Visual Data Combination for Object Detection and Localization for Autonomous Robot Manipulation Tasks

Luis A. Morgado-Ramirez, Sergio Hernandez-Mendez, Luis F. Marin-Urias, Antonio Marin-Hernandez, Homero V. Rios-Figueroa

Department of Artificial Intelligence, Universidad Veracruzana, Sebastian Camacho No. 5, 91000, Xalapa, Ver., Mexico
lamr22@gmail.com, smendez234@hotmail.com, luisfelipe@ieee.org, {anmarin,hrios}@uv.mx

Abstract. For mobile robot manipulation, autonomous object detection and localization is at the present still an open issue. In this paper is presented a method for detection and localization of simple colored geometric objects like cubes, prisms and cylinders, located over a table. The method proposed uses a passive stereovision system and consists on two steps. The first is colored object detection, where it is used a combination of a color segmentation procedure with an edge detection method, to restrict colored regions. Second step consists on pose recovery; where the merge of the colored objects detection mask is combined with the disparity map coming from stereo camera. Later step is very important to avoid noise inherent to the stereo correlation process. Filtered 3D data is then used to determine the main plane where the objects are posed, the table, and then the footprint is used to localize them in the stereo camera reference frame and then to the world reference frame.

1 Introduction

Automatic and robust detection and spatial localization of objects under no specific or controlled conditions, i.e. unlike of assembly lines or structured environments, is one of the most challenging tasks for computer vision. There are many problems where the use of these sort of solutions can improve the performance of automatic systems, e.g. in autonomous mobile robot manipulators or unknown scene analysis. On this work we focus mainly on mobile robot manipulation scenarios but the approach presented can be adapted to other complex environments.

For autonomous mobile robot manipulators is really a challenge to locate and detect objects on scenes. The applications for these kinds of task vary from folding clothes to environment mapping and so on. Typically the problem of mobile manipulators is decomposed on three stages: a) object detection and localization, b) approach planning and c) finally correct grasping. On this work we focus on the first of the three mentioned stages.

Commonly to detect and localize objects in such scenarios active sensors are used, e.g. active stereovision systems or time of flight (TOF) sensors. In active stereovision

the use of a monocular video camera and a known laser pattern (e.g. straight lines or a matrix of points) is used to recover 3D information from images captured from the camera [1]. On the other side, the use of sensors like LIDARs or TOF cameras, avoid the computation of 3D information as it is returned directly from the sensor and are a very good and robust source of data. As active sensors the consumption of energy is high and it is important to avoid having multiple sensors of the same kind on the same environment. For example having many robots doing their respective tasks on the same environment could be a problem if they use active sensors.

The use of passive stereovision is an alternative to deal with such a problem, however the accuracy of the 3D reconstruction is sometimes reduced. Accuracy on stereovision systems depends on many factors, e.g. a good calibration, the selection of internal parameters of the stereovision process (maximal disparity range, size of disparity window, etc.), the correct choice of lenses for specific tasks, etc.

The main purpose of this work is to allow an autonomous mobile robot to locate simple geometric objects as the ones showed on Fig. 1 (child's toys), in order to be able, on a continuation of this work, to manipulate them.

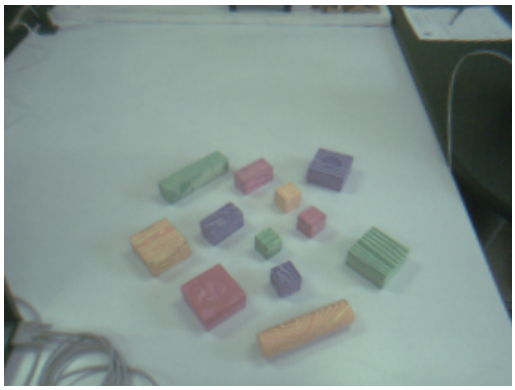


Figure 1. Simple Geometric textured objects used in our experiments, in a) image

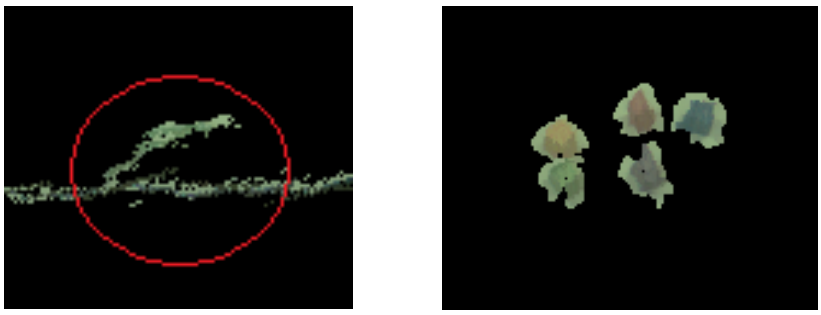
To reach mentioned objects the robot should be at a closer distance of them. Condition that poses some problems, for example, when the objects are closer to the minimal disparity plane, the stereovision correlation process induces many errors (Fig. 2). Moreover, if the size of the objects relative to the image is small, then the correlation window should be reduced to avoid finding false correlation values, however, this compromise the stereo vision accuracy.

On Fig. 2b is showed the results of a 3D visualization process of the recovered 3D information from a single cube in the image. As we can see on Fig. 2a, the results of disparity plane induce errors from which is not possible to recover 3D shape with the desired accuracy to induce form, in this case perpendicular planes belonging to the object. The main factor that induces these errors is the position of the camera relative to the plane where are posed the objects, in this case the table. A tilted camera produce that disparity planes, that are parallel to the image plane, will be not enough

to describe the planes of objects that are not parallel to these planes. Then errors are induced creating larger deformations.

On Fig. 2b are showed some segmented objects where it is possible to see the problem relative to correlation process, as we can see many wrongly matching pixels on the neighborhood of the objects are considered as part of them.

When the objects in the scene are bigger these errors can be neglected, however when it is desired that mobile robots could manipulate commonly humans objects this is not the case. Bigger objects will be more difficult to manipulate by only one anthropomorphous arm.



(a) (b)

Figure 2. Stereovision common problems, a) form induced from the correlated data does not match the form of the real object, b) some mismatched pixels create false data, which interpreted with truthful data deforms the objects.

This paper is organized as follow; on next section we analyze some recent related works. On section three and four we present the proposed approach, beginning with the object detection, followed by the object localization method, respectively. On section five we present the results, and finally on section six our conclusions and future work.

2 Related Work

The problem of object detection and localization for autonomous mobile robot is an active research field. As a direct result there are many approaches to deal with such a problem. As it has been exposed previously, many of these works use active sensors, together with CAD models in order to detect known objects and localization. For example in [2] CAD models are used in combination with an efficient hierarchical search computed offline to guide the exhaustive pose estimation.

In [3] it is presented a method for object detection using GPU computation in order to accelerate commonly object detection algorithms usually very slow and impractically to be used on mobile robot manipulation. There are some other approaches that use shapes descriptors, as for example in [4] where these descriptors are used in combination with a TOF sensor in order to recover 3D localization of the

objects detected. In [5] shape descriptors are also used to detect transparent objects detection.

In [6] passive stereo data has been used also in combination with CAD models in order to match known objects. Ulrich et al. in [7] propose a method for 3D model selection from a database over Internet.

There are some approaches to deal with object recognition when 3D points acquire from different sensors. For example in [8] Boyer et al. propose a robust sequential estimator to adjust surfaces of noisy data that included outliers. To eliminate outliers, the detected edges and smooth regions extracted and adjusted the areas determined by the AIC criterion. Takeda and Latombe in [9] considered a special case where the problem can be reduced to a problem in one dimension of a line as a set and compute a maximum likelihood solution.

In this paper, we present an alternative to geometric object detection and localization based only on visual characteristics, without the use of active sensors. The proposed method works at 18 Hz, speed which fast enough to deal with mobile robot manipulation tasks and allowing dynamic object localization, very important for visual servoing used in accurate grasping.

3 Object detection

The methodology proposed for object detection combines two kinds of visual information. On one side a color segmentation procedure is applied to the original images in order to get the regions where the colored objects are. On a second step borders in image are obtained. Both, visual information are merged together as a characteristics level fusion to get a robust object detection method.

3.1 Color Segmentation.

Image color segmentation is the decomposition of an image on the colored component parts with the same color attributes. For human beings is a very easy task, however for computer vision systems robust color segmentation is still a challenge. Human beings are able to combine different sources of visual information as texture, gradients, or different tonalities to recover the dominant color of a given object.

For autonomous computer vision systems are not always easy to define the way that all that information could be merged. For example in Fig. 1, are showed some objects that we want to be detected, as we can see, they have texture, and different color tonalities.

The first step in the proposed approach is to convert image to a normalized color space. The result of this process is to eliminate color variations from reflections, shadows or not equal illumination as it is showed on Fig. 3a. However at this stage is very difficult to make a good color segmentation, because as a result of the normalize color process the color space is reduced.

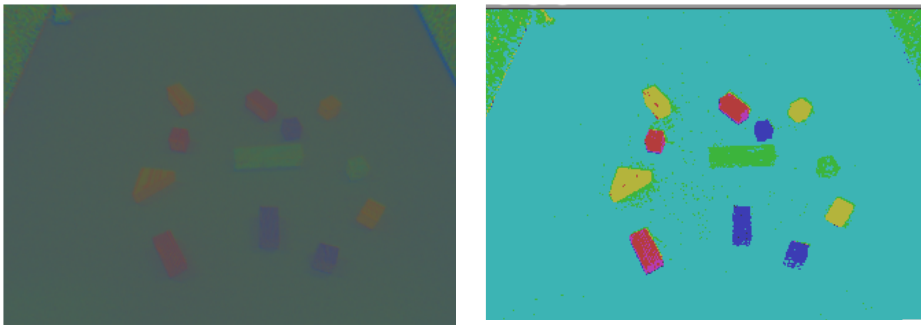
In order to detect easily colors a luminance increase is applied to the resulting image as it is showed on Fig 3b.

A color space reduction is then applied to the processed image in order to cluster similar colors as humans do easily, that is, many green tonalities are grouped together as only green color (Fig 4).

Image color space reduction is obtained applying the following formula:

$$RS_c = \frac{C}{D_v^2}, \text{ with } C \in \{R,G,B\} \tag{1}$$

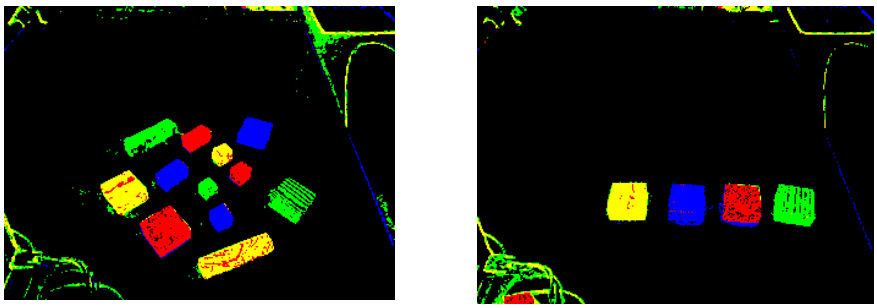
with RS_c the new color component and D_v the reduction factor for the cubic color space.



(a) (b)

Figure 3. a) Normalized color space and b) luminance adjustment to separate colors

In figure Fig. 4a are shown the results for the color space reduction where it is applied a color labeling in order to distinguish more easily the colors. As we can see in this figure, there are still some holes in the objects as well as some colors that do not belong correctly to the segmented object.



(a) (b)

Figure 4. Two images where reduce space color segmentation is applied

3.2 Edge Detection.

Combination of color segmentation with edge detection allow us to delimit object as well as assign colors to the regions where color has not been detected or it has been wrongly detected.

We use a simple canny edge detector that was the one that gives us best results as it is showed on figure 5. Even so, edge detectors can be used isolated to delimited objects, mainly edge detectors because edge detectors are sensible to illumination conditions given different edges in a sequence of images which correspond most of the time to shadows that are more or less perceive by the camera.

(a) (b)

Figure 5. Canny edge detector applied to the original images

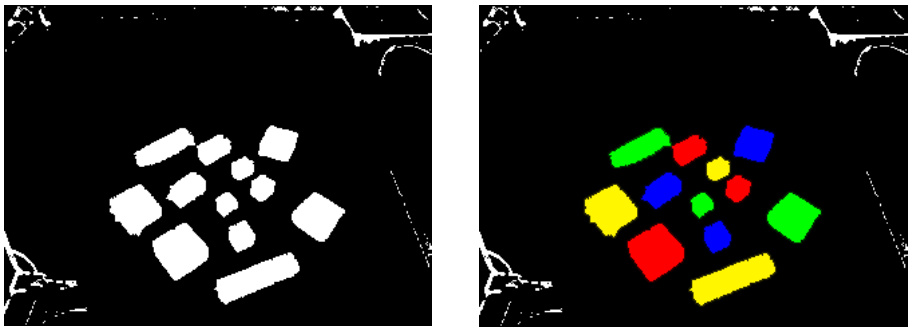
3.3 Object detection data fusion

Data fusion at this stage is done with the following procedure.

Be *C* the color label of a given pixel and *P* the percentage of similar colored neighbor pixels, then:

```
Add all the edges to a list L.
While L is not empty do
  If the pixel u in L belongs to a given color label C then
    Add u to a region C
    For each neighbor pixel v of u do
      If v belong to the same region C insert v in L
      If not, label as visited pixel.
```

With this procedure we obtain a binary mask that determines the colored objects as the images showed on Fig. 6a. Finally, dominant label on the object is assigned to the entire segmented region as it is showed on Fig. 6b.



(a) (b)

Figure 6. a) Binary mask obtained by the fusion of color segmentation and edge detection, and in b) Color object labeled

At this stage we have detected colored objects in the scene, now is necessary to localize these objects on the camera reference frame.

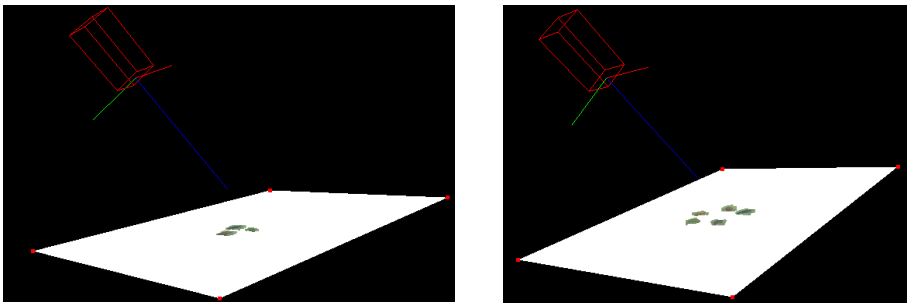
4 Object Localization

In order to avoid the problems with the noise produced by the stereovision correlation process as is has been showed on Fig. 2b, we use the object segmentation mask fig. 6 to recover 3D information only on the regions where the objects has been detected Fig. 7a. However, as the objects are small, a very near the minimal disparity plane the shape of the 3D form does not correspond to the real object form as has been described in Fig. 2a.

In order localize objects; we assume that all objects are over a table and as it has been said, all these objects are very simple geometric forms. So, the shape of the footprint in the surface of the table, give us information about the position and orientation of a given object. The problem here is the to find the plane equation corresponding to the table in order to project all the 3D segmented points on this surface to recover their footprint and then their position in the camera reference frame, and then to the world or manipulator reference frame.

As we have seen on Fig. 2a, disparity data has a lot of noise, so is not possible to recover flat surfaces belonging to the up side plane of the objects. In order to deal with this problem we have applied a RANSAC algorithm to find the best plane that fits to our objects and then with the bounding box of the 3D points belonging to the objects in the recently calculated plane, we have move the plane to the bottom of this bounding box.

Then the 3D points belonging to the objects are projected over this plane that correspond to the table plane Fig. 7b.

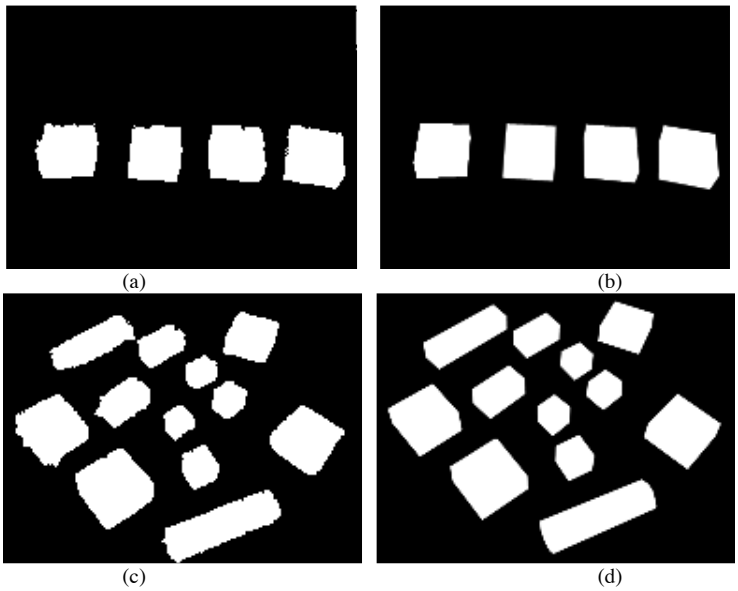


(a) (b)

Figure 7. Projection of 3D seemed points over a fitting plane that correspond to the table plane.

5 Results

In order to get a confidence measure we have compared the results of the proposed methodology for color image segmentation with the ground truth (Fig. 8) of the coor segmented objects and we have obtained a rate of 90% of matched pixels with an average of 2% of false positives and a 8% of true negatives matches.



(a) (b)

(c) (d)

Figure 8. a) and c) examples of color object segmentation mask and b) and d) corresponding ground truth.

We have estimated a maximal error of 1 cm on the projection of 3D footprint center considered as the location of the objects. Many tests under different light conditions were evaluated to validate the results of the proposed methodology without varying the described accuracy.

6 Conclusions

We have presented a method for simple geometric colored object detection based on the fusion of different visual characteristics. The proposed works at a frame rate of 18hz, making it ideal for robotics applications where computing time is crucial. The mask obtained by the color segmentation process is used as a noise filter in order to avoid errors on the stereo correlation process that produces bad objects localization estimations. The method can be applied continuously in order to give to a manipulation robot the active state of the world.

Fusion data structure will be used in future works in order to match edges from both images coming from the stereo camera in order to get a more robust object detection and localization.

References

1. Quigley M., Batra S., Gould S., Klingbeil E., Le Q. V., Wellman A., Ng A. Y.: High-accuracy 3d sensing for mobile manipulation: Improving object detection and door opening," in IEEE International Conference on Robotics and Automation, 2009
2. Ulrich M., Wiedemann C., Steger C.: Cad-based recognition of 3D objects in monocular images. in International Conference on Robotics and Automation (2009) 1191–1198.
3. Coates A., Baumstarck P., Le Q. V., Ng A. Y.: Scalable learning for object detection with gpu hardware," in *IROS* (2009) 4287–4293.
4. Marton, Z.; Pangercic, D.; Blodow, N.; Kleinehellefort, J.; Beetz, M.: General 3D modeling of novel objects from a single view. Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on (2010) 3700 – 3705.
5. Fritz M., Darrell M., Black M., Bradski G., Karayev S.: An additive latent feature model for transparent object recognition," in *NIPS*, S. for Oral Presentation, Ed., 12/2009 (2009)
6. Hillenbrand U.: Pose clustering from stereo data. in Proceedings VISAPP International Workshop on Robotic Perception – RoboPerc, (2008).
7. Klank, U.; Zia, M. Z., Beetz, M.: 3D model selection from an internet database for robotic vision. Robotics and Automation, 2009. ICRA '09. IEEE International Conference on. (2009) 2406 – 2411.
8. K.L Boyer, M.J. Mirza, and G. Ganguly, The robust sequential estimator: a general approach and its application to surface organization in range data, IEEE Trans. Pattern Anal. Machine Intell., vol. 16, no.10 (1994) 987-1001.
9. H. Takeda and J-C. Latombe, Maximum likelihood fitting of a straight line to perspective range data, IEICE Trans., vol.J77-D-II, no.6 (1994) 1096-1103.