EURON
IST-2000-26048
European Robotics Network

KA 1.10 Benchmarks for Robotics Research

Rüdiger Dillmann
University of Karlsruhe

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

24th April 2004

Table of Contents

1 Introduction

Today's robots are systems with a very high degree of complexity. Their function is a cooperation of their separate components, as there are actuators, controllers, sensors, computer hardware & software, interactive components, and so on. Obviously, it is not a trivial task to evaluate such a complex system and compare it to others.

One useful tool for such an evaluation are *benchmarks*. Hanks briefly describes benchmarks as "precisely defined, standardized tasks" [21]. This short definition contains three essential aspects of benchmarks:

1. Task: the robot has to perform a given mission, e.g., it actually has to do something.
2. Standard: the benchmark is accepted by a significant set of experts in the field.
3. Precise Definition: the task is described exactly, especially the execution environment, the mission goal, and limiting constraints.

Unfortunately, this definition lacks one important feature of benchmarks, which is a numerical evaluation of the performance. Without that, it is only possible to decide whether or not a given system is able to perform a mission. What we need in fact is to "develop performance metrics" [20] for a given application. With such a score we are able to evaluate systems that only partially accomplished the mission, or decide, how well the mission was finally accomplished.

Furthermore, benchmarks must have the following features: repeatability, independency, and unambiguity. It must be possible to perform a benchmark test with reasonable resources, and the expected outcome stays more or less the same. Any benchmark has to produce a score for the tested system that is independent of the observer and unambiguous.

Additional features that are highly desirable are relation to reality, a widespread acceptance and use by a relevant user group, and the applicability to problems of the real world [19]. It is clear that the mission and the constraints should reflect reality on a certain degree. If that is not completely possible, then the design of the task should at least cover some important facts of the real world which make the results transferable to a particular amount. At last, a benchmark should be accepted by a majority of the users, otherwise it will be useless.

Development and design of benchmarks is a controversial issue. Each party has her own visions and expectations for their system, which most often differ from those of the other parties. It is necessary that experts agree on one and the same benchmark. In many cases, recognized authorities develop standards and benchmarks, which will then be accepted by the many.

Benefits of the introduction and application of benchmarks are the comparability of very complex systems. But on the other side of the coin there are also disadvantages connected to the introduction of benchmarks. As soon as benchmarks enter the field and are widely respected, researchers and manufacturers are likely to compare and optimize their products to the benchmarks rather than to the real application areas. Whenever there exists a gap between the benchmark and the real world, optimization towards the benchmark test will not necessarily improve the system's performance in the real application.

We identified two different aspects of how to categorize benchmarks (cf. to illustration). One way to classify benchmarks is by method. The analytical method observes the system and evaluates only by observation of the system its performance. The functional method will probe the system on a specific problem and generates from the performance on that problem the benchmark score.

Another way to classify benchmarks is by focus. Does the benchmark consider the system as a whole or as a sum of its components respectively its separate qualities? With these two categories in mind, there are 4 types of benchmarks: analytical benchmarks that consider components, analytical benchmarks that consider complete systems, functional benchmarks that consider components, and functional benchmarks that consider complete systems. There will be references to this classification later in this report.

analytical		
functional		
	component	system

Fig. 1 The benchmark classification diagram

The next sections will cover the benchmark topic in industry as well as in research. The following section discusses open problems and our visions to the benchmark problem. Finally, we will draw conclusions from the collected results.

2 Benchmarks in Industry

Benchmarks in industry are established in various areas for quite some time. There are even organizational structures, which take care of creating and maintaining reliable tests. Industry vendors have the highest level of interest in developing credible benchmarks. Without good evaluation tools, vendors would not be able to do valid system comparisons when developing new products, or gain recognition from the trade media and public for significant technology advances. Therefore these organizations often do not publish benchmarks in a void - they develop the benchmarks based on interaction with user groups, publications, developers and others. Contrarily to some beliefs, "vendor-driven" benchmarks are probably the most objective, as they are not subject to personal biases. The competitive nature of vendors provides a natural system of checks and balances that help ensure objective, repeatable benchmarks. Exemplarily there are the following organizations:

- SPEC
  The *Standard Performance Evaluation Corporation* is a non-profit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers. SPEC develops suites of benchmarks and also reviews and publishes submitted results from member organizations and other benchmark licensees [1].
- BAPCo
  The *Business Applications Performance Corporation* is a non-profit consortium to

develop and distribute a set of objective performance benchmarks based on popular computer applications and industry standard operating systems [5].

- EEMBC
  The *Embedded Microprocessor Benchmark Consortium* was formed in 1997 to develop meaningful performance benchmarks for the hardware and software used in embedded systems. Through the combined efforts of its members, EEMBC® benchmarks have become an industry standard for evaluating the capabilities of embedded processors, compilers, and Java implementations according to objective, clearly defined, application-based criteria [2].
- CIS
  The *Center for Internet Security* mission is to help organizations around the world effectively manage the risks related to information security. CIS provides methods and tools to improve, measure, monitor, and compare the security status of Internet-connected systems and appliances. A main focus of this organization is to develop internet security benchmarks available for widespread adoption [3].
- NAFEMS
  The *National Agency for Finite Element Methods and Standards* was founded as a special interest group in 1983 with a specific objective namely: "To promote the safe and reliable use of finite element and related technology". At the time when this mission statement was written the engineering community was concerned primarily with the accuracy of stress analysis codes, which were predominantly based on the finite element method. A lot of efforts were done on developing standard 'Benchmarks' against which codes could be tested [3].
- TPC
  The *Transaction Processing Performance Council* is a non-profit corporation founded to define transaction processing and database benchmarks and to disseminate objective, verifiable TPC performance data to the industry.

In the following it will be shown the current state of benchmarking in industry with the help of some well-chosen examples which fit into introduction’s classification (Fig. 1).

2.1 Processor Benchmarks

A very early method of evaluating processor performance were “millions of instruction per seconds (MIPS)” and “millions of floating point operations per second (MFLOPS)” ratings. These were commonly used until the late 1980’s. However, once RISC processors appeared on the market, the main weakness of these ratings became readily apparent; instruction and floating point operation are not clearly defined. It was soon realized that processor performance is determined by three factors: the number of instructions, the average clocks per instructions, and the clock frequency. Trying to evaluate performance using a subset of these features leads to meaningless results. Processors must be evaluated using real world applications.

Early popular benchmark programs were small toy programs such as the popular Dhrystones and Wheatstones benchmarks. The fact that these programs were easy to understand and their behavior easy to analyze led some people to exploit the benchmarks for marketing purposes. For example, DEC used a C compiler flag with a special DHRYSTONE flag. This flag would turn on some optimizations in the compiler which in general would reduce the efficiency of the generated code, but would improve performance dramatically on the Dhrystone benchmark.

These shortcomings led a number of companies to form the SPEC group in 1999. The SPEC CPU benchmark consists of part from eight real applications ranging from Neural Net simulation to the GNU C compiler [6].

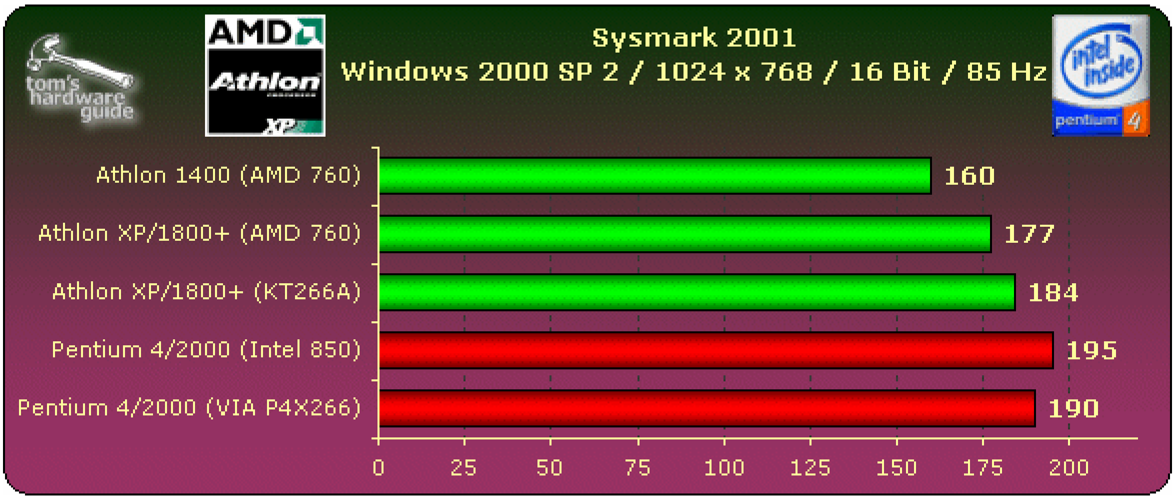

Fig. 2 A typical benchmarking result: two processor families from Intel and AMD have been compared by the BAPCo's Sysmark 2001 benchmark [9]

2.2 Database Benchmarks

In the field of database systems it is common to investigate the performance in terms of how many transactions a given system and database can perform per unit of time, e.g., transactions per second or transactions per minute.

The term transaction is often applied to a wide variety of business and computer functions. Looked at as a computer function, a transaction could refer to a set of operations including disk read/writes, operating system calls, or some form of data transfer from one subsystem to another. A transaction as it is commonly understood in the business world is regarded as a commercial exchange of goods, services, or money. A typical transaction would then include the updating to a database system for such things as inventory control (goods), airline reservations (services), or banking (money). Well known benchmarks are from TPC for example [7]:

- The TPC Benchmark™C (TPC-C) simulates a complete computing environment where a population of users executes transactions against a database. The benchmark is centered on the principal activities (transactions) of an order-entry environment. These transactions include entering and delivering orders, recording payments, checking the status of orders, and monitoring the level of stock at the warehouses. While the benchmark portrays the activity of a wholesale supplier, TPC-C is not limited to the activity of any particular business segment, but, rather represents any industry that must manage, sell, or distribute a product or service. TPC-C involves a mix of five concurrent transactions of different types and complexity either executed on-line or queued for deferred execution
- The TPC Benchmark™H (TPC-H) is a decision support benchmark. It consists of a suite of business oriented ad-hoc queries and concurrent data modifications. The queries and the data populating the database have been chosen to have broad industry-wide relevance. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions. The performance metric reported by TPC-H is called the TPC-H Composite Query-per-Hour Performance Metric (QphH@Size), and reflects multiple aspects of the capability of the system to process queries. These aspects include the selected database size against which the queries are executed, the query processing power when queries are submitted by a single stream and the query throughput when queries are submitted by multiple concurrent users. The TPC-H Price/Performance metric is expressed as $/QphH@Size.

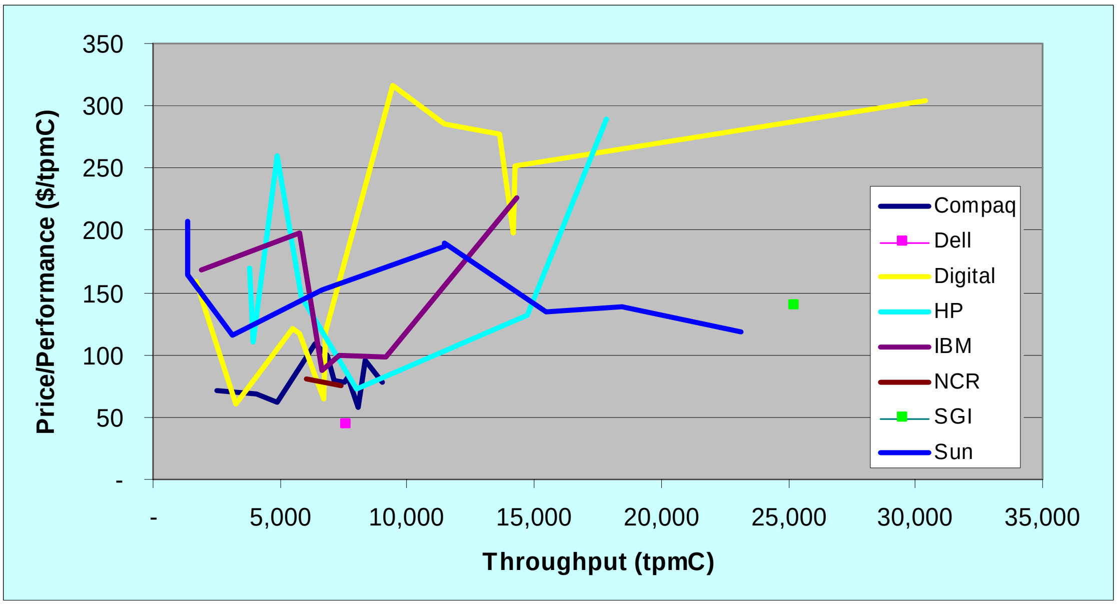

Fig. 3 A typical diagram for benchmark comparison of databases from several vendors [8]

2.3 Industrial Robots

The situation in the field of industrial robots is completely different from that in the computer domain as presented above. Industrial robots exist in much smaller numbers and they are often manufactured and delivered for a special purpose. Usually the manufacturer of a robot and the customer jointly develop a clear specification of the tasks the robot must perform, i.e. certain performance features and properties are guaranteed by the manufacturer. The decision is finally based on the overall concept.

There are, however, common performance indicators such as workcycle time, throughput, energy consumption, mean-time between failures (MTBF) etc. that play an important role when different robot systems are to be compared. Referring to one of these indicators, manufacturers often declare the performance of their system to be *the* benchmark, i.e. the reference value competitors have to compare to. These benchmarks in the literal sense may be right, but they are usually not independently verified and confirmed as it is the case with processors or databases.

Industrial robots for different applications are hardly comparable in a way to base a purchase decision upon. Within one certain application domain though, manufacturers often work together with selected customers in order to evaluate the overall performance of the installation. Especially when new products are introduced, the results of such an evaluation may serve as benchmark for potential customers.

In summary it may be said that in the field of industrial robots commonly accepted benchmarks from independent organizations virtually do not yet exist. However, as robots become more and more standardized, independent performance benchmarks may be defined in the future.

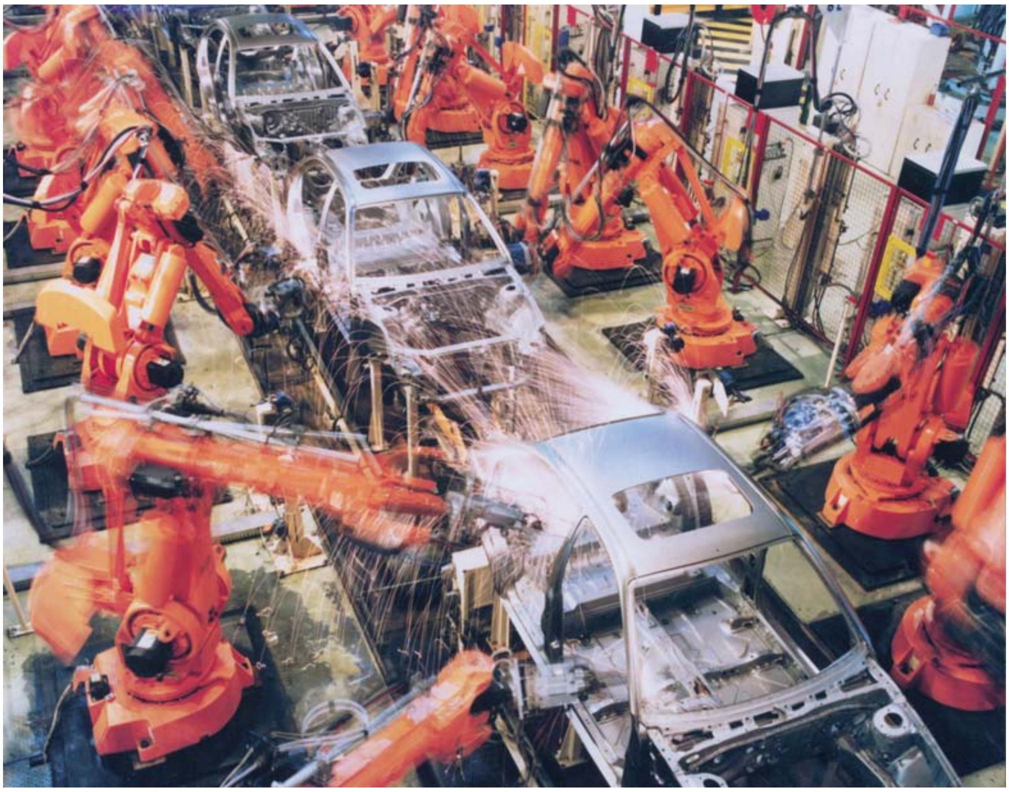

Fig. 4 Assembly-line with robots: High throughput and synchronized work cycles are essential

3 Benchmarks in Robotics Research

As mentioned previously a benchmark approach can be analytically or functionally, i.e. evaluation by proof vs. evaluation by tests. Another categorization is to have a look at the whole system or only small part, a component, of the robot system. In research it can be found examples for nearly all combinations: analytically - robot system, analytically - robot component, functionally - robot system, functionally - robot component.

3.1 Analytical Benchmarks for a Robotic System

An analytical benchmark has the aim of evaluating a robot system with mathematical means. However, in general, it is difficult to create such benchmarks. Many assumptions must be made and some assumptions may not be true in robots' physical world. Analytical approaches are in general only applicable in small units of a robot system, e.g. the classical verification calculus of computer science weakest precondition calculus or the verification by the means of loop invariants. And even there, usual algorithms are too complex to be evaluated by this approach.

analytical		▨
functional		
	component	system

3.2 Analytical Benchmarks for a Robotic Component

Advantage of analytical benchmarks is the realization without a lot of technical effort, flexible usage and the ability to perform extreme tests which might not be accomplished in reality due to costs or safety reasons (e.g. testing with maximum velocity). Analytical benchmarks require an exact system model which is hard to derive for complete systems. Therefore, this kind of benchmarking is mainly used to test single system components.

analytical		
functional		▨
	component	system

Simulation tools are used for system design, e.g. to try and verify different control strategies according to desired needs like maximum overshoot or response time. Standard control problems like “Floating Ball” or “Inverse Pendulum” could be defined as benchmarks in this field, but problems are usually very complex and task oriented so that benchmarks have to be defined individually.

Algorithms for data analysis, control or planning are suited for analytical benchmarking through simulation. To test and judge algorithms for motion planning, a benchmark was defined in [23], representing simple manipulation tasks. The goal of “Alpha Puzzle” is to combine or to separate two alpha-shaped tubes; the “Pentamino Puzzle” is used to test disassembling methods by extracting parts in the right order out of a cube, both shown in Fig. 5.

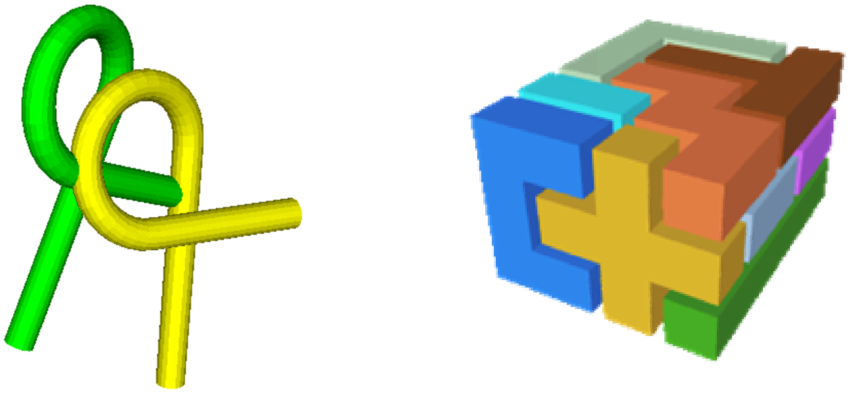

Fig. 5 Benchmark for motion planning algorithms; "Alpha Puzzle" (left) and "Pentomino Puzzle" (right) [20]

Benchmark results gathered from simulation are often just a first clue and are usually confirmed with functional benchmarks.

In the publication of Knotts et al. [24], an approach to define benchmarks for indoor navigation is presented which rates algorithms for planning and navigation of mobile platforms. The benchmark contains the following tests, representing most important abilities for mobile systems:

- Robust dynamic obstacle avoidance
- Path replanning in case of obstruction
- Automatic avoidance of dangerous regions (staircase)
- Mapping for new buildings

Implementation of this benchmark is again possible in simulation or in real experiments.

3.3 Functional Benchmarks for a Robotic Component

This section discusses benchmarking and problems of functional benchmarking for robotic components. It is divided into two subsections regarding hardware and software components.

analytical		
functional	▨	
	component	system

Hardware components

In the range of hardware components, no established benchmarks are known. Benchmarks are mainly restricted to comparisons with other known components. For example Zelinsky et. al. compare their newly built pan-tilt-unit with other pan-tilt-units from different institutes. They take into consideration maximum speed, maximum acceleration, maximum payload, number of saccades per second and repositioning accuracy. Results can be found in [20].

Software components

In the field of benchmarks for software components much research has been done in the field of computer vision which plays an important role for robotic systems. This section will discuss some facets of benchmarks mainly in the fields of face and gesture recognition, object recognition and scene analysis including the understanding of dynamic scenes. Benchmarks for vision systems mainly consist of image data bases which are provided to research groups in order to test their methods on known images. Together with theses images, a detailed description, often acquired by hand, gives a ground truth of the contents.

Face and gesture recognition

Face and gesture recognition is an area of great interest from many research groups as can be seen by the existence of an IEEE conference (FGR, “IEEE Int’l Conference on Face and Gesture Recognition”). Face recognition can be divided into two subtasks:

- Recognition of faces in images (“Where is a face in the image?”) and
- Identification of faces/persons in images (“Who is there?”)

Gesture recognition on one hand deals with defining and detecting gestures in order to instruct robotic systems and on the other hand tries to understand human sign language, especially the American Sign Language (ASL).

The American NIST (National Institute of Standards and Technologies) is organizing benchmarks for face recognition systems on regular basis. These are called “Face Recognition Vendor Test (FRVT)” and the last one was done in 2002 as a large-scale evaluation, details can be found on their website [20] (http://www.frvt.org). It was opened to all interested researchers and developers, including academia, research laboratories and commercial companies. The primary objective was to provide performance measures for assessing the ability of automatic face recognition systems. Therefore three tasks had to be fulfilled which were: verification, identification and watch list tasks. The purpose of watch list tasks is the following: Given a list of persons, the system is presented an unknown image of person. It has to decide if the person on the image is part of that list and, if yes, identify that person.

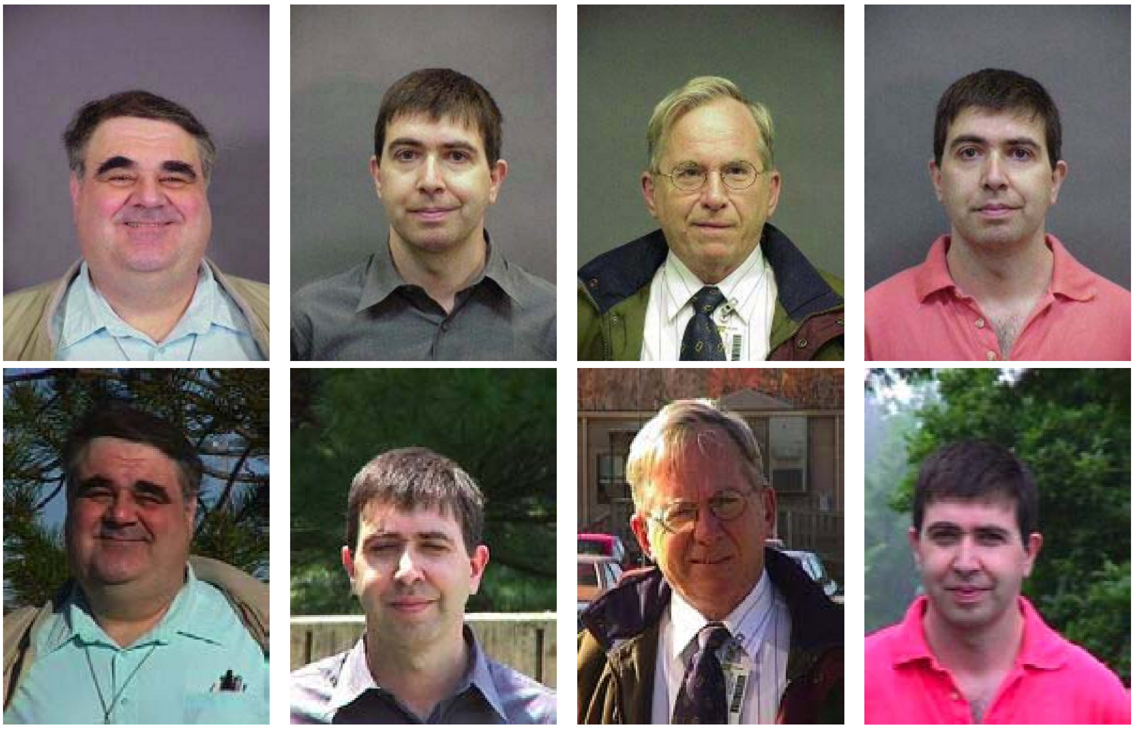

Fig. 6 Images used for evaluation. The top row shows indoor captures whereas the bottom row shows images of the same person taken outdoors.

As mentioned before, image databases are a good way to provide test conditions for computer vision applications which are equal to all applicants. The images of the database must be chosen carefully, because the database heavily influences the quality and the significance of the benchmarks to be performed. Therefore a database was set up with a vast amount of images of different persons. In total it contains 121589 images of 37437 different persons, each image containing exactly one person. There were at least three different images of each person in the database which were taken under different conditions. These were mainly different lighting conditions, indoor and outdoor captures (see Fig. 6 for a list of pictures as an example) and different age of the same person. The difference of the age of a single person was up to three years. Another important characteristic of the database is its distribution of the persons. Beside male and female persons, attention was paid to the age of the persons in order to achieve a wide distribution of different ages. Tests revealed that males are easier to identify than females. Additionally the identification of young persons is more difficult than that of older ones. A complete description of the test results can be found in the FRVT Evaluation Report [20], but already these two results about sex and age show, that the database has to be designed carefully to measure the quality of the applied algorithms.

Another benchmark was introduced by the Computer Science Department of the Colorado State University at the International Conference on Computer Vision Systems 2003 in Graz. They investigated face identification systems and the effects which influence identification performance. Images were taken from different persons in different conditions. The conditions varied in a way that the images were taken indoors and outdoors as well as different appearances of the same person. Influencing effects of a person's appearance comprise different hair styles, wearing glasses or not, using make-up or not, having a moustache and others. A detailed description can be found in [12]. For their tests they used four common approaches implemented at their institute, but as it is a benchmark the database is open to every interested researcher and they offer to test everyone algorithms.

Conferences and workshops being held in conjunction with these conferences are another way of benchmarking computer vision applications. One of these workshops targeted at benchmarking face recognition and scene understanding is the workshop on Performance Evaluation of Tracking and Surveillance (ICVS-PETS) [13]. The workshop provided a set of images of a conference scenario that could be used to test the researchers' algorithms against. Beside face recognition, scene understanding could be performed. Typical scenes of the conference scenario were constructed like standing up, walking to the blackboard, and so forth. The data sets could be acquired by every interested researcher. The following picture (see Fig. 7) shows two typical images of the data set. An example for understanding not only still images is the detection of human gestures especially pointing gestures. These gestures can be done using the hand and arm but also through pointing the head in the desired direction. Example data sets can be found in [14].

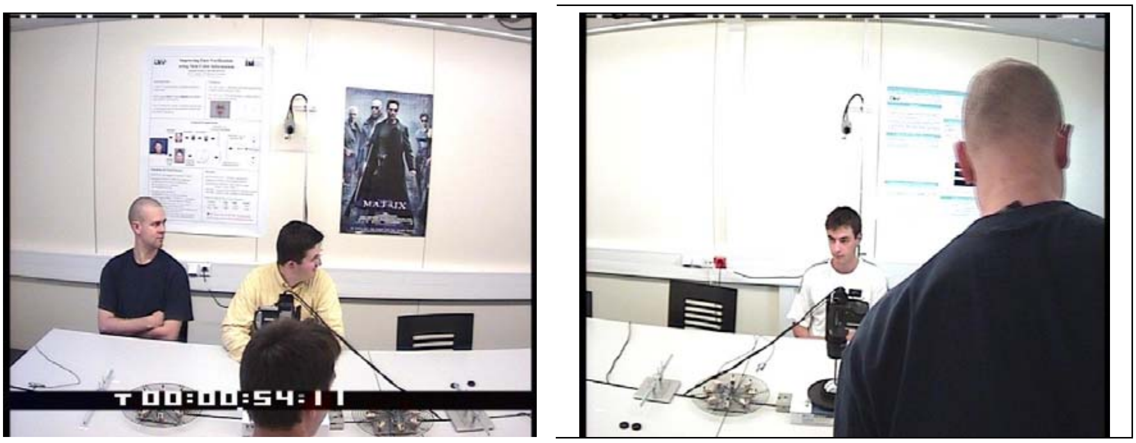

Fig. 7 Two test images from the ICVS-PETS workshop

Object recognition

Object recognition is another important research area for which benchmarks have been established over the last years. Similar to face recognition tasks most benchmark are performed using a standardized set of images. Two popular image data bases are the COIL-20 and the COIL-100, see [16] and [17]. COIL stands for Columbia Object Image Library and has been used in many publications for testing the developed methods. The first one consists of 20 objects where each object was capture in grey-scale and the second one of 100 objects with colored images. 72 images were taken per object at every 5 degree resulting total in a turn of 360 degree. Fig. 8 shows the images of all the objects provided by the two COIL databases.

Several tasks can be benchmarked with these data sets. Firstly, an exact object recognition can be done, that means, the object has to be identified exactly (i.e. by name). Another possible task could be the categorization of an object, like fruit, puppet, toy or something else. A third possibility is the estimation of the pose of the object. Because the objects were captured at different rotational angles, it is therefore feasible to estimate its rotation angle.

Several other image data sets can be found throughout the world-wide-web, a vast list of links to data sets can be found in [18].

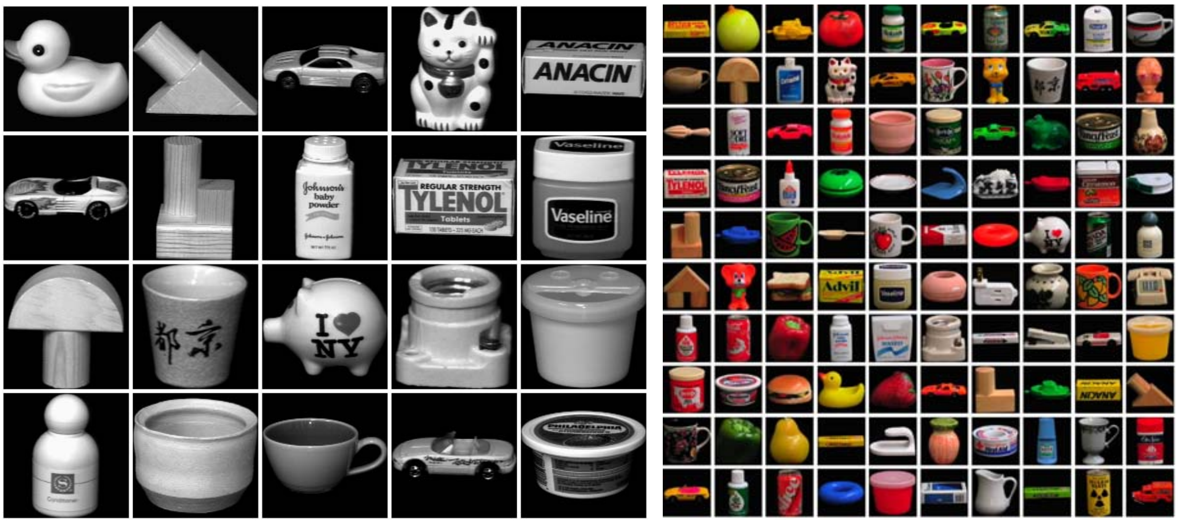

Fig. 8 The COIL-database. Left: COIL-20, Right: COIL-100

3.4 Functional Benchmarks for a Robotic System

The most famous examples for functional benchmarking of entire systems in research are robot soccer championships. In September 1995, Professor Jong-Hwan Kim of KAIST University founded an International Organization Committee (IOC) for Micro-Robot World Soccer Tournaments (MiroSot) [25].

analytical		
functional		▨
	component	system

The aim was to set up a multi-purpose testing ground for learning and application of high-tech technology. Competition leads to intensified research in the fields of image analysis, artificial intelligence, perception and control. A meeting of 30 teams in September 1996 was called to have a common basis by defining the following play rules:

- Field size 150 x 130 cm
- Two teams play against each other with 3 robots each, size is not allowed to exceed 7,5 cm in height, width and length
- Points are given by placing the ball in the other team's goal
- Robots are not allowed to be remote-controlled; all actions have to be initiated by artificial intelligence of a computer

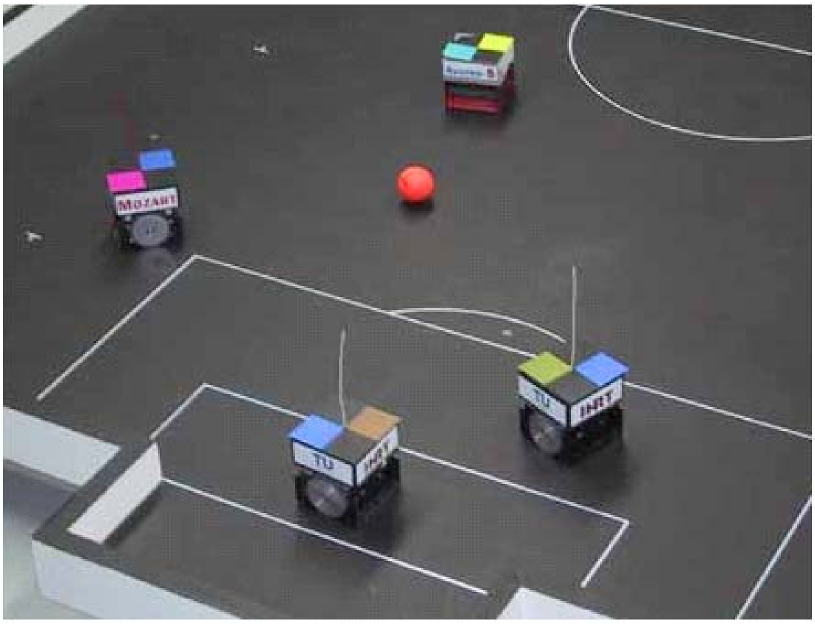

Fig. 9 MiroSot Robot Soccer Tournament

These rules set a standard in robot soccer tournaments to compare two systems. A first world championship in robot soccer followed in 1996 (Fig. 9). During the FIRA (Federation of International Robot-Soccer Association) competition in 1999, an additional benchmark contest was held, including the following tests (out of [25]):

- Ball striking: To control a single robot to move from a given initial position to strike a stationary ball. This benchmark runs three times, each time with a different initial position. The robot has one minute to complete the task.
- Goal scoring: same than above with the requirement of scoring
- Passing between players and shooting: To control two robots starting at given initial positions such that Robot one strikes the ball once, and Robot two strikes the moving ball once to make the ball pass over the goal line. This benchmark runs twice, each time with a different initial position. The robots have unlimited time to complete the task.

Benchmarks above were arranged to lower the complexity of the main task, and to not only have the final score as a result to judge robots and teams. Now it was easier to draw conclusions about robot abilities and different methods of resolution for single tasks. Unfortunately, only four teams took part in this benchmark contest. This could be a sign that some researchers might be afraid of disgracing themselves in such direct tests, pretended to be “easy” which not the case is. In a more complex benchmark like an entire game, the overall strategy is not easy to discover and therefore hard to judge. Maybe, another reason for the few teams taking part was the fact that they prepared for the game itself and single methods for goal scoring or ball passing could not be excluded out of the robot’s behavior easily.

After the great success of Robocup Soccer, a second project called Robocup Rescue was started [26]. The goal of the project is to accelerate research activities and development for robots serving in search and rescue operations in disaster environments. A forum for technical discussions and competition in simulation and disaster scenario are used to exchange information and to evaluate research progress.
In Robocup Rescue Simulation Project, intelligent agents are used in a virtual disaster environment to resemble fire fighters, commanders, victims and re-enact the entire scene. In these virtual scenes, strategies are tested which then could be used as decision support in real disasters.

In Robocup Rescue Robot League, real robots are used in a standard destroyed building (Fig. 10). The task is to develop maps with positions of victims, identify signs of life of victims and to avoid touching these persons. According to the fulfillment of these tasks, points are added (or subtracted) to benchmark robots and rescue strategies.

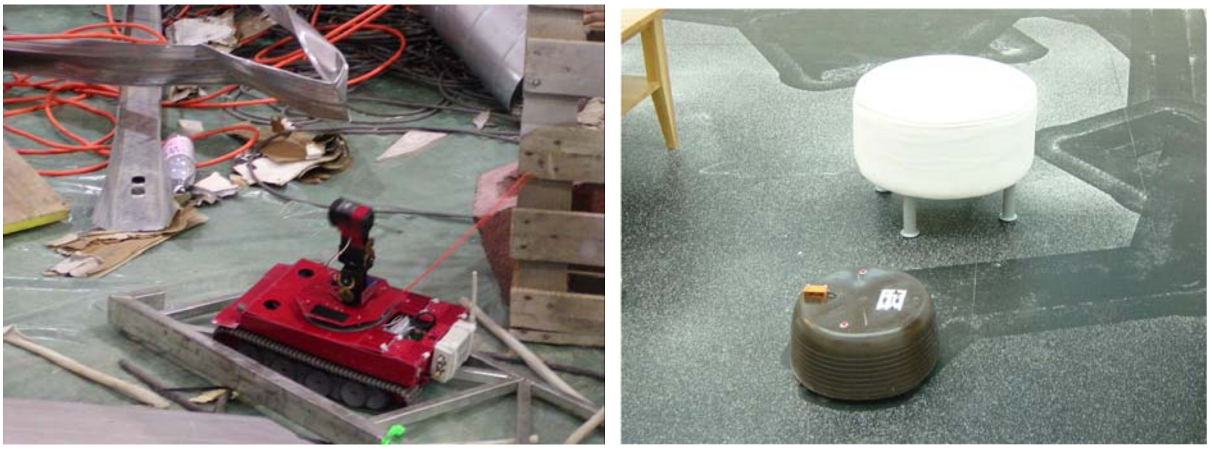

Fig. 10 RoboCup Rescue Project: Robot in destroyed building (left) and and Robot performing Robot Cleaning Contest (right)

Another example for system benchmarking are cleaning contests. In Lausanne, the first cleaning contest took part in October 2002, including functional benchmarks like floor and window cleaning, and also an analytical idea contest for housekeeping robots.

In the floor cleaning contest, a robot is placed in a certain environment where it has to clean the floor from dirt (sugar or confetti). After 10 minutes, the clean area is rated according to size and quality of cleaning. With changing environment and performing several runs, the benchmark results are more independent from situations and therefore more representative.

The window cleaning contest also consists of measuring time, clean area and quality of a defined test window. These two competitions have to be seen as benchmarks between industry and research. As cleaning is an important part of the service sector and several products are already available or coming soon, such benchmarks will be used more intensively from companies to advertise their products.

The DARPA Grand Challenge [28] is an example of a complex test for robot systems in military. The competition specifications were to drive in a certain time (10 hours) with an unmanned vehicle a specific route from Los Angeles to Las Vegas (350 km total distance). Due to little limitations and restrictions of this contest, the high number of possible solutions to combine perception, planning and control techniques and the high technical challenge, this resembles a very complex benchmark.

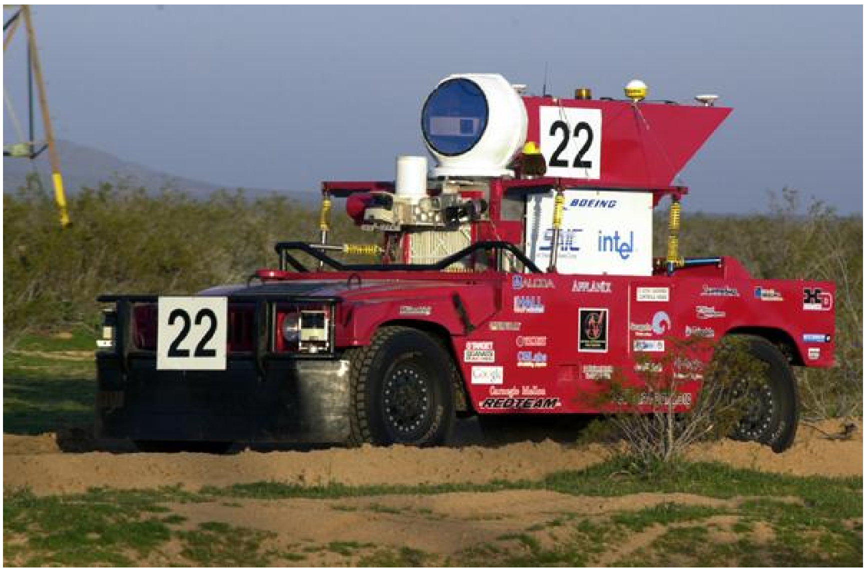

Fig. 11 Vehicle of CMU Red Team, taking part in DARPA Grand Challenge

The research department of the US military arranged this contest to push the development of new unmanned vehicles and raise interest in public. Therefore it was not surprising that the focus of organization was on good event advertising this first time. The results “Finish not reached” or “Finish reached in xx hours” are quite fuzzy to really judge and compare vehicles and different control approaches, at least in this early state in which nobody crossed the finish line. Pre-testings of the competition, e.g. driving the vehicles in a parcour and avoiding obstacles were much more informative and would serve as better benchmarks for actual systems.

4 Open Problems and Future Development

The following chapters give an overview on problems to be addressed and future research trends in the field of robotics benchmarks.

4.1 Open Questions

Robotics research is currently characterized by two facts:

- *Details:* Each research group has its own, specific focus which is emphasized by them. This centre of attention may be influenced by the group and their work, by political or economical factors and by historical coherence.
- *Diversity:* Nowadays we have a wide heterogeneity of existing robotics systems. This fact covers hardware, software and application area.

As an initial conclusion of the current state of robotics research one can say that it is very difficult to design an overall robotics benchmark. Existing robotics systems are too different, and different research groups focus on dissimilar areas and details. In detail, the following open problems can be identified which have to be discussed and solved:

- *Sense:* The initial and most important question is: How do benchmarks have to look like in order to be useful? One has to design and to implement benchmarks in such a way that they really provide us useful information. A very important aspect here is the choice of a certain metric or, more generally speaking, the method we use for measuring and comparing the performance of different robotics approaches.

- *Goals:* Before we start with the realization of a benchmark, we have to think about the intention of that work. Each benchmark has to be designed according to the specific aim we want to achieve.
- *Scenario:* As soon as the goal has been defined, one has to choose an adequate test scenario for the benchmark. The scenario should be a realistic one, it should allow for reproducible testing and it should be only as complicated as necessary.
- *Universality:* A very important question is the transferability of test results of, more generally speaking, and the interpretation of the test results. One may think of benchmarks which test on very specific tasks and on robots which are able to fulfill exactly that task very well – but only that task. Such a benchmark will not allow for estimation of the quality of a specific robotics approach.
- *Realism:* The test scenario and the overall benchmark have to be as realistic as possible. This may in some cases be difficult, but only realistic situations allow for a realistic appraisal of actual developments.
- *Trade-off:* The design of a benchmark always includes an implicit benchmark on generality and specialty. If a benchmark is too general, it is no longer possible to compare systems with each other. If, on the other hand, a benchmark is too specialized, it is very easy to tune a robotics system to achieve good results in that benchmark.
- *State of the art:* From a critical point of view, one may ask whether current robotics research is advanced enough to undergo realistic benchmarks. Most existing systems are not designed for neither used to realistic test scenarios.

All these topics will have to be considered in solved within the following years.

4.2 Towards Bottom-Up Benchmarking

One proposal to achieve objective, reproducible benchmarks is the formalization of tests. The philosophy of this approach is to confront the robotics system in a black-box manner with requirements of a “rational customer”. The performance of the system is then characterized in an objective way by the outcomes of a test-suite tailored to this scenario [20].

As an application area for this new concept, grasp & place tasks within a domestic assistance scenario are considered. A formalized single test description consists of a clear and reproducible specification of the robot’s task and the full context. The test results are displayed as a number of figures.

Each test scenario in this proposal has a workspace of about 2,5 m x 2,5 m x 2 m which may later be extended. The hardware of the robotics system shall not be limited in any way. In principle, also distributed solutions are allowed. The general idea is now to specify a not too large, but sufficiently representative set of test tasks. It should not cover all possible situations but include a random choice (“N out of a set of M tasks”) and a sufficient repetition of the test. Tests may also be clustered to test suites in order to assess the degree of “universality” achieved (Fig. 12). All relevant aspects of a test should be quantifiable and measurable.

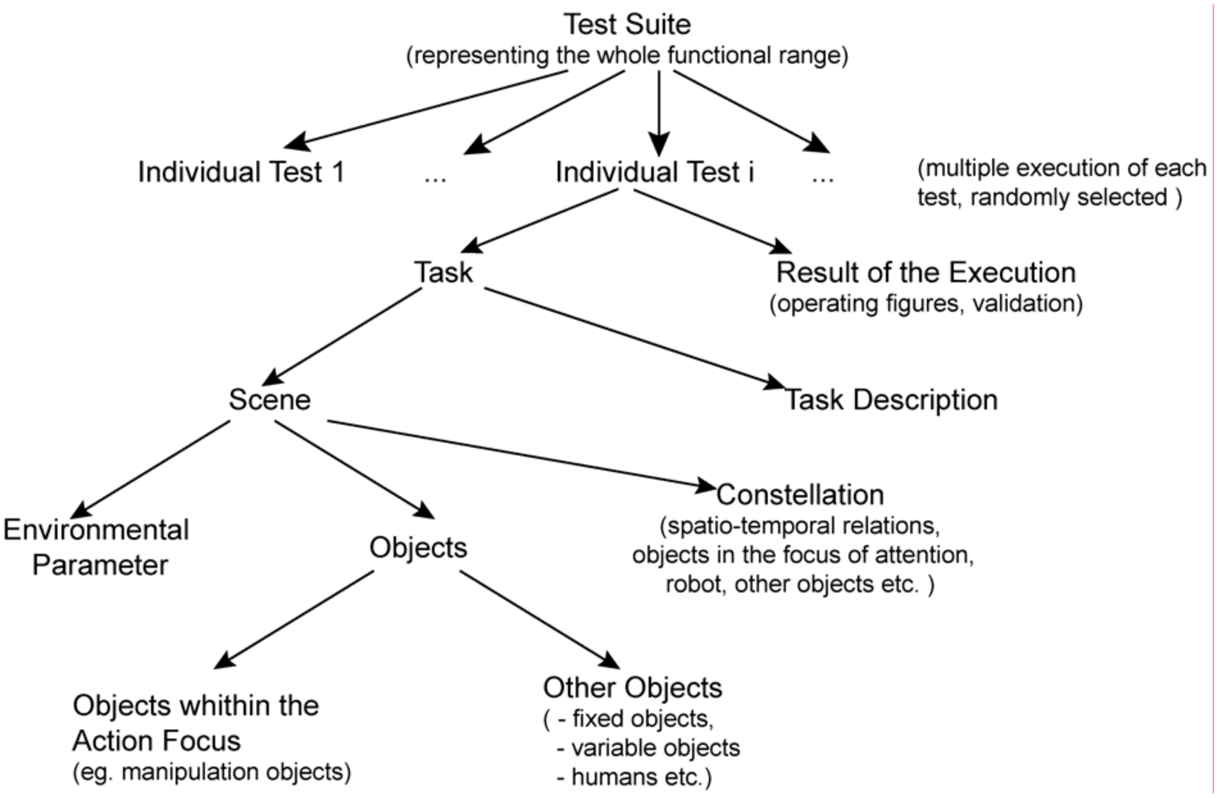
Test Suite
(representing the whole functional range)
Individual Test 1
...
Individual Test i
...
(multiple execution of each test, randomly selected)
Task
Result of the Execution
(operating figures, validation)
Scene
Task Description
Environmental Parameter
Objects
Constellation
(spatio-temporal relations, objects in the focus of attention, robot, other objects etc.)
Objects whithin the Action Focus
(eg. manipulation objects)
Other Objects
(- fixed objects,
- variable objects
- humans etc.)

Fig. 12 Test Description Graph

According to [20] benchmarking includes two aspects:

- Confronting a system with some specified tasks which are considered to be both challenging and representative for a certain range of applications
- Detecting and eliminating own weaknesses by systematic comparison with the best (competitors) in the field

However, the authors assume that the time has not come yet for designing benchmarks that are representative for a target area and have a chance to be really accepted by the community. Therefore they suggest to first achieve clear and complete specifications of tests and reproducibility of results. Afterwards, in a second step, good benchmarking may be achievable.

5 Conclusion

The results of this report show that the field of benchmarks for robotic research is still far away from being solved, especially the transition from component to full systems must be considered more carefully. Therefore, more efforts must be done in the future to get more objective comparisons between different robotic systems. This will cause more competition in this field but this will lead to more reliable systems with an increasing quality.

6 References

[1] Standard Performance Evaluation Corporation (SPEC), http://www.spec.org

[2] Embedded Microprocessor Benchmark Consortium (EEMBC), http://www.eembc.hotdesk.com

[3] National Agency for Finite Element Methods and Standards (NAFEMS), http://www.nafems.org

[4] Center for Internet Security, http://www.cisecurity.org

[5] Business Applications Performance Corporation (BAPCo), http://www.bapco.com

[6] J. Baltes, "A Benchmark Suite for Mobile Robots", IEEE International Conference on Intelligent Robots and Systems (IROS'00), Takamatsu, Japan, 2000

[7] Transaction Processing Performance Council, http://www.tpc.org

[8] C. Levine, "TPC-C: The OLTP Benchmark", SIGMOD '97, Tucson, USA, 1997

[9] Tom's Hardware Guide, http://www.tomshardware.com

[10] Face Recognition Vendor Test (FRVT) 2002, http://www.frvt.org

[11] Face Recognition Vendor Test 2002, Evaluation Report, http://www.frvt.org/DIs/FRVT_2002_Evaluation_Report.pdf

[12] David S. Bolme, J. Ross Beveridge, Marcio Teixeira and Bruce Draper, "The CSU Face Identification Evaluation System: Its Purpose, Features, and Structures", Third Int'l Conf. On Computer Vision Systems 2003 (ICVS 03), April 2003, Graz, Austria, pp. 304-313

[13] Workshop on Performance Evaluation of Tracking and Surveillance, http://petsicvs.visualsurveillance.org/

[14] Pointing Gestures Workshop 2004, http://www-prima.inrialpes.fr/Pointing04/datasets.html

[15] Andrew Dankers and Alexander Zelinsky, " A Real-World Vision Systems: Mechanism, Control, and Vision Processing", ", Third Int'l Conf. On Computer Vision Systems 2003 (ICVS 03), April 2003, Graz, Austria, pp. 223-235

[16] S. A. Nene, S. K. Nayar and H. Murase, "Columbia Object Image Library (COIL-100)", Technical Report CUCS-006-96, Columbia University, February 1996

[17] S. A. Nene, S. K. Nayar and H. Murase, "Columbia Object Image Library (COIL-20)", Technical Report CUCS-005-96, Columbia University, February 1996

[18] CMU Computer Vision Test Images, http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/cil/ftp/html/v-images.html

[19] J. C. Mogul, "Brittle Metrics in Operating Systems Research", HOTOS99

[20] A. Jacoff, E. Messina, J. Evans: "Reference Test Arenas for Autonomous Mobile Robots", 14th International FLAIRS Conference, May 2001

[21] S. Hanks, M. E. Pollack,P. R. Cohen, "Benchmarks, Test Beds, Controlled Experimentation, and the Design of Agent Architectures", AI Magazine Volume 14, Issue 4, Winter 1993

[22] I. Iossifidis, G. Lawitzky, S. Knoop, R. Zöllner, "Towards Benchmarking of Domestic Robotic Assistants", accepted for publication in "Springer Tracts in Advanced Robotics (STAR) Series", Springer-Verlag, Heidelberg, ISSN: 1610-7438

[23] Motion Planning Puzzles, Texas A&M University, http://parasol-www.cs.tamu.edu/groups/amatogroup/benchmarks/

[24] R. Knotts, I. Nourbakhsh, R. Morris, NaviGates, “A Benchmark for Indoor Navigation”, Proceedings of Third International Conference and Exposition on Robotics for Challenging Environments. ASCE. Reston, VA, 1998.

[25] Yujin Robotics Co., Ltd., http://www.yujinrobot.com/yujin-e.htm

[26] RoboCup Federation, http://www.rescuesystem.org/robocuprescue/

[27] Cleaning Robot Contest, Oct 2002, Lausanne, Switzerland, http://www.service-robots.org/cleaningrobotscontest/index.php

[28] Darpa Grand Challenge, http://www.darpa.mil/grandchallenge/