '88., IEEE Conference Proceedings*, 1988, pp. 220 –224.

[14] F. H. Snz, J. Q. C, G. H. Bakr, C. E. Rasmussen, and M. O. Franz, "Learning depth from stereo," in *In Pattern Recognition, Proc. 26th DAGM Symposium*. Springer, 2004, pp. 245–252.

[15] G. Di Leo, C. Liguori, and A. Paolillo, "Covariance propagation for the uncertainty estimation in stereo vision," *Instrumentation and Measurement, IEEE Transactions on*, vol. 60, no. 5, pp. 1664 –1673, 2011.

[16] D. Gonzalez-Aguirre, T. Asfour, and R. Dillmann, "Robust image acquisition for vision-model coupling by humanoid robots," in *IAPR Conference on Machine Vision Applications*, 2011.

[17] R. F. Lyon, "Prism-based color separation for professional digital photography," in *PICS*, 2000, pp. 50–54.

[18] I. Foveon, "Foveon X3 direct image sensor," http://www.foveon.com, 2011, [Online; accessed 1-July-2011].

[19] Point Grey Research, Inc., *Dragonfly Technical Reference Manual*, 5-8 2008.

[20] K. Hirakawa and T. Parks, "Joint demosaicing and denoising," *Image Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2146 –2157, aug. 2006.

[21] D. Gonzalez-Aguirre, T. Asfour, and R. Dillmann, "Eccentricity edge-graphs from hdr images for object recognition by humanoid robots," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, 2010, pp. 144 –151.

[22] L. Noskovicova and R. Ravas, "Subpixel corner detection for camera calibration," in *MECHATRONIKA, 2010 13th International Symposium*, 2010, pp. 78 –80.

[23] P. Azad, T. Asfour, and R. Dillmann, "Combining harris interest points and the sift descriptor for fast scale-invariant object recognition," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009, pp. 4275 –4280.

[24] K. Okada, M. Kojima, S. Tokutsu, T. Maki, Y. Mori, and M. Inaba, "Multi-cue 3d object recognition in knowledge-based vision-guided humanoid robot system," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, 2007, pp. 3217 –3222.

[25] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of Graphics, GPU, and Game Tools*, vol. 12, no. 2, pp. 13–21, 2007.

[26] Karlsruhe Institute of Technology (KIT), "The Integrating Vision Toolkit," http://ivt.sourceforge.net/index.html, 2011, [Online; accessed 1-July-2012].

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[28] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.