Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition:

Advancing Technologies

Vijay Kumar Mago
Simon Fraser University, Canada

Nitin Bhatia
DAV College, India

Information Science
REFERENCE

Managing Director: Lindsay Johnston
Senior Editorial Director: Heather Probst
Book Production Manager: Sean Woznicki
Development Manager: Joel Gamon
Development Editor: Mike Killian
Acquisitions Editor: Erika Gallagher
Typesetters: Lisandro Gonzalez
Cover Design: Nick Newcomer, Lisandro Gonzalez

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Cross-disciplinary applications of artificial intelligence and pattern recognition : advancing technologies / Vijay Kumar Mago and Nitin Bhatia, editors.
p. cm.
Summary: “This book provides a common platform for researchers to present theoretical and applied research findings for enhancing and developing intelligent systems, discussing advances in and applications of pattern recognition technologies and artificial intelligence”-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-61350-429-1 (hardcover) -- ISBN 978-1-61350-430-7 (ebook) -- ISBN 978-1-61350-431-4 (print & perpetual access) 1. Pattern recognition systems. 2. Artificial intelligence. I. Mago, V. K. II. Bhatia, Nitin, 1978-
TK7882.P3C66 2012
006.3--dc23
2011046541

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

Chapter 1
From Object Recognition to Object Localization

Rigas Kouskouridas
Democritus University of Thrace, Greece

Antonios Gasteratos
Democritus University of Thrace, Greece

ABSTRACT

Recognizing objects in a scene is a fundamental task in image understanding. The recent advances in robotics and related technologies have placed more challenges and stricter requirements to this issue. In such applications, robots must be equipped with a sense of location and direction with a view to the efficient accomplishment of navigation or demanding pick and place tasks. In addition, spatial information is required in surveillance processes where recognized targets are located in the working space of the robot. Furthermore, accurate perception of depth is mandatory in driver assistance applications. This chapter presents several recently proposed methods capable of first recognizing objects and then providing their spatial information in cluttered environments.

INTRODUCTION

Computer vision generally interferes with recognizing patterns and targets. A wealth of research is devoted to the building of algorithms capable of either detecting simple blob-like structures or recognizing complicated patterns. Generally, the efficiency of a pattern recognition technique depends on its ability to decode, with as much accuracy as possible, vital visual information contained in the natural environment. During the past few years, remarkable efforts were made to build new algorithms for robust object recognition in difficult environments. To this end, researchers emphasized in developing recognition paradigms based on appearance features with local potency (Nister, D., & Stewenius H. 2006, Sivic, J., & Zisserman, A. 2003). Algorithms of this field extract features with local extent that are invariant to possible illumination, viewpoint, rotation and scale changes.

Another aspect that has received much attention in the literature is to exploit the data derived during recognition with a view to provide objects' spatial

DOI: 10.4018/978-1-61350-429-1.ch001

information. Apart from its identity, several other object-related characteristics, such as its distance to the camera or its pose (orientation relative to the camera's plane), could be obtained (Thomas, A., et al. 2009, Sandhu, R., et al. 2009, Ekvall, S., 2005). As a result, assigning spatial attributes to recognized objects provides solutions to numerous technical problems. In robotics applications robots must be equipped with a sense of location and direction with a view to the efficient accomplishment of navigation or demanding pick and place tasks (Kragic, D., et al. 2005, Wong, B., & Spetsakis, M. 2000). In addition, spatial information is required in surveillance processes, where recognized targets are located in the working space of the robot. Furthermore, accurate perception of depth is mandatory in driver assistance applications (Borges, A. P. et al. 2009). Quality control procedures of industrial production frameworks demand accurate acquisition of enhanced spatial information in order to reject faulty prototypes. To sum up, the ultimate challenge for computer vision society members is the building up of advanced vision systems capable of both recognizing objects and providing their spatial information in cluttered environments.

This chapter is mainly devoted to two major and heavily investigated aspects. Initially, the current trend in recognition algorithms suitable for spatial information retrieval is presented. Several recently proposed detectors and descriptors are analytically presented along with their merits and disadvantages. Their main building blocks are examined and their performance against possible image alterations is discussed. Furthermore, the most known techniques emphasizing in the estimation of pose and location of recognized targets are presented. The remainder of the chapter is structured as follows: In Section 2, we give an overview of the current trend in object recognition techniques. A review of recently proposed pose estimation and 3D position calculation algorithms is presented in Section 3. Furthermore, at the last part of the section a comparison study of the presented pose estimation schemes is illustrated. Finally, the chapter concludes with a discussion and an outlook to the future work.

LOCAL APPEARANCE-BASED OBJECT RECOGNITION AND MULTI-CAMERA SYSTEMS

State-of-the-art object recognition frameworks rely on local appearance-based features extracted and organized by detectors and descriptors respectively. Generally, the main idea behind interest location detectors is the pursuit of points or regions containing exceptional information about an object or a scene. These spots or areas hold data that distinguish them from others in their local neighborhood. Thus, regions in a scene that enjoy solitary quality and quantity of information can be easily detected. It is apparent that, detector's efficiency is directly related to its ability to locate as many distinguishable areas as possible in an iterative process. Harris and Stephens (Harris, C., & Stephens M. 1988) were the first to implement an interest point detector, known as Harris Corner detector. Due to the fact that it provides significant repeatability, many recent proposed studies (Rothganger, F.et al.2006, Schmid, C., & Mohr, R. 2006) have adopted it in order to perform demanding object recognition tasks. Furthermore, several variations of Harris Corner detector, such as Harris-Laplace (Mikolajczyk, K., & Schmid, C. 2004) and Harris-affine (Mikolajczyk, K., et al. 2005), were presented due to the fact that they provide significant efficiency. In turn, another profitable detector, the Maximal Stable Extremal Regions Detector (MSER), was recently proposed in (Matas, J.et al. 2004). In short, regions darker or brighter than their surroundings are detected, while the efficiency of the algorithm relies on the trade-off between pixels' intensity value on the center of the mask and those on the local neighborhood. In turn, the "Features from Accelerated Segment Test (FAST)" feature detector constitutes

the most recently interest point extractor and it was proposed in (Rosten, E., & Drummond, T. 2006). It incorporates the Bresenham's circle theory into a window-based interest point pursuit. Moreover, biological inspired vision systems aim at the human's visual cortex simulation and, as a result, to the development of human-based interest point detector. For this purpose Kadir's Salient Detector proposed in (Kadir, T., & Brady, M. 2001) extracts spots in an image that contain important information which distinguish them from others in their local neighborhood. The Salient Detector is based on the probability density function (PDF) of intensity values over an elliptical region, whilst for each pixel an entropy extremum is estimated.

On the other hand, a descriptor organizes the information collected from the detector in a discriminating manner. Thus, high dimensional feature vectors corresponding to locally sampled feature descriptions are produced. In other words, an object (or parts of it) located in a scene is represented by descriptors. Furthermore, the final object depiction is accomplished by stacking descriptors in a logical coherence. Generally, by organizing descriptors from multiple patterns into large databases demanding multiple-object recognition tasks can be achieved. These databases, that play essential role in all the recently proposed recognition approaches, are structured in a vocabulary-tree format. In (Lepetit, V., & Fua, P. 2006), simple real-time recognition of a single object that is based on multiple randomized trees is presented. One of the most profitable vocabulary-tree formats was presented in (Nister, D., & Stewenius H. 2006), where each descriptor from training set represents a single leaf of the tree. In (Lazebnik, S., & Ponce, J. 2005), a new invariant descriptor, called Spin Image, that outperforms Gabor filter was presented. The most important drawback of this detector is its inefficiency to directly transpose spin image descriptor to region-based information. On the contrary, complex filters that were proposed in (Schaffalitzky, F., & Zisserman, A. 2002) and (Baumberg, A. 2000), were applied to the descriptor matching process since it composes the prerequisite for every object recognition method. Moreover, complex filters, that use either Gaussian or polynomial derivatives, generate kernels for the purposes of average intensity estimation of a region.

The complexity of the visual information contained in a scene is described by generalized moment invariants as it is proposed in (Van Gool, J., & Moons, T., & Ungureanu, D. 1996). These moments can be easily computed for arbitrary order and degree, whilst they also characterize the distribution of shape and intensity over a region in an image. In turn, local and gradient histograms were efficiently utilized as adequate feature descriptors. Local Energy based Shape Histogram (LESH), which was introduced in (Sarfraz, MS., et al. 2008), encodes the shape of an object by accumulating local energy of the underlying signal. Finally, this local energy is organized into 128-dimensional spatial histogram. Gradient Location and Orientation Histogram (GLOH) that was proposed in (Mikolajczyk, K., & Schmid, C. 2005), produces descriptor histograms that are calculated on a fine circular grid. The final outcome corresponds to 272-dimensional histogram that efficiently represents visual spatial information over a region in an image. Techniques comprising of a detector and a descriptor are referred in the literature as two-part approaches. Currently, the two most popular approaches that implement both a detector and descriptor are the Scale Invariant Feature Transform (SIFT) (Lowe, D.G. 2004) and Speeded up Robust Features (SURF) (Bay, H., et.al 2008).

Scale Invariant Feature Transform (SIFT)

In order to detect interest point locations, that are invariant to scale change of the image, the usage of a scale space function is mandatory. In (Lindeberg, T. 1994), it has been shown that the only possible scale space mechanism, which can

be used for interest point detection, is the Gaussian function. Thus, SIFT's detector requires that the image is convolved with the variable-scale Gaussian function for the production of the scale-space image:

$$L\left(x, y, \sigma\right) = G\left(x, y, \sigma\right) * I(x, y)$$

with

$$G\left(x, y, \sigma\right) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

L(x, y, σ) represents the function of scale space of an image, G(x, y, σ) the variable-scale Gaussian and I(x, y) the input image. Afterwards, stable key-point locations are detected by using scale-space extremum in the difference-of-Gaussian (DoG) function convolved with the image D(x, y, σ). The later is computed from the difference of two nearby scales separated by a constant multiplicative factor κ:

$$\begin{aligned} D\left(x, y, \sigma\right) &= \left(G\left(x, y, \kappa\sigma\right) - G\left(x, y, \sigma\right)\right) * I\left(x, y\right) \\ &= L\left(x, y, \kappa\sigma\right) - L(x, y, \sigma) \end{aligned}$$

In (Lindeberg, T. 1994) it was shown that, the DoG function provides a close approximation to the scale-normalized Laplacian of Gaussian, $\sigma^2\nabla^2 G$, needed for the true scale invariance of a key-point location. The local maxima and minima estimation of D(x, y, σ) is executed with the procedure shown in Figure 1. Each sample point is compared to its eight neighbors in the current image and nine in the scale above and below. Finally, it is selected only if it is greater than all these neighbors or smaller than all of them. After the efficient key-point location assignment by the detector, information around a feature point is exploited for the needs of the descriptor. Initially, a consistent orientation to each key-point based on local image properties is estimated. For each image sample, L(x, y), the gradient magnitude m(x, y), and orientation, θ(x, y), is computed using pixels' intensity values differences:

$$m\left(x, y\right) = \sqrt{\begin{aligned}&(L\left(x+1, y\right) - L(x-1, y))^2 \\ &+(L\left(x, y+1\right) - L(x, y-1))^2\end{aligned}}$$

$$\begin{aligned}\theta\left(x, y\right) = tan^{-1}((L\left(x, y+1\right) - L\left(x, y-1\right)) \\ /(L\left(x+1, y\right) - L\left(x-1, y\right)))\end{aligned}$$

These gradient orientations of every sample point within a 16x16 region around the key-point are used to form an orientation histogram. The later contains 36 bins representing the 360 degree range of orientations, whilst each added sample is weighted by its gradient magnitude. Afterwards, a 4x4 descriptor matrix containing vectors with magnitude and orientation relative to the contents of the orientation histogram is produced. The final descriptor representation is a 4x4x8=128 element feature vector with magnitude and orientation derived from the algebraic sum of the

Figure 1. The marked pixel (with X) is compared to its 26 neighbors in 3X3 regions at the current and adjacent scales

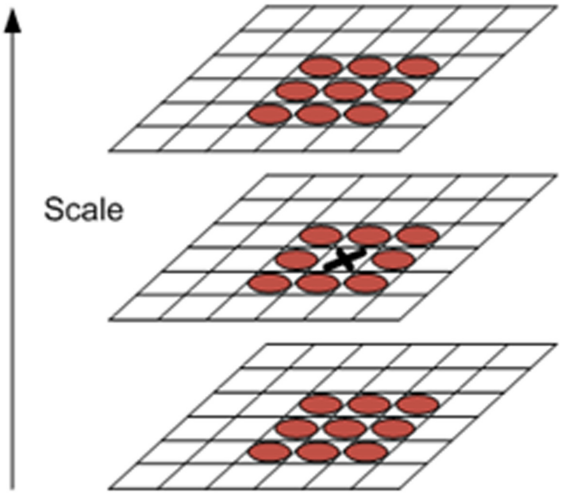
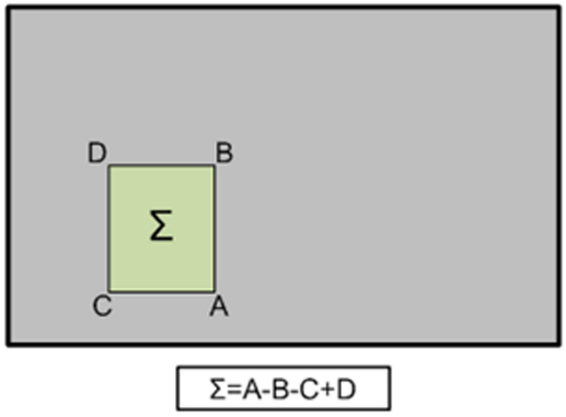


Figure 2. The matching process of the SIFT algorithm along with the estimated feature correspondences



orientation histogram contents for every key-point. The matching process of SIFT is illustrated in Figure 2, where the feature correspondences are extracted over the surface of the object.

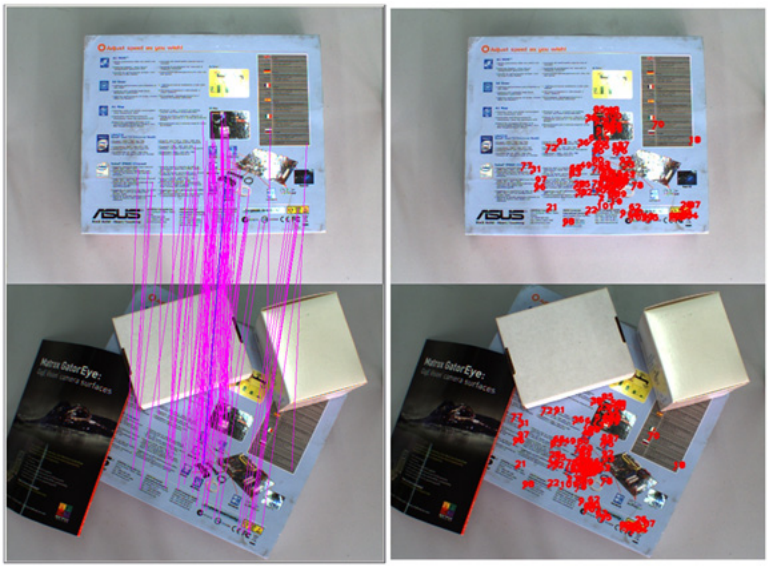
Speeded Up Robust Features (SURF)

Interest point detection is performed by using the basic Hessian matrix approximation, and as a result, the usage of integral images, which were proposed in (Viola, P., & Jones, M. 2001), is mandatory. The most important advantage of integral images is that they reduce the computation time drastically by allowing box type convolution filters. The record of an integral image $I_{\Sigma}(\chi)$ at a location $\chi = (x, y)^T$ corresponds to the sum of all pixels' intensities of the input image I within a rectangular region formed by the origin and χ.

$$I_{\Sigma}(\chi) = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le y} I(i, j)$$

The sum of the intensities over any upright, rectangular area is calculated after three additions and four memory accesses. Thus, the computation time is directly related to the size of the rectangular region. The main structure of an integral image is shown in Figure 3. For the needs of the

Figure 3. The sum of intensities inside the rectangle area Σ is taken into account for the needs of integral images



efficient detection of blob-like structures, SURF's detector is based on the Hessian matrix because of its good performance in accuracy. A key-point is found where the determinant of the Hessian matrix becomes maximum. Given a point χ= (x, y) in an image I, the Hessian Matrix H(χ, σ) in χ at scale σ is defined as follows:

$$H\left(\chi,\sigma\right)=\begin{bmatrix}L_{xx}(\chi,\sigma) & L_{xy}(\chi,\sigma)\\ L_{yx}(\chi,\sigma) & L_{yy}(\chi,\sigma)\end{bmatrix}$$

where $L_{xx}(\chi, \sigma)$ represents the convolution of the Gaussian second order derivative with the image I at point χ, and similarly for $L_{xy}(\chi, \sigma)$ and $L_{yy}(\chi, \sigma)$.

The accurate key-point localization in the image and over scales is accomplished by applying non-maximum suppression in a 3X3X3 neighborhood, as it was proposed in (Neubeck, A., & Van Gool, L. 2006). The final estimation of interest's point location is fulfilled by the interpolation of the maxima of the Hessian's matrix determinant in scale and image space. The construction process of SURF's descriptor is divided into two phases. In the first stage, and with a view to descriptor's invariance to a possible image rotation, a reproducible orientation of the interest points is estimated. Especially, the Haar wavelet responses in x and y direction within a circular neighborhood of radius 6s around the interest point are calculated. Here s corresponds to the scale at which the key-point was detected. Only six operations are needed to compute the response in x and y direction in any scale. Finally, for every interest point an orientation, that is estimated by calculating the sum of all Haar wavelet responses within a sliding window of size $\pi/3$, is assigned.

In the second phase, a square region with size of 20s, whilst centered around the key-point and oriented along the orientation extracted in the previous stage, is constructed. This region is split up regularly into smaller 4X4 square sub-regions, for each of them Haar wavelet responses at 5X5 sample points are computed. Afterwards, for each sub-region a vector, that is produced by the sum of Haar responses in x and y, is estimated. The final descriptor representation is a feature vector with 64 elements that is extracted by taking into account the vectors of each sub-region.

OBJECTS' POSE AND POSITION ESTIMATION IN THE 3D SPACE

In the last few years, a tendency to introduce local appearance-based features into pose estimation procedures is discerned. Matched features from 2D images are combined in order to produce the 3D model of a pre-recognized object. In (Wu, C. et al. 2008) a method that is able to compute camera poses from single query images and to efficiently search for 3D models in a city-scale database is presented. It employs Viewpoint Invariant Patches (VIP) that are based on the creation of ortho-textures for the 3D models and on the detection of local features, e.g. SIFT or SURF, on them. In turn, Time-of-Flight (ToF) cameras can be used for the precise 3D environment mapping, as it was shown in (May, S., et al 2008). Camera's pose was estimated using visual odometry, whilst SIFT was employed, among other feature extraction mechanisms, during the registration process.

Additionally, in (Andreopoulos, A., & Tsotsos, J. 2009) a model for actively searching for a target in a 3D environment, which incorporates both multi-view and single view recognition and detections schemes, is presented. However, the results stand only in theory when sophisticated vision systems required realistic and practical measurements of targets' location in the 3D space. On the contrary, in (Husler, G., & Ritter, D. 1999) it was shown that it is possible to estimate objects' location in a scene by combining 3D-based target models with information derived by a single 2D image. The final 6 DOF localization is obtained by accumulating the correspondences between extracted features (edges, corners or centers of

ellipses) and the 3D target on a Hough table. On the contrary, the main drawbacks of the method include both its deficiency to recognize and estimate the location in the 3D space of non-rigid objects and the extraction only low-level features.

Due to the large amount of pose estimation methods available in the literature, this chapter's section is devoted to the presentation of those techniques that exploit information derived through object recognition procedures based on local features. As a result, in the following paragraphs respective methods for objects' pose estimation or localization are presented. At the last part of the section, the merits and the disadvantages of the referred methods are noted.

Objects' Depth Estimation Using Any Two-Part Recognition Technique (Kouskouridas, R., et al. 2010)

The main idea underlying this algorithm is the maintenance of any two-part approach's properties whilst it exploits spatial information to derive depth. The method initiates with the construction of a large database containing images from several objects. With a view to database's enrichment, these objects are photographed from different viewpoints and distances from the camera. Moreover, by taking into account SIFT's and SURF's matching sub-procedures an on-line scene search engine is built. Estimations derived from this engine are taken into consideration for the position estimation task.

The algorithm is divided into five discriminative stages, when during the first one the detector mechanism is applied to the scene's and object's image, in order to estimate the features position in both of them. This stage could be apprehended as the training session of the algorithm. In this phase, for each image in the database the key-point features are extracted using the detector mechanism. Images of each object are captured at different distances from the camera and the measured depth d_O is stored. This process is performed while the system remains off-line, thus, execution time is not taken into account. The results are stored for further use at the next phases.

During the second phase, the N matching features in the two images are obtained by applying the matching sub-procedure of the two-part algorithm. Define as (X_{Si}, Y_{Si}), i=1,…,N the positions of the N features in the scene image and (X_{Oi}, Y_{Oi}), i=1,…,N the positions of the N features in the object image. Specifically, in this stage the matching sub-procedure of the two-part algorithm is performed. Especially, descriptors that are common in both images (scene and object) are extracted. It is apparent that, one image representing the scene is compared with several others, representing the object from different viewpoints. Furthermore, the locations of the common features are stored for further use. Afterwards, define as (X_{Sc}, Y_{Sc}) and (X_{Oc}, Y_{Oc}) the features' centers of mass for both images. This is accomplished by estimating the mean values of the features positions in the two images:

$$X_{Sc} = \frac{1}{N}\sum_{i=1}^{N} X_{Si} \text{ and } Y_{Sc} = \frac{1}{N}\sum_{i=1}^{N} Y_{Si}$$

$$X_{Oc} = \frac{1}{N}\sum_{i=1}^{N} X_{Oi} \text{ and } Y_{Oc} = \frac{1}{N}\sum_{i=1}^{N} Y_{Oi}$$

During the third stage, the position estimation sub-procedure takes place till the conclusion of the algorithm. Moreover, at this phase the features' centers of mass in both images are calculated. The last are obtained by estimating the mean values of features locations in both representations. As a next step, calculate the mean Euclidean distance (in pixels) of each feature from the corresponding center of mass that is extracted in the previous stage. Set as E_S and E_O the mean Euclidean distances in the scene and object image, respectively. The following relations are used:

$$E_S = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(X_{Si} - X_{Sc})^2 + (Y_{Si} - Y_{Sc})^2}$$

$$E_O = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(X_{Oi} - X_{Oc})^2 + (Y_{Oi} - Y_{Oc})^2}$$

During this phase, the distance of each key-point from the centre of mass is calculated. This is measured in pixels with the use of Euclidean Distance. By the end of this sub-routine, significant spatial information of an object in a scene is collected. This is accomplished by simply estimating the distribution of trained features around their center of mass. Finally, estimate d_S that correspond to the ratio of the two mean distances E_S and E_O. Furthermore, the pre-computed depth d_O is introduced, which is obtained during the training session and while the object is captured alone. The final object's distance (Z) from the camera is obtained by multiplying this ratio with the respective sought object's distance from the sensor (d_O):

$$Z = d_O \cdot d_S = d_O \cdot \frac{E_O}{E_S}$$

During the last phase of the algorithm, object's distance from the camera is computed. The pre-computed depth d_O, measured during the training session of the first phase, is taken into account. The ratio d_S is used to measure the proportion of object's features to those found in the scene. Concluding, after the necessary training session and the database construction at the first stages of the method, an on-line search engine takes over. It is responsible for querying in the scene for objects contained in the trained database. After an object is found, the scene is compared to the image of the object, providing the majority of common matches. Finally, features' information from both images is interpolated with a view to object's position allocation. This approach excels in simplicity and computational cost, whilst its database can be easily modified for the needs of multiple-object recognition and location assignment.

Real-Time 3D Object Pose Estimation and Tracking for Natural Landmark Based Visual Servo (Choi, C., et al 2008)

This method combines scale invariant features matching with optical flow based tracking, KLT tracker, for real-time 3D pose tracking. SIFT matching process makes the system robustly estimate an initial pose of a target object due to the fact that it is invariant to scale, rotation, illumination, and partial change in 3D viewpoint. To overcome SIFT's computational burden this difficulty, this approach adopts KLT tracker to calculate 3D pose consecutively from initially estimated pose. This tracker is already well known method, whilst there are several implementations available (Bouguet, J.-Y. 2000). The advantages of the KLT tracker are in that it is free from prior knowledge and computationally cheap. On the other hand, possible illumination changes or occlusions directly affect the efficiency of the method. In addition, KLT tracking points are likely to drift due to the aperture problem of optical flow. The system excels in removing outliers to get accurate pose results.

The method is able to estimate objects' pose utilizing both mono images and stereo pairs, whilst the whole process is divided into two separate phases. The first one is devoted to the initial pose estimation of the object when the second emphasizes in calculating targets' orientation locally. The system initiates by taking several images of the target and calculating 3D points from a structured light system. This off-line data is taken into account and accompanied with the POSIT algorithm (Intel 2006) produce the initial pose hypothesis. As a next step, KLT tracking points are built around those of SIFT and 3D reference

features are stored for further exploitation at the later stages. The most important issue that arises is that the tracked points are generated only inside the convex hull of a set of the matched SIFT features with the respective criteria proposed in (Jianbo, S., & Tomasi, C. 1994).

In turn, during the second phase, the correspondences between the 3D reference points and those of the KLT are taken into account. As a result, the system is able to both quickly track a possible target and estimate its pose. However, this process produces a large number of outliers that affect directly the efficiency of the algorithm. To this end, with a view to overcome this problem, the RANSAC algorithm is utilized and, as a result, the outliers are eliminated. The algorithm executes until the number of the remained tracking points are less than a threshold. In cases where the number of key-points available is not adequate, the system returns to the first phase in order to execute the initial pose estimation process, which in turn, restarts SIFT's matching procedure globally.

Robust Pose Estimation with 3D Textured Models (Gall, R.B.J., & Seidel, H.P. 2006)

This method implies that assuming a prior knowledge of object's 3D model, the final pose estimation is mainly based in correspondences between some 2D features located in the images and their actual positions on the 3D model. Moreover, the system is enhanced by simply incorporating texture information over the targets' surface, which in turn, provides more accurate and reliable correspondences. The system initiates by assuming a prior knowledge of the object's pose for frame t-1, where a textured 3D model is generated. The latter is formed in the same world coordinate system used for the calibration of the cameras, where the calibration matrices are converted to the model-view and projection matrix representation of OpenGL. It is important to keep in mind that image sequences are undistorted by hand and a number of initial views of the 3D model (by rotating and storing the respective extracted features) are rendered. This phase terminates by projecting the 3D model onto the image plane according the calculated calibration matrix.

During the second stage, at frame t, numerous features from both the objects' rendered images and those depicting the scene are extracted using PCA-SIFT (Ke, Y., & Sukthankar, R. 2004). Afterwards, the reliable correspondences needed for the accurate pose estimation are established by assuming that each 2D point lies inside or on the border of a projected triangle of the 3D mesh. Thus, considering an affine transformation, the respective triangle for a point is efficiently determined by a look-up table comprising of information concerning the color index and the vertices of each triangle.

The next phase is devoted to the accurate estimation of object's orientation relative to the one at frame t-1. The contour based philosophy of the method is mainly responsible to the large amount of outliers contained in the feature correspondences. The latter are minimized by adopting a least squares method, which involves the estimation of the object's motion. In cases where not enough feature correspondences are extracted by PCA-SIFT an auto-regression model is utilized for the adequate estimation of the target's pose.

Finally, an image segmentation technique that involves color and texture information (Brox, T., et al. 2003) is utilized for the extraction of the object's contour. Thus, by simply matching the latter with the projected contour of the model (using the closest point algorithm (Zhang, Z. 1994)) generates new feature correspondences between the 3D model and the 2D image. The latter correspondences accompanied with those extracted from the PCA-SIFT are used for the final object's pose estimation at frame t.

Robot-Vision Architecture for Real-Time 6-DOF Object Localization (Sumi, Y., & Ishiyama, Y., & Tomita, F. 2007)

This method proposes a robot - vision architecture, called Hyper Frame Vision (HFV), which is able to both detect moving objects and estimate their 6-DOF motion. The recognition scheme is mainly dependent on a feature extractor based on information derived from edges. Practically, the presented detection framework implements the same procedure followed at the SIFT's detector. The system consists of calibrated stereo cameras for 3D sensing and stereo-vision-based identification (Sumi, Y., et al. 2002) and tracking algorithms. The latter require 3D edge segments (known to be robust against possible illumination changes) that are provided by the stereo cameras in terms of extracted features for object localization. Moreover, in order to achieve real time execution, both algorithms are implemented as independent software modules requiring stereo image sequences as input.

The stereo cameras used in the HFV system continuously capture a stereo image sequence, which is buffered into the large frame memory. In turn, the object recognition task is performed by the identification module, which chooses the latest frame of stereo images that includes the object. Furthermore, this module is responsible for the model matching procedures and the final 6-DOF localization of the object. Targets' tracking is adequately fulfilled by the respective algorithm that exhibits frame-by-frame operations, which, in turn, result in the accurate objects' motion estimations.

Generally, the method is divided into three discriminative steps. The first one is responsible for the object localization, which is accomplished by adopting the identification algorithm at frame t_{st}. During the task, all input stereo images are buffered in the frame memory. Define p_{id} as the processing time of the identification task, whilst T(i) as a 3x4 transformation matrix representing the 6-DOF object's position in the i^{th} frame. It is apparent that, the task terminates at time $t_{st} + p_{id}$, whilst an object is identified using the respective confidence value. The latest process of this first step involves the data transferring between the identification module and the tracking one.

During the second phase, the tracking module takes place that requires as input the buffered image sequence provided by the first stage. When the tracking module receives the identification result $T(i_0)$ as the initial position of the object, it starts a new tracking task from the frame i_0+1 to measure the object's motions $T(i = i_0+1,i_0+2,\ldots)$ and to calculate the confidence values. Finally, the tracking frame catches up with the input frame (at frame k in the time diagram). It is important to keep in mind that the only restriction of the tracking module is the frame-by-frame period, p_f, which is mainly depended on the working computer.

The final step emphasizes in the real-time object motion tracking, where the tracking algorithm is made to wait for the latest images. As a result, if p_{tr} is the average processing time per frame of the tracking process, then $p_{tr} < p_f$. In instances where the latter is not satisfied the HFV system repeats the process presented in step 2, which in turn invokes re-estimation of object's motion and confidence values. Furthermore, if a tracking failure is detected then the system terminates and returns to the initial step 1.

Combination of Foveal and Peripheral Vision for Object Recognition and Pose Estimation (Bjorkman, M., & Kragic, D. 2004)

This method presents a biologically inspired one, capable of adequately accomplish object recognition and pose estimation tasks. A real-time vision system that integrates a series of algorithms for object recognition, tracking and pose estimation tasks is presented. Furthermore, both monocular and binocular cues are taken into account by using

one set of stereo cameras for object recognition tasks and one camera responsible for object's tracking and pose estimation. SIFT is mainly used for the recognition purposes, although, its' spatial attributes are exploited during the pose estimation procedure. The most important and interesting issue constitute the fact that this method involves a stereo-head where the sensors are positioned respectively.

The system is divided into four individual modules that are responsible for specific tasks. The first sub-system is called Visual Front-End and is devoted to the figure-ground segmentation, in order to obtain constant flow of reliable data from the surrounding environment. Moreover, this part of the system extracts metric information, i.e. sizes and distances concerning objects and obstacles, by adopting a stereo-based philosophy. Since most efficient methods for dense disparity estimation assume the image planes to be parallel, rectification is performed using epipolar geometry's attributes. Fragmentally, this module could be apprehended as a three-step process, which includes epipolar geometry estimation, image rectification and calculation of dense disparity maps.

The second subsystem emphasizes in generating a number of hypotheses about the objects in the scene that may be relevant to the task in hand. Practically the purpose of this component is to derive qualified guesses of where a requested object might be located in the current scene. In turn, the third module corresponds to the object recognition framework. For the adequate accomplishment of this task, the system adopts the SIFT algorithm and an appearance based module that relies on color histograms. Co-occurrence Color Histograms are utilized in a classical learning framework that facilitates a winner-takes-all strategy across scales.

The last module of the system is referred as Action Generation and is responsible for triggering visual tracking and pose estimation. The last require input derived from the aforementioned modules, including recognition and hypotheses generation. As far as the tracking is concerned, it is based on the accumulation of several visual cues including motion, colors and gradients, whilst the framework incorporates a "voting" procedure. On the other hand, pose estimation is adequately accomplished by taking into account information derived through SIFT and by adopting the technique presented in (Kragic, D., & Christensen, H. I. 2003).

Comparison

The following section is devoted to the qualitative and quantitative results that are extracted from both surveys, i.e. the one concerning the object recognition algorithms and the other over the pose estimation methods. In the first case, the amount of recognition algorithms available generates numerous applicable solutions to several target detection problems. The most accurate and reliable comparison studies available in the literature are presented in (Mikolajczyk, K., & Schmid, C. 2004, Mikolajczyk, K. et al. 2005, Mikolajczyk, K., & Schmid, C. 2005), where several detectors and descriptors are experimentally evaluated through numerous tests. Fragmentally, in general the SIFT algorithm outperforms almost any two-part approach when its disadvantages constitute its computational cost and its descriptors' complexity. On the other hand, and remaining at the two-part strategies, the SURF method can be adopted in procedures that require real time execution due to the fact that provides adequate feature correspondences accompanied with low computational burden. Moreover, SIFT extract features that are distributed more over the surface of the object in cases where SURF only detects points that mostly lay at the center of the image. As far as the non two-part approaches are concerned, several detectors have been proposed, whereas all of them provide adequate fulfillment of the feature extraction process. Depending on the task, a computer vision researcher is able to select several solutions from a large deposit of available ones.

In turn, the task of finding an object located in a scene only constitutes the first step of a larger scale framework that emphasize in providing more information concerning the object. Especially, computer vision researchers aim at providing a sense of location for any potential application, such as robotic platforms or automated driving vehicles. To this end, state of the art research is devoted to the accurate estimation of objects' – targets' orientation relative to a pre-defined coordinate system. Through this chapter several methods that produce effective solutions to this problem have been presented. Moreover, these methods and their main building blocks have been experimentally assessed, whilst numerous qualitative and quantitative results have been extracted. The latter are summed in Table 1. The estimation error actually corresponds to each method's efficiency, whilst the "dealing with occlusions" topic examines algorithms' solutions to the respective problem. Furthermore, execution time refers to the computation time and it is, obviously, related to the computational burden of the framework. Finally, the pose estimation schemes are collated according to their feasibility. Despite the fact that an algorithm have been shown to provide remarkable results, it is possible that its implementation constitutes its adoption prohibitory.

CONCLUSION AND FUTURE WORK

Throughout this chapter two major topics of the computer vision society were assessed, i.e. the

Table 1. Accumulated qualitative and quantitative results concerning the pose estimation methods already presented in Section 3

Method	Estimation Error	Dealing with occlusions	Execution time	Implementation – feasibility
Kouskouridas, R. et al. 2010	< 9%	Partial occlusions reduce algorithm's efficiency to 70%	~ 2 seconds per object	Very easy to implement since it requires only a two-part recognition approach
Choi, C. et al 2008	< 15%	Partial occlusions affect directly the efficiency of the algorithm (the system fails to estimate the object's pose)	Real-time	Requires either stereo or mono image sequences captured with Bumblebee camera. (use of KLT feature tracker)
Gall, R.B.J., & Seidel, H.P. 2006	< 50% in cases where only the contour of the 3D model is taken into account	Partial occlusions affect directly the efficacy of the algorithm (failure to estimate 3D model's contour)	Very time demanding operation with high computational burden	Almost infeasible since it involves 3D object modeling in OpenGL and enough time devoted to the off-line training of the system
Sumi, Y., et al. 2007	< 10% (The confidence value alters drastically depending on the constant rotational velocity)	The system fails to initially track and afterwards estimate target's 6-DOF location in cases where the latter is occluded spontaneously	Real-time	Feasible enough although it requires a stereo camera and a large amount of memory buffer available
Bjorkman, M., & Kragic, D. 2004	< 25% (Possible object's rotation over 20 degrees leads to efficiency reduction)	This is the most heavily influenced by occlusions method, since information derived through SIFT are lost in such cases	~ 6 sec per object	CCH's require large amount of time dedicated to off-line training of the system

object recognition and pose estimation. As far as the recognition process is concerned, the current trend implies the usage of local appearance-based detection schemes that rely on locally sampled descriptions of the target. The detectors and descriptors constitute the two main mechanisms underlying any recently proposed recognition framework. Moreover, SIFT and SURF that are referred as two-part approaches, are constantly adapted to numerous of robotics applications providing remarkable solutions to several vision – based problems.

On the other hand, a pose estimation procedure aims at making one step further from a typical recognition scheme in terms of providing spatial information to the sought targets. All the presented methods are based on feature correspondences between 2D image features and the respective 3D ones located in the real world. Moreover, the frameworks presented utilize local appearance based recognition techniques (e.g. SIFT), whilst the pose estimation task is fulfilled by exploiting the extracted features' attributes. Generally, the overall estimation error mainly depends on general assumptions made during the first stages of the algorithms or, in the best case, on the re-projection error of the camera.

Finally, the ultimate goal of the chapter was to both present the state-off-the-art in the field of object recognition and pose estimation algorithms, and to actually, point out possible solutions in real problems. Towards this end, in this chapter the most important methods are analytically presented. During the last few years, a tendency to introduce autonomous vehicles into domestic environments is discerned. Thus, due to the fact that only the visual sense provides the majority of the information available, computer vision researchers have to outcome the challenge of making simple and easy to build solutions to everyday tasks.

REFERENCES

Andreopoulos, A., & Tsotsos, J. (2009). A theory of active object localization. In *The Proceedings of the International Conference on Computer Vision, Poster Session*.

Baumberg, A. (2000). Reliable feature matching across widely separated views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* (pp. 774-781). Hilton Head, USA.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *International Journal of Computer Vision and Image Understanding*, *110*(3), 346–359. doi:10.1016/j.cviu.2007.09.014

Bjorkman, M., & Kragic, D. (2004) Combination of foveal and peripheral vision for object recognition and pose estimation. In *Proceedings of the IEEE International Conference on Robotics and Automation,* (pp. 5135-5140).

Borges, A. P., Ribeiro, R., Avila, B. C., Enembreck, F., & Scalabrin, E. E. (2009). A learning agent to help drive vehicles. In *Proceedings of the International Conference on Computer Supported Cooperative Work in Design*, (pp. 282-287).

Bouguet, J.-Y. (2000) *Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm.* Technical Report, Intel Corporation, Microprocessor Research Labs, OpenCV documentation.

Bradski, G. R., & Kaehler, A. (2008). *Learning OpenCV* (1st ed.). Sebastopol, CA: O'Reilly Media, Inc.

Brox, T., Rousson, M., Deriche, R., & Weickert, J. (2003). Unsupervised segmentation incorporating colour, texture, and motion. In *Computer analysis of images and patterns*, (LNCS 2756, pp. 353-360).

Choi, C., Baek, S. M., & Lee, S. (2008). Real-time 3D object pose estimation and tracking for natural landmark based visual servo. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems,* (pp. 3983-3989). Nice, France.

Ekvall, S., Kragic, D., & Hoffmann, F. (2005). Object recognition and pose estimation using color cooccurrence histograms and geometric modeling. *International Journal of Image and Vision Computing*, *23*(11), 943–955. doi:10.1016/j.imavis.2005.05.006

Gall, R. B. J., & Seidel, H. P. (2006). *Robust pose estimation with 3D textured models*. Lecture Notes in Computer Science.

Harris, C., & Stephens, M. (1988). A combined corner and edge detection. In *Proceedings of the Fourth Alvey Vision Conference* (pp. 147-151). Manchester, UK.

Husler, G., & Ritter, D. (1999). Feature-based object recognition and localization in 3D-space, using a single video image. *International Journal of Computer Vision and Image Understanding*, *73*(1), 64–81. doi:10.1006/cviu.1998.0704

Intel. (2006). *Open source computer vision library*. Retrieved from http://www.intel.com/research/mrl/research/opencv/

Jianbo, S., & Tomasi, C. (1994) Good features to track. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition,* (pp. 593-600).

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105. doi:10.1023/A:1012460413855

Ke, Y., & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (pp. 506-513).

Kouskouridas, R., Badekas, E., & Gasteratos, A. (2010). (in press). Evaluation of two-parts algorithms for objects' depth estimation. *IET Computer Vision*.

Kragic, D., Bjorkman, M., Christensen, H., & Eklundh, J. (2005). Vision for robotic object manipulation in domestic settings. *International Journal of Robotics and Autonomous Systems*, *52*(1), 85–100. doi:10.1016/j.robot.2005.03.011

Kragic, D., & Christensen, H. I. (2003). Confluence of parameters in model-based tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Taipei, Taiwan.

Lazebnik, S., & Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1265–1278. doi:10.1109/TPAMI.2005.151

Lepetit, V., & Fua, P. (2006). Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1465–1479. doi:10.1109/TPAMI.2006.188

Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *International Journal of Applied Statistics*, *21*(2), 414–431.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. doi:10.1023/B:VISI.0000029664.99615.94

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *International Journal of Image and Vision Computing*, *22*(10), 761–767. doi:10.1016/j.imavis.2004.02.006

May, S., Droeschel, D., Holz, D., Wiesen, C., Birlinghoven, S., & Fuchs, S. (2008). 3D pose estimation and mapping with time-of-flight cameras. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS), 3D Mapping Workshop*, Nice, France.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86. doi:10.1023/B:VISI.0000027790.02288.f2

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630. doi:10.1109/TPAMI.2005.188

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., & Schaffalitzky, F. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, *65*(1-2), 43–72. doi:10.1007/s11263-005-3848-x

Neubeck, A., & Van Gool, L. (2006). Efficient non-maximum suppression. In *Proceedings of the International Conference on Pattern Recognition,* (pp. 850-855).

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. (pp. 2161-2168). New York, USA.

Rosten, E., & Drummond, T. (2006). *Machine learning for high-speed corner detection* (pp. 395–430). Lecture Notes in Computer Science.

Rothganger, F., Lazebnik, S., & Ponce, J. (2006). 3D object modeling and recognition from photographs and image sequences. In *Proceedings toward Category-Level Object Recognition* (pp. 105-126).

Sandhu, R., Dambreville, S., Yezzi, A., & Tannenbaum, A. (2009). Non-rigid 2D-3D pose estimation and 2D image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 786-793). Miami, USA.

Sarfraz, M. S., Hellwich, O., Yilmaz, U., Bellmann, A., Rodehorst, V., & Erten, E. (2008). Head pose estimation in face recognition across pose scenarios. In *International Conference on Computer Vision Theory and Applications* (pp. 235-242). Funchal, Portugal.

Schaffalitzky, F., & Zisserman, A. (2002). *Multi-view matching for unordered image sets, or How do I organize my holiday snaps?* (pp. 414–431). Lecture Notes in Computer Science.

Schmid, C., & Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 530–535. doi:10.1109/34.589215

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision* (pp. 1470-1477). Nice, France.

Sumi, Y., Ishiyama, Y., & Tomita, F. (2007). Robot-vision architecture for real-time 6-dof object localization. *International Journal of Computer Vision and Image Understanding*, *105*(3), 218–230. doi:10.1016/j.cviu.2006.11.003

Sumi, Y., Kawai, Y., Yoshimi, T., & Tomita, F. (2002). 3D object recognition in cluttered environments using segment-based stereo vision. *International Journal of Computer Vision*, *46*(1), 5–23. doi:10.1023/A:1013240031067

Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., & Van Gool, L. (2009). Shape-from-recognition: Recognition enables meta-data transfer. *International Journal of Computer Vision and Image Understanding*, *113*(12), 1222–1234. doi:10.1016/j.cviu.2009.03.010

Vaish, V., Levoy, M., Szeliski, R., Zitnick, C., & Kang, S. (2006). Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the International Conference on Pattern Recognition*.

Van Gool, J., Moons, T., & Ungureanu, D. (1996). Affine/photometric invariants for planar intensity patterns. In *Proceedings of the European Conference on Computer Vision (*pp. 642-651). Cambridge, UK.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 511-518).

Wong, B., & Spetsakis, M. (2000). Scene reconstruction and robot navigation using dynamic fields. *International Journal of Autonomous Robots*, *8*(1), 71–86. doi:10.1023/A:1008992902895

Wu, C., Fraundorfer, F., Frahm, J., & Pollefeys, M. (2008). 3D model search and pose estimation from single images using VIP features. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 1-8).

Zhang, J., McMillan, L., Yu, J., & Hill, U. (2006) Robust tracking and stereo matching under variable illumination. In *Proceedings of the International Conference on Pattern Recognition.*

Zhang, Z. (1994). Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, *7*(3), 119–152. doi:10.1007/BF01427149

Zwicker, M., Vetro, A., Yea, S., Matusik, W., Pfister, H., & Durand, F. (2007). Resampling, antialiasing, and compression in multiview 3-D displays. *IEEE Signal Processing Magazine*, *24*(6), 88–96. doi:10.1109/MSP.2007.905708

ADDITIONAL READING

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Wiley-Interscience.

Forsyth, D. A., & Ponce, J. (2002). Computer vision: A modern approach (US ed.). Prentice Hall Professional Technical Reference

Hartley, R. I., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511811685

Paragios, N., Chen, Y., & Faugeras, O. (2005). *Handbook of mathematical models in computer vision*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Steger, C., Ulrich, M., & Wiedemann, C. (2008). *Machine vision algorithms and Applications*. Wiley VCH.

Treiber, M. A. (2010). *An introduction to object recognition: Selected algorithms for a wide variety of applications (Advances in pattern recognition)*. Springer.

KEY TERMS AND DEFINITIONS

3D Position Estimation: The task of identifying and estimating the absolute position of an object in the 3D space. At least two different views are required, whilst multi-camera systems provide the majority of the information needed.

Computer Vision: The science devoted to the design and implementation of process that

emphasize in making machines capable of sensing what is visually perceived. It is directly related to the extraction of information contained in images, whilst algorithms of this field try to decode vital visual information contained in natural environments.

Depth Calculation: The procedure of calculating a target's distance from the capturing device (camera). Stereo-vision algorithms have been proven to provide the most efficient solutions to this problem.

Descriptor: Organizes the information collected from the detector in a discriminating manner. Therefore, high dimensional feature vectors corresponding to locally sampled feature descriptions are produced. In other words, an object, or parts of it, located in a scene are represented by these vectors, namely the descriptors.

Detector: A mechanism contained in advanced object recognition algorithms. The main idea behind interest location detectors is the pursuit of points or regions in a scene containing unique information. These spots or areas hold data that distinguish them in their local neighborhood from any other.

Object Manipulation: The task of handling of objects usually via a robotic arm. In order to adequately accomplish manipulation tasks, computer vision algorithms emphasize in estimating the necessary spatial information of the target along with the accompanied grasping positions of the object.

Object Recognition: The process of querying an image for a specific target. Illumination circumstances (resulting in shadows) along with possible object occlusions affect directly the efficiency of the respective algorithms.

Pose Estimation: The task of estimating recognized object's orientation and position relative to a given coordinate system. Generally, the goal of this process is to calculate the 6 Degrees of Freedom (rotation and translation matrixes) of an object relative to a specific frame. Information extracted is utilized in either manipulation tasks or obstacle avoidance ones.