# Chapter 1

# Perception and Computer Vision

The wish to build artificial and intelligent systems results in the expectation that they are placed in our typical environments. Hence, the expectations on their perceptual capabilities are high. Perception refers to the process of becoming aware of the elements of the environment through physical sensation, which can include sensory input from the eyes, ears, nose, tongue, or skin.

In this Chapter we focus on visual perception, which is the dominant sense in humans and has been used from the first days of building artificial machines. Two early examples are Shakey, a mobile robot with range finder and camera to reason about its actions in a room with a few objects [58], and FREDDY, a fixed robot with a binocular vision system controlling a two-finger hand [2] (also refer to Section 13 on Robotics).

The goal of computer vision is to understand the scene or features in images of the real world [6, 29]. Important means to achieve this goal are

the techniques of image processing and pattern recognition [28, 30]. The analysis of images is complicated by the fact that one and the same object may present many different appearances to the camera depending on the illumination cast onto the object, depending on the angle from which it is viewed, the shadows it casts, the specific camera used, if object parts are occluded, and so forth.

Nevertheless, today computer vision is sufficiently well advanced to detect specific objects and object categories in a variety of conditions, to enable an autonomous vehicle to drive at moderate speeds on open roads, a mobile robot to steer through a suite of offices, and to observe and understand human activities.

The objective of this Chapter is to highlight the state of the art in computer vision methods that have been found to operate well and that led up to the above mentioned capabilities. After a short discussion of more general issues, we summarise work structured into five key topics: object recognition and categorisation, tracking and visual servoing, understanding human behaviour, and contextual scene understanding. We conclude with a critical assessment of what computer vision has achieved and what challenges remain open.

## 1.1 Computer vision paradigms & principles

Computer vision is a heterogeneous field that embraces a large spectrum of methods as well as scientific perspectives. This starts with the physical understanding of the plenoptic function that describes how the light gets refracted, reflected, scattered, or absorbed with regard to a scene (Fig. 1.1). The plenoptic function is a theoretical construct that specifies the illumi-

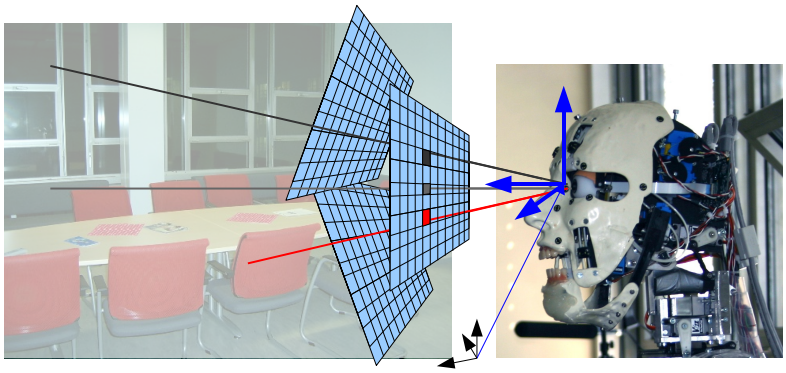


Figure 1.1: An image is given by a 2D pixel array where each pixel measures the amount of light traveling along a ray. The plenoptic function would specify this for each possible viewing point and viewing angle.

nation for each possible ray of light in the scene. However, this function is typically not known nor is the scene behind it. Computer vision aims at reconstructing relevant aspects of both in order to solve tasks from the visual measurements of a camera. In this regard, computer vision solves the inverse problem of computer graphics. A second perspective on computer vision is to mimic biological vision in order to get a deeper understanding of involved processes, representations, and architectures. Here, it is becoming more and more obvious that the fundamental questions and open problems in computer vision are at the cutting edge of cognition research. They cannot be solved in isolation but concern the fundamental basis of cognition itself. A third perspective understands computer vision as an engineering discipline that aims at the solution of practical vision tasks. But instead of a systematic methodological approach, the current state-of-the-art is mainly dominated by heuristics and knowledge from experience. All three perspectives cannot be separated and deeply influence each other, which – together with an

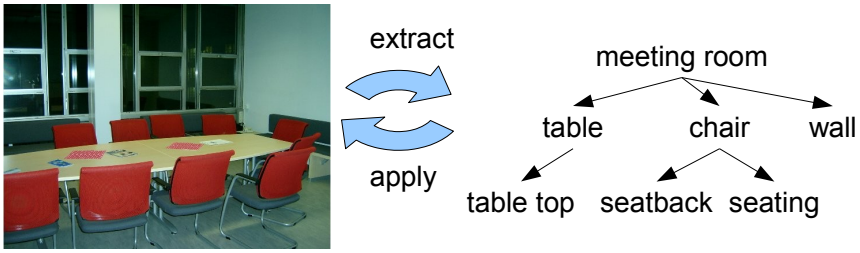



Figure 1.2: Computer vision as a knowledge engineering task

immense technical progress – has made computer vision a highly dynamic field over the last 50 years. In order to solve specific computer vision tasks, different design decisions need to be made. Some of these are pointed out in the following.

**What kind of knowledge is needed?** In order to understand the content of an image, relevant parts of it need to be linked to semantically meaningful concepts. For the scene of a meeting room (Fig. 1.2) the knowledge base might include that it consists of a large table and a couple of chairs positioned around it, that a table has a table top, etc. The knowledge base can be used in two different ways. In a bottom-up process it guides the construction of higher-level concepts from primitive parts, or it is exploited in a top-down process in order to predict structures expected in the image. This has led to a bunch of interesting work in the 70s and 80s [6, 16, 70].

**How to represent scene geometry?** Scene geometry is an important intermediate representation in the interpretation process of an image. It can be dealt with either in 2D or 3D. In Fig. 1.3 a depth image is generated, first. This can be computed from pairs of stereo images or directly be measured by e.g. Time-of-Flight sensors. Because the representation is still

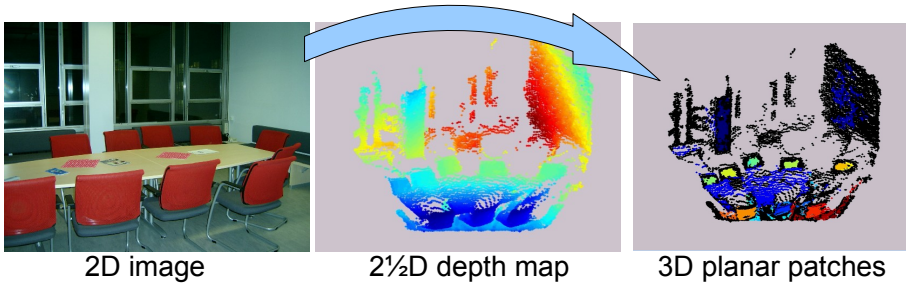



Figure 1.3: 3D scene geometry: the image in the middle is showing the reconstructed depth coded in colors; the right image shows 3D points that have been grouped to planar patches in the same color coding.

view-dependent, it is also called $2\frac{1}{2}$D. In a next step, 3D geometric primitives are fitted into the scene providing a view independent, object-centered representation. Such kind of approach was already suggested by David Marr (1982) who also looked at the concepts of human vision known at his time [52]. However in many cases, the extraction of 3D geometry is too fragile, so that more stable geometric representations are directly extracted from the 2D image. Therefore, images are typically analyzed with regard to spatial discontinuities in the gray-level or color-surface. Representations either focus on homogeneous image patches (regions) or on edges (border lines) (Fig. 1.4). Both provide a basis for further interpretation processes. The extraction of such geometric primitives is a problem of digital image processing [30].

**What are appropriate features?** In order to match a geometric or image representation to a semantic concept, like "table", "chair", or "meeting room", one needs to specify a decision function that decides for or against a membership of a class $\omega$. This is a *classification problem* that is intensively dealt with in the area of pattern recognition. A pattern is represented by

Figure 1.4: 2D scene geometry: the left image shows a region segmentation based on a color clustering; in the middle a Difference of Gaussian (DoG) filter has been applied to the original image; the right image shows a contour segmentation based on the DoG image.

a high-dimensional feature vector $x$ and the decision function $d(x)$ is typically trained on a large set of training examples $\{(x_i, \omega_i)\}_{i=1...N}$ where the corresponding class has been annotated (typically by hand). In Fig. 1.5 a simplistic example is given. The image is divided into 6 parts and for each sub-image a color histogram is computed. The concatenated histograms provide a feature vector that can be used, e.g., for classification of specific meeting rooms. The invention and design of appropriate visual features is a long standing discussion and had always a deep impact on the whole field of computer vision, like the use of color histograms [72], Eigenfaces [74], Hair-like features [79], or the scale invariant feature transform (SIFT) [50].

**How to control the acquisition process?** Biological vision is not a passive interpretation process, nor should it be for autonomous artificial systems. The movement of an agent in the real world basically determines the perception problem it has to solve. Vision is understood as an active process that includes the control of the sensor and is tightly coupled to the successful accomplishment of a decision or action [5]. This has certain consequences on

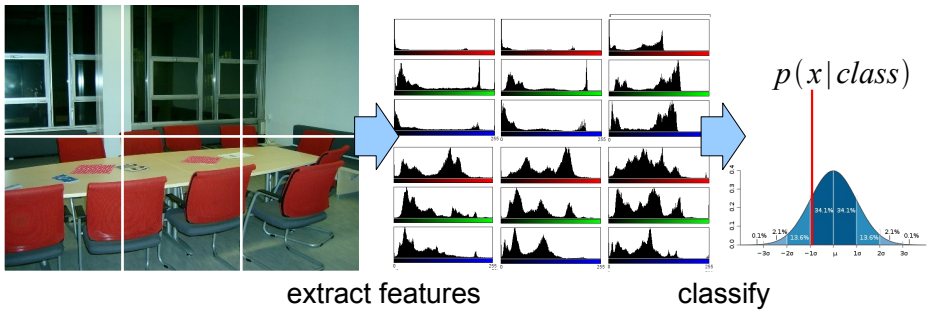
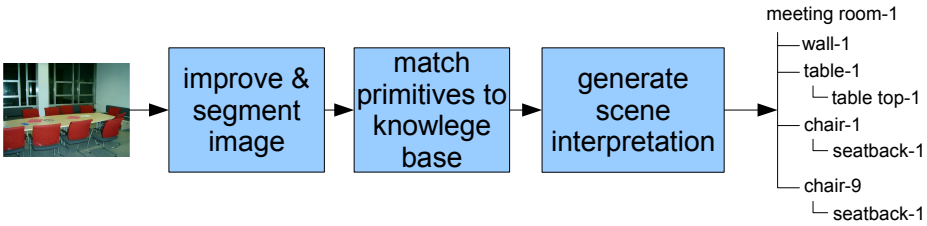



Figure 1.5: Pattern classification: a decision function for a specific class (e.g. meeting room) could be based on a maximum likelihood principle. Here the feature vector is defined using color histograms.

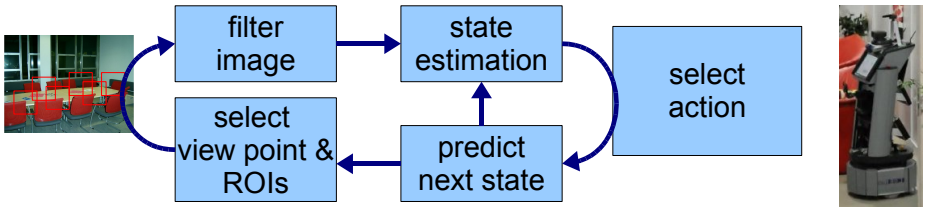
the design of computer vision systems that have already been noted in the early 90's [23]: (i) Instead of modeling an isolated image interpretation process, the system is *always running* and controls its own behaviour using an image stream. (ii) The overall goal of visual processing is not image understanding. Instead, the vision system work as a *filter* that extracts information relevant for its task. (iii) The system responses within a fixed time delay in order to be useful for its current task that needs to performed in (soft) *real time*. (iv) Instead of processing the complete image, the system focuses on a *region of interest* in order to meet the performance goals. The different perspectives are shown in Fig. 1.6. The first aims at a complete interpretation of the image, the second extracts relevant information for action selection and state prediction. More details can be also found in [44].

## 1.2 Object recognition and categorisation

Object Recognition can be seen as the challenge to determine the "*where*" and "*what*" of objects in a scene. A whole bunch of different techniques have




(a) Image understanding: Extraction of explicit meaningful descriptions




(b) Active vision: Control of the image acquisition and interpretation process with regard to a tasks

Figure 1.6: Two different perspectives on computer vision.

been proposed, that all have their own pros and cons. Given an application scenario, one has to carefully select an appropriate object recognition technique that fulfils the anticipated set of constraints. Techniques also differ in the precise problem that they solve.

Many of them are **object detectors** that post a yes/no question regarding the presence of an object class. The image is typically scanned by some kind of filter method that matches a kind of template model to a sub-image. Each different object parametrization needs a separate scan. More sophisticated approaches efficiently perform multiple passes on different scales and apply filters that are learnt from large sets of labelled images. A good example is the Viola-Jones detector [79], that has been widely used for face detection.

**Segmentation**-based techniques first extract a geometric description of

an object by a bottom up grouping process that defines the object's extension in an image. In a second step, they compute an invariant feature set for recognizing its object class or a set of generic primitives from which the objects are constructed. A classical example is given by Brooks [16] for the interpretation of aerial images. Modern techniques interleave or combine both steps in order to deal with over- and under-segmentation problems.

**Alignment** methods use parametric object models that are fitted to the image data [34] This can be performed top-down by some energy minimization technique or bottom-up by discrete voting techniques like the generalized Hough transform [7] and its variants [50].

All three approaches provide different information about objects in images and assume different kinds of pre-knowledge available.

### 1.2.1 2-D modelling

Most objects in the real world are inherently 3D. Nevertheless, many object recognition techniques stick to 2D representations with a significant success. The reasons for this are multifaceted: (1) *Easy accessibility:* We get 2D image information nearly for free using a standard camera equipment. (2) *Fast computation:* Features can directly be calculated from image pixel data and do not involve a search for complex geometric primitives. (3) *Simple model acquisition:* Models are typically learned from example images. (4) *Robustness to noise:* Features have a low degree of abstraction from pixel data. The detection of more abstract primitives typically involves segmentation issues that are error prone with regard to clutter and noise. (5) Furthermore, many interesting objects have quite characteristic 2D views, e.g. cover pages, traffic signs, side views of motor bikes or cars, front views of faces.

The price to pay for ignoring the 3D characteristics of objects are typically

over- or under-constrained models because there are a number of perspective variations that cannot be systematically dealt with. A typical case of under-constrained approaches are bag-of-feature models or histogram techniques which ignore the spatial distribution of features. Instead they discretize the feature space into bins and compute feature statistics [72]. Over-constraint models need multiple representations in order to deal with different part configurations or rotations of objects. Good examples are the template-based methods mentioned before. Additionally, we need to cope with a more challenging segmentation problem. Typically, 3D information provides a much stronger segmentation cue which has a much weaker correspondence in luminance values of 2D images.

The dominant class of 2D object recognition techniques are appearance-based approaches. Instead of using a view invariant object-centered representation, they represent different aspects of an object. Compact representation are provided by aspect-graphs [42, 46] that relate different two-dimensional appearances to each other in an efficient data structure. Secondly, appearance-based approaches drop an intermediate geometric representation level by computing features directly from pixel values. This has certain consequences on the kind of object classes that can be distinguished and the with-in class variations that can be covered. A well established method to encounter statistical variations in object appearance are parametric Eigenspaces that are applied for recognition of individual objects [56]. In the last years, the robustness of the method has been significantly improved [48].

So far, the discussed methods deal with variations of rotation, lighting, noise, and small distortions of an object's shape. They mostly assume that objects are solid, approximately rigid, have similar textures or colors, and are

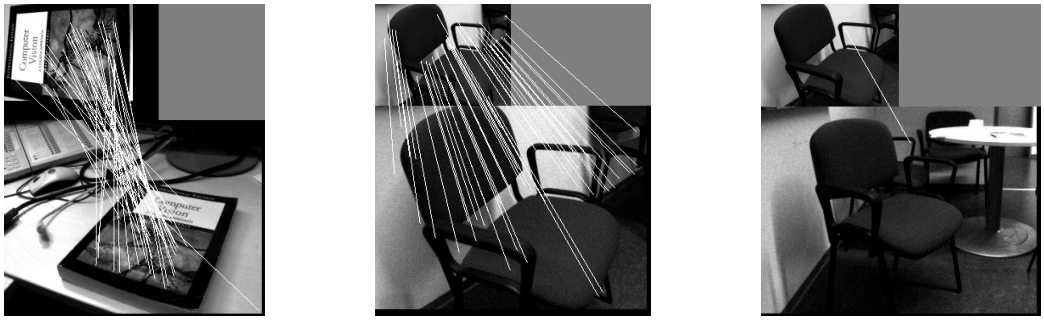


Figure 1.7: Matching result based on local descriptors (here SIFT [50]): First, salient points are computed on different scales. Then, the corresponding local descriptors are matched to a model database (given by the small image). Left is an ideal example of planar object that is highly textured. Middle and right examples show that the approach breaks down for less textures 3D objects if the perspective only slightly changes. In the right image only a single feature correspondence is found.

occluded to a minor degree. Further variations are covered by local descriptor approaches. Here, the main idea is to detect salient points in an image that provide a partial feature description instead of a complete appearance model. These approaches gained attention in the last 10 years and have reached a performance unachieved before. By relying on local descriptors (typical examples are SIFT or SURF features) these methods are able to cope with occlusion and local variations as they occur in real world settings [50, 38]. In Fig. 1.7 an example of such an approach is given. Pioneering work has already been conducted by Cordelia Schmid et al.[64] in 1996, who introduced scale and rotation invariant gray level features for image comparison.

### 1.2.2 3-D modelling

2D Colour or intensity images do not directly encode depth or shape information. Consequently object recognition and localization is a difficult problem and in general ill-posed [1]. To overcome these problems the 3D shape of objects can be directly recovered from range images. Range images can be obtained through various methods ranging from laser scanning over structured light approaches to stereo, which is the solution following the human example but the accuracy of active depth measurements is considerably higher.

The main question in computer vision is how to model or represent the object for detection in depth data. One way is parses shapes into component parts [66] and define their spatial relationships. In computer vision parts are useful for two reasons. First many objects are articulated and the part-based description allows to decouple the shapes of the parts from the spatial relationships. And second, not all parts of objects are seen but parts are often sufficient to recognise the object, e.g. a cup from either body or handle.

A key aspect of part-based representations is their number of parameters. In the past decade much work has been made describing range data with rotational symmetric primitives (sphere, cylinder, cone, torus) [53]. Generalized cylinders can be created by sweeping a two-dimensional contour along an arbitrary space curve [11]. The contour may vary along the curve (axis). Therefore, definitions of the axis and the sweeping set are required to define a generalized cylinder. An often cited early vision system that applied generalized cylinders is the ACRONYM system to detect aeroplanes [16]. However, parameterization and fitting are complicated.

Superquadrics became popular because a small set of parameters can describe a large variety of different basic shapes. Solina et al. pioneered work

in recovering single Superquadrics with global deformations in a single-view point cloud [67] and demonstrated that the recovery of Superquadrics from range data is sensitive to noise and outliers, in particular from single views as given in applications such as robotics. [36] summarises the recover and select paradigm for segmenting a scene with simple geometric objects without occlusions. This method aims at a full search with an open processing time incompatible to most applications such as robotics. Recently [45] shows the recovery of a known complex objects from their parts in a scene with occlusions and [10] shows that this can be done in real time. Finally, there exist several models with an open number of parameters such as implicit polynomials [40] and spherical harmonic surfaces [69]. They can adapt to arbitrary shapes and find usage in medical imaging or describing free-form surfaces. The advantage is that locally very different shapes can be described. Such local shape characteristics can be also used to recognize objects in range images [18].

Lately, stereo data is used more often to obtain 3D data. Since data is in general not as good as from laser scans, statistical methods rather than direct shape methods are employed. An example is the detection of chairs using spherical harmonics descriptor [83] shown in Fig. 1.8.

In summary, the recovery of Superquadrics has been investigated most. Open problems are to handle sparse data due to one-view scans of the scene and to cope with the typical laser and camera shadows and occlusions in cluttered scenes and the uncertainty of stereo images.

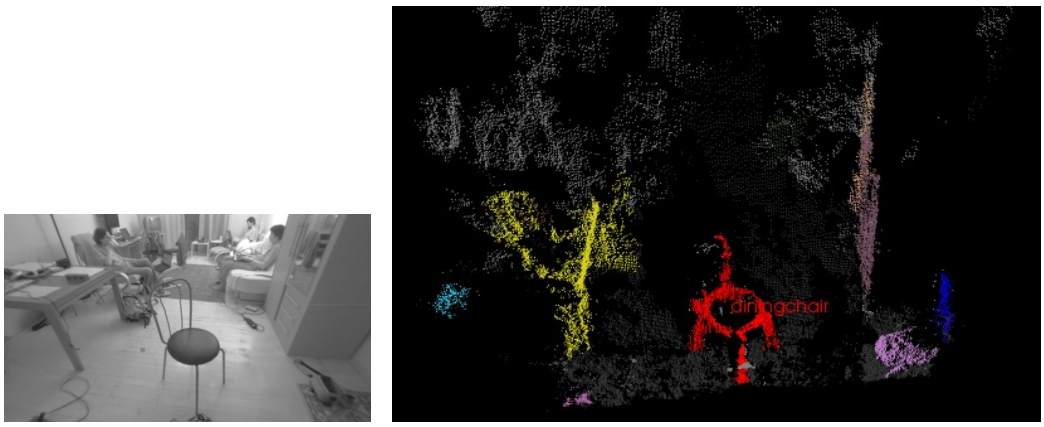



Figure 1.8: Detection results for a dining chair in a home scene [83]): Left the left image, right the stereo point cloud and colour coded for segmented region and the detected chair annotated red (best viewed in color).

## 1.3 Tracking and visual servoing

Another typical task humans perform is to detect and follow the motion of objects. When grasping an object the relative motion is observed. When walking the motion of the environment is monitored. The technique of visually tracking an object and determining its location is used particularly in surveillance and robotics tasks. In the former the paths of cars or persons are estimated to recover the ongoing activities and react accordingly (also see Section 1.4.1 below). In robotics the goal is to track the relative position between the mobile robot and its environment or to steer the robotic hand towards an object. The continuous feedback control of the position of the robot is referred to as visual servoing [19].

First successes in autonomous car driving and air vehicle guidance indicate the use of visual servoing [27, 3]. However, there are still two major roadblocks for further use in real-world scenarios [78, 19].

1. Efficient Tracking Cycle: vision and control must be coupled to assure good dynamic performance. Fast motions are needed to justify the use of visual servoing in real robotic applications.

2. Robust target detection: vision must be robust and reliable. Perception must be able to evaluate the state of the objects and the robot to enable a reaction to changes and to assure the security of the robot and its environment.

The tracking control problem has received a lot of attention in the literature (e.g., [33]) but robust visual tracking is just as critical and only recently receives more and more attention. The following sections summarise the state of the art with respect to these two criteria.

### 1.3.1 The Tracking Cycle

The goal of visual servoing is to consider the entire system and its interfaces. The basic control loop is depicted in Fig. 1.9. It contains three major blocks: the Vision System, the Controller and the Mechanism or robot or vehicle. The vision system determines the error between the command location and the present location of the target. First the result is expressed as an error in the image plane. The controller converts the signal to a pose or directly into command values for the axes of the mechanism and transfers the values to the robot. The robot or vehicle commonly uses a separate controller to control the motors at axes level.

The structure of the loop in Fig. 1.9 derives from the fact that the target motion is not directly measurable. Therefore the target motion is treated as a non-measurable disturbance input [21].

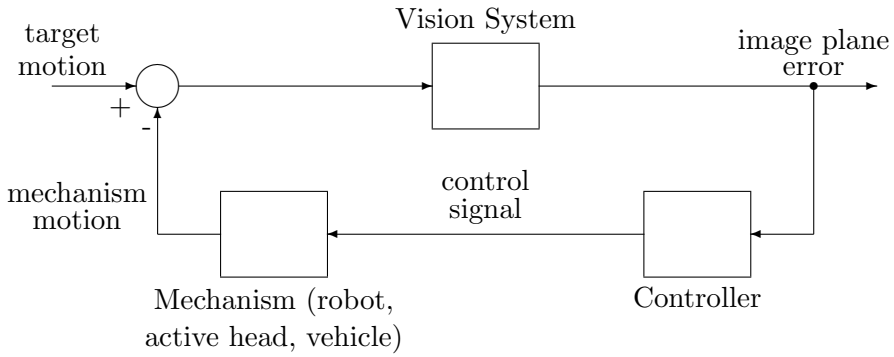



Figure 1.9: *Basic block diagram of visual servoing.*

The objective is to build the tracking system such that the target is not lost. A limit is given by the field of view of the camera. Hence it is useful to investigate tracking of the highest possible target velocity (or acceleration). The relevant property is the delay (or latency) of the feedback generated by the vision system [65, 77]. The two main factors to take care of are then (1) the *latency* or delays in one cycle from obtaining the image and (2) the part or window of the image that is actually processed.

Latencies accumulate from the cameras, today fire-wire cameras produce images at typically 25 or 30 $Hz$. Additionally all times to transfer data to the controller and as biggest factor the time to process the image needs to be considered. While it seems intuitive that latencies delay tracking, the second factor, image processing, is often not respected. If the full image is calculated, this might take much longer than the frame time of the camera. Hence images are lost. If a small window is used, for example around the location where the target has been seen in the last image, it is possible to exploit every image. The optimum is reached when the window size is selected such that processing is as fast as acquiring images [77]. This means it is optimal to operate a tracking system with a latency of two cycles of the

frame rate for cameras. Classically Kalman or other filters then take care of this delay [19].

It is interesting to note that the human eye exhibits space-variant tessellation with a high resolution fovea in the centre and a wide field of view at logarithmically decreasing resolution. The effect is that all the image should always be processed [77] and humans can react to motion in the periphery while actual recognition only works in the fovea which is rotated to the target and tracks it.

### 1.3.2 Robust Target Detection

Robustness of tracking is of major concern to the continuous operation in applications. Robustness indicates the quality of a method to degrade smoothly when input data is noisy or erroneous. A common denominator of techniques to improve robustness is the exploitation of redundancy by using multiple cameras, multi-resolutions, temporal constraints intrinsic to tracking, models, and the integration of several cues or features.

A minimal form of redundancy is inherent in a **stereo vision** system and exploited as the epipolar constraint [31]. Nevertheless the correspondence problem (finding the same scene point in both images) remains and successful stereo applications are rare. Today systems to calculate a depth image from two stereo images are commercially available, e.g., Videre Design. The correspondence problem of stereo vision is reduced by using **three or more cameras**. Using this technique a depth image can be generated, e.g., TRICLOPS (Point-Grey Research). A system to steer cars at high speeds exploits three cameras with different fields of view [27].

The idea of merging information from different **levels of resolution** has been exploited in scale space approaches, e.g., [49]. Consistency is approved

over the levels to obtain a measure of the reliability of edge detection. Recently, the SIFT features [50] exploit this to select the most robust local scale of a gradient point. However, the use of space variant image resolutions has not been sufficiently exploited, yet.

The redundancy of a series of images can be exploited by considering the **temporal consistency** of the detected features, also referred to as temporal data association [8, 27]. From the perspective of control theory filtering and prediction are the common methods to improve robustness [33]. Today the most common approach to cope with this uncertainty is Kalman or particle filtering [73], where several hypotheses aid in adapting to uncertainties of motion or measurements. Already the dynamic vision approach [27] exploited the temporal evolution of geometric features to build a model of the perceived world. Object behaviour is used to predict image positions. Tracking is then used to confirm or update motion behaviours.

The work in [27] is also a reference work of **model-based vision**. The model is commonly a CAD representation of the target, which is used for model prediction and local feature processing. Mobile robots hold (or build up) a representation of the building and use landmarks, such as walls or pillars, for navigation (e.g., [73]). A recent summary is given in [26]. Fig. 1.10 gives an example where model projection and probabilistic tracking is combined and the code is freely available [55].

In humans the **integration of cues** has been found as a likely source of the excellent ability to cope with changing conditions [82]. While the recognition process is understood only partially, it has been found that cortex areas V1-V5 report back results to each other and integration has been proposed as a first model. Active vision research was the first to utilise **cue integration for tracking**. E.g., [43] demonstrates that weighted consensus


Figure 1.10: An example scene where objects have been recognised based on their main structure (displayed) and are then tracked using the superimposed texture taken from the actual objects [55].

voting of five cues highly improves the performance of view-based tracking with a fuzzy fusion method.

To obtain the object pose for robotic applications, **contour tracking** is required. Edge contours can be followed with the active contour model (snake) [39]. A probabilistic extensions are able to follow faster motions, e.g., [35, 85], since the probabilistic sampling simulates a space-variant tessellation.

In **summary**, a plethora of approaches to tracking exist. Most approaches are either robust or fast. While tracking based on regions or interest points is more robust in textured environments, edge based tracking schemes provide best input for visual servoing in robotics or augmented reality systems [75, 20]. With the steady increase in computing power the idea to integrate cues will go further. The integration of more cues, knowledge and context (stereo, levels of resolution, temporal consistency, cue integration,

colour constancy, model-knowledge, task-knowledge) has been achieved only partially.

## 1.4 Understanding human behavior

### 1.4.1 Visual surveillance

Smart rooms, human-machine interfaces, and safety and security applications require work to recognise activities of humans, also summarised under the notion of visual surveillance. For recent reviews see [17, 76]. The annual PETS (Performance Evaluation of Tracking and Surveillance) workshop series is an excellent resource of ongoing work.

Typically surveillance systems operate from fixed cameras. This enables to use the technique of background subtraction to detect changes in the image. For a review see [61]. The main task is to cope with the varying illumination, which changes the appearance of the image and might hide the changes due to moving foreground objects. The result of this change detection are image regions as indications of objects. As next step these blobs are tracked over the image sequence, where data association methods are used to find consistently moving object and to detect erroneous regions generated. Hidden Markov Models and Bayesian networks are preferred approaches [17, 76].

Surveillance systems often work in two phases: a learning phase and a run-time phase. In the learning phase the system is initialised to a scene and models are either adapted or learned from observations. These models contain data about normal activities, e.g., lanes of cars, entrance points, or typical human gestures. In the run-time phase the data streams are compared

to the model data to come up with interpretations and reactions. At present systems can detect and recognise the behaviour of a few persons up to larger groups of people, e.g., [25]. In traffic scenes processing is mostly bottom up, while newer systems exploit domain knowledge in a top down way, e.g., [84]. An example is to use object models and expected activity models to monitor at airport aprons [14].

In the domain of robotics the object to human relation has been studied in approaches such as Programming by Demonstration (PbD), where the task is to interpret user commands to teach a robot [4]. In recent work activities of hand and objects are interpreted and stored using natural language expressions in an Activity Plan. An Activity Plan is a concise account of the scenario specifying the relevant objects and how they are acted upon, e.g., [62].

With the decrease of camera costs present direction of work is towards camera networks surveying large areas. Detailed models of a human and typical activities yield finer gesture interpretation in less constraint settings [76].

### 1.4.2 Human machine interaction

Proceeding from visual observation techniques to a vision-based interactive human computer interface seems to be a small step. It opens up a full range of new applications where computers, monitors, and input devices like keyboard and mouse disappear into the everyday environment. However, as attractive this step might be, its realization includes several technical and conceptual pitfalls that need to be addressed: (1) *Reactivity:* A system needs to react in soft real time to user activity. Otherwise a user is distracted, frustrated, and lost with regard to the communicative state. Appropriate techniques have

been developed for face recognition, gaze detection, or gesture recognition and define a field of active research [41, 59]. (2) *Robustness:* High false positive detection rates would imply a system behaviour that is unwanted by a user and leads to inconsistencies to his/her expectations. This is especially problematic as not all user behaviour is directed to the system. Here, *joint attention* is an important concept [15]. This includes that both communication partners are attending to the same thing and that they are aware of each other's attention. In human-robot interaction, for instance, the robot needs to detect when a user is facing it. At the same time, the robot's head and eyes will track the user's face in order to re-assure the established communication. (3) *Reliability:* User activities partly missed by the system could corrupt the whole user input to the system. Thus, there needs to be a notion if the input is well-formed or not. This is a difficult learning and recognition problem because humans typically perform tasks with a large variability and they are not aware of the system's limits. One interesting research direction is proceeding towards shaping the human-machine interaction by mixed-initiative dialogue strategies [51]. (4) *Situativity:* The interpretation of most human behaviour is context specific. Therefore, many systems are designed for a very specific scenario or application domain. In order to overcome these limitations, *context awareness* is an important concept. This term was introduced in the mobile computing community [63]. For computer vision, it has been operationalised, e.g., by Crowley et al. who present an ontology for context and situation [22].

As a consequence of the discussion above, research towards vision-based human machine interaction is always system-oriented research and a highly interdisciplinary task. Most systems in this area tightly constrain the communicative setting. Early work has been done by Bolt et al. [13] in his

"Put That There" system. Today's systems range over a wide spectrum of techniques and applications. The SafetyEYE developed in industry research estimates the action radius of an industrial manufacturing robot and stops it in case of human machine interference. The MIT Kidsroom provides an interactive narrative play space for children [12]. It is based on visual action recognition techniques that are coupled with the control of images, video, light, music, sound, and narration. Crowley et al. [24] describe an interactive Magic Board based on the tracking of fingers and a perceptual window that scrolls by detecting head movements. In the last years, body tracking has become a hot commercial topic for game consoles, like Sony's play station or Microsoft's X-box. A different focus was set in the VAMPIRE system [81]. It provided assistance to people in everyday tasks by leading them step-by-step through a recipe. This was demonstrated in a drink mixing scenario and used object recognition, tracking, localization, and action recognition techniques in order to achieve a user assistance based on augmented reality techniques. Much work has been conducted in order to bridge the communication gap between humans and personal robots, examples are the PR2 from Willow Garage, STAIR from Stanford, Care-O-Bot 3 from Fraunhofer IPA, GRACE from CMU, Jijo-2 from AIST, or BIRON and BARTHOC from Bielefeld.

Compared to human-human communication, human-machine interaction is still brittle and in its infancy. Today's research concentrates on mimicking certain aspects of it in order to address the four challenges named before.

## 1.5 Contextual scene understanding

Most approaches in computer vision do not interpret entire images, but selective parts of it. They aim at extracting foreground objects from background

clutter. Then, each object is classified in isolation. Background is ignored and viewed as irrelevant distracting data or simply as noise. Contextual scene understanding is somehow the dual process. It recycles the data ignored before, i.e. background clutter and relational information, in order to infer possible interpretations for foreground objects. Thus, these techniques aim to incorporate scene context into the classification process.

Pioneering work has been conducted by Strat and Fischler [70] who define context sets that govern the invocation of the system's processing steps. They identify four different kinds of criteria that comprise context sets: (1) *Global contexts* are attributes of an entire scene like daytime or landscape. (2) *Location* characterizes the spatial configuration of a scene like touching the ground, coincidence with other object types. (3) *Appearance* of neighbouring objects may be similar like neighbouring trees. (4) *Functionality* describes the role of an object in a scene like supporting another object or bridging a stream. From the control point of view, Strat and Fischler employ three kinds of context-driven operations: (1) Hypothesis generation, (2) hypothesis validation, and (3) hypothesis ordering, that guide the scene interpretation process. During search consistent cliques are constructed that represent partial interpretations of a scene. The main drawback of this kind of approach was the huge knowledge engineering task in coding the contextual knowledge of the system. However, the general types of contexts introduced and the different kinds of control principles designed are still valid for the current state-of-the-art.

Later work adapted probabilistic models for contextual interpretation which capture causal relationships in directed dependencies (Bayesian networks), spatial relationships in undirected dependencies (Markov Random Fields), and temporal relationships in dynamic models (Bayesian filter). All

these variants are unified under the theoretical framework of *graphical models* [60, 47, 37]. In order to name a few examples that reflect more recent trends in contextual interpretation, we shortly discuss the the general context types raised by Strat and Fischler:

**Global contexts** have been used by Murphy et al. [57]. They compute a holistic image representation – the so called image gist – in order to classify semantic places and relate them to object hypotheses by graphical models.

**Local constraints** are applied by Hoiem et al. [32]. They related object detections to an overall 3-d scene context and judge the scale and location with regard to the estimated scene geometry.

**Functional aspects** are used by Moore, Essa and Hayes [54] who relate human actions and objects by Bayesian networks. They introduce the concept of *object spaces* that link both kinds of information in space and time. In a different approach to functionality is mapped to 3D shapes that are extracted from range images [71].

**Linguistic contexts** refer to additional information given by parallel text or speech. These kind of bi-modal data frequently occurs in catalogues, newspapers, magazines, web pages, broadcasting news, movies, or human-machine interaction dialogues. The verbal information principally includes all three types of contextual information. Global information was used by Barnard et al. [9] in order to learn models for automatic image annotation. They employed a hierarchical mixture model by Hofmann for describing an image on a course topic level as well as on a detailed object level. Local constraints have been extracted from image captions by Srihari & Burhans

[68] for labelling human faces in newspaper photographs. Wachsmuth & Sagerer incorporate spatial information from spoken human-robot dialogues in order to robustify the understanding of visual scenes [80]. Functional contexts are provided by verbs which have been incorporated to a less degree in scene understanding.

## 1.6 Summery and Conclusion

Agents, human or artificial, need to perceive their environment for operating and surviving in it. Visual perception is the strongest human sense and work in the field of computer vision sets out to provide the required capabilities. In this chapter we summarised main achievements. We started with reviewing trends and perspectives and then highlighted a few areas.

Today it is possible to learn and then recognise objects from 2D images up to around 1000 and the number increases continuously. It is however constraint to databases of images where size of objects or typical scenes are similar. In open environments, such as a search task in homes, variations in illumination, view point, or occlusion still pose challenges. When using 3D images, e.g., using laser scanners, shape of the objects can be acquired and exploited to control industrial processes such as robotic grasping or spray painting.

Tracking of objects or interest points over longer video sequences can be done in real time given sufficient texture. Rules on how to exploit the image information and predict and search efficiently in subsequent images are established and visual servoing methods to control robot arms available.

The real-time performance and robustness achieved by today's computer vision techniques for hand tracking, human body tracking, face recognition,

etc., leads to a new quality of vision-based human-machine interaction. In this chapter, we have discussed several challenges in this new field that merges the areas of HCI and CV. Over the last years, several new workshop series have been established like CV4HCI or human-centered CV. And we expect that this marriage will provide further fruitful influences on the field – taking two perspectives: How to design CV systems for users and how to effectively include the user in the visual processing loop?

One of the challenges pointed out in the HMI section was situatedness: When and with what information should the user be bothered? The same question could be asked for the vision system. Not all information is important, not all detection results are valid. The notion of context provides a concept of a global consistency on the one hand and a frame of meaning on the other hand. Even with quite sophisticated and high performance recognition techniques, context will keep its role when we talk about computer vision systems that need to act in real world environments.

Computer vision systems need to combine computer vision techniques for application purposes. This is the core of CV as an engineering discipline. However, it has been proven over the years that general integration architectures are hard to define. Some approaches have shown their applicability in successful multi-partner projects (e.g. ActIPret, VAMPIRE, or CogX). Real progress is hard to achieve on the theoretical side and needs to be proven by the practical realization of systems. For more details about vision for robotic systems further reading of [44] is suggested.

While these results indicate the advance of the field, several challenges lie ahead. Examples are: recognising classes of objects has been started but is limited to very few salient classes such as wheels or aeroplanes. Detecting grasp points on arbitrary object needs to be extended from planar to full 3D

object locations. Or the function of an object as indicated by its design and shape cannot yet be deduced from imaging it. Nevertheless, the hope is that computer vision is used more often in combination with other AI methods to see how to build more complete systems.

## 1.7 Further Reading

- Dana Ballard and Christopher Brown. it Computer Vision. Prentice Hall, Inc., Englewood Cli?s, New Jersey, 1982. The basic book on methods in computer vision. Available on-line: http://homepages.inf.ed.ac.uk/rbf/BOOKS/B

- S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr. *Object Categorization: Computer and Human Vision Perspectives, Cambridge University Press, 2009.* Excellent overview of approaches to object recognition including a historical perspective. A must to get started in this direction.

- Richard Szeliski. *Computer Vision: Algorithms and Applications, Springer, 2010.* An excellent lecture book for the introduction and more depth study of computer vision It has an emphasis on techniques that combine computer vision and graphics, but covers also modern techniques for object recognition, segmentation, and motion estimation. Available on-line: http://szeliski.org/Book/

- David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach, Prentice Hall, 2003.* A broad collection of computer vision techniques that is a very good reference for the advanced study of computer vision.

- Danica Kragic and Markus Vincze. *Vision for robotics. Foundations*

*and Trends in Robotics, 1(1):178, 2009.* An overview of the specific needs of robotics to computer vision methods plus an overview of applications.

- R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2003. An deep coverage of geometrical aspects in computer vision for the advanced reader.

# Bibliography


[1] Y. Aloimonos. *Active Perception.* Lawrence Erlbaum, Hillsdale, NJ, 1993.

[2] A.P. Ambler, H.G. Barrow, C.M. Brown, R.M. Burstall, and R.J. Popplestone. A versatile computer-controlled assembly system. In *Proc. Third Int. Joint Conf. on AI*, pages 298–307, Stanford, California, 1973.

[3] O. Amidi, T. Kanade, and R. Miller. *Vision-based Autonomous Helicopter Research at CMU.* in [78], 2000.

[4] T. Asfour, P. Azad, and et al. Toward humanoid manipulation in human-centred environments. *Robotics and Autonomous Systems 56(1)*, pages 54–65, 2007.

[5] Ruzena Bajcsy. Active perception. *Proc. of the IEEE*, 76(8):996–1005, 1988.

[6] Dana Ballard and Christopher Brown. *Computer Vision.* Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1982.

[7] D.H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[8] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, 1991.

[9] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[10] Georg Biegelbauer and Markus Vincze. Efficient 3d object detection by fitting superquadrics to range image data for robot's object manipulation. In *IEEE ICRA'07 Int. Conf. on Robotics and Automation*, pages 1086–1091, 2007.

[11] T. Binford. Visual perception by a computer. In *Proceedings of the IEEE Conference on Systems and Control*, pages 116–123, 1971.

[12] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schtte, and A. Wilson. The kidsroom: A perceptually-based interactive and immersive story environment. *PRESENCE: Teleoperators and Virtual Evironments*, 8(4):367–391, August 1999.

[13] R.A. Bolt. Put that there: Voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270, 1980.

[14] M. Borg, D. Thirde, J. Ferryman, F. Fusier, V. Valentin, F. Bremond, and M. Thonnat. A real-time scene understanding system for airport apron monitoring. In *IEEE International Conference on Computer Vision Systems ICVS06*, 2006.

[15] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, 8(1):49–74, 2000.

[16] R. Brooks. Model-based 3d interpretation of 2d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983.

[17] H. Buxton. Learning and understanding dynamic scene activity: A review. *Vision Computing*, 21:125–136, 2003.

[18] R.J. Campbell and P.J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Image Understanding*, 81:166–210, 2001.

[19] F. Chaumette and S. Hutchinson. Visual servo control i: Basic approaches. *IEEE Robotics and Automation Magazine*, 13(4):82–90, 2006.

[20] A.I. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):615–628, 2006.

[21] Peter I Corke. *Visual Control of Robots: High Performance Visual Servoing*. Research Studies Press, John Wiley, 1996.

[22] J. L. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual components for context aware computing. In *UBICOMP 2002, International Conference on Ubiquitous Computing*, Goteborg, Sweden, September 2002.

[23] James L. Crowley and Henrik I. Christensen, editors. *Vision as Process*. Springer, 1995.

[24] L. Crowley, J. Coutaz, and F. Bérard. Things that see: Machine perception for human computer interaction. *Communications of the A.C.M.*, 43(3):54–64, March 2000.

[25] F. Cupillard, F. Bremond, and M. Thonnat. Behaviour recognition for individuals, groups of people and crowd. In *IEE Proc. of the IDSS Symposium - Intelligent Distributed Surveillance Systems*, 2003.

[26] G.N. Desouza and A.C. Kak. Vision for mobile robot navigation: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):237–267, 2002.

[27] Ernst D. Dickmanns. *Dynamic Vision for Perception and Control of Motion*. Springer London, 2007.

[28] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[29] D.A. Forsyth and J. Ponce. *A Modern Approach*. Prentice Hall, 2002.

[30] Rafael C. Gonzales and Richard E. Woods. *Digital Image Processing (2nd edition)*. Prentice Hall, 2002.

[31] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[32] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[33] S. Hutchinson, G.D. Hager, and P. Corke. Visual servoing: A tutorial. *IEEE Trans. RA*, 12(5), 1996.

[34] D.P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, November 1990.

[35] M. Isard and A. Blake. Condensation - conditional density propagation for visual servoing. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[36] A. Jaklic, A. Leonardis, and F. Solina. *Segmentation and Recovery of Superquadrics*. Kluwer Academic Publishers, 2000.

[37] M. I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[38] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[39] M. Kass and A. Witkin. Snakes: Active contour models. *Int. Journal of Computer Vision*, 1:321–331, 1987.

[40] D. Keren, D. Cooper, and J. Subrahmonia. Describing complicated objects by implicit polynomials. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):38–53, 1994.

[41] Branislav Kisacanin, Vladimir Pavlovic, and Thomas S. Huang, editors. *Real-Time Vision for Human-Computer Interaction*. Springer, 2005.

[42] Jan J. Koenderink. An internal representation for solid shape based on the topological properties of the apparent contour. In Whiteman Richards and Shimon Ullman, editors, *Image understanding 1985-86*, pages 257–285. Ablex Publishing Corp., Norwood, NJ, USA, 1987.

[43] D. Kragic and H.I. Christensen. Cue integration for visual servoing. *IEEE Transactions on Robotics and Automation*, 17(1):18 –27, 2001.

[44] Danica Kragic and Markus Vincze. Vision for robotics. *Foundations and Trends in Robotics*, 1(1):1–78, 2009.

[45] J. Krivic and F. Solina. art-level object recognition using superquadrics. *Elsevier; Computer Vision and Image Understanding*, 95(1):105–126, 2004.

[46] Aldo Laurentini. Introducing the reduced aspect graph. *Pattern Recognition Letters*, 16(1):43–48, January 1995.

[47] S.L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, United Kingdom, 1996.

[48] Ales Leonardis and Horst Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding: CVIU*, 78(1):99–118, 2000.

[49] Tony Lindeberg. Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30(2):79–116, 1998.

[50] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[51] Ingo Lütkebohle, Julia Peltason, Lars Schillingmann, Christof Elbrechter, Britta Wrede, Sven Wachsmuth, and Robert Haschke. The curious robot - structuring interactive robot learning. In *International Conference on Robotics and Automation*, Kobe, Japan, 14/05/2009 2009. IEEE, IEEE.

[52] David Marr. *Vision*. W. H. Freeman and Company, San Francisco, 1982.

[53] D. Marshall, G. Lukacs, and R. Martin. Robust segmentation of primitives from range data in the presence of geometric degeneracy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):304–314, 2001.

[54] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *Proceedings of IEEE International Conference on Computer Vision*, Corfu, Greece, March 1999.

[55] T. Mrwald, J. Prankl, A. Richtsfeld, M. Zillich, and M. Vincze. Blort - the blocks world robotic vision toolbox. In *Best Practice in 3D Perception and Modeling for Mobile Manipulation (in conjunction with ICRA 2010)*, 2010.

[56] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int.J. Computer Vision*, 14(1):5–24, 1995.

[57] K. Murphy, A. Torralba, and W.T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Conference on Neural Information Processing Systems*, 2003.

[58] N.J. Nilsson. Mobile automaton: An application of artificial intelligence techniques. In *Technical Note 40, AI Center, SRI International, also presented at IJCAI*, 1969.

[59] Vladimir Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.

[60] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Palo Alto, 1988.

[61] M. Piccardi. Background subtraction techniques: a review. In *IEEE International Conference on Systems, Man and Cybernetics, Volume 4*, 2004.

[62] K.H. Sage, A.J. Howell, and H. Buxton. Recognition of action,activity and behaviour in the actipret project. *KI Special Issue*, 2005.

[63] B. Schilit, N. Adams, and R. Want. Context aware computing applications. In *First international workshop on mobile computing systems and applications*, pages 85–90, 1994.

[64] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. *Proc. Int Conf. Computer Vision and Pattern Recognition*, 1996.

[65] P.M. Sharkey and D.W. Murray. Delays versus performance of visually guided systems. In *IEE Proc. Control Theory & Applications 143(5)*, pages 436–447, 1996.

[66] T. Shipley and P. Kellman. Advances in psychology: Form fragments to objects. *Elsevier Science B.V.*, 130, 2001.

[67] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):131–147, 1990.

[68] Rohini K. Srihari and Debra T. Burhans. Visual semantics: Extracting visual information from text accompanying pictures. In *AAAI*, pages 793–798, 1994.

[69] L. Staib and J. Duncan. Model based deformable surface finding for medical images. *IEEE Transactions on Medical Imaging*, 15(5):720–731, 1996.

[70] Thomas M. Strat and Martin A. Fischler. Context-based vision: Recognizing objects using information from both 2d and 3d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, October 1991.

[71] Melanie A. Sutton, Louise Stark, and Ken Hughes. Exploiting context in function-based reasoning. In *Sensor Based Intelligent Robots*, volume 2238 of *Lecture Notes In Computer Science*, pages 357–374. Springer-Verlag, London, UK, 2000.

[72] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, november 1991.

[73] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005.

[74] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuro Science*, 3(1):71–86, 1991.

[75] L. Vacchetti, V. Lepetit, and P. Fua. Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1385–1391, 2004.

[76] M. Valera and S.A. Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings- Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[77] M. Vincze. On the design and structure of artificial eyes for tracking tasks. *Journal of Advanced Computational Intelligence and Intelligent Informatics JACIII*, 9(4):353–360, 2005.

[78] M. Vincze and G.D. Hager. *Robust Vision for Vision-based Control of Motion*. IEEE Press, 2000.

[79] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[80] Sven Wachsmuth and Gerhard Sagerer. Bayesian networks for speech and image integration. In *Eighteenth national conference on Artificial intelligence*, pages 300–306, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[81] Sven Wachsmuth, Sebastian Wrede, and Marc Hanheide. Coordinating interactive vision behaviors for cognitive assistance. *Computer Vision and Image Understanding*, 108:135–149, 2007.

[82] B.A. Wandell. *Foundations of Vision*. Sinouer Publishers, 1996.

[83] W. Wohlkinger and M. Vincze. 3d object classification for mobile robots in home-environments using web-data. In *IEEE 19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD)*, pages 247–252, 2010.

[84] T. Xiang and S.G. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[85] Rathi Yogesh, Vaswani Namrata, Tannenbaum Allen, and Yezzi Anthony. Tracking deforming objects using particle filtering for geometric active contours. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(8):1470–1475, 2007.