Statistical Methods and Models for Video-Based Tracking, Modeling, and Recognition
Statistical Methods and Models for Video-Based Tracking, Modeling, and Recognition

Rama Chellappa
UMIACS, University of Maryland
College Park, MD
rama@cfar.umd.edu

Aswin C. Sankaranarayanan
Rice University
Houston, TX
saswin@rice.edu

Ashok Veeraraghavan
Mitsubishi Electric Research Laboratory
Cambridge, MA
veerag@merl.com

Pavan Turaga
UMIACS, University of Maryland
College Park, MD
pturaga@cfar.umd.edu

now
the essence of knowledge
Boston – Delft
Foundations and Trends® in
Signal Processing
Volume 3 Issues 1–2, 2009
Editorial Board

Editor-in-Chief:
Robert M. Gray
Dept of Electrical Engineering
Stanford University
350 Serra Mall
Stanford, CA 94305
USA
rmgray@stanford.edu

Editors

Abeer Alwan (UCLA)          Jelena Kovacevic (CMU)
John Apostolopoulos (HP Labs)    Jia Li (Pennsylvania State
Pamela Cosman (UCSD)         University)
Michelle Effros (California Institute
of Technology)               B.S. Manjunath (UCSB)
Yonina Eldar (Technion)       Urbashi Mitra (USC)
Yariv Ephraim (George Mason
University)                  Thrasos Pappas (Northwestern
University)
Sadaoki Furui (Tokyo Institute
of Technology)               Michael Unser (EPFL)
Vivek Goyal (MIT)             P.P. Vaidyanathan (California
Institute of Technology)
Sinan Gunturk (Courant Institute)      Rabab Ward (University
of British Columbia)
Christine Guillemot (IRISA)    Susie Wee (HP Labs)
Sheila Hemami (Cornell)        Clifford J. Weinstein (MIT Lincoln
Laboratories)
Lina Karam (Arizona State
University)                  Min Wu (University of Maryland)
Nick Kingsbury (Cambridge
University)                  Josiane Zerubia (INRIA)
Alex Kot (Nanyang Technical
University)

Full text available at: http://dx.doi.org/10.1561/2000000007
Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- classification and detection
- estimation and regression
- tree-structured methods

Information for Librarians
Foundations and Trends® in Signal Processing, 2009, Volume 3, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.
Statistical Methods and Models for Video-Based Tracking, Modeling, and Recognition

Rama Chellappa¹, Aswin C. Sankaranarayanan², Ashok Veeraraghavan³ and Pavan Turaga⁴

¹ Department of Electrical and Computer Engineering, Center for Automation Research, UMIACS, at University of Maryland, College Park, MD, rama@cfar.umd.edu
² Department of Electrical and Computer Engineering, Rice University, Houston, TX, saswin@rice.edu
³ Mitsubishi Electric Research Laboratory, Cambridge, MA, veerarag@merl.com
⁴ Department of Electrical and Computer Engineering, Center for Automation Research, UMIACS, at University of Maryland, College Park, MD, pturaga@cfar.umd.edu

Abstract

Computer vision systems attempt to understand a scene and its components from mostly visual information. The geometry exhibited by the real world, the influence of material properties on scattering of incident light, and the process of imaging introduce constraints and properties that are key to interpreting scenes and recognizing objects, their structure and kinematics. In the presence of noisy observations and other uncertainties, computer vision algorithms make use of statistical methods for robust inference. In this monograph, we highlight
the role of geometric constraints in statistical estimation methods, and how the interplay between geometry and statistics leads to the choice and design of algorithms for video-based tracking, modeling and recognition of objects. In particular, we illustrate the role of imaging, illumination, and motion constraints in classical vision problems such as tracking, structure from motion, metrology, activity analysis and recognition, and present appropriate statistical methods used in each of these problems.
Contents

1 Introduction 1
1.1 Goals 5
1.2 Outline 6

2 Geometric Models for Imaging 9
2.1 Models of Surface Reflectance 9
2.2 Camera Models 13
2.3 Motion 22

3 Statistical Estimation Techniques 27
3.1 Static Estimation 28
3.2 Robust M-Estimators 30
3.3 Performance Evaluation of Statistical Methods 34
3.4 Dynamical Systems for Estimation 35

4 Detection, Tracking, and Recognition in Video 43
4.1 Detection 43
4.2 Tracking 48
4.3 Multi-View Metric Estimation 50
4.4 Behavioral Motion Models for Tracking 60
4.5 Simultaneous Tracking and Recognition 67

Full text available at: http://dx.doi.org/10.1561/2000000007
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Statistical Analysis of Structure and Motion Algorithms</td>
<td>75</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>75</td>
</tr>
<tr>
<td>5.2 Feature-Based Methods</td>
<td>77</td>
</tr>
<tr>
<td>5.3 Flow-Based Methods</td>
<td>87</td>
</tr>
<tr>
<td>6 Shape, Identity, and Activity Recognition</td>
<td>97</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>97</td>
</tr>
<tr>
<td>6.2 Shape Representations</td>
<td>100</td>
</tr>
<tr>
<td>6.3 Manifold Representation of Shapes</td>
<td>102</td>
</tr>
<tr>
<td>6.4 Comparing Sequences on Manifolds</td>
<td>110</td>
</tr>
<tr>
<td>6.5 Applications</td>
<td>112</td>
</tr>
<tr>
<td>7 Future Trends</td>
<td>127</td>
</tr>
<tr>
<td>7.1 New Data Processing Techniques: Non-linear Dimensionality Reduction</td>
<td>127</td>
</tr>
<tr>
<td>7.2 New Hardware and Cameras: Compressive Sensing</td>
<td>129</td>
</tr>
<tr>
<td>7.3 Mathematical Tools: Analytic Manifolds</td>
<td>133</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>139</td>
</tr>
<tr>
<td>References</td>
<td>141</td>
</tr>
</tbody>
</table>
The goal of computer vision is to enable machines to see and interpret the world. Computer vision algorithms use input from one or more still images or video sequences that are related in a specific manner. The distribution of intensities and their spatial and temporal arrangements in an image or a video sequence contain information about the identity of objects, their reflectance properties, scene structure, and objects in the scene. However, this information is buried in images and video and that makes it a challenging task. One of the fundamental reasons for this difficulty occurs because mapping from the 3D scene to 2D images is generally non-invertible. Most traditional computer vision algorithms make appropriate assumptions about the nature of the 3D world and acquisition of images and videos, so that the problem of inferring scene properties of interest from sensed data becomes recoverable and analytically tractable.

Within this context, reasonably accurate, yet simple geometric models of scene structure (planar scene, etc.), scene illumination (point source), surface properties (Lambertian, Phong, etc.), imaging structure (camera models) serve critical roles in the design of inference algorithms. Moreover, images and video sequences obtained using imaging devices are invariably corrupted by noise. Common noise
sources in the imaging system are due to shot noise, thermal noise, etc. Inference in this noisy environment is further complicated by the inherent errors in physical modeling. Real surfaces are never truly Lambertian, real cameras are never truly perspective, illumination in a scene is never a point light source, nevertheless inference algorithms make these assumptions in order to make the problem tractable. In addition, motion of objects in a scene could complicate the recovery of scene and object properties due to blur, occlusion, etc. Therefore, it becomes important that the developed inference algorithms can cope with varying sources of error.

To illustrate these sources of error, let us consider the following simple illustration. Suppose we are interested in designing a robot that can localize and identify the entrances to buildings (see Figure 1.1(a)). To begin, we first define a ‘model’ of an entrance. For computational tractability, we assume the edges of the entrance form a rectangle. Now, given an image containing the entrance, we might choose to use an edge detector or a corner detector to extract features. Due to image-noise, occlusions, and shadows, the features may not exactly correspond to edge locations. With these noisy feature locations, we proceed to fit two sets of parallel lines, where the lines from different sets are perpendicular to each other. Consider the edge figure in Figure 1.1(b). Finding the set of points corresponding to the entrance and grouping them into a rectangle comprises a combinatorial optimization problem. Suppose

![Fig. 1.1 Fitting a rectangle to an entrance. Various sources of error arise here – feature points are noisy, grouping of the points into a rectangle is a challenge, and a rectangle is not an accurate model for the entrance.](image-url)
we obtain a solution to this optimization problem, perhaps by using
the Hough transform. The final error in fit would have occurred due to
noisy measurements, the difficulty in solving the constrained optimiza-
tion problem, and the error in modeling itself, since the entrance does
not appear as a rectangle due to perspective effects. The error would
become even worse when the viewing angle moves further from frontal,
or if shadows are present or the entrance is partially occluded, etc.

As this example illustrates, computer vision algorithms involve the
interplay between geometric constraints that arise from models of the
scene. Inference makes assumptions about the imaging devices and
about appropriate statistical estimation techniques that can contend
with varying sources of error. This tutorial attempts to re-examine
and present several computer vision techniques accordingly.

The acceptance of statistical methods in computer vision has been
slow and steady. In the early days of the field, the understanding of the
geometrical aspects of the problem was given much attention. When
uncertainties due to noise and other errors had to be taken into account,
and when prior information and massive sensor data became available,
the infusion of statistical methods was inevitable. Statistical models
and methods entered into computer vision through image models. Non-
causal models were first introduced in the analysis of spatial data by
Whittle [222]. Subsequently, in the 1960s and 1970s, Markov random
fields (MRFs) were discussed in statistical [16] and signal processing
literature [223]. In the 1980s, statistical methods were introduced
primarily for image representation; thus MRFs [36] [47] [110] and other
non-causal representations [37] [110] [111] were suggested for images. This enabled the formulation of problems such as image estimation
and restoration [75], and texture analysis (synthesis, classification, and
recognition) [43] [50] as maximum a posteriori estimation problems.
Appropriate likelihood expressions and prior probability density func-
tions were used to derive the required posterior probability density.
Nevertheless, the maximum of the posterior probability density func-
tions did not always yield a closed form expression, requiring techniques
such as simulated annealing [75] [118].

The introduction of simulated annealing techniques could be con-
sidered a seminal moment as it opened a whole new class of sampling
approaches for synthesis and segmentation of textured images [137] and other early vision problems. Simulated annealing techniques were followed by techniques such as mean field annealing [19], iterated conditional mode [18], and maximum posterior marginal [138]. These techniques are now part and parcel of computer vision algorithms. It is worth noting that MRFs and conditional random fields are making a strong resurgence in graphics and machine learning literature.

Applications of Monte Carlo Markov chain techniques for non-linear tracking problems have also been studied [79]. Since the introduction of the CONDENSATION algorithm in 1996 [95], numerous papers have discussed appearance, shape, and behavior-encoded particle filter trackers. Robust estimation methods offer another statistical area that has received attention in the computer vision literature. Many problems such as fitting lines, curves, and motion models relied on least square fitting techniques which are quite sensitive to the presence of outliers. Since the early 1980s, the use of RANSAC [67, 139], M-estimators [93], and least median square estimators [170] has become valuable in all model fitting problems, including fitting moving surfaces and objects to the optical flow generated by them. Discussions of robust estimation with applications in computer vision can be found in Meer et al. [139].

One of the recurring issues in the development of computer vision algorithms is the need to quantify the quality of the estimates. Haralick and his co-workers [108] pioneered this area. In the classical problem of estimating the 3D structure of a scene from motion cues, which is termed as the ‘structure from motion’ (SfM) problem, one would like to compute the lower bounds on the variances of the motion and structure estimates. Similar needs arise in camera calibration, pose estimation, image alignment, tracking, and recognition problems. A time-tested approach in statistics — the computation of Cramer–Rao bounds [166] and their generalizations — has been adopted for some computer vision problems.

The exploitation of statistical shape theory for object recognition in still images and video sequences has also been studied intensively since the 1980s. In particular, Cooper and collaborators [27, 26] have developed several algorithms based on Bayesian inference techniques.
for the object recognition problem. Introduction of statistical inference techniques on manifolds which host various representations used in shape, identity, and activity recognition problems is garnering a lot of interest.

Kanatani pioneered statistical optimization under the constraints unique to vision problems. His books explore the use of group theoretical methods and statistical optimization in image understanding and computer vision. In particular, Kanatani explores parametric fitting under relationships such as coplanarity, collinearity, and epipolar geometry, with focus on the bounds on the estimate’s accuracy. Kanatani also explored the idea of geometric correction of data to make them satisfy geometric constraints.

Finally, we will be grossly remiss, if we do not acknowledge Prof. Ulf Grenander, who created the area of probabilistic and statistical approaches to pattern analysis and computer vision problems. His series of books, and the recent book with Prof. Mike Miller, have laid the foundations for much of what has been accomplished in statistical inference approaches to computer vision. Prof. Julian Besag’s contributions to the development of spatial interaction models and Monte Carlo Markov chain techniques are seminal. Many researchers have developed statistical approaches to object detection and recognition in still images. In particular, Professors David Mumford, Don Geman, Stu Geman, Yali Amit, Alan Yuille, Mike Miller, Anuj Srivastava, Song-Chun Zhu and many others have made significant contributions to statistical approaches to still image-based vision problems. As we are focusing on video-based methods and have page constraints, we are unable to provide detailed summaries of outstanding efforts by the distinguished researchers mentioned above.

1.1 Goals

In this monograph, we will examine several interesting video-based detection, modeling, and recognition problems such as object detection and tracking, structure from motion, shape recovery, face recognition, gait-based person identification, and video-based activity recognition. We will explore the fundamental connections between these different
problems in terms of the necessary geometric modeling assumptions used to solve them, and we will study statistical techniques that will enable robust solutions to these problems. Of course, a host of other image processing applications exist where statistical estimation techniques have found great use. The goal of some of these applications, such as image denoising, image deblurring, and super-resolution, is to recover an image, not ‘understand’ the scene captured in the image. We therefore will not delve in detail about these applications in this tutorial. An in-depth discussion of some statistical techniques applied to image processing may be found in [77, 97].

Writing this tutorial presented a great challenge. Due to page limitations, we could not include all that we wished. We simply must beg the forgiveness of many of our fellow researchers who have made significant contributions to the problems covered here and whose works could not be discussed.

1.2 Outline

We begin the monograph with an in-depth coverage of the various geometric models that are used in imaging in Section 2. Light from illumination sources interacts with materials, reflects off them, and reaches the imaging system. Therefore, it is important to study the reflectance properties of materials. We describe popular models of reflectance, such as the Lambertian and Phong models, and indicate vision applications where such reflectance models find use. Next, we describe popular models for the imaging sensor — the camera. In particular, we provide a brief description of the perspective projection model and some of its variants. Image sequences obtained from video cameras are related through scene structure, camera motion, and object motion. We also present models for both image motion (optical flow) and object/camera motion and describe how scene structure, motion, and illumination are coupled in a video.

In Section 3, we describe commonly used statistical estimation techniques such as maximum likelihood and maximum a posteriori estimators. We also describe robust estimators such as M-estimators. We state
the problem of Bayesian inference in dynamical systems and describe two algorithms — the Kalman filter and particle filters — that can perform Bayesian inference with applications to object tracking and recognition in video sequences.

In Section 4, we develop models for detection, tracking, and recognition in surveillance applications, highlighting the use of appearance and behavioral models for tracking. Section 5 describes an important fundamental problem in computer vision — structure from motion (SfM). SfM techniques study the relationship between the structure of a scene and its observability given motion. In the section, we highlight various approaches to explore this relationship, then use them to estimate both the structure of the scene and the motion. We also discuss Cramer–Rao bounds for SfM methods based on discrete features and optical flow fields.

Section 6 discusses some applications in vision where the parameters of interest lie on a manifold. In particular, we study three manifolds, the Grassmann manifold, Stiefel manifold, and the shape manifold, and show how several vision applications involve estimating parameters that live on these manifolds. We also describe algorithms to perform statistical inference on these manifolds with applications in shape, identity, and activity recognition. Finally, in Section 7, we conclude the monograph with a discussion on future trends.
References


References


References


References


References


References


Full text available at: http://dx.doi.org/10.1561/2000000007
References


References


References


References


References


References


References


References


