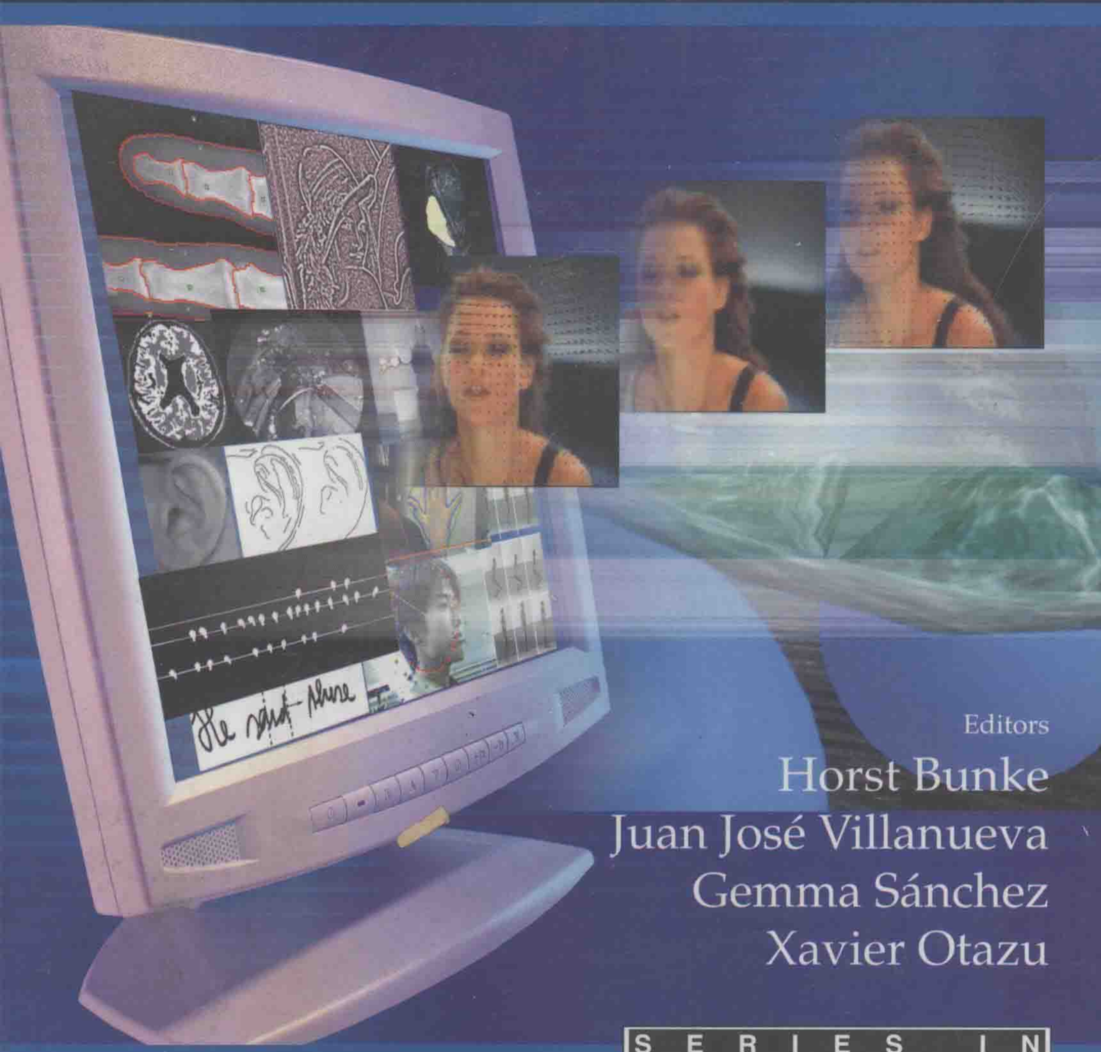
PROGRESS IN COMPUTER VISION AND IMAGE ANALYSIS



Editors

Horst Bunke
Juan José Villanueva
Gemma Sánchez
Xavier Otazu

SERIES IN MACHINE PERCEPTION & ARTIFICIAL INTELLIGENCE
Volume 73

World Scientific

Series in Machine Perception and Artificial Intelligence – Vol. 73

PROGRESS IN COMPUTER VISION AND IMAGE ANALYSIS

Editors

Horst Bunke
University of Bern, Switzerland

Juan José Villanueva
Gemma Sánchez
Xavier Otazu
Computer Vision Center/UAB, Spain

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

PROGRESS IN COMPUTER VISION AND IMAGE ANALYSIS
Series in Machine Perception and Artificial Intelligence — Vol. 73

ISBN-13 978-981-283-445-4
ISBN-10 981-283-445-1

Printed in Singapore by World Scientific Printers

PROGRESS IN
COMPUTER VISION
AND
IMAGE ANALYSIS

SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*

Vol. 58: Computational Web Intelligence: Intelligent Technology for Web Applications
(Eds. *Y. Zhang, A. Kandel, T. Y. Lin and Y. Yao*)

Vol. 59: Fuzzy Neural Network Theory and Application
(*P. Liu and H. Li*)

Vol. 60: Robust Range Image Registration Using Genetic Algorithms and the Surface Interpenetration Measure
(*L. Silva, O. R. P. Bellon and K. L. Boyer*)

Vol. 61: Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications
(*O. Maimon and L. Rokach*)

Vol. 62: Graph-Theoretic Techniques for Web Content Mining
(*A. Schenker, H. Bunke, M. Last and A. Kandel*)

Vol. 63: Computational Intelligence in Software Quality Assurance
(*S. Dick and A. Kandel*)

Vol. 64: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications
(*Elżbieta Pękalska and Robert P. W. Duin*)

Vol. 65: Fighting Terror in Cyberspace
(*Eds. M. Last and A. Kandel*)

Vol. 66: Formal Models, Languages and Applications
(Eds. *K. G. Subramanian, K. Rangarajan and M. Mukund*)

Vol. 67: Image Pattern Recognition: Synthesis and Analysis in Biometrics
(Eds. *S. N. Yanushkevich, P. S. P. Wang, M. L. Gavrilova and S. N. Srihari*)

Vol. 68: Bridging the Gap Between Graph Edit Distance and Kernel Machines
(*M. Neuhaus and H. Bunke*)

Vol. 69: Data Mining with Decision Trees: Theory and Applications
(*L. Rokach and O. Maimon*)

Vol. 70: Personalization Techniques and Recommender Systems
(Eds. *G. Uchyigit and M. Ma*)

Vol. 71: Recognition of Whiteboard Notes: Online, Offline and Combination
(Eds. *H. Bunke and M. Liwicki*)

Vol. 72: Kernels for Structured Data
(*T Gärtner*)

Vol. 73: Progress in Computer Vision and Image Analysis
(Eds. *H. Bunke, J. J. Villanueva, G. Sánchez and X. Otazu*)

*For the complete list of titles in this series, please write to the Publisher.

PREFACE

An image is worth more than ten thousand words - and for that reason Computer Vision has received enormous amounts of attention from several scientific and technological communities in the last decades. Computer Vision is defined as the process of extracting useful information from images in order to be able to perform other tasks.

An image usually contains a huge amount of information that can be utilized in various contexts. Depending on the particular application, one may be interested, for example, in salient features for object classification, texture properties, color information, or motion. The automated procedure of extracting meaningful information from an input image and deriving an abstract representation of its contents is the goal of Computer Vision and Image Analysis, which appears to be an essential processing stage for a number of applications such as medical image interpretation, video analysis, text understanding, security screening and surveillance, three-dimensional modelling, robot vision, as well as automatic vehicle or robot guidance.

This book provides a representative collection of papers describing advances in research and development in the fields of Computer Vision and Image Analysis, and their applications to different problems. It shows advanced techniques related to PDE's, wavelet analysis, deformable models, multiple classifiers, neural networks, fuzzy sets, optimization techniques, genetic programming, among others. It also includes valuable material on watermarking, image compression, image segmentation, handwritten text recognition, machine learning, motion tracking and segmentation, gesture recognition, biometrics, shadow detection, video processing, and others.

All contributions have been selected from the peer-reviewed international scientific journal ELCVIA (http://elcvia.cvc.uab.es). The contributing authors (as well as the reviewers) are all established researchers in the field and they provide a representative overview of the available techniques and applications of this broad and quickly emerging field.

The aim of this book is to provide an overview of recent progress in methods and applications in the domains of Computer Vision and Image Analysis for researchers in academia and industry as well as for Master and PhD students working in Computer Vision, Image Analysis, and related fields.

H. Bunke
J.J. Villanueva
G. Sanchez
X. Otazu

CONTENTS

CHAPTER 1

AN APPEARANCE-BASED METHOD FOR PARAMETRIC VIDEO REGISTRATION

Xavier Orriols, Lluis Barceló and Xavier Binefa

Computer Vision Center, Universitat Autònoma de Barcelona
08193 Bellaterra, Spain
E-mail: xavier.binefa@uab.es

In this paper, we address the problem of multi-frame video registration using an appearance-based framework, where linear subspace constraints are applied in terms of the appearance subspace constancy assumption. We frame the multiple-image registration in a two step iterative algorithm. First, a feature space is built through and Singular Value Decomposition (SVD) of a second moment matrix provided by the images in the sequence to be analyzed, where the variabilities of each frame respect to a previously selected frame of reference are encoded. Secondly, a parametric model is introduced in order to estimate the transformation that has been produced across the sequence. This model is described in terms of a polynomial representation of the velocity field evolution, which corresponds to a parametric multi-frame optical flow estimation. The objective function to be minimized considers both issues at the same time, i.e., the appearance representation and the time evolution across the sequence. This function is the connection between the global coordinates in the subspace representation and the parametric optical flow estimates. Both minimization steps are reduced to two linear least squares sub-problems, whose solutions turn out to be in closed form for each iteration. The appearance constraints result to take into account all the images in a sequence in order to estimate the transformation parameters. Finally, results show the extraction of $3D$ affine structure from multiple views depending on the analysis of the surface polynomial's degree.

1.1. Introduction

The addition of temporal information in visual processing is a strong cue for understanding structure and $3D$ motion. Two main sub-problems appear when it comes to deal with motion analysis; *correspondence* and *reconstruction*. First issue (correspondence) concerns the location analysis of which elements of a frame correspond to which elements in the following images of a sequence. From elements correspondence, reconstruction corresponds to $3D$ motion and structure recovery of the observed world. In this paper, we focus on the first issue, and, more specifically, the problem is centered on the observed motion in static scenes onto the image plane which is produced by camera motion: *ego-motion*. In previous work, dense[1,2] and sparse[3–5] methods to estimate the motion field have been used to this end. Sparse methods strongly rely on the accuracy of the feature detector and not all the information available in the image is employed. Dense methods are based on optical flow estimation which often produces inaccurate estimates of the motion field. Moreover the analysis is instantaneous, which means that is not integrated over many frames. Many authors[6–10] focus on this registration problem in terms of $2D$ parametric alignment, where the estimation process is still between two frames. Thus, taking into account that the second step, *reconstruction*, requires that all the transformations must be put in correspondence with a certain frame of reference, the accumulation error can be present in these computations.

Authors in[11] introduce the notion of *subspace constancy assumption*, where visual prior information is exploited in order to build a views+affine transformation model for object recognition. Their starting point is that the training set has to be carefully selected with the aim of capturing just appearance variabilities; that is, the training set is assumed to be absent of camera (or motion) transformations. Once the learning step is performed, the test process is based on the computation of the affine parameters and the subspace coefficients that map the region in the focus of attention onto the closest learned image. However, in this paper, the topic that we deal with has as input data the images of a sequence that include a camera (or motion) transformations.

In this paper, we address the problem of multi-frame registration by means of an *eigenfeatures* approach, where linear subspace constraints are based on the assumption of constancy in the appearance subspace. We frame the multiple-image registration in a two-step iterative algorithm. First, a feature space is built through and SVD decomposition of a second moment matrix provided by the images in the sequence to be analyzed. This technique allows us to codify images as points capturing the *intrinsic degrees of freedom* of the appearance, and at the same time, it yields compact description preserving visual semantics and perceptual similarities.[12–14]

Second, a parametric model is introduced in order to estimate the transformation that has been produced across the sequence. This model is described in terms of a polynomial representation of the velocities field evolution. Polynomial coefficients are related with $3D$ information. For instance, in the specific case of affine transformations of a planar surface, the linear terms (0 and 1 degree) will contain information about its translations and rotations, the quadratic terms will explain the projective behavior, and so forth. Each step is utilized as the input entry to the next step; that is, once the eigen-subspace is computed, we show how the transformations are estimated, therefore, images are registered according to these estimates and again the eigen-subspace is built with the registered images in the previous step. These two step are iterated until the error function converges under a certain degree of tolerance.

The outline of the paper is as follows: section 2 frames the idea of using the eigenfeatures approach and its relation with the parametric model of transformations. More specifically, we analyze how such an appearance subspace is built according to a previously selected frame of reference. Therefore, a polynomial model is introduced in order to link the appearance constraints to the transformations that occurred across the sequence. In the experimental results, section 3, we show a new manner of encoding temporal information. We point out that when parallax is involved in the problem of video registration, the temporal representation gives a visual notion of the depth in the scene, and therefore it offers the possibility of extracting the affine $3D$ structure from multiple views. The relation between the surface polynomial's degree and $3D$ affine structure is also illustrated. In section 4, the summary and the conclusions of this paper are shown.

1.2. Appearance Based Framework for Multi-Frame Registration

In this section, we present an objective function which takes into account appearance representation and time evolution between each frame and a frame of reference. In this case, temporal transformations estimation is based on the fact that images belonging to a coherent sequence are also related by means of their appearance representation.

Given a sequence of F images $\{I_1, \ldots, I_F\}$ (of n rows and m columns) and a selected frame of reference I_0, we can write them in terms of column vectors $\{y_1, \ldots, y_F\}$ and y_0 of dimension $d = n \times m$. Both pictures *pixel-based* I_i and *vector-form* y_i of the i-th image in the sequence are relevant in the description of our method. The first representation I_i is useful to describe the transformations that occurred to each pixel. The vector-form picture is utilized for analyzing the underlying appearance in all the sequence.

Under the assumption of brightness constancy, each frame in the sequence I_i can be written as the result of a Taylor's expansion around the frame of reference I_0:

$$I_i(\vec{x}) = I_0(\vec{x}) + \nabla I_0(\vec{x})^T \vec{\omega}_i(\vec{x}) \quad (1.1)$$

This is equivalent, in a vector-form, to:

$$y_i = y_0 + t_i \tag{1.2}$$

where t_i is the vector-form of the second summand $\nabla I_0(\vec{x})^T \vec{\omega}_i(\vec{x})$ in eq. (1.1). First description is exploited in section 1.2.2, where the parametric polynomial model to describe the velocity field estimates is applied. The vector-form description in eq (1.2) is employed in the following section 1.2.1 to develop the appearance analysis respect to a chosen reference frame.

1.2.1. *Appearance Representation Model*

First of all, we need to define a space of features where images are represented as points. This problem involves finding a representation as a support for analyzing the temporal evolution. To address the problem of appearance representation, authors in[12–14] proposed Principal Component Analysis as redundancy reduction technique in order to preserve the semantics, i.e. perceptual similarities, during the codification process of the principal features. The idea is to find a small number of causes that in combination are able to reconstruct the appearance representation.

One of the most common approaches for explaining a data set is to assume that causes act in linear combination:

$$y_i = W\xi_i + y_0 \tag{1.3}$$

where $\xi_i \in \Re^q$ (our chosen reduced representation, $q < d$) are the causes and y_0 corresponds to the selected frame of reference. The q-vectors that span the basis are the columns of W ($d \times q$ matrix), where the variation between the diferents images y_i and the reference frame is encoded.

With regard to equation (1.2), and considering the mentioned approximation in (1.3), we can see that the difference t_i between the frame of reference y_0 and each image y_i in the sequence is described by the linear combination $W\xi_i$ of the vectors that span the basis in W. Notice that in the usual PCA techniques y_0 plays the role of the sample mean. In recognition algorithms this fact is relevant, since there is assumed that each sample is approximated by the mean (ideal pattern) with an added variation which is given by the subspace W. However, in our approach, each image y_i tends to the frame of reference y_0 with a certain degree of variation, which is represented as a linear combination of the basis W.

Furthermore, from eq. (1.1), the difference t_i, that relies on the linear combination of the appearance basis vectors, can be described in terms of the parametric model which defines the transformation from the reference frame y_0 and each image y_i. This parametric model is developed in the following section 1.2.2.

Besides, from the mentioned description in terms of a subspace of appearance, we can see the form that takes the objective function to be minimized. Indeed, the idea is to find: a basis W, a set of parameters $\{p_1, ..., p_r\}$, (that model the temporal transformations), and a set of registered images where the squared distance between the difference obtained through the taylor's expansion t_i and the projected vector in the appearance subspace $W\xi_i$ is minimum, i.e.:

$$\mathcal{E}(W, \ldots, p_1^i, \ldots, p_r^i, \ldots) = \sum_{i=1}^{F} \mid t_i(p_1^i, \ldots, p_r^i) - W\xi_i \mid^2 \tag{1.4}$$

The minimization of this objective function requires of a two-step iterative procedure: first it is necessary to build an appearance basis, and therefore, to estimate the parametric transformations that register the images in the sequence. In the following sections introduce closed forms solutions for each step.

1.2.2. *Polynomial Surface Model*

In this section we present a polynomial method to estimate the transformation between de reference frame I_0 and each frame I_i in the sequence. To this end we utilize the pixel-based picture. From equation (1.1) we can see that the difference between a frame I_i and the frame of reference I_0 relies on the velocities field $\vec{\omega}_i(\vec{x})$. A s-degree polynomial model for each velocity component can be written as follows:

$$\vec{w_i}(\vec{x}) = \mathcal{X}(\vec{x})\vec{P_i} \tag{1.5}$$

where $\mathcal{X}(\vec{x})$ is a matrix that takes the following form:

$$\mathcal{X}(\vec{x}) = \left[\begin{array}{c|c} \Omega(\vec{x}) & 0 \\ \hline 0 & \Omega(\vec{x}) \end{array}\right]$$

with

$$\Omega(\vec{x}) = \begin{bmatrix} 1 \; x \; y \; xy \; x^2 \ldots (x^l y^k) \ldots y^s \end{bmatrix}$$

where $\Omega(\vec{x})$ is a $d \times 2r$, $(r = (s+1)(s+2))$, matrix that encodes pixel positions, and $\vec{P_i}$ is a column vector of dimension $r = (s+1)(s+2)$, which corresponds to the number of independent unknown parameters of the transformation. In matrix language $\mathcal{X}(\vec{x})$ is a matrix $2d \times r$, $\vec{P}$ has dimensions $r \times 1$, and the velocities corresponding to each pixel can be encoded in a matrix $\vec{w_i}(\vec{x})$ of dimensions $2d \times 1$. The gradient expression in the linear term of the taylor's expansion (1.1) can be written in a diagonal matrix form as follows:

$$G_x = \begin{bmatrix} g_x^1 & 0 & \dots & 0 \\ 0 & g_x^2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & g_x^d \end{bmatrix} G_y = \begin{bmatrix} g_y^1 & 0 & \dots & 0 \\ 0 & g_y^2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & g_y^d \end{bmatrix}$$

Stacking horizontally both matrices we obtain a matrix G of dimensions $d \times 2d$: $G = [G_x \mid G_y]$. Therefore, according to the vector-form in eq (1.2), the difference t_i between the i-th frame y_i and the frame of reference y_0, is expressed in terms of the polynomial model through:

$$t_i(\vec{x}, \vec{P}_i)_{d\times 1} = G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}\vec{P}_i \mid_{r\times 1} \tag{1.6}$$

Given that the term $G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}$ is computed once for all the images in iteration, we re-name it as $\Psi_{d\times r} = G_{d\times 2d}\mathcal{X}(\vec{x})_{2d\times r}$. Notice that even when images are highly dimensional, (e.g. $d = 240 \times 320$), the computation of Ψ can be perfomed easily in *Matlab* by means of the operator ".*", without incurring in an out of memory.

1.2.3. *The Algorithm*

Given the parametric model for the transformations of the images in a sequence, the objective function (1.4) can be written explicitly in terms of the parameters to be estimated:

$$\mathcal{E}(W, \vec{P}_1, \dots, \vec{P}_F) = \sum_{i=1}^{F} \mid \Psi\vec{P}_i - W\xi_i \mid^2 \tag{1.7}$$

In order to minimize this objective function, we need a two step procedure: first given a set of images, the subspace of appearance W is computed, and secondly, once the parameters $\vec{P}_i$ that register each frame y_i to the frame of reference y_0 are obtained, the images are registered in order to build again a new subspace of appearance.

a. Appearance Subspace Estimation. Consider an intermediate iteration in the algorithm, thus, the set of registered images to be analyzed are: $\{\phi_1(y_1, \vec{P}_1), \dots, \phi_F(y_F, \vec{P}_F)\}$. From this set and the reference frame y_0, the appearance subspace can be performed by means of an Singular Value Decompo-