# Adaptive Sensor-Fusion of Depth and Color Information for Cognitive Robotics

Denis Klimentjew, Jianwei Zhang
*Dept. of Informatics, University of Hamburg, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany*
{*klimentjew, zhang*}*@informatik.uni-hamburg.de*

***Abstract*— The presented work goes one step further than only combining data from different sensors. The corresponding points of an image and a 3D point cloud are determined through calibration. Color information is thereby assigned to every voxel in the overlapping area of a stereo camera system and a laser range finder. Then we analyze the image and search for the locations, which are especially susceptible to errors by both sensors. Depending on the ascertained situation, we try to correct or minimize errors. By analyzing and interpreting the images as well as removing errors we create an adaptive tool which improves multi-sensor fusion. This allows us to correct the fused data and to perfect the multi-modal sensor fusion or to predict the locations where the sensor information is vague or defective. The presented results demonstrate a clear improvement over standard procedures and show that other progress based on our work is possible.**

## I. INTRODUCTION

Multi-sensor fusion is a well-known topic in computer science. For several decades now this technology has been successfully used in different research areas like topography, robotics, environment reconstruction, the building of virtual worlds or object recognition. The manner as well as the used sensors depend mostly on the scenario. In the following we present some typical examples and explain them briefly.

The work in [1] presents a method for collision avoidance based on the fusion of camera and sonar sensors for mobile robots. Objects are recognized through the camera data and edge detection algorithms compared/completed with the sensor data. The estimation or/and calculation of the distance to the found object is based on the position of the object in the image and on sonar data. The performance is better than that of single sensors.

For the same approach the authors in [2] use the fusion of a laser range finder and sonar sensors. The optimal path is used as an initial solution to avoid nearby obstacles. A triangular area in front of the robot which is guaranteed to be free of obstacles is computed and used to search for the next drive commands. The characteristic attribute of this algorithm is its quickness, which is essential for its application in the RoboCup.

The use of sensor fusion for Simultaneous Localization and Mapping (SLAM) is also widespread. This is possible with the combination of many kinds of different sensors. For example, in work [3] several laser scans are merged to permit not only a SLAM estimate, but to reconstruct objects as well.

Furthermore, for humanoid robotics sensor fusion is an indispensable tool. For balance and stability control the fusion of information from gyroscopic and/or acceleration sensors is needed. In [4] the input of two gyroscopic sensors is used for active balancing. In [5] the gyroscopic and acceleration sensor data are fused to ensure stable robot control.

In geodesy, the fusion of the 3D point cloud and camera images has become very popular after the development of 3D laser scanners. Mostly it is only required for understanding what the point cloud represents or for taking the texture and adopting it to the point cloud. A similar estimate is also presented in [6]. The correspondence between the voxels of a terrestrial laser scanner and image pixels is found through a geometrical model of each sensor.

The examples named above present different kinds of multi-sensor fusion, like cooperative, complementary or redundant. Nevertheless, no interpretation of the data takes place. Moreover, the data are not qualitatively compared or evaluated.

Of course there were several estimates to compare the acquired information. In our group [7] multi-sensor fusion was realized by the use of fuzzy rules. Thus, for example, tables can be found in an unknown environment. Besides, the different data are weighted with regard to their quality or significance and afterwards are fused. The use of the Dempster-Shafer theory is also possible and was successfully realized in [8].

Most presented methods allow for large tolerances concerning the accuracy of the results, for example in outdoor scenarios. For indoor scenarios and/or grasping is is not possible to accept the same range of tolerances. In this paper we present multi-sensor fusion with high accuracy. The method opens possibilities for object detection, recognition and grasping. The presented approach is evaluated with a typical office table scene. The authors know of no other research groups that analyze the data in the step after the fusion and find potential error sources or locations. This would make it possible to interpret the data in the next step and to improve multi-modal sensor fusion or to predict the locations where sensor information is vague or defective. With the application to dynamic surroundings multi-sensor fusion could automatically adapt itself to external conditions.

## II. CHARACTERISTICS OF SINGLE SENSORS

The most important sensors for environment perception are the camera and laser range finder systems. The following list summarizes the problem areas of both sensors.

**Camera:**

- Homogeneous surfaces,
- Strong reflexions,
- Changing of lighting conditions.

**Laser scanner:**

- Invalid values (during and after the "warm-up" time),
- Black surfaces,
- Object orientation related to the incident laser beams,
- Tiny objects (the beam size is bigger than the object),
- Strong reflexions.

Of course, the another important difficulty of both sensors is the occlusion.

## III. FUSION

One of the essential assumptions of our work is the using of the same model for 3D point clouds as for an image. So an image $I$ can be seen as the sum of several components:

$$F_{i,f}(\overrightarrow{x}) = \alpha \cdot BG_{i,f}(\overrightarrow{x}) + N_{i,f}(\overrightarrow{x}) + T_{i,f}(\overrightarrow{x}) \quad (1)$$

or simplified

$$F_i = T_i + BG_i + N_i \quad (2)$$

where $F$ denotes the image, $T_i$ the foreground region, $BG_i$ the background, and $N_i$ the camera noise at frame number $i$. The advantages of this approach are the retaining of the relation between neighboring voxels and the possibility of using the know-how of 2D image processing and the many existing pertinent algorithms.

Before fusion we use our own method for autonomous 3D exploration of indoor environments. We developed a system that lets a robot move through an indoor environment, take 3D scans of its surroundings, and finally build a 3D map of the environment. For this purpose the perception system presented in fig. 1 is used.

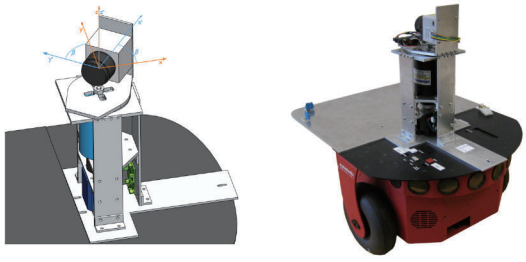


Fig. 1. The used robot plattform. The engineering drawing of the environment exploration platform on the left, and the realized platform mounted on the mobile robot on the right.

The biggest advantages of the platform are the facts that a permanent rotation of the laser scanner is possible and a $180^o$ rotation of the laser scanner delivers a full scan of the surroundings. In the resulting 3D map, shown in fig. 2, parallel to the ground planar surfaces are detected with the standard segmentation algorithms. These surfaces, like tables, build the regions of interest (ROI), because the probability of finding some kinds of objects on these surfaces is very high.

With further segmentation the walls, ceiling and ground can be recognized and removed as a background; in the end the tables are separated. After this the platforms as shown in

(a) (b)

Fig. 2. The left image shows the result obtained from the SLAM algorithm. The right frame shows direct visualization of the recorded data points and the visualization based on a surface reconstruction algorithm.

figures 7(a) and 7(b) can be used to perceive this location with more voxel density and color information from the cameras. To reduce the data traffic the sensor fusion takes place only for the detected ROI's.

Both kinds of sensors have different characteristics and perception areas. Fusion is possible only for the overlapping regions. Fig. 3 shows an example of the perception areas as well as the overlapping region of the camera and a laser scanner.

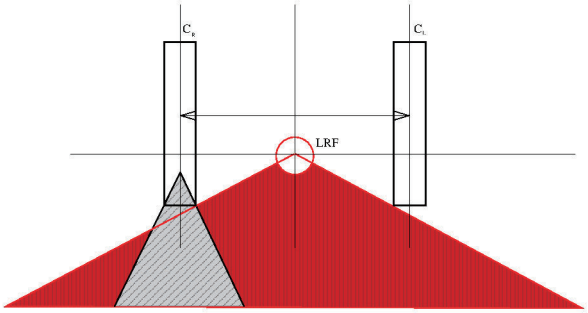
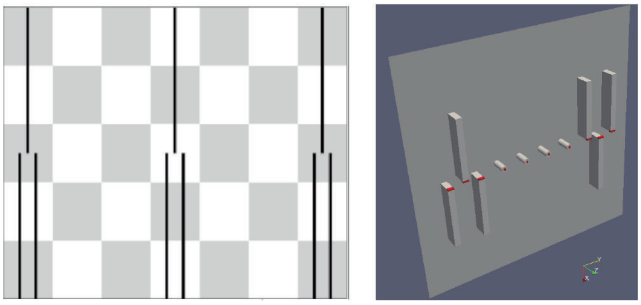



Fig. 3. The perception areas as well as the overlapping region of the camera and laser range finder.

The sought transformation between laser scanner and camera is based on the extrinsic parameters of both sensors, it is independent of the arrangement of the scene or distance to the objects. The cameras are first calibrated alone and then together. After the stereo calibration we transform the image of one camera to the other camera and calculate the disparity. In this case for the calibration between the stereo camera



(a) (b)

Fig. 4. The planned calibration pattern on the left and the model of the 3D calibration pattern resulting from several experiments on the right.

system and the laser range only the transformation from one camera to the laser scanner is needed, as shown in fig. 3.

The original idea was to combine the calibration patterns for the camera and the laser scanner. Basically we use the typically checkered pattern which is extended with a 3D structure. Because of the difficulties of most laser scanners with black surfaces, as already mentioned, the checkered pattern was renounced in the design of the calibration pattern.

The 3D calibration pattern consists of a planar surface, two opposite so-called Y structures and some pins of different length. The Y structures permit the ideal alignment of the laser scanner. Their surfaces as well as those of the pins are used later for the localization of the corresponding points for the laser scanner. We use the color information for the localization of corresponding pixels in the image (red lines at the Y-structure and points at the front of the pins).

The model of the resulting 3D calibration body is shown in fig. 4(a), the lateral view of the calibration body can be seen in fig. 7(a). The pseudo code of the implemented sensor fusion method is presented in algorithm 1, for more detail see our publication in [9].

**Algorithm 1** The sensor fusion algorithm

1: **procedure** FUSION($LRF$, $SCS$)
2: The laser scanner is moved by the pan-tilt unit and scans the environment in coarse steps.
3: Through changes in the depth information, our system localizes the region of interest (ROI) and the changeover from one to two teeth (or vice versa) in the Y-structure and rescans it in finer steps to find the desired position.
4: Detection of the characteristic pattern for the correct position of the laser scanner in relation to the calibration body and moves the pan-tilt unit to that position.
5: Stopping the platform and acquiring a camera image.
6: Applying the color threshold value to the camera image.
7: Calculating of corresponding points.
8: Rectification of both lines in relation to each other (laser scanner and an image).
9: Calculation of transformation matrix between a laser range finder and a camera (is independent of scene structure).
10: Calculation of the overlapping area with help of known geometrical parameters and computed fundamental matrix (the transformation matrix can be used in the overlapping area only, otherwise the transformation would cause an error and produce wrong correspondences).
11: Adopt the matrix to the point cloud and an image.
12: **return** partially colored point cloud
13: **end procedure**

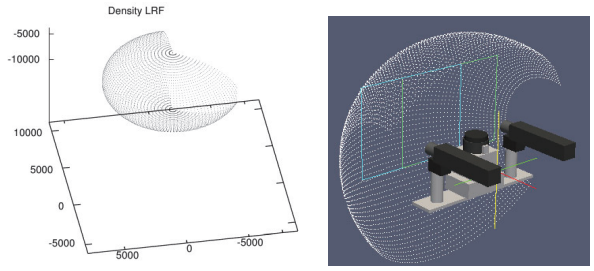
When the parameters have been determined, the overlapping area has been calculated and the transformation is unambiguous, adaptive sensor fusion can be accomplished.

The best and most accurate results were achieved with the described calibration method by using a 3D calibration body. Nevertheless the authors tested other calibration methods. For example the Iterative Closest Point algorithms (ICP). The difficulties are the different sizes of the resulting point clouds and quality of the stereo camera depth results. The algorithm temporarily delivers practical transformation, but often there are no or inaccurate results. Repeatability was not given in many cases, the calculation of meaningful transformation cannot be guaranteed. The improvement of the algorithms by merging characteristic local features from the information of both sensors like edges or corners sounded very promising and is under examination. The same applies to previously segmented statical surfaces or segmentation from motion. The authors are convinced that the use of a 3D calibration body yields an accurate fusion result. The comparison with well-known approaches in 2D image processing confirmed this assumption.

## IV. DATA INTERPRETATION

As already mentioned, the resulting depth information in the overlapped areas is redundant and can be used to improve the data of the early multi-sensor fusion. The point density depends on the physical properties of the single components and is shown for the laser scanner and its combination with the stereo camera in fig. 5.

Therefore we employ the depth information to improve the fusion data. To this end, the camera images are used for detecting locations which are particularly susceptible to errors of both sensors as describe before. The location can be determined with standard algorithms of image processing like a Median filter or a value comparison in the HSV color space. For example, the retrieval of homogeneous regions can be accomplished with Mean-Shift-segmentation or Similarity-Measure-algorithms.




(a) (b)

Fig. 5. Point density for the simulated 3D laser range finder based on the 2D laser scanner and pan-tilt unit as well as their combination with a stereo camera system.

The following table shows the choice of the sensor for the depth information depending on the ascertained problems.

| Problems | LRF | SCS | Avail. inform. |
|---|---|---|---|
| Homogeneous surfaces | × | – | $d_{LRF}$ |
| Invalid values (LRF) | – | × | $d_{SCS}$+c+t |
| Tiny objects | – | × | $d_{LRF}$+$d_{SCS}$+c+t |
| Black surfaces | – | × | $d_{SCS}$+c+t |
| Lighting conditions | × | – | $d_{LRF}$ |
| Strong reflexions | – | – | – |
| – | × | × | $d_{LRF}$+$d_{SCS}$+c+t |

Thereby LRF is an acronym for the laser scanner, SCS for the stereo camera system, $d$ for depth, $c$ for color and t for texture information. The depth information from × marked sensors is preferred in the ascertained situation. Otherwise the more exact data of the laser scanner are used. Using the knowledge

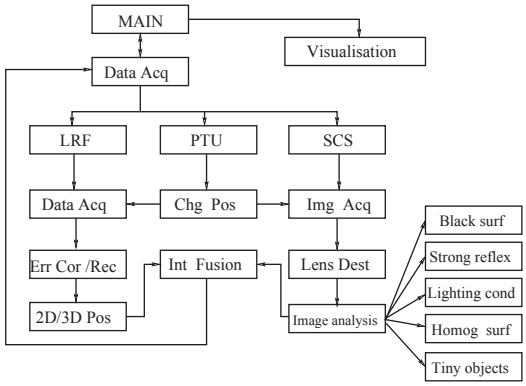



Fig. 6. Simplified flow-chart of our 3D colored reconstruction system.

of the previous table, the implemented architecture is based on the presented assumptions. Fig. 6 illustrates a simplified flow-chart of our 3D colored reconstruction system.

The more sensor information about the surrounding exists the more the accuracy and safety of the client applications can be guaranteed. Consequently one of the best strategies for the robot motion can be to find the position and orientation related to the ROIs, that deliver all possible sensor information with minimal errors. The problem is comparable to the Next Best View problem in image processing.

Another approach can be the integration of further sensor information, if the sensor information is exact, continuable and reliable, like laser scanner data. For example, the second laser scanner of our main platform is mounted on the arm. There are two possibilities to integrate this data, one is the presented calibration method, the another one is the registration process. All fused data in this work could be registered to the base frame of TASER. During the sensor fusion the data from the camera system transform to the frame of the laser scanner as described below. The laser beams are registered to the coordinate system in the bottom of the pan-tilt unit, see eq. 3.

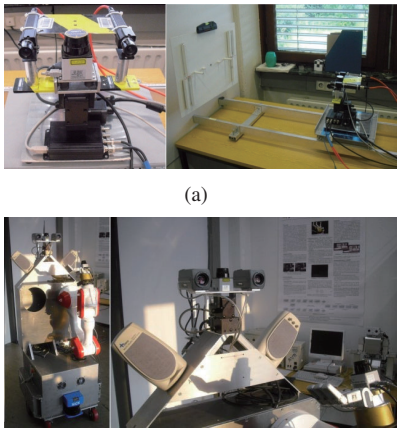
$$\begin{bmatrix} c(\varphi)c(\theta) & -s(\theta)c(\varphi) & s(\varphi) & -d_z s(\frac{\varphi}{2}) \\ s(\theta) & c(\theta) & 0 & 0 \\ -c(\theta)s(\varphi) & s(\theta)s(\varphi) & c(\varphi) & -d_z s(\frac{\varphi}{2})c(\frac{\varphi}{2}) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3)$$

where $c$ and $s$ are $cos$ and $sin$ respectively, $\theta$ is the deflection angle of the laser beams (multiple of angular resolution) and $\varphi$ is the current angle of the pan-tilt unit, $d_z$ is a vertical vector to the coordinate origin.

With two further simple translations this frame can be transformed to the base coordinate system of TASER.

## V. EXPERIMENTAL RESULTS

For our experiments a perception platform developed by us is used. The main platform consists of a 2D laser scanner, a pan-tilt unit and a stereo camera system and is shown in fig. 7(a). The right image of fig. 7(a) shows our calibration



(a) (b)

Fig. 7. The upper image shows a laser scanner and a stereo camera system mounted on the pan-tilt unit. The right image shows our calibration arrangement. The lower image shows the service robot TASER, the main platform of our group on the left. On the right the developed environment perception system.

arrangement. The stereo camera system and the laser scanner are both mounted on a pan-tilt unit with a displacement between the optical axes (baseline) of approximately 0.08 m. The 2D laser scanner together with the movable platform constitute the simulated 3D laser range finder. The setup is similar to the platform mounted on TASER, the service robot of our group. The fact that the robot is equipped with a manipulator offers the possibility of not only recognizing objects, but also manipulating them.

(a) (b)

Fig. 8. The left image shows the original image of the table scene used for our initial experiments. The right image shows a robot arm as a moving platform over a table scene.

For the second platform (see fig. 7(b)) we use the robot arm Mitsubishi PA10-6C. This manipulator is a part of

TASER, our institute's service robot. The arm has 6 degrees of freedom and moves along a straight line in Cartesian space. The position exactness of the manipulator lies within +/- 0.1 mm. The results are presented as a separate table scene in fig. 9. Through the known transformation between laser scanner, arm and base of the platform the data from the laser scanner mounted on the arm can be registered in the same coordinate system as the fused sensor information. Thereby it is possible to have more data and especially another viewpoint. Consequently the integration of further sensors is possible through the calibration or if the transformations are known via the registration process.

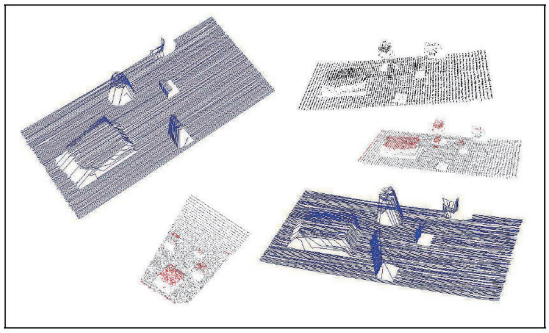


Fig. 9. Separated table scenes with a 2D laser scanner mounted on and moved through the 6 DOF robot arm.

Both systems permit an exact and variable control so that different scan distances and positions can be realized. The advantage of this system is the fact that the laser scanner can still be used in 2D without any physical changes.

For the demonstration of the results we use the table scene presented in fig. 8(a). We present the results of the adaptive sensor fusion between a laser range finder and a camera system mounted on a pan-tilt unit in fig. 10.

The results are partially colored point clouds. The perspective was changed for the purpose of a better view. The view area of the laser range finder is limited in respect to the construction. The resulting algorithm is implemented in C/C++ and visualized with OpenGL.

After the adaptive sensor fusion the Ball-pivoting algorithm

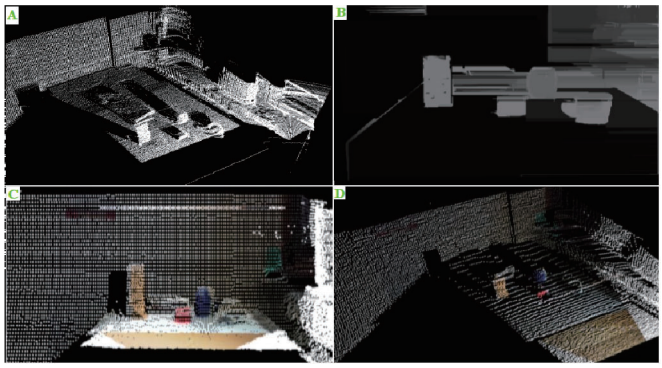



Fig. 10. a) The result of the 3D point cloud of the laser scanner. b) Disparity image of the stereo camera system. The images c) and d) show the early sensor fusion of depth and color information. Two different perspectives are shown.

is again used for surface reconstruction, this time for the partially colored point cloud. The result can be seen in fig. 11. Only the fused area is shown. The color for the recumbent voxels outside the fused area is set to black.

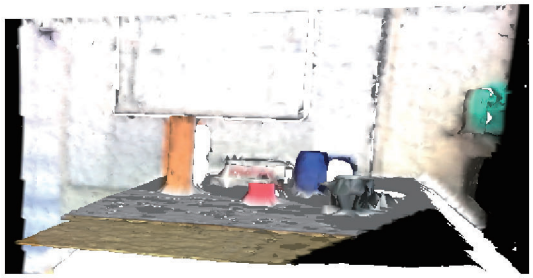


Fig. 11. The resulting reconstructed scene with the Ball-pivoting algorithm from the colored point cloud includes the interpretation step. Only the fused area is presented.

The implemented application combines the early sensor fusion and interpretation components, so it allows us to achieve better results than with only the data fusion. The analysis of the images permits taking an adaptive decision for sensor data consolidation.

For example, the black cup in our table scene (fig. 8(a)) is not clearly identifiable in the 3D laser scanner image. Moreover, it is invisible after surface reconstruction. The same behavior can also be observed after an early fusion, shown in images c) and d) in fig. 10. The interpretation steps deliver more suitable results, as presented in fig. 11. The representation of the cup is stable and clearly recognizable.

The counter-example shows the table surface. The camera finds no correspondences and therefore no depth information is available, either, as shown in image b) of fig. 10. After the interpretation the surface is clearly recognizable. Of course, more extensive testing is needed.

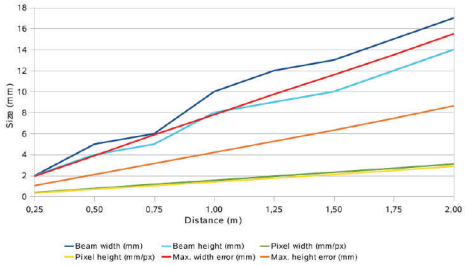



Fig. 12. The propagation of the laser beam and pixel size in relation to the distance as well as the maximal resulting error in $mm$.

The developed system inherits the errors and properties of their individual components, therefore the results of the adaptive sensor fusion are difficult to evaluate. The assessment of a laser scanner was already carried out in section II, so this issue will not be addressed further. The size of the laser beams increases with the distance, besides, the shortest measured distance or an average value of the measured distances will be delivered. At the same time the size of the surface references to an image pixel increases, however, the image pixel still has

only one value per color channel. Another source of errors is the Parallax effect which increases with the distance between both sensors. An error is thereby produced the minimum of which lies in the middle area and increases outwardly. In contrast, the effect of Parallax decreases with an increasing distance.

During the empiric experiments we assess a maximal error of approximately 5 pixels in the horizontal and 3 pixels in the vertical direction respectively. Besides, several table scenes with differently placed objects were examined. The propagation of the laser beam and pixel size in relation to the distance as well as the maximal error resulting from it in $mm$ are summarized in fig. 12.

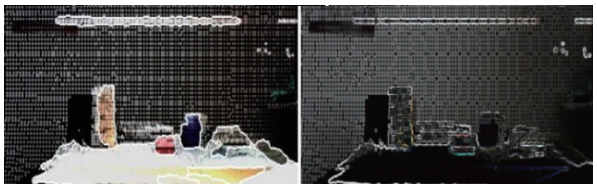


Fig. 13. Application of two common segmentation algorithms to the fused images. A color segmentation algorithm (JSEG) is applied on the left and a Sobel edge detector on the right.

The aim of the fusion was not only embellishment, but to permit the use of more information for robotic tasks like object recognition. In this sense we have applied two common segmentation algorithms for object recognition to the resulting images, color segmentation and edge detection. The results are shown in fig. 13.

We use JSEG [10] and the Sobel [11] algorithm for color segmentation and edge detection respectively. The quality of the results is satisfactory and is absolutely comparable with the results for the original images. After applying the mentioned 2D segmentation algorithms and enhancement of 3D the separated objects can even be used for possible grasp calculation. Fig. 14 illustrates the grasp calculation and simulation for the reconstructed simplified model of a blue barrel in the original image. The color information was removed after the segmentation for the better performance of the grasp calculator. For more information about the grasp calculator please see [12].

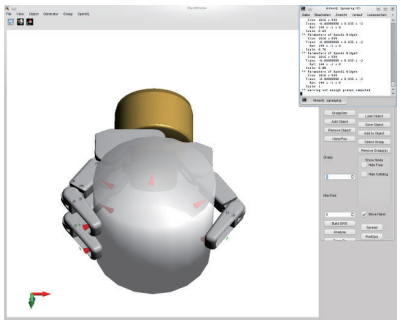


Fig. 14. The grasp calculation and simulation for the reconstructed simplified model of a blue barrel in the original image. The color information was removed after the segmentation for the better performance of the grasp calculator.

The temporal performance of the whole system is linear, directly proportional to the resolution of the pan-tilt unit and the number of the 3D points.

## VI. CONCLUSION AND DISCUSSION

In this paper, we proposed a perception system which permits the adaptive fusion of a 3D laser scanner and a stereo camera system. The result is an improved 3D-colored point cloud.

The presented adaptive sensor fusion and interpretation opens up enormous possibilities for robotics. First of all the method is interesting for object recognition, but also for the other robotics areas. The combination of a precise distance measurement and color information permits the application of several methods and data fusion on a high level. However, not only perception but also the interaction with the environment is possible. Due to the lower-error susceptibility and strong decrease of external influences like lighting conditions, even the safe interaction in human surroundings and with people becomes realizable.

## REFERENCES


[1] T. C. H. Heng, Y. Kuno, and Y. Shirai, Active Sensor Fusion for Collision Avoidance, 0-7803-4119-8, in Proc of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, vol. 3, pp. 12441249, 1997.

[2] S. Jacobs, A. Ferrein, S. Schiffer, D. Beck, and G. Lakemeyer, Robust Collision Avoidance in Unknown Domestic Environments, Springer Berlin / Heidelberg, ISSn 0302-9743 (Print) 1611-3349 (Online),vol. 5949, pp. 116-127, 2010.

[3] P. Tokekar, V. Bhatawadekar, D. Fehr, and N. Papanikolopoulos, Experiments in Object Reconstruction Using a Robot-mounted Laser Range-finder, 17th Mediterranean Conference on Control & Automation Makedonia Palace, Thessaloniki, Greece, June 24 - 26, 2009.

[4] J. Baltes, S. McGrath, and J. Anderson, Active Balancing Using Gyroscopes for a Small Humanoid Robot, 2nd International Conference on Autonomous Robots and Agents, Palmerston North, New Zealand, December 13-15, 2004.

[5] K. Nishiwaki, J. Kuffner, S. Kagami, M. Inaba, and H. Inoue, The experimental humanoid robot H7: a research platform for autonomous behaviour, Phil. Trans. R. Soc. A (2007) 365, 79107, doi:10.1098/rsta.2006.1921, 2006.

[6] D. Schneider, H.-G. Maas, Integrated Bundle Adjustment With Variance Component Estimation Fusion of Terrestrial Laser Scanner Data, Panaramic and Central Perspective Image Data, IAPRS Volume XXXVI, Part 3 / W52, 2007.

[7] M. Weser, S. Jockel, J. Zhang " Fuzzy Multisensor Fusion for Autonomous Proactive Robot Perception", In Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ 2008) at IEEE World Congress on Computational Intelligence (WCCI), Hong Kong, China, pp. 2262-2267, June 1-6, 2008.

[8] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, Sensor Fusion Using Dempster-Shafer Theory, IEEE Instrumentation and Measurement Technology Conference, Anchorage, AK, USA, 21-23 May 2002.

[9] D. Klimentjew, N. Hendrich, J. Zhang "Multi Sensor Fusion of Camera and 3D Laser Range Finder for Object Recognition", In Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI) Fort Douglas, University of Utah, Salt Lake City, USA, 2010.

[10] Y. Wang, J. Yang, Ni. Peng, Unsupervised color-texture segmentation based on soft criterion with adaptive mean-shift clustering, Pattern Recognition Letters, vol. 27, nr. 5, issn 0167-8655, 2005.

[11] I. Sobel, G. Feldman, A 3x3 Isotropic Gradient Operator for Image Processing, In Pattern Classification and Scene Analysis, R. Duda and P. Hart, John Wiley and Sons,'73, 1968.

[12] T. Baier, J. Zhang, "Reusability-based Semantics for Grasping Evaluation in Context of Service Robotics", IEE International Conference on Robotics and Biomimetic, Kunming, China, 2006.