NISTIR 7923

Ground Truth Systems for Object Recognition and Tracking

Afzal Godil
Roger Eastman
Tsai Hong Hong

http://dx.doi.org/10.6028/NIST.IR.7923

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7923

Ground Truth Systems for Object Recognition and Tracking

Afzal Godil
Information Access Division
Information Technology Laboratory

Roger Eastman
Tsai Hong Hong
Intelligent Systems Division
Engineering Laboratory

http://dx.doi.org/10.6028/NIST.IR.7923

March 2013

U.S. Department of Commerce
Rebecca Blank, Acting Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

ACKNOWLEDGEMENTS

The authors wish to give special thanks to Stacy Bruss of the NIST Research Library who provided the initial set of reference papers.

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

1. Abstract:

In this report we discuss different types of ground-truth systems used for evaluation of object recognition and tracking systems in industrial manufacturing environments. We discuss four main ways of acquiring ground truth for object recognition and tracking: 1) physics-based simulation ground-truth systems, 2) manual annotation or labeling of ground-truth systems, 3) platform-based ground-truth systems, and 4) physically-based ground-truth systems. We also include a separate discussion of motion capture systems, a special case of physically-based ground-truth systems focused on human tracking. We discuss previous efforts and discuss the different physical quantities used for ground-truth systems. Currently, there are no solid, universal solutions for ground truth measurements for human and object detection and tracking. There are a number of partial solutions suitable for specific applications, with varying drawbacks.

2. Introduction

Object recognition and tracking are among the most common and challenging tasks that a robotic perception system has to accomplish to support more complex perception tasks such as identifying meaningful events and activities. In our study, an object can be a robot, an Automated Guided Vehicles (AGV), a person or limb, a queue of people, an assembly, a component, a part, a product on an assembly line, or any other object which may need to be identified in an industrial environment. The goal of object recognition is to correctly identify objects that are present in a 3D scene from sensor data. This is a challenging task since the scene could have clutter, objects could occlude each other, and there could also be illumination or viewpoint variations.

Object recognition and localization is important in many practical applications such as manufacturing automation, navigation, part inspection, and computer aided design /computer aided manufacturing (CAD/CAM), among others. Our main interest is in object recognition for manufacturing applications in a dynamic indoor factory environment. We assume for this report we are monitoring people, robots, AGVs, conveyer belts, parts, semi-completed assemblies, and other typical objects in a fixed, indoor work cell. We want to monitor people and these objects to improve the performance and safety of automation, so robots and similar systems can interact with their environment. We emphasize the recovery of position, motion, pose, and classification data, so we can determine if a person is walking across the cell, or if a particular type of engine part is on its side on a table. We are less concerned with the identification of individual people or parts (so the recovery of a name of a person, or serial number of a part, are not a priority but may be possible with the systems we consider) and we are not at all concerned with recovery of shape data. For the latter, we assume we have a priori knowledge of the shape of objects, such a CAD model of parts. The following scenarios are those for which we would like to capture ground truth data:

- Human detection and tracking for safety
- Articulated human motion tracking
- Action and activity recognition
- Human robot collaboration
- Detection and tracking of parts and assemblies
- Pose determination of parts for robot grasping

The basis with which to evaluate the strengths of different object recognition algorithms is to compare algorithm results for a set of tasks with known ground-truth based on standardized performance metrics, such as identification accuracy, geometric position accuracy, or robustness to scene complexity. The tasks, the ground truth data, and different performance metrics should allow researchers to fully understand the strengths and limitations of different approaches and are an essential step towards establishing the credibility of object recognition for real time manufacturing applications.

To reliably evaluate the performance of systems that localize and recognize objects, the ground truth system needs to be more accurate than the system under test. Typically the ground truth system measurements should be an order of magnitude greater than those obtained by the system being evaluated. Since the system is to be used in a dynamic environment, the spatial and temporal resolution should be high enough to resolve the motions of the objects to avoid motion blur. We also want the ground truth system to have adequate spatial resolution to resolve the locations of the objects. There are a number of issues that need to be resolved for a successful evaluation, such as synchronization, latency issues, and time drift between the ground truth system and the system under test. Some of these issues are specific to a particular system and/or to the problems they solve.

For systems that only recognize objects, the ground truth system needs to know the identities of all the objects with a very high probability of correctness compared to the system under test. Ideally, ground truth identity should be known perfectly. For some of the stationary object recognition tests, the ground truth identity is exact, since the experimenter places known objects in the scene. For some tests with moving objects or sensors, if we know the identities and locations of all the objects at the

start, then if we can track them accurately we will also know their identities throughout the test. There are some more complex scenarios where we might not know a priori the identities of parts, such as when parts appear randomly on a conveyer-belt or are randomly dropped into a bin for testing. In these cases we need to rely on ground truth identification systems. An advantage of autonomous identification systems is the ability to use them to conduct large-scale randomized tests.

3. Ground-truth systems

We organize ground-truth systems for object recognition and tracking into four categories: 1) physics based simulation, 2) annotation/label based systems, 3) platform-based systems, and 4) physically-based ground-truth systems, with the special case of motion capture systems broken out for separate review. The following subsections discuss each category, with emphasis on the physical space systems, which we conclude are most appropriate for the evaluation of real-time object recognition and localization systems for robotics and manufacturing applications.

3.1. Physics Based Simulation Ground Truth Systems

The factory floor environment including people, robots, conveyer belts, and other equipment in different scenarios (tasks) can be simulated with a physics-based simulation and modeling program, such as the one shown in Figure 1 [81]. Based on the simulation, realistic synthetic images can be rendered with real-time or offline rendering programs [1] (e.g., OpenRT, real-time Raytracing, Renderman, Renderer 2.0, Blender). Since we simulated the environment, we know the exact spatial and temporal location of every part/sub-part and hence the exact ground truth is known. On the other hand, sensor data generated from a simulation does not exactly match real world sensor data despite advances in algorithms for realistic graphics generation. The performance of object recognition and tracking algorithms on synthetic images will differ from that on real images because sensor noise types and levels are not yet modeled with sufficient fidelity. English et al. [19] showed that for verification and validation processes, the best solution to the evaluation problem may be through synthetic data generated by physics-based simulation. To evaluate a pose estimation algorithm applied to range data in a bin picking algorithm, Park et al. [80] generated synthetic images by dropping a large number of parts with a physics-based simulation. The authors used the simulated data both to generate a priori statistics for likely part orientation, and to evaluate their pose algorithm. They also used data from real world experiments in their evaluation, but with limited metrics since they had no ground truth.

Physics based simulation can be applied to any scenarios where suitably accurate physical models exist. There are three ways in which it can have an advantage over monitoring in real environments:
1) Usually there is cost savings associated with physics-based simulation, since there is less need of expensive equipment for experiments and experiments are often time consuming to conduct; initially many problems can be modeled and simulated. However, effective simulations can be expensive to create – the payoff may come with standardized simulation systems that support many scenarios and many different objects.
2) Complex scenes with many objects including occlusion and clutter as in the case of bin picking can be simulated by dropping multiple objects randomly, whereas it might be more difficult and expensive to capture the ground truth locations and poses of all these objects in a real environment, as shown in Figure 2.
3) Human safety can be assured, and there are no issues with institutional human subjects review.

However, physics based modeling can be limited in producing accurate sensor data, as those models are not yet mature.

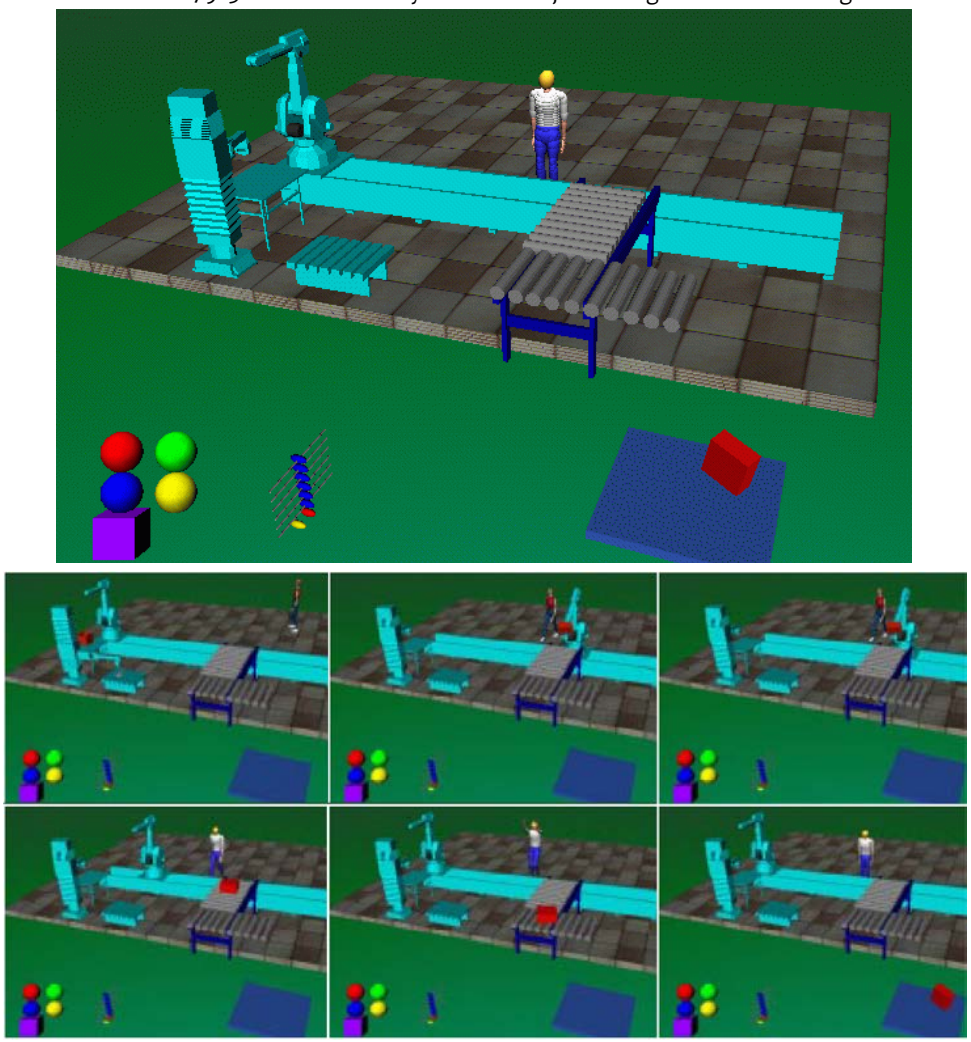

Figure 1a and 1b. show physics based modeling and simulation of a factory floor environment. [Permission granted by the author [81] to use the Figures]

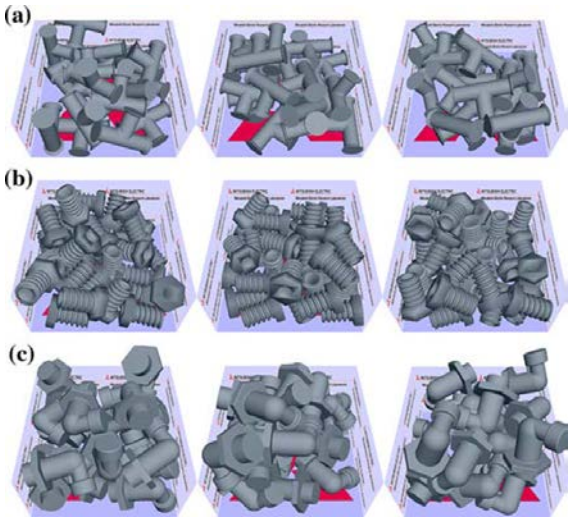

Figure 2. Synthetically generated scene images with physics based simulation and modeling. Three bins with: a) 20 T-pipe models, b) 30 bolt models, c) 20 elbow pipe models in each bin
[Permission granted by the author [80] to use the Figure]

3.2. Annotation/Label Based Ground-Truth Systems

One popular way to create ground truth for object detection and tracking is by human annotation of images (including video and depth images). One of the commonly-used desktop tools for video annotation is VIPER-GT [2] as shown in Figure 3, where users can annotate by drawing a bounding box around an object, indicating its identity and providing detailed spatial and temporal information. A survey paper on video annotation tools is the paper by Dasiopoulou et al. [3] which discusses tools such as: SWAD, an Resource Description Framework (RDF) based image annotation tool [4], Caliph, an MPEG-7 based image annotation tool [5], Anvil [6], a tool that supports audiovisual content annotation, SVAS [7], a Semantic Video Annotation Suite, and many others.

LabelMe [8] is a web-based tool for image annotation and has been widely used for creating very large imaging benchmarks [79], including for the Pascal visual object classes challenge [77] and TinyImages [78]. Figure 4 shows a screenshot of LabelMe in use. It allows the annotation boundaries to be drawn along the actual boundaries of the object instead of drawing bounding boxes as in other annotation tools. Crowd-sourced annotation of images and video can be accomplished at low cost through platforms such as Amazon Mechanical Turk (MTurk). This has revolutionized the annotation of static and dynamic image datasets and has led to massive image and video datasets. Figure 5 and Figure 6 show a crowdsourcing annotation tool for video that is described in [9]. Human detection and tracking is an essential step for pose estimation and articulated human motion analysis. A large number of datasets based on annotation are available for evaluating people detection and tracking algorithms [11, 12, 13, 14, 15, 16, 17, 18]. However, these datasets typically only provide a bounding box around a person as ground truth and so are of limited use for pose estimation and articulated human motion analysis. Finally, in the KITTI Vision Benchmark Suite for Autonomous Driving [10], the 3D laser range data are annotated by drawing 3D bounding boxes around cars in the data as shown in Figure 7.

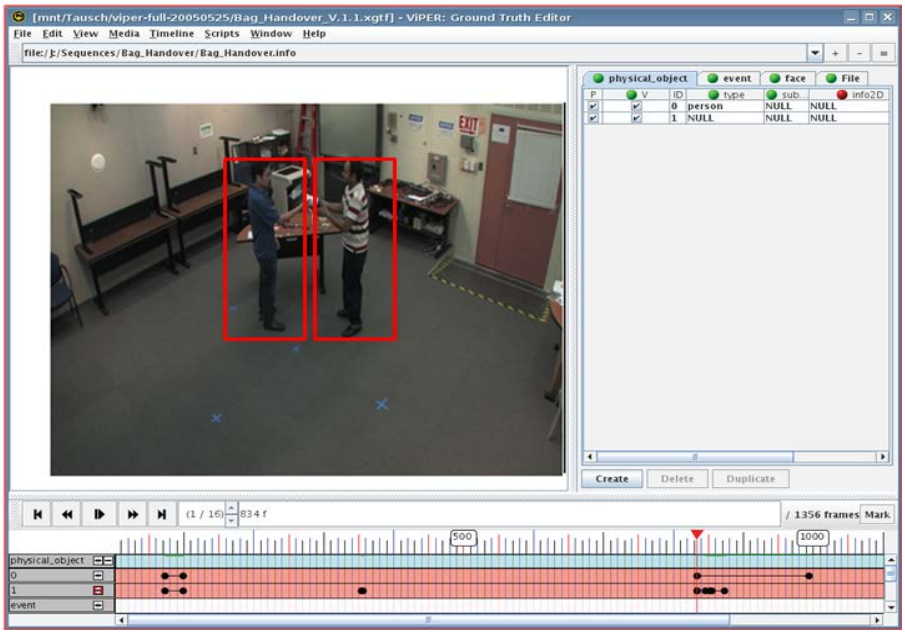

Figure 3. Shows annotation of video using ViPER-GT on a desktop. The two human subjects have been outlined by a bounding box. [Permission granted by the author [2] to use the Figure]

Figure 4. The LabelMe tool is shown in use. The user has the option of annotating any object by dragging the mouse along the boundary of the object and indicating its identity. They can annotate as many objects as they want. [Permission granted by the author [8] to use the Figure]

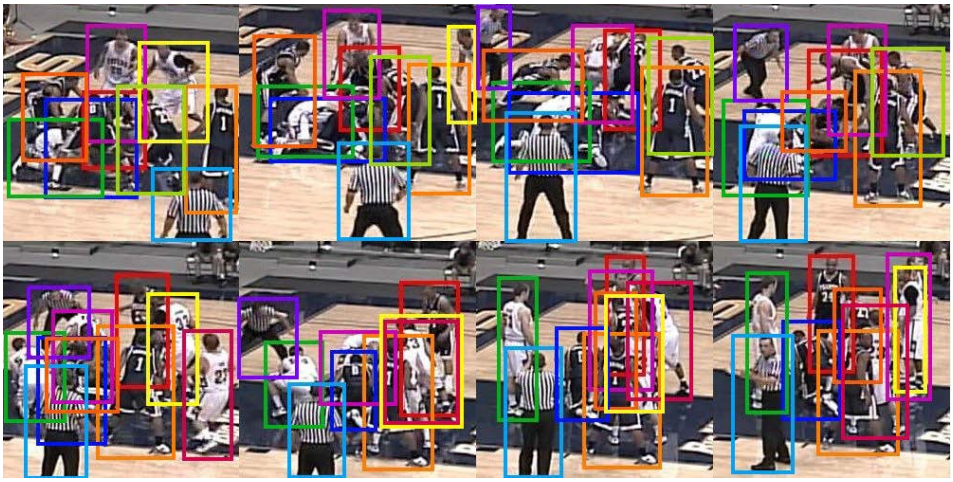

Figure 5. Shows the annotation of video by crowd sourcing.
[Permission granted by the author [9] to use the Figure]

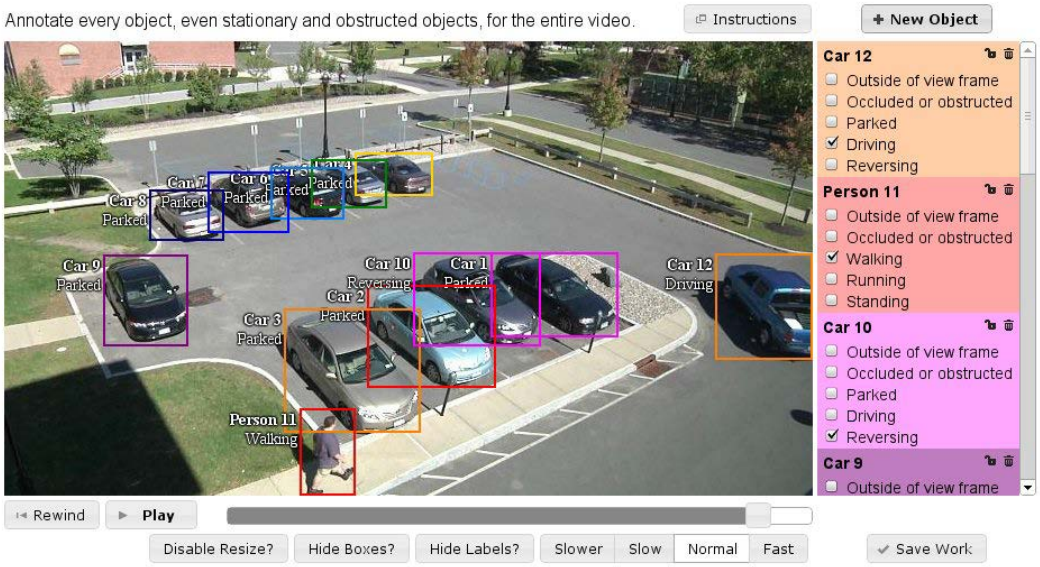

Figure 6. Shows the user interface of a crowd-sourcing tool for interactive video annotation. Users can play the video, draw bounding boxes around objects of interest, and track each object throughout the time line. [Permission granted by the author [9] to use the Figure]

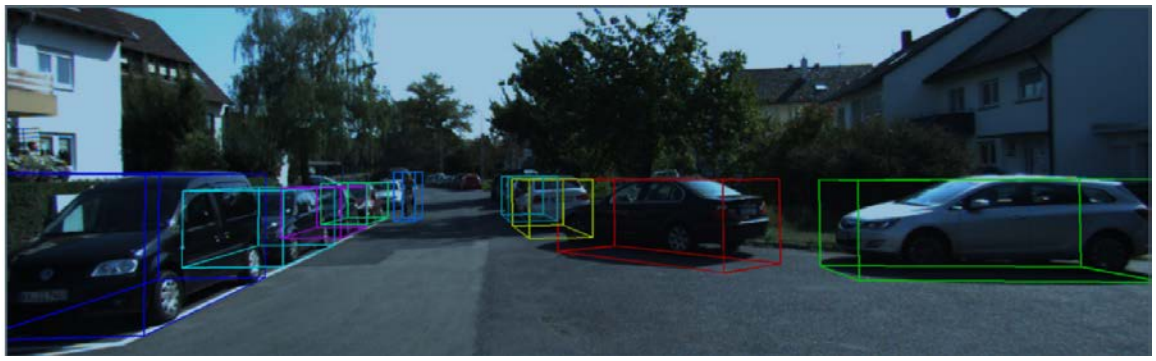

Figure 7. Annotation of 3D range data by drawing a 3D bounding box around cars and people. [Permission granted by the author [10] to use the Figure]

Annotation can be done automatically, producing a dataset to be distributed to researchers but not offering real-time ground truth. A video dataset with ground truth data for human tracking is the Carnegie Mellon University (CMU) Motion of Body (MoBo) Database [20]. Initially developed for gait analysis, it has also proved useful in analyzing the performance of articulated tracking algorithms [20, 21, 22]. While the initial dataset contains an extensive collection of walking motions, the CMU Multi-Modal Activity Database (CMU-MMAC) database [23] contains multimodal data of subjects performing cooking and food preparation tasks. The data were collected with three high definition video cameras while the ground truth data were collected with a Vicon motion capture (MoCap) system and Inertial Measurement Units. (Motion capture systems for real-time ground truth are discussed again in Section 3.5.)

The synchronized video and motion capture dataset for evaluation of articulated human motion (HumanEva) dataset [28, 29] is the most widely used dataset data set where actions of a single person have been captured by multi camera video together with marker-based motion capture data obtained using the Vicon MoCap system. Synchronization of the sensors for the HumanEva data was done by software in [28] and by a hardware trigger in [29]. The main drawback of some of the datasets is that there is only one person in the environment at a time, so there is no person-to-person occlusion. Other datasets do include multiple people. One is the Utrecht Multi-Person Motion (UMPM) Benchmark [30] that includes a collection of multi-person video recordings together with ground truth based on Vicon MoCap data. It is intended to be used for assessing the quality of

methods for pose estimation and articulated motion analysis of multiple persons using video data from single or multiple cameras.

Annotation-based approaches are typically applied to scenarios where a scene is monitored by an image or video sensor suitable for human interpretation, have the following advantages over real-time monitoring:
a) Complex scenes and behaviors can be annotated by hand when effective algorithms do not exist.
b) The software is often free, allowing a low-cost entry into the project.
c) The resulting annotated data supports analysis by multiple groups using multiple algorithms, so repeatability is good and cross-comparisons easily made.

Disadvantages include the labor cost of performing annotation, the variable and often unknown accuracy and reliability of the labels, and the fact that the annotations are mainly based on the images and recorded in sensor-based coordinates instead of 3D world coordinates.

3.3. Platform-based Ground Truth Systems

Platform-based systems give ground truth for object pose by placing the object on a platform that fixes the pose in advance of a test. There is no need to sense object position unless greater precision is needed than the fixture can provide. Object identity is also known in advance.

In Marvel et al. [39], the authors discuss multiple platform-based systems developed to estimate poses of objects in 6 degree of freedom (6DOF) Cartesian space (X, Y, and Z coordinates plus roll, pitch, and yaw) to determine performance and measurement accuracy of different systems. They describe three 6DOF ground truth systems utilized at the National Institute of Standards and Technology (NIST) (as shown in Figure 9): a laser-tracker-based system for pose measurement while the object is moving, an aluminum fixture-based system that can be used to fix the pose of an object while it is static, and a modular, medium-density fiberboard (MDF) fixturing system for location and pose, for static objects. The authors provide descriptions, characterizations, and measured accuracies of these systems. The NIST systems were inspired in part by previous work (Radu et al. [40]) using a pan-tilt mechanism to control a platform.

The systems share the characteristic that a single object is held on a platform so that its pose is precisely known. The first system uses a robot arm, the second a rotatable metal stage, and the third a fiberboard fixture. In that order, the three have decreasing accuracy – the robot arm with laser tracker is accurate to about 1/100 of a millimeter in position and 3/100 of a degree in rotation; the metal platform is accurate to about a half a millimeter in position and 1/6 of a degree in rotation; and the fiberboard accurate to about 2/3 mm in position and 1/6 of a degree in rotation. Similarly, the three decline in precision – the robot arm can be oriented with repeatability to several decimal places (in degrees), the rotating platform has repeatable fixed stops at limited orientations, and the fiberboard is constructed with stops at 15° intervals. The first two can be made to move dynamically, while the fiberboard must be configured for each test. If the robot arm is precise enough in its movements, it alone can be used to define a ground truth pose. In the NIST system, the known robot position is further refined by a laser tracker.

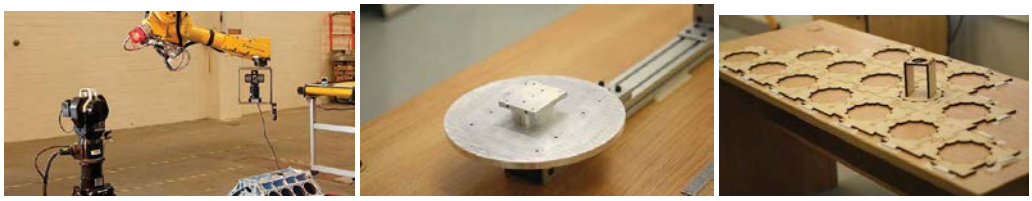

Figure 9. Shows the three systems developed and tested at NIST. [Permission granted by the author [39] to use the Figure]

Marvel at el. [41] presented the results and performance metrics of the 2011 Perception Challenge. The artifacts used in the challenge and the ground truth rig with the sensor are shown in Figure 10.

Figure 10. Shows the artifacts used in the challenge and the ground truth rig with the sensor mounted at top left. [Permission granted by the author [41] to use the Figure]

Platform based systems have the following advantages over other options:
a) There is a potential of low cost, although increasing precision and accuracy can be expensive.
b) The desired pose and location can be set in advance, although possibly in limited increments.

Motion can be included, if a suitable conveyer or robot can be set to the required motion and accuracy. There is a potential to use robot arms with known repeatability to simulate the motion of a part. The disadvantages of the approach include the limited position options for the low cost systems, the artificiality of the scenario, and the difficulty of integrating multiple objects, clutter, and complex backgrounds. This approach is best for measuring pose and location of smaller individual parts.

3.4. Physically-Sensed Ground Truth Systems

In these systems, some physical element of the object is remotely sensed to determine pose and/or identity. The main physical quantities that have been identified for localization are radio frequency, optics (photonic), sound, and less frequently inertial data and contact or touch. We review these technologies by physical quantity, and then discuss a set of metrics by which the performance of different localization systems can be evaluated, such as static and dynamic accuracy, static and dynamic precision, scalability, range, and cost. We will then review the main technologies and classify them based on the physical quantity. There are a number of important performance metrics for identification and localization including accuracy, precision, update rate, degrees of freedom, dynamic vs. static data acquisition, number of tracking objects, latency, work volume, cost, and time to identify the object. In this paper we will not address all of these issues, as our intent is an initial survey of feasible methods. There are a number of review papers on localization [42-46, and 84].

Some ground truth systems can identify and localize objects, while others can mainly identify objects and others can mainly localize objects. It is also possible to have hybrid systems effective for both identification and localization. Since at time t=0 we know the identities and locations of all the objects, and if we can track the objects very accurately then we also know the identities of these objects at time t> 0.

Another issue to clarify is the nature of any required markers. To clarify terminology, we will define active markers or targets as those that broadcast or communicate with the detection system; compliant markers or targets are those that passively reflect a detector's signal; and markerless systems are those that work without the need of markers.

Optical positioning systems are currently the dominant ground truth technique that cover a wide area of applications at all levels of accuracy and cost, with its main application in the sub-millimeter domain for close range (one meter or less). The achievements of optical methods originate from improvements including cost reduction, miniaturization of actuators, and particularly with better performance of the detectors (Charge-Coupled Device (CCD)). A comprehensive survey of optical positioning systems can be found in [43, 46]. Optical systems provide high accuracy, but they require line of sight to an object (hence are affected by occlusion) and are impacted by interference of light signals from fluorescent lights and sunlight.

The alternatives to optical positioning systems are typically radio based, such as Radio Frequency Identification (RFID), Bluetooth, or Ultra-Wide Band (UWB). Similar in purpose to the optical internal GPS (Global Positioning System) systems,

wireless indoor positioning systems are widely used to locate people and objects within an indoor environment. These systems are very good at identification but are not very accurate in positioning. These systems are used in different application domains, such as locating products stored in a warehouse, positioning workers on a factory floor, positioning a first responder in a building, or locating medical personnel or patients in a hospital. There are a number of good survey papers on wireless positioning systems [42] [44]. Wireless systems are widely used because they can cover very large areas.

3.4.1. Indoor GPS systems

This subsection covers five systems that require either larger active targets, or sensors, on the object and so are only suitable for larger objects that can carry the weight. Some are designed for navigation of mobile robots.

The Nikon iGPS system [83] as shown in Figure 11 is a 3D high precision commercial measurement system based on laser transmission and receivers to determine the 3D pose of static or moving objects. It is modular, suitable for large or small volumes, and can enable factory-wide localization of multiple objects with high accuracy. It is widely used by industrial manufacturers both for positioning and tracking applications and for robotic control. An iGPS system consists of two or more static transmitters Figure 11., which continuously send out two rotating fan-shaped laser beams and a reference infrared pulse. Based on the Time Difference of Arrival (TDoA) between the three sources, the relative positions of the receivers with respect to the transmitters are determined.

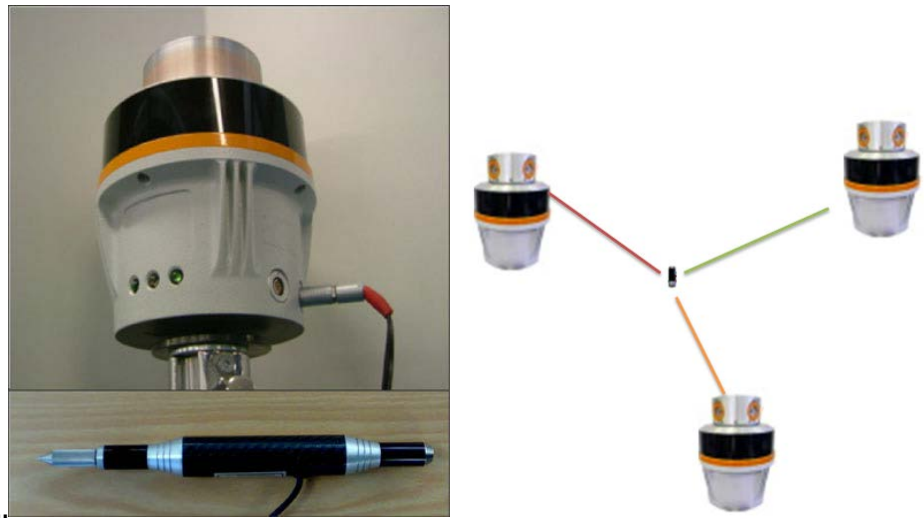

Figure 11. Shows an iGPS transmitter and two sensor receiver bars. [Permission granted by the company [83] to use the Figure]

The manufacturer specified static accuracy of 3D positions measured using the iGPS is 0.2 mm and the measurement rate is 40 Hz. A typical measurement volume based on four to eight transmitters is 1200 m^2. A detailed system analysis is presented by Schmitt et al. [47] and Mosqueira et al. [48]. Wang et al. [49] showed that the tracking accuracy is similar for speeds below 10 cm/s. However, as speed of an object is increased, the tracking accuracy goes down, to where at 1 m/s, the mean tracking error can be in the order of 3 mm to 4 mm. In another study, dynamic results presented are much better; Depenthal [50] showed that when tracking objects at velocities of 3 m/s, the 3D position deviation is less than 0.3 mm. Depenthal also described the experimental comparison of the dynamic tracking performance between an iGPS and a laser tracker and showed that the iGPS performed well under dynamic conditions. Depenthal also proposed a novel method for dynamic repeatability comparisons of tracking systems, by using four iGPS transmitters arranged around the rotating arm and a scale bar used for bundling as shown in Figure 12. The iGPS can be effective for collecting very accurate localization ground truth for moving objects in manufacturing environments.

Figure 12. The rotating arm and the scale bar, for dynamic repeatability comparisons of tracking systems, by using four iGPS transmitters arranged around the rotating arm and a scale bar used for bundling. [Permission granted by the author [50] to use the Figure]

Tilch and Mautz [51] of the Institute of Geodesy and Photogrammetry, ETH (Swiss Federal Institute of Technology) Zurich have developed the CLIPS (Camera and Laser based Indoor Positioning System) to determine the pose of an object using a mobile camera with respect to a laser hedgehog as shown in Figure 13 (right side). The hedgehog emits laser-beams from a virtual central location, which is like an inverse camera. By viewing the bright laser spots that are projected on any surface (e.g., ceiling, walls) without knowing any specific structure of the scene as shown in Figure 13 (left side), the relative positions and orientations of the camera and the laser hedgehog can be computed. Point tracking is achieved at frame rates of 15 Hz and the reported accuracy of the camera position is sub-millimeter. This type of system can be very effective for collecting very accurate localization ground truth for moving objects in manufacturing environments. The system is a prototype, the reported accuracy [51, 84] is 0.5 mm, the cost is 1000 Euro, the frame rate is 30 Hz, and the coverage area is (35 x 35) m.

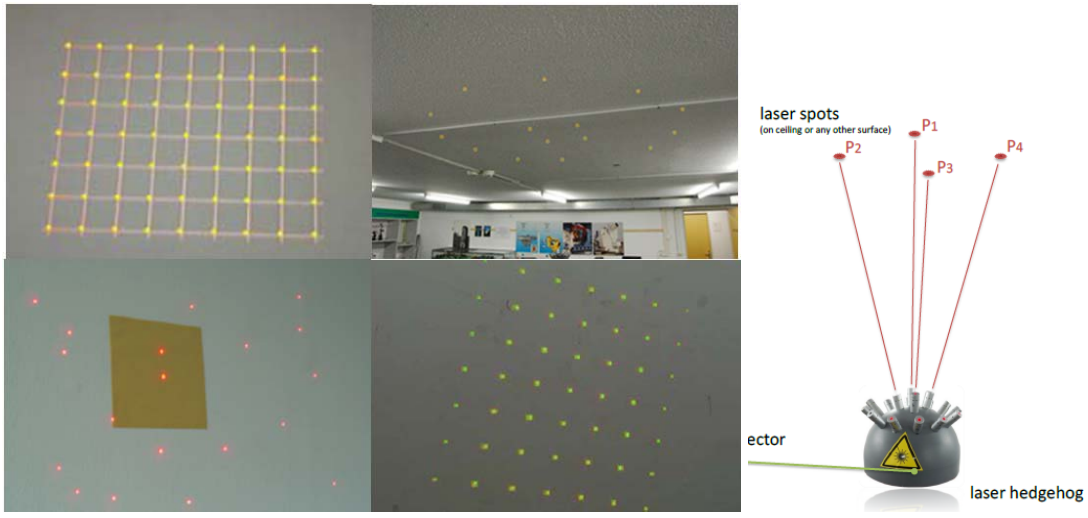

Figure 13. The CLIPS laser hedgehog projects laser spots on the ceiling. [Permission granted by the author [51] to use the Figure]

Evolution Robotics [52] developed the NorthStar indoor localization system for navigation of shopping carts or robotic vacuum cleaners, which could be used for ground truth location of AGVs or other mobile platforms. The location and pose of a mobile

robot is determined based on triangulation from infrared light spots, emitted from one or more infrared LEDs (Light Emitting Diodes). Each mobile unit can be equipped with an infrared detector and projector that allow determination of the relative orientation between mobile devices. The reported positioning accuracy is on the order of a few centimeters and the system has a 10 Hz update rate. Many configurations are possible where the projector or the detector is stationary as shown in Figure 14. This system can be used for collecting localization ground truth for moving objects in manufacturing environments. The system reported accuracy [52, 84] is on the order of centimeters and decimeters, the cost is US $1500, the frame rate is 10 Hz, and the coverage area is 35 m^2.

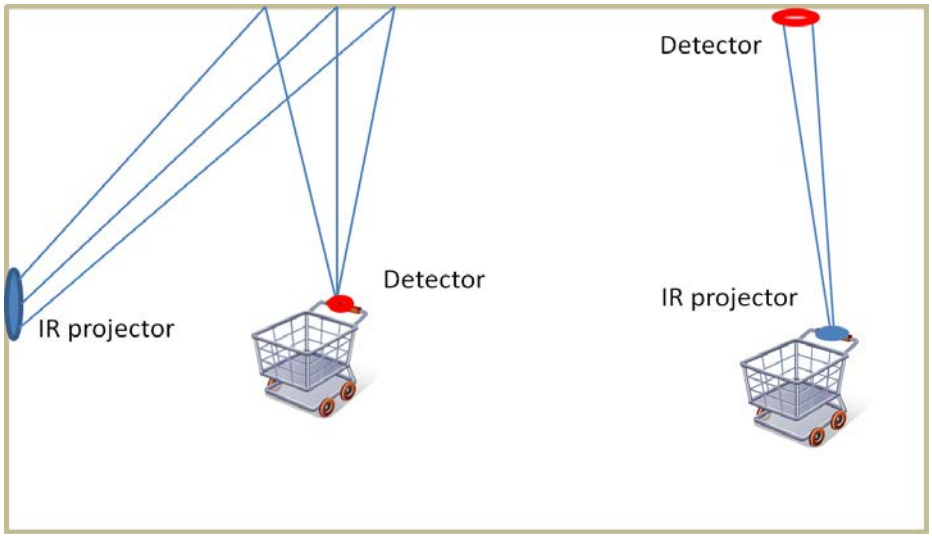

Figure 14. Different possible configuration of the NorthStar system. [Permission granted by the company [52] to use the Figure]

The Sky-Trax system [53] is an optical navigation system for forklift trucks in warehouses and for tracking the locations of materials and equipment inside warehouses and factories. Coded reference markers as shown Figure 15, are installed on the ceilings along the paths. On the top of each forklift, an optical camera takes pictures that are sent to a central server where they are processed and the position is determined based on triangulation. The position accuracy is reported to be +/- a few centimeters. The reported accuracy of the system is [53, 84] is 2 cm to 30 cm and the coverage area is scalable.

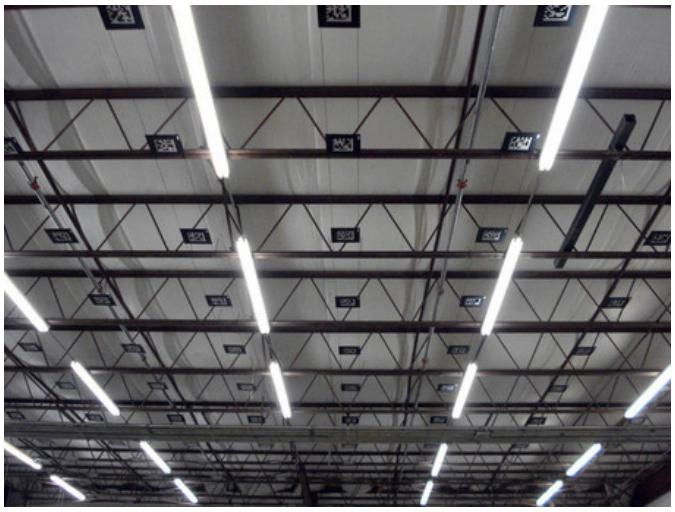

Figure 15. The coded markers as used by Sky-Trax system are shown [Permission granted by the company [53] to use the Figure]

The StarGazer system [54] is specifically designed for robot positioning and is based on retro reflective targets mounted on the

ceiling. An infrared camera views different point patterns that are actively illuminated by an infrared light source. Based on the points uniquely arranged on a 3 × 3 or 4 × 4 grid on the targets, a room can be identified and the pose of the roving camera can be determined within decimeter level accuracy. The system has a reported accuracy [54, 84] on the order of centimeters to decimeters, costs US $1000, has a frame rate of 20 Hz, and the coverage area is scalable. This system can also be used for collecting localization ground truth for moving objects in manufacturing environments, if the object can support the mounted camera.

The five optical GPS systems share the characteristics of being designed for indoor use, working over large, scalable volumes, using relatively heavy targets or sensors on the object being tracked, and being off the shelf commercial systems. Their advantages are the large working volume, the known performance, and the commercial availability. They are useful for larger objects in larger, open rooms.

3.4.2. Optically Based Augmented Reality Tags

Visual fiducials [55], passive compliant markers used for optical tracking, have been used in many applications for a number of years. They can be simple circles, points, or squares, and used for identification and pose estimation. Simple, circular markers may need to be used in sets of three or more to determine pose, while other shapes like squares support pose from a single marker. Recently, the most vigorous research has focused on their use in Augmented Reality (AR) applications [56, 57], leading to the term AR tags and the popularization of the square format. Although related to other 2D barcode systems, visual fiducials have a small information payload and are designed to be automatically detected and localized. Visual fiducial systems provide relative position and orientation of a tag and can detect a number of tags in an image. The orientation estimation of a single tag is usually poor. This problem has been addressed by a novel method, based on variable moiré patterns [58]. Some of these patterns are shown Figure 16. These tags can be used to obtain both localization and identification ground truth for stationary and moving objects in manufacturing environments. Some of these systems are commercially available systems and some are prototypes. The expense can be minimal since open-source systems exist. Since the system performance is dependent on the resolution of the camera, the speed of the computer, and the specific tag system used, the performance is not easily quantifiable. The reported data in [58] is (10 to 15) frames per second angular accuracy of about 1° about all three axes in a limited range of front-facing angles, position accuracy of about a centimeter, and the working range up to about 3 m.

Visual fiducals are a special case of barcodes that provide pose and location data. For identification alone, 1D and 2D bar codes can be used [85]. They are already widely used in manufacturing and logistics for material tracking and inventory. Military standard 1D LOGMARS (Code 39) is required by the U.S. Defense Department for inventory labeling in some contracts, and the standard is open for general industrial use. Matrix, or 2D, codes can carry more information on a small scale making them better for smaller parts.

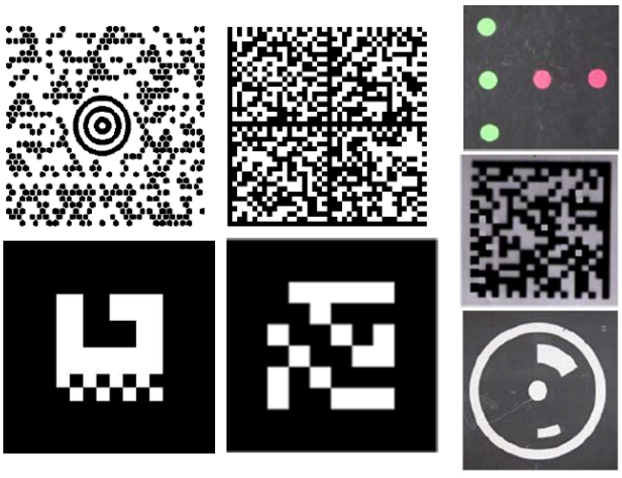

Figure 16. Different types of existing planar markers (coded targets). (Art-toolkit, Art-tag, Maxi code, Data matrix symbols, Concentric Rings, etc.). [Permission granted by the authors [55, 56, 57, 58]] to use the Figure]

If we generalize from fixed barcodes, any known and fixed texture, such as logos on commercial packaging, can be used for product identification as well as pose and location. There are commercial and open source systems based on visual pattern recognition technology that provide reliable and robust vision solutions for particular categories of objects [75, 76]. As in the NIST Perception Challenge discussed previously, the objects should have areas of flat texture.

In a second generalization, active optical systems illuminate one LED or multiple LEDs and powering each LED marker successively in phase with the capture system provides a unique identification of each. These are in effect temporal bar codes. The ability to identify each marker in this manner is useful in real-time applications and for collecting data from large numbers of objects and people [68]. Time modulated active markers based on high-speed cameras, and systems used for tracking [69], also provide unique identification. The main drawbacks of these optical systems are that line of sight is required and the targets have small information payload.

Visual fiducials and barcodes are inexpensive, widely used technologies that can provide ground truth for identification, pose, and location. Cost, performance, and work volume vary. If used to provide ground truth for testing optical systems, these targets have the disadvantage of being highly visible so they can give systems under test unwanted information. Also, they typically require flat surfaces, but can be attached to a base or fixture that accompanies an object. Line of sight can also be a problem.

3.4.3. Radio based systems

Radio frequency based systems [61, 62] are widely used for identification, and less so for localization. They can provide a large payload of information about the detected objects. Unlike optical systems, these systems do not require line of sight and can be embedded in the tracking object, and can be invisible to optical tracking systems.

Ultra wide band (UWB) is based on sending ultra-short pulses [59, 60] over multiple bands of frequencies simultaneously. Each receiver has a unique identification and is used in outdoor and indoor applications [25]. UWB is robust and provides higher precision indoor positioning compared to other wireless technologies. These systems can cover a very large area compared to other technologies and have been successfully used for collection human tracking data [25, 26]. The UWB systems are commercially available and have reported accuracies [59, 84] of 0.5 m. They cost about US $10,000 dollar and the frame rate is 30 Hz.
For example, Bodt at al. [25, 26] evaluated multiple algorithms for real-time detection of pedestrians using Laser Detection and Ranging (LADAR) and video sensor data taken from a moving platform by evaluating it with an ultra-wideband (UWB) technology as their ground truth system.

Radio Frequency Identification (RFID) is very attractive for identification because of the reasonable system price, and reader reliability. RFID is a wireless non-contact system that uses radio frequencies to transfer data from a RFID tag for the purpose of identification and tracking. It has recently been also used for localization [61, 62]. There are two types of RFID tags, passive (not powered) and active (self-powered) and the accuracy is directly proportional to the tag density and the deployment and reading ranges. The reported range of active RFID commercial systems is 3 m to 70 m and the accuracy is 1.5 m. For passive RFID tags, the range is 2m and the accuracy is on the order of decimeters to meters. These systems can be used for collecting ground truth data when identification is more important than the localization error. RFID tags need to be selected for the application, considering part composition (metal parts will reflect some RFID frequencies), distance between reader and part, and whether the tag is passive or active.

A number of people and companies have developed localization techniques based on Wireless Fidelity (WIFI) data [63] and Bluetooth data [64]. These systems are cheap and very easy to deploy; but they are not very accurate. These systems can be used to collect ground truth data where gross localization of objects is sufficient. There are a number of commercial products and the reported accuracies [63, 64, 84] range from 1 m to 10 m, the frame rate is 30 Hz, and the coverage area is scalable.

Radio based systems have the previously mentioned advantages of not requiring line of sight and can therefore be embedded in the tracking object which makes it invisible to optical systems under test. They differ in their range and their performance can

be affected by the part/object materials. Thus, the selection of the type of system requires care and consideration of the conditions in particular scenarios.

3.5. Motion Capture systems

Motion capture (MoCap) [65] refers to a wide category of methods for recording the motion of objects and people, and for capturing the articulated motion of a whole human body and/or robotic arm. This technology was introduced earlier in the discussion of human tracking and is discussed in more detail in this section because while used in a number of applications, the technology is typically specialized for articulated human motion tracking. It is widely used for ground truth data collection for validating the performance of computer vision systems and also for applications such as military, entertainment, sports, and gait/medical applications and robotic control. In some domains, motion capture is called motion tracking and when the capture includes gestures and expressions based on faces and finger motion, it is sometimes referred to as performance capture. Motion capture systems are based on physical sensors such as optical, acoustic, inertial, LED, magnetic, or reflective markers, or hybrid systems where markers [Figure 17] near each joint are attached to identify the motion from the positions of and angles between the markers. These systems can be very effective for the collection of ground truth data of the articulated human motion for human-robot collaboration, human-object interaction, human activity, and human-human interaction applications in manufacturing environments. In the paper by Cloete [31] the authors benchmark different types of motion capture systems based on their accuracy and suitability for clinical gait analysis. Noonan at el. [33] use a motion capture system to validate a computer vision system for minimally invasive surgery.

Optical mocap systems exploit data captured from multiple cameras to triangulate the 3D position of a marker. These systems produce data with three degrees of freedom for each marker, and rotational information must be inferred from the relative orientation of three or more markers. Some of the newer hybrid systems combine optical sensors with inertial systems and a RF positioning system to reduce occlusion, increase the number of trackable objects, and improve the ability to track. The optical systems are also based on different types of markers, which are described in the next subsection.

Figure 17. Reflective markers are attached to the surface of a human or object to identify and capture the 3D motion. (Image from http://www.flickr.com/photos/funksoup/88138710/ Creative commons))

3.5.1. Passive, Compliant markers

Two systems based on markers coated with reflective material to reflect light are OptiTrack [66] and Vicon MoCap[67]. Both support tracking multiple single points, from which pose and identity can be inferred. Balan [27] evaluated the 3D pose of human motion obtained from synchronized, multi-camera video against 3D ground truth poses acquired with a Vicon motion capture system [67], i.e., the Vicon was used as ground truth.

The paper by Wienke [34] describes a framework for acquisition of multimodal human-robot interaction datasets (shown in Figure 18). The dataset was captured using three High Definition (HD) cameras; the ground truth data for the human and Nao robots (a type of autonomous, programmable, and humanoid robots) was captured by Vicon MoCap with additional data captured by the Nao robot, including video, audio, audiometry, joint angles, and audio. The paper describes the framework in detail and explains the acquisition of data sets and issues involved which include system level information.

Figure 18. Figure shows the Noa robot and the Vicon MoCap framework for acquisition of multimodal human-robot interaction datasets. [Permission granted by the author [34] to use the Figure]

The Korean University Gesture (KUG) dataset [37] was created for human tracking and gesture recognition research by recording motion capture data along with video from multiple cameras. The Georgia Tech human identification at a distance database [38] was created for gait recognition and has motion capture data along with videos of 20 walking subjects.

3.5.2. Active markers

Active optical systems illuminate one LED at a time very quickly, or multiple LEDs in known relative positions, and use software to identify them by their relative positions. Rather than reflecting light, these markers emit their own light, which can increase the distances and volumes over which the systems can be used. By powering each LED marker successively in phase with the capture system, a unique identification of each marker is provided at each time frame. The ability to identify each marker in this manner is useful in real time applications and for collecting data from a large number of objects and people. PhaseSpace [68] manufactures a commercial system based on active LED markers. Asteriadis et al. [36] extracted ground truth data to be used as a benchmark for head pose estimation using three LEDs placed on the face and extracting the directional information at each frame.

3.5.3. Time modulated active markers

In Raskar et al. [69] motion tracking is done by multiple small, active photodetector units on the object or person. The receiver units detect time-encoded signals broadcast by multiple LEDs, in a form of optical GPS where the LEDs multiplex by time delay. Each receiver wirelessly returns the strength of the detected signal for each unique broadcasting LED and from these messages the system computes the location of receiver unit (identity being given by the receiver id.) Since the receiving units are small, are self-powered, and do not broadcast, they can be hidden on a person or object of suitable size and profile. The system performance can be high speed (500 Hz or higher), indoor or outdoor, in natural scenes, with location and orientation resolution depending on the configuration. In the experiments performed for the article by Raskar, the working distance was around 5 m and the accuracy near 2 cm. The system uses low-cost components but is not currently commercially available and no apparent development has taken place since the initial publication of the technology in 2007.

3.5.4. Markerless human motion capture

Two recent markerless motion capture systems are under development at Stanford [70] and the University of Maryland [71]. These systems are based on finding the silhouette, or visual hull, of the human body in multiple cameras and matching a model to the hull. A commercial markerless system, based on a similar approach is Organic Motion [72]. This system requires strong lighting, no occlusion, and a room with simple white walls, so will not be applicable to many manufacturing scenarios. It can track up to four actors at 60 frames per second (fps) in a (5 x 5) meter volume, with 25 ms to 50 ms latency and millimeter resolution of a highly articulated human model.

Baat et al., [35] analyzed and evaluated the performance of a markerless motion capture system compared to inertial motion capture sensors. The TUM (Technische Universität München) Kitchen Data Set [24] provides Motion Capture data extracted from videos using their own markerless full-body MeMoMan tracker [24]. These markerless motion capture systems can provide reliable data compared to other systems for some of the cases.

3.5.5. Inertial MoCap System

Inertial sensors [73] have been used for motion tracking and motion capture. These systems have become popular with the availability of low-cost, miniature inertial sensors. These sensors mainly use the integral of angular velocity measured by gyroscopes. These systems are used for tracking pedestrians and first responders using shoe-mounted sensors and can also generate trajectories based on a dead reckoning method. Inertial sensors are also popular for motion capture as shown Figure 19. Xsen [73] is a commercial version whose specifications claim operation up to 150 m in clear indoor or outdoor space, or 50 m in an indoor office environment. The onbody system weighs 1930 g (4.2 lbs). The orientation accuracy is about 0.5°, and it handles accelerations up to 180 m/s^2, but location accuracy is not given in the technical reference. Yang [32] presents a simple and accurate method for evaluating the similarities between human postures from inertial motion capture data. The advantages of these systems are that they can be used both indoors and outdoors, and in any light condition, and since line of sight is not required, can capture motion data from multiple peoples. The disadvantage of these systems is that they suffer from drift so there needs to be some correction mechanism.

Figure 19. Shows an inertial motion capture system for tracking and motion capture. [Permission granted by the company [73] to use the Figure]

3.5.6. Mechanical MoCap Systems

Mechanical motion capture systems record motion directly with an exo-skeletal suit attached to a person (see Figure 20). As the person moves the relative joint angles and motion are captured and sent to a server for processing. Since only relative motion is captured, another type of sensor is required for absolute positioning. The advantages of these systems are that their cost is much lower compared to other motion capture systems and the systems can be used both indoors and outdoors, and in any light condition. The disadvantages of these systems are that they are not very accurate and not comfortable to wear because of the heavy exoskeleton.

Figure 20. Shows a mechanical motion capture system. [Permission granted by the company [82] to use the Figure]

3.5.7. Magnetic MoCap systems

Magnetic systems capture position and orientation based on magnetic fields on both the transmitter and receiver. Their output is a 6DOF pose. The systems can be based on "AC" and/or "DC" magnetic fields. The systems are not affected by nonmetallic objects but are susceptible to magnetic and electrical interference from metal objects in the scene. The main drawbacks of these systems are limited range, the influence of metal objects, and that they require frequent re-calibration.

3.5.8. Motion Capture system summary

Motion capture systems, distinguished here because they originated for articulated human tracking, are a well-established technology with good potential for tracking people and objects in manufacturing test scenarios. They excel relative to other systems at tracking a full human skeletal model, and are therefore useful for scenarios for human task and gesture recognition. However, they can be intrusive, with highly visible markers or bulky suits, so would be hard to use for extended periods, for testing optical or other tracking systems where the appearance or bulk of the suits would interfere. Lighter weight Inertial Measurement Unit (IMU) systems can be mounted invisibly and could mitigate those disadvantages. Similarly, markerless tracking would not have these disadvantages, but have requirements for room lighting and configuration that would prevent its use in a general, cluttered manufacturing environment. Optical MoCap systems based on isolated passive markers can be used for multiple purposes, such as tracking a single point, the pose and location of an object, or full human tracking.

4. Conclusions

Currently, there do not seem to be universal solutions for acquiring ground truth for human and object detection and tracking. There are a number of partial solutions suitable for particular applications, with each solution having various drawbacks.

Physical simulations offer a number of advantages, but a critical drawback is that the synthetic data they generate is not suitable for testing perception systems. It would be useful to better understand when this data might be adequate, what can be done to improve the data, and how physical simulations can contribute at least in part to perception system testing.

Annotation systems are very useful for sensor-centric ground truth for tracking and identification, and for creating datasets for wide distribution. They can be labor-intensive to use, and are thereby limited and hard to rapidly adapt, but can serve well for initial validation of approaches.

For uncluttered, static tests of object pose and identification, a platform-based system can provide adequate pose information, and identification is inherent in the setup. Limited motion can be introduced if the platform supports it. If not automated by a computer-controlled mechanism, the testing process can be slow, and it is difficult to integrate into cluttered scenes or use with humans. Still, the testing setup can be realistic for manufacturing scenarios that require detecting and locating single objects against controlled backgrounds.

For uncluttered, dynamic tests for object pose and identification, there are a number of potential ground truth sensing technologies for identifying and locating objects. The best technology for an application can be dependent on the system to be tested and the required precision. If the system under test is optical, it is preferable that the ground truth systems not require visual markers that could interfere with the test validity.

For cluttered scenes, both static and dynamic, where optical line of sight cannot be assured, the options are fewer. Longer reading RFID tags can be used for identification, while the options for pose detection are highly dependent on the details of the scenario. Each of these techniques needs to be validated before use as ground truth, so their resolution and accuracy can be verified and their robustness tested.

5. Bibliography

[1] Software rendering, http://en.wikipedia.org/wiki/Software_rendering. December 15, 2012.
[2] Mihalcik, David, and Doermann, David. The design and implementation of ViPER. Technical report, 2003.
[3] Dasiopoulou, Stamatia, Giannakidou, Eirini, Litos, Georgios, Malasioti, Polyxeni, and Kompatsiaris, Yiannis. A survey of semantic image and video annotation tools. Knowledge-driven multimedia information extraction and ontology evolution 196-239, 2011.
[4] Miller, Michael. and McCathieNevile, Charles., Semantic web tools to help authoring: A semantic web image annotation tool. In: SWAD-Europe Deliverable 9.3, 2001.
[5] Lux, Mathias, Becker, Jutta and Krottmaier, Harald. Caliph & emir: Semantic annotation and retrieval in personal digital photo libraries. In Proceedings of CAiSE, vol. 3, pp. 85-89, 2003.
[6] Kipp, Michael. Anvil - a generic annotation tool for multimodal dialogue. In: in Proc. 7th European Conf. on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, 2001.
[7] Schallauer, Peter, Sandra Ober, and Helmut Neuschmied. Efficient semantic video annotation by object and shot re-detection. In Posters and Demos Session, 2nd International Conference on Semantic and Digital Media Technologies (SAMT), Koblenz, Germany, 2008.
[8] Russell, Bryan C., Torralba, Antonio, Murphy, Kevin P. and Freeman, William. LabelMe: a database and web-based tool for image annotation. International journal of computer vision 77, no. 1, pp. 157-173, 2008.

[9] Vondrick, Carl, Ramanan, Deva and Patterson, Donald. Efficiently scaling up video annotation with crowd sourced marketplaces. Computer Vision–ECCV 2010, pp. 610-623, 2010.
[10] Geiger, Andreas, Lenz, Philip and Urtasun, Raquel. Are we ready for autonomous driving? the kitti vision benchmark suite. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, pp. 3354-3361, 2012.
[11]CAVIAR. Context aware vision using image-based active recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, December 15, 2012.
[12] i-LIDS. Image library for intelligent detection systems.http://scienceandresearch.homeoffice.gov.uk/hosdb2/physical-security/detection-systems/i-lids/, December 15, 2012.
[13] CLEAR, Classification of events, activities and relationships—evaluation campaign and workshop. http://www.clear-evaluation.org/, December 15, 2012.
[14] CREDS. Call for real-time event detection solutions (creds) for enhanced security and safety in public transportation. ,http://www.visiowave.com/pdf/ISAProgram/CREDS.pdf, December 15, 2012.
[15] ETISEO. Video understanding evaluation. http://www-sop.inria.fr/orion/ETISEO/ , December 15, 2012.
[16] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), http://pets2007.net/, December 15, 2012.
[17] VACE. Video analysis and content extraction. December 15, 2012, http://www.perceptual-vision.com/vt4ns/vace brochure.pdf
[18] Dollár, Piotr, Wojek, Christian, Schiele, Bernt and Perona, Pietro. Pedestrian detection: A benchmark. In Computer Vision and Pattern Recognition, IEEE Conference on, pp. 304-311, 2009.
[19] English, Chad, Okouneva, Galina , Saint-Cyr, Pierre Choudhuri, Aradhana and Luu, Tim. Real-time dynamic pose estimation systems in space: Lessons learned for system design and performance evaluation. Special issue on Quantifying the performance of intelligent systems, International Journal of Intelligent Control and Systems 16, pp. 79-96, 2011.
[20] Gross, Ralph, and Shi, Jianbo. The cmu motion of body (mobo) database. (2001).Technical Report CMU-RI-TR-01-18, Robotics Institute, 2001.
[21] CMU GLMC, http://mocap.cs.cmu.edu/, December 15 2012.
[22] Guerra-Filho, Gutemberg and Biswas, Arnab . The human motion database: A cognitive and parametric sampling of human motion. Image and Vision Computing, 30, 3, pp. 251-261, 2012.
[23] Hodgins, Fernando, De la Torre, Jessica, and Macey, J.. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. CMU-RI-TR-08-22, 2009.
[24] Tenorth, Moritz, Bandouch, Jan and Beetz, Michael. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In Computer Vision Workshops (ICCV Workshops), 2009.
[25] Bodt, Barry, Camden, Richard, Scott, Harry , Jacoff, Adam, Hong, Tsai, Chang, Tommy, Norcross, Rick, Downs, Tony and Virts, Ann. Performance measurements for evaluating static and dynamic multiple human detection and tracking systems in unstructured environments, PerMIS '09: Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems, 2009.
[26] Bodt, Barry and Hong, Tsai UGV SAFE OPERATIONS CAPSTONE EXPERIMENT, Army Science Conference Paper, 2010.
[27] Balan, Alexandru, Sigal, Leonid and Black, Michael J..A Quantitative Evaluation of Video-based 3D Person Tracking, Proceedings 2nd Joint IEEE International Workshop on VS-PETS, Beijing, 2005.
[28] Sigal, Leonid, Balan, Alexandru O. and Black, Michael J.. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision 87, no. 1: 4-27, 2010.
[29] Sigal, Leonid, and Black, Michael J. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Brown Univertsity TR 120, 2006.
[30] Van der Aa, Nico, Luo, Xinghan, Geert-Jan , Tan, Robbby T. and Veltkamp, Remco C.. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In Computer Vision Workshops (ICCV Workshops), 2011.
[31] Cloete, Teunis, and Scheffer, Cornie. Benchmarking of a full-body inertial motion capture system for clinical gait analysis. In Engineering in Medicine and Biology Society, 2008.
[32] Yang, Wei-Ting , Luo, Zhiqiang , Chen, I-Ming and Yeo, Song Huat. A Method for Comparing Human Postures from Motion Capture Data, ROMANSY 18 Robot Design, Dynamics and Control, CISM International Centre for Mechanical Sciences, Volume 524, Chapter VI, 441-448, DOI: 10.1007/978-3-7091-0277-0_52, 2010.
[33] Noonan, David, Mountney, Peter, Elson, Daniel , Darzi, Ara, and Yang, Guang-Zhong. A Stereoscopic Fibroscope for Camera Motion and 3D Depth Recovery During Minimally Invasive Surgery. In proc ICRA 2009, pp. 4463-4468, 2009.
[34] HUMAVIPS Project: http://vernissage.humavips.eu/related.html, Johannes Wienke, David Klotz, Sebastian Wrede, A Framework for the Acquisition of Multimodal Human-Robot Interaction Data Sets with a Whole-System Perspective, LREC Workshop on Multimodal Corpora for Machine Learning: How should multimodal corpora deal with the situation, 2012.

[35] Baak, Andreas, Helten, Thomas, Müller, Meinard , Pons-Moll, Gerard, Rosenhahn, Bodo and Seidel, Hans-Peter. Analyzing and Evaluating Markerless Motion Tracking Using Inertial Sensors. ECCV, 2010.
[36] Asteriadis, S., Soufleros, D., Karpouzis, K., and Kollias, S. (2009, November). A natural head pose and eye gaze dataset. In ACM 2009.
[37] Hwang, Bon-Woo, Kim, Sungmin and Lee, Seong-Whan. A full-body gesture database for automatic gesture recognition. In IEEE Automatic Face and Gesture Recognition, 2006.
[38] Georgia Tech Human Identification at Distance Database. http://www.cc.gatech.edu/cpl/projects/hid/.
[39] Marvel, Jeremy A., Falco, Joe and Hong, Tsai-Hong. Ground truth for evaluating 6 degrees of freedom pose estimation systems. In ACM Proceedings of the Workshop on Performance Metrics for Intelligent Systems, pp. 69-74, 2012.
[40] Rusu, Radu Bogdan, Bradski, Gary , Thibaux, Romain and Hsu, John. Fast 3d recognition and pose using the viewpoint feature histogram. In Intelligent Robots and Systems (IROS), 2010.
[41] Marvel, Jeremy A., Hong, Tsai-Hong and Messina, Elena. 2011 solutions in perception challenge performance metrics and results. In ACM Proceedings of the Workshop on Performance Metrics for Intelligent Systems, pp. 59-63, 2012.
[42] Torres-Solis, Jorge, Falk, Tiago H. and Chau, Tom. A review of indoor localization technologies: towards navigational assistance for topographical disorientation. Ambient Intelligence: pp. 51-84, Bloorview Research Institute & University of Toronto (intech Chapter), 2010.
[43] Gu, Yanying, Lo, Anthony and Niemegeers, Ignas. A survey of indoor positioning systems for wireless personal networks. Communications Surveys & Tutorials, IEEE 11, no. 1: 13-32, 2009.
[44] Liu, Hui, Darabi, Houshang, Banerjee, Pat and Liu, Jing. Survey of wireless indoor positioning techniques and systems. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 37, no. 6, pp. 1067-1080,2007.
[45] Al Nuaimi, Klaithem, and Kamel, Hesham. A survey of indoor positioning systems and algorithms. In Innovations in Information Technology (IIT), IEEE 2011 International Conference on, pp. 185-190, 2011.
[46] Mautz, Rainer, and Tilch, Sebastian. Survey of optical indoor positioning systems. In Indoor Positioning and Indoor Navigation (IPIN), IEEE 2011 International Conference on, pp. 1-7, 2011.
[47] Schmitt, Robert, Nisch, S., Schönberg, A. , Demeester, F. , and Renders, S.. Performance evaluation of iGPS for industrial applications. In IEEE Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on, pp. 1-8., 2010.
[48] Mosqueira, G., Apetz, J., Santos, K. M. , Villani, E., Suterio, R. and Trabasso, L. G.. Analysis of the indoor GPS system as feedback for the robotic alignment of fuselages using laser radar measurements as comparison. Robotics and Computer-Integrated Manufacturing 28, no. 6: 700-709, 2012.
[49] Wang, Zheng, Mastrogiacomo, Luca , Franceschini, Fiorenzo and Maropoulos, Paul. Experimental comparison of dynamic tracking performance of iGPS and laser tracker. The International Journal of Advanced Manufacturing Technology 56, no. 1,pp. 205-213, 2011.
[50] Depenthal, Claudia. Path tracking with IGPS. In Indoor Positioning and Indoor Navigation (IPIN), IEEE 2010 International Conference on, pp. 1-6. ETH Zurich, Switzerland, 2010.
[51] Tilch, Sebastian, and Mautz, Rainer. CLIPS – A Novel Optical Indoor Positioning System, IPIN Conference 2010, Zurich, pp. 15-17, 2010.
[52] Evolution Robotics, NorthStar Navigation system, The Company was bought by iRobot, www.irobot.com, December 15, 2012.
[53] The Sky-Trax system , http://totaltraxinc.com/index.php/smart-forklift-solutions/forklift-tracking/sky-trax, December 15, 2012.
[54] StarGaze robotic localization, http://www.robotshop.com/hagisonic-stargazer-localization-system-3.html , December 15 2012.
[55] Xu, Anqi, and Dudek, Gregory. Fourier Tag: A Smoothly Degradable Fiducial Marker System with Configurable Payload Capacity. In IEEE Computer and Robot Vision (CRV), 2011.
[56] Wagner, Daniel, Reitmayr, Gerhard Mulloni, Alessandro, Drummond, Tom and Schmalstieg, Dieter. Pose tracking from natural features on mobile phones. In Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 125-134, 2008.
[57] Fiala, Mark. Automatic projector calibration using self-identifying patterns. In Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, pp. 113-113, 2005.
[58] Tanaka, Hideyuki, Sumi, Yasushi and Matsumoto, Yoshio. Avisual marker for precise pose estimation based on lenticular lenses. In Robotics and Automation (ICRA), 2012.
[59] Smith, John, UWB - a context-aware system perspective, Workshop on Positioning, Navigation and Communication, Ubisense Limited, 2004.

[60] Gezici, Sinan, Tian, Zhi , Giannakis, Georgios B. , Kobayashi, Hisashi, Molisch, Andreas F. , Poor, H. Vincent and Sahinoglu, Zafer. Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. IEEE *Signal Processing Magazine,* 22, no. 4, pp. 70-84, 2005.

[61] Finkenzeller, Klaus. RFID–Handbuch. Hanser Fachbuch. Also available in English as RFID Handbook: Radio-Frequency Identification Fundamentals and Applications, JohnWiley & Sons, 2000.
[62] Hahnel, Dirk, Burgard, Wolfram , Fox, Dieter, Fishkin, Ken and Philipose, Matthai. Mapping and localization with RFID technology. In Robotics and Automation, 2004. Proceedings. ICRA'04. IEEE International Conference on, vol. 1, pp. 1015-1020, 2004.
[63] Biswas, Joydeep, and Veloso, Manuela. Wifi localization and navigation for autonomous indoor mobile robots. Robotics and Automation (ICRA), 2010 IEEE International Conference on, 2010.
[64] Bargh, Mortaza S., and Groote, Robert de. Indoor localization based on response rate of bluetooth inquiries. Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments, 2008.
[65] Moeslund, Thomas B., Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104.2: pp. 90-126, 2006.
[66] Optitrack (motion capture), http://www.naturalpoint.com/optitrack/, December 15 2012.
[67] Vicon motion capture system, www.vicon.com/, December 15, 2012.
[68] PhaseSpace motion capture system, December 15 2012, http://www.phasespace.com/impulse_motion_capture.html
[69] Raskar, Ramesh, Hideaki Nii, Bert Dedecker, Yuki Hashimoto, Jay Summet, Dylan Moore, Yong Zhao et al. Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. In ACM Transactions on Graphics (TOG), vol. 26, no. 3, pp. 36, 2007.
[70] Corazza, S., Mündermann, L. , Chaudhari, A. M. , Demattio, T. , Cobelli, C. , and Andriacchi, T. P.. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. Annals of Biomedical Engineering 34, no. 6, pp. 1019-1029, 2006.
[71] Sundaresan, Aravind, and Chellappa, Rama. Markerless motion capture using multiple cameras. In IEEE Computer Vision for Interactive and Intelligent Environment, pp. 15-26, 2005.
[72]Organic motions capture, http://www.organicmotion.com/, December 15 2012.
[73] Xsens motion capture systems, http://www.xsens.com/en/general/mvn, December 15 2012.
[74] Chu, Chung-Hua, Yang, De-Nian and Chen, Ming-Syan. Image stablization for 2D barcode in handheld devices. In Proceedings of the ACM, 15th international conference on Multimedia, pp. 697-706, 2007.
[75] Evoretail LaneHawk, http://www.evoretail.com/lanehawk/, December 15, 2012.
[76] StopLift checkout vision systems, www.stoplift.com/, December 15, 2012.
[77] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88, no. 2: pp. 303-338, 2010.
[78] Torralba, Antonio, Fergus, Rob and Freeman, William T.. 80 million tiny images: A large data set for nonparametric object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30, no. 11: pp. 1958-1970, 2008.
[79] Yao, Benjamin, Yang, Xiong and Zhu, Song-Chun. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer Berlin/Heidelberg, 2007.
[80] In Kyu Park, Germann, Marcel , Breitenstein, Michael D. and Pfister, Hanspeter Fast and automatic object pose estimation for range images on the GPU, Machine Vision and Applications, DOI 10.1007/s00138-009-0209-8, 2009.
[81] Ressler, Sandy , Godil, Afzal , Wang, Qiming, Seidman, Gregory. A VRML integration methodology for manufacturing applications. ACM Proceedings of the fourth symposium on Virtual reality modeling language, pp. 167-172, 1999.
[82] Gypsy, http://www.animazoo.com/, December 15, 2012.
[83] Nikon iGPS, December 15, 2012, http://www.nikonmetrology.com/en_US/Products/Large-Volume-Applications/iGPS/iGPS
[84] Mautz, Rainer. Indoor Positioning Technologies. PhD diss., Habil. ETH Zürich, 2012.
[85] Osman, Keith A., and Furness, Anthony. Potential for two-dimensional codes in automated manufacturing. Assembly Automation 20, no. 1: pp. 52-57, 2000.