Technical Sciences
Academy of Romania
www.jesi.astr.ro

Journal of Engineering Sciences and Innovation
Volume 2, Issue 1 / 2017, pp. 22-34

A. Mechanics, Mechanical and Industrial Engineering, Mechatronics

Received ***8 November 2016*** *Accepted* ***7 March 2017***
Received in revised from ***22 January 2017***

SpatialVISION – software module for objects identification before robotic gripping

IONEL STARETU*, CATALIN MOLDOVAN

Transilvania University of Brasov, 29 Eroilor Bd., Brasov, Romania

Abstract: In this paper we present a software module for object recognition before robotic gripping with an anthropomorphic gripper. To this effect, recognizing the shape of an object using digital image processing algorithms, falling in the area of digital sequences processing, is important. This area has evolved a lot lately, in particular due to the progress of computing resources, involving parallel processing of each basic unit of an image. The paper summarizes the theoretical background and the results related to the software developed for this purpose, called SpatialVISION. To implement the SpatialVISION application, the implementation of the SIFT algorithm was used, which is a personal approach in C #, using EmguCV library for working with images. To use the SpatialVISION application, a tool called SVAdminUtility was also created, which is used to populate a database with three-dimensional objects that will later be recognized. In structural terms, the SVAdminUtility application has four functional components namely: SVAdminUtility component; Collada component; Viewer component and XSD2NET component. To assess the SpatialVISION application, an experiment was conducted, whose results are presented in the paper, which consists in measuring the average time of execution of the object recognition algorithm using two types of images captured by two types of cameras: a WEB camera and a Canon PowerShot camera.

Keywords: software, shape recognition, image processing, object, robotic gripping.

1. Introduction

Object shape recognition using digital image processing algorithms falls in the area of digital sequences processing. This area has evolved in recent years due to the progress of computing resources, involving parallel processing of each basic unit of an image.

The basic unit of a digital image is the pixel. A digital image can be represented by a MxN matrix like this [1]:

$$F = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0, M-1) \\ f(1,0) & f(1,1) & \dots & f(1, M-1) \\ \vdots & \vdots & & \vdots \\ f(N-1,0) & f(N-1,1) & \dots & f(N-1, M-1) \end{bmatrix} \tag{1}$$

In the matrix 1, f (x, y) is a function of two spatial variables x and y, associated with each pixel, parameters characterizing the intensity of the pixel in question.

The purpose of processing digital images can be divided into three main categories, namely [2] : **image processing**: it is a form of image processing, applying a transformation to the function f (x, y) resulting in a new matrix; **image analysis**: refers to the way in which meaningful information can be extracted from images by digital image processing techniques; **image understanding**: is the process by which the results of actions performed after processing based on analysis and processing algorithms, can be interpreted and possibly decisions can be made based on this information.

To achieve an analysis module, and to understand objects, we aim at developing a processing component of video sequences, taken from an image capture device. For a captured video sequence, processing is performed for each frame separately, and consists of initial application of filters, to improve image quality, to identify regions of interest in the image, and subsequently, to apply artificial intelligence algorithms for object recognition.

In this paper, we describe the first component of the command and control strategy of an anthropomorphic gripper, which consists in a software module for understanding objects and for estimating their size. This software module is made of two components, namely: the first component is a database of known objects, together with their physical characteristics, which can also be extended by adding new models of three-dimensional CAD objects, and the second component is the image recognition module, which also estimates the size of objects.

The algorithm for understanding objects is based on analysis of interest spots, and it is a variation of the SIFT algorithm, first set by Lowe [3] created based on Lindenberg's research [4]. It uses spots of interest to match unknown and known objects and compares to originally recorded objects' spots in an objects' database.

For object recognition in images, the following two steps were implemented: locating key spots: setting interest spots, on the one hand, in the image captured by an image capturing device, and on the other hand, setting interest spots in all the images stored in the database of three-dimensional objects; and key spots stabilizing: is the process that removes previously set interest spots that are not stable in changes of direction or light intensity.

The paper describes the practical component that estimates physical properties of objects and the generation of the three-dimensional model based on image analysis.

The result is the SpatialVISION application, able to estimate, in the first instance, the shape of the objects, and then, based on image analysis, to detect the physical properties of the objects and to generate their three-dimensional model (using the XML – COLLADA format). A final step is providing information to the gripping functional simulation component to achieve then physical gripping [5], [6].

2. Characteristics of SpatialVISION software component

The first component of the SpatialVISION application uses an architecture that allows the extension of the communication possibility with the video device so as we can use any data capture device (RGB camera provides depth data). For the data captured we apply object recognition algorithms, to estimate physical properties and to generate 3D models.

Data analysis is possible using Microsoft technologies for the management of external devices connected to a computer and image processing techniques. The frames' analysis is performed in four stages, namely: in the first phase images are filtered as follows: color changes or improvement of their quality, in the second phase analysis techniques are used, such as recognizing objects, and in the third phase, an estimation algorithm is performed, to estimate the object size, and shape and other physical characteristics that will have implications in the gripper's configuration, and in the final phase, the three-dimensional model of the recognized object, is generated.

The video content analysis to obtain spatial information represents a new approach in robotics, in terms of the possibility to equip a robot with senses, especially the sense of vision. Initially, the video content has been captured and transmitted using an analog method, but advanced research in digital integrated circuits area led to a revolution in the video content compression and communication, which currently can be analyzed and processed. Thus, a new area of research came to light, based on applications in robotics of the result of processing the digital video content. Currently, reference to robotics and visual senses is not possible without referring to video image capture and digital processing.

Each file or video stream comes with a number of physical characteristics that can increase or decrease the size in kilobytes of each frame captured separately. With the size increasing of each frame, the number of the details that can be analyzed increases too, and in parallel, the processing power necessary to apply certain filters also increases.

3. Standards and software libraries used in obtaining the SpatialVISION component achievement

To implement the first component of the command and control system defined in this paper, we used universally accepted standards for defining three-dimensional models and software libraries for the digital image processing detailed below.

3.1. The EmguCV library

EmguCV library is a software package created using the OpenCV library. EmguCV is written entirely in .NET and contains implementations of the most popular algorithms for digital image processing. The EmguCV library can be used in many programming languages, namely C #, VB.NET, C ++ or Iron Python [7].

From the architectural point of view, EmguCV is divided into two levels, namely: at the first level there are functions, structures, and enumerations that relate to the implementation of the ones in OpenCV; at the second level there are classes that add the advantages of the implementation in .NET. The connection between the functions EmguCV and OpenCV is possible using the CVInvoke class. Each method in this class corresponds to a class in OpenCV, bearing the same name, for example, to create an image in the memory, we use:

IntPtr image = CvInvoke.cvCreateImage(new System.Drawing.Size(400, 300), CvEnum.IPL_DEPTH.IPL_DEPTH_8U, 1);

which in OpenCV is equivalent to:

IplImage* image = cvCreateImage(cvSize(400, 300), IPL_DEPTH_8U, 1);.

Through EmguCV, we can use the OpenCV functionality for loading, saving and acquisition of video images. The EmguCV library contains many implementations of algorithms working with image files. Among the most important functionalities offered by the EmguCV library, there are [1] and [7]: manipulating an image data (allocation, deallocation, copying, saving, conversion); images and video stream Input / Output (both at file and device level - web-cam, camcorders, digital cameras); support for operations with matrix and vectors (scalar product, vector product, eigenvectors, etc.); various dynamic structures (lists, queues, trees, graphs); structural analysis (connected components, contours processing, distance transformations, variable moments, assembling molds / templates, type Hough transformation, polygonal approximation, approximation of areas of interest through lines, approximation of areas of interest through ellipses, Delaunay triangulation); calibration of cameras (calibration by determining and tracking patterns, calibration, fundamental matrix estimation, stereo correspondence); motion analysis (motion tracking in a video sequence, motion segmentation); object recognition (methods based on eigenvalues, HMM (Hidden Markov Model)); Simple interface GUI - Graphics User Interface (displaying images / video at the window, delegates for handling events received from the mouse and keyboard, scroll bar); possibility to overlay layers and images (lines, polygons, text), etc.

The EmguCV library is divided into several modules, grouped by the way they affect an image. These modules are [7]: Emgu.Util.dll and Emgu.CV.dll representing components with basic functionality of the library EmguCV; Emgu.CV.Util.dll – are OpenCV utility elements; Emgu.CV.UI.dll - is a functionality that allows the creation of graphical user interfaces; Emgu.CV.GPU.dll - is a functionality that allows working with NVIDIA graphics plates, containing CUDA accelerators [8]; Emgu.CV.ML.dll - is a functionality that allows the use of artificial intelligence algorithms. The EmguCV library provides a number of tools for working with images or frames of video streams in a structured manner. For each frame from the video stream, we allocate an IntPtr type structure. In this mode, the SpatialVISION application has access to the capabilities of EmguCV library, working with images being performed very easy compared to other libraries.

3.2. XML - COLLADA standard

COLLADA standard [9] is a file format used in applications that enable working with models. This standard is managed by the Khronos Group and is available. COLLADA standard is defined as an open XML standard used for exchanging digital assets between applications for graphic processing. COLLADA standard is defined as an open XML standard used for exchanging digital assets between applications for graphic processing. Documents using the COLLADA standard are called Collada instance documents and have the extension .dae (Digital Asset Exchange). For describing COLLADA files content, their structure and semantics, we can use XML language. An XML document contains elements which are blocks of information with a start and end label.

4. The software SpatialVISION

Locating and classifying objects in a scene automatically is a difficult and topical problem. A robot, if it benefited from the ability to classify unknown objects in object classes, it would be very useful in engineering or domestic applications. Currently, there are many ways of object modeling and recognition techniques.

In many applications currently applied to recognize objects, we use the primitive concept [10], [11], but this concept does not allow the recognition of complex objects. Because of this, algorithms based on features search in images [11] and [12] or classifiers [13] were developed.

For objects recognition, in this paper we use algorithms based on features search in images. We also use the research result of [14] research, to obtain invariance at noise by running a smoothing filter on image, and the result of [15] research, to obtain invariance at lighting conditions using image variations reported to the level of grey.

To implement SpatialVISION application, we used the SIFT algorithm implementation. SIFT algorithm implementation presented in this paper is our approach in C # using EmguCV library for working with images.

To use SpatialVISION application usage, we also created a tool called SVAdminUtility, which is used to populate a database with three-dimensional objects that will later be recognized.

4.1. SVAminUtility tool

The software component for the database management of objects that can be recognized by SpatialVISION application is called SVAdminUtility. With this component, a knowledge database can be populated initially and subsequently enriched with new three-dimensional models and with different physical properties.

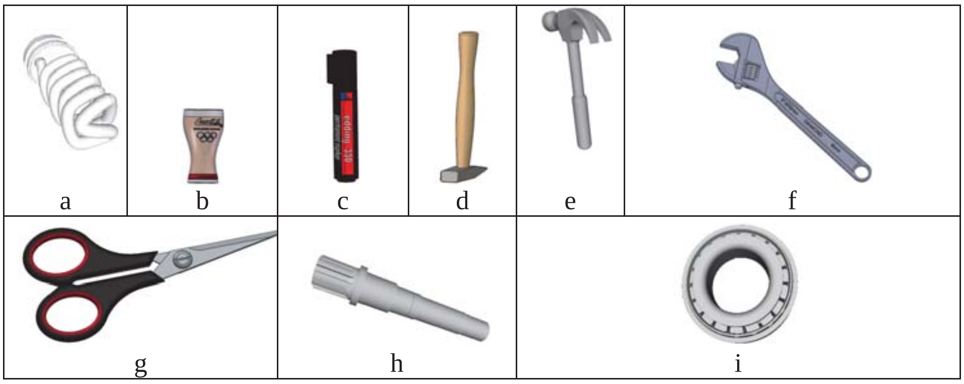
All the data, the object, the three-dimensional mode, the images and the points are stored in the knowledge database. In the diagram UML (Unified Modeling Language) of the database developed within the application, three problems are solved: adding objects and assigning images to an object; adding the three-dimensional model of an object association; adding lists of physical properties to each object.

The database originally attached to the SVAdminUtility application is populated by 9 different objects from different classes of objects. Table 1 shows objects from various classes to which we attached three-dimensional models. All the objects were chosen randomly and a three-dimensional model was originally added.

Because the SpatialVISION application make comparisons between images, for each model we will generate a number of images from different perspectives (in Table 1 a small number of perspectives are exemplified). We noted that the small number of perspectives is the main cause of algorithms' instability. They detect features in scenes from video streams in real time [16] - [18].

From the structural point of view, the application SVAdminUtility has four functional components namely: **Component SVAdminUtility**: the main component used to manage objects, classes and their properties (it contains the entry point into the application, graphical interfaces and classes for working with database); **Collada component**: .NET component used to working with COLLADA files; **Viewer Component**: component used to view 3D objects in COLLADA format; **XSD2NET component**: component used to map the scheme to define the COLLADA standard for .NET.

Table 1. Objects in the database



The SVAdminUtility application allows adding physical properties besides the three-dimensional model of an object. For each model stored in the database because the SIFT algorithm is partially invariant to rotation, images were generated in different perspectives, with SIFT points for each image individually. For each object that is added to generate images from different perspectives (Table 2) we use the

OpenGL framework. For every image separately, we search, using the SIFT classification method, the points of interest.

Table 2. Images generated based on a three-dimensional model from different perspectives

4.2. Defining the objects' properties

To define the objects' properties, the SVAdminUtility application will use the window illustrated in Fig. 1. The properties of each object are organized by types of properties (color, weight, etc.). The window in Fig. 1 allows processing of properties' list in a unified way. When a property is selected, a text dialog opens allowing the editing of that property.

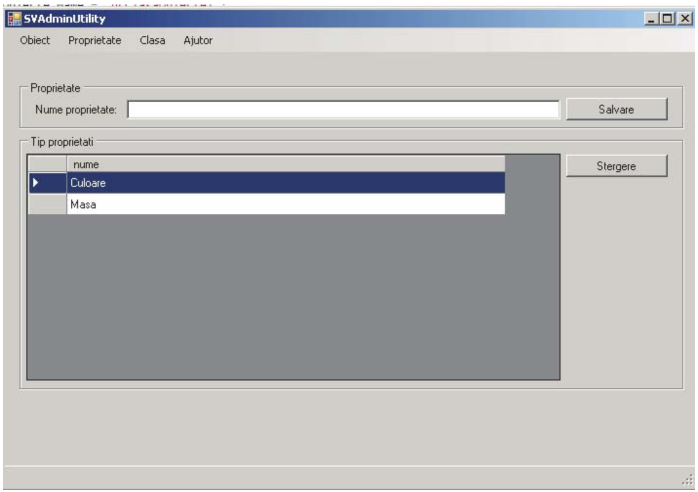


Fig. 1. The image that represents the way of adding physical properties to an object and their mapping with physical properties specific to the COLLADA format.

At the run time of the detection and objects recognition algorithm, for each object detected, we will load its three-dimensional model (along with its physical properties) in the functional simulation component of gripping. There are two types of objects that can be loaded into the functional testing component of gripping, i.e.

static objects (obstacles) and dynamic objects (objects that can be gripped). Static objects don't participate in the gripping dynamics, but they provide an area of collision for dynamic objects. In some cases, dynamic objects can be used as static objects. For a dynamic object, the following physical properties can be translated into the application of functional simulation of gripping:

- **Geometry**: it describes the form of the object to be gripped. The geometry of the object is stored in a file similar to VRML. The structure can contain 2D polygons that define an area or surface knots for primitives such as: cubes, spheres, cylinders, cones;

- **The material** of the object affects the friction force at contact between the virtual hand and the object. For each material we can define a friction coefficient using the graphical interface shown in Fig. 2.

	Without friction	Glass	Metal	Plastic	Wood	Rubber
Without friction	0	0	0	0	0	0
Glass	0	0.2	0.2	0.2	0.3	1
Metal	0	0.2	0.2	0.2	0.3	1
Plastic	0	0.2	0.2	0.3	0.4	1
Wood	0	0.3	0.3	0.4	0.4	1
Rubber	0	1	1	1	1	2

Fig. 2. Graphic interface that allows manual editing of friction coefficients.

- **Gravity:** in case of functional simulation of self-gripping, if an object has mass, it will be gripped automatically. The gripping quality differs if gravity is enabled or not (Fig. 3);

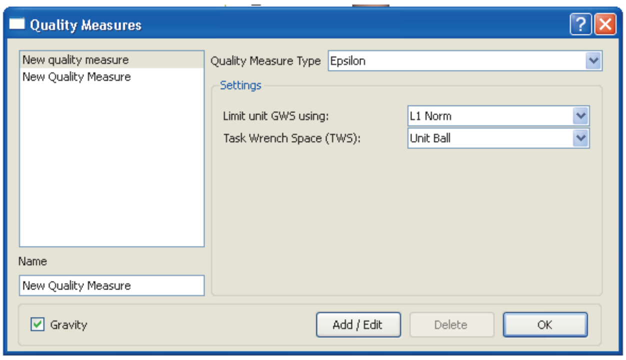


Fig. 3. Functional simulation of gripping.

- **The mass** of an object in grams (Fig. 4);

Fig. 3. Editing object mass.

- **Center of mass:** is a three-dimensional position relative to the coordinates system, which is used when gravity is enabled;

- **Inertia tensor:** it represents the distribution standard of 3X3 matrix of mass relative to a coordinates system.

Three-dimensional models are stored in the database using UTF-8 encoding. For each record in the Object table, we will store the in three-dimensional model in XML- COLLADA standard, with images generated from different perspectives for viewing the three-dimensional model.

The SpatialVISION application, presented in this paper, is a software module with the following specific objectives: recognizing objects in images or video stream based on image processing algorithms implemented; estimating the physical properties of the objects recognized; creating a virtual three-dimensional object based on the object recognized in the image. The application allows loading models or parts of an object, having the possibility to specify physical features that may influence interaction with other objects. For storing and transferring objects with physical properties, we use the three-dimensional model of XML - COLLADA files [9].

4.3. Detecting points of interest in images

The name SIFT comes from "Scale invariant feature transform" and represents invariant features in scalar transformations. The SIFT algorithm is used in digital image processing applications to detect and describe specific features of an image. This algorithm was first published in 2004 by David Lowe [20]. Lowe's research started from identifying the location, in the picture, of spaces that are invariant under translations, scaling, and rotating, and are not quite affected by noise in the image, and minor changes in orientation.

In Fig. 5, there is the number of key points detected using the SIFT algorithm on a range of two octaves.

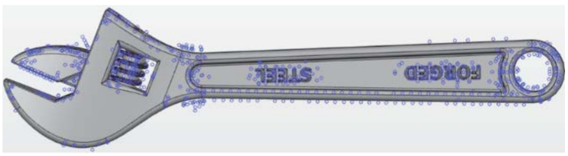


Fig. 5. Key points detection

4.4. Object class identification and calculation of limits

As search through all objects in the database is very costly in terms of time, it was decided to split the objects in object classes. The search space is significantly reduced using this mode. For each item, which is initially inserted into the database, we define a maximum and a minimum limit. Based on the minimum and the maximum limit, the search space decreases. For example, if for an object that wants to be recognized, there are 15,000 points of interest, if the limit is +/- 10%, then the search is performed only between objects that have between 13,500 and 16,500 points of interest.

4.5. Search of data in the CAD database

To find the three-dimensional model associated with each image, we use the limit points calculated on the basis of the above definition. SIFT points of interest for each image model I_m must be within the limit defined. For each captured image, SIFT points of interest of an image I_t will be searched in the images database. Thus, in the first stage, the number of possibilities is used to generate the collection of model images I_m based on the calculated limit.

5. SpatialVISION application assessment

To assess the SpatialVISION application, we conducted an experiment to measure the average execution time of the object recognition algorithm. The recognition method implemented was tested using two types of images captured by two different camera types. The two cameras used are: WEB with maximum resolution 1 MP and Canon PowerShot with 15 MP resolution. The reason for a comparison between the execution of the algorithm for the two types of images is based on the fact that we tried to find an optimal ratio between the image quality and the execution time of the recognition algorithm. This second test is very important because, if the algorithm is executed and the result is acceptable using the image captured by the camera with lower resolution, the execution speed is considerably improved.

The test was performed for 12 images, which are captured using the two cameras. For each image I_t separately there are points of interest detected and limits are calculated for points of interest. The points of interest of an object in the database must be must be within these limits.

Following the execution of the test, the points of interest detected in images captured with the WEB camera are shown in Table 3, and the points of interest detected in images captured with the camera Canon PowerShot are represented in Table 4.

Table 3. Points of interest found in captured images using WEB camera

Object	Image size	SIFT points detected	Time - milli-seconds	Limits calculated (min and max)
Screw spanner	800X600	246	294	*221 – 270*
Hammer	800X600	91	249	*81 – 100*
Markers	800X600	186	276	*167 – 204*
Hammer 2	800X600	103	258	*92 – 113*
Scissors	800X600	202	350	*181 – 222*
Energy saving bulb	800X600	192	276	*172 – 211*
Glass	800X600	150	267	*135 – 165*
Axis	800X600	232	295	*201-255*
Bearing	800X600	179	211	*152-196*

Table 4. Points of interest found in captured images using Canon PowerShot camera

Object	Image size	SIFT points detected	Time - milli-seconds	Limits cal-culated (min and max)
Screw spanner	4000X3000	2229	6588	*2006 - 2451*
Hammer	4000X3000	453	4657	*407 – 498*
Markers	4000X3000	950	5002	*855 - 1045*
Hammer 2	4000X3000	235	4630	*211 – 258*
Scissors	4000X3000	1620	5589	*1458 - 1782*
Energy saving bulb	4000X3000	1296	5271	*1166 - 1425*
Glass	4000X3000	820	4782	*738 - 902*
Axis	4000X3000	1789	4568	*1611-1969*
Bearing	4000X3000	1432	4872	*1289-1575*

As can be seen from the analysis of the two tables, the points of interest grow in number with increasing image resolution, but the execution time also increases. Using the points of interest and the approximation of the number of points, the images will be selected from the CAD database using the process model shown in Fig. 6. The recognition test was performed on a computer Intel Core 2 Duo with 2.39 GHz, and 1.94 GB RAM, with Windows XP operating system, version 3.06. In the case of algorithm execution on captured images using the WEB cam, the average number of points of interest is 133 points detected on average in 271 milliseconds, and in the case of images of resolution 15 megapixels, the average number of points of interest is 864 points in 5025 milliseconds.

As a conclusion, we can say that the minimum and maximum limits are useful only for images with high resolution. For images with low resolution, the number of points of interest detected is very low, so the minimum and maximum limits include either all objects with low number of points of interest or no object at all.

Fig. 6. Steps used to identify a three-dimensional object.

SpatialVISION application software has been achieved the control system of an anthropomorphic gripper with five fingers, which is shown in Fig. 7. After object gripping identify with SpatialVISION software, information is transmitted in the virtual environment where grip is correct simulation and testing using the component HandSIM, and after data is transmitted to real gripper RoboHAND, which makes catching the object.

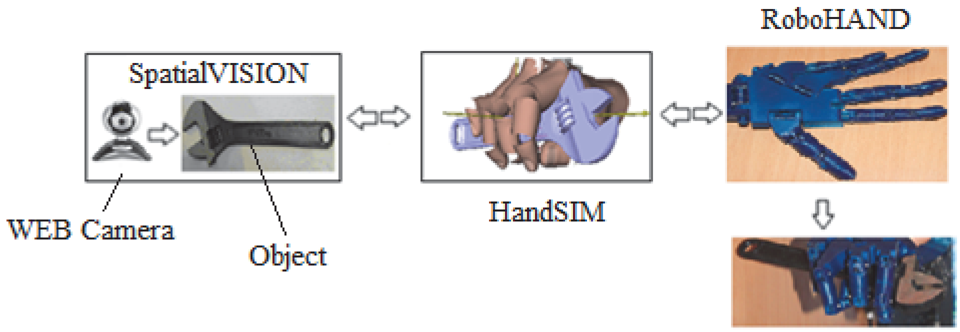


Fig. 7. Simplified structure of the control system developed.

More details on this control system of Fig. 7 are shown in [22] and other details will be presented in the next papers.

6. Conclusions

The algorithms based on local characteristics are widely used in research applications in the field of digital image processing or in processing video streams. With these types of algorithms, certain tasks, such as recognition of classes of objects, matching templates, object tracking or three-dimensional reconstruction, can be easily performed in a repeatable manner [21]. We can say that SIFT algorithm is a very effective method of matching objects to a class comparing points of interest.

One of the most important aspects of algorithms based on local characteristics is that they are invariant to changes in lighting conditions, orientation of objects or change in the field of view. These algorithms are executed in effective manner, even if the physical characteristics of the objects are slightly modified.

Local descriptors detection algorithms locate covariant regions in transformation classes. These regions are then used for the detection of invariant descriptors, which are finally used for object recognition.

Following objects identification we can achieve an integrated robotic gripping system, using a real anthropomorphic gripper, which will be obtained through subsequent research and achievements.

References

[1] Itu A.M., *Grasping Strategies Contributions in Real and Virtual Environments Using a Three Fingered Anthropomorphic Gripper*, PhD Thesis, Transylvania University of Brasov, Romania, 2010.

[2] Shapiro L.G., Stockman G.C., *Computer Vision*, Prentice Hall, 2001.
[3] Lowe D.G., *Object recognition from local scale invariant features*, Proceedings of International Conference on Computer Vision, Corfu, Greece, 1999, p. 1150-1157.
[4] Lindeberg T., *Scale-space theory: A basic tool for analyzing structures at different scales*, Journal of Applied Statistics, 21, 2(1994) 224–270.
[5] Staretu I., *Gripping Systems*, Derc Publishing House, Tewksbury, Massachusetts, SUA, 2011.
[6] Staretu I., Itu, A., Moldovan C., *Processing and analyzing object images with the view of object grasping*, The Annals of „Dunarea de Jos" University of Galati, 1 (2011) 5-10.
[7] Shi S., *Emgu CV Essentials*, Packt Publishing, 2013.
[8] Moldovan C.C., Staretu I., *Hardware used for realistic rendering in virtual world of industrial virtual models with physical properties*, AGIR Bulletin, 2 (2013) 19-24.
[9] COLLADA – *Digital Asset Schema Release 1.5.0*, Technical Specifications, Sony Corporation, 2008.
[10] Carr P., Sheikh, Y., Matthews, I., *Monocular Object Detection Using 3D Geometric Primitives*, ECCV'12 Proceedings of the 12th European conference on Computer Vision, Vol. 7572, 2012, p. 864-878.
[11] Berner A. et al., *Combining contour and shape primitives for object detection and pose estimation of prefabricated parts*, Proceedings of IEEE International Conference on Image Processing (ICIP), Melbourne, VIC, Australia, 2013, p. 3326-3330.
[12] Vondrick C. et al., *Inverting and Visualizing Features for Object Detection*, Technical Report, 2013.
[13] Shen C. et al., *Training Effective Node Classifiers for Cascade Classification. International Journal of Computer Vision*, 103, 3(2012) 326-347.
[14] Maania R., Kalrab S., Yanga, Y., *Noise robust rotation invariant features for texture classification*, Pattern Recognition, 46, 8(2013) 2103-2116.
[15] Maeno K. et al., *Light Field Distortion Feature for Transparent Object Recognition*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, p. 2786-2793.
[16] Zhang S. H., Liu, J. X., *A Self-Occlusion Detection Approach Based on Depth Image Using SVM*, International Journal of Advanced Robotic Systems, Vol. 9, 6 (2012) 1-8.
[17] Mahmood F., Kunwar F., *A Self-Organizing Neural Scheme for Door Detection in Different Environments*, International Journal of Computer Applications, 60, 9 (2012) 13-18.
[18] Zia Z.M., Stark M., Schindler K., *Explicit Occlusion Modeling for 3D Object Class Representations*, Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, 2013, p. 3326-3333.
[19] Wohlkinger W., Aldoma A., Rusu R.B., Vincze M., *3DNet: Large Scale Object Class Recognition from CAD Models*, Proceedings of IEEE International Conference on Robotics and Automation, Saint Paul, Minnesota, USA, 2012, p. 5384-5391.
[20] Lowe D.G., *Distinctive Image Features from Scale Invariant Key-points*, International Journal of Computer Vision, 60, 2(2004) 91-110.
[21] Moldovan C.C., Staretu I., *Analyze of the SIFT Feature descriptors algorithm implementation used for object pieces recognition*, The Annals of "Dunarea de jos" University of Galati, 2(2012) 83-86.
[22] Staretu I., Moldovan C., *Leap Motion Device Used to Control a Real Anthropomorphic Gripper*, International Journal of Advanced Robotic Systems, Vol. 13, article number: 113, Jun 6 (2016) 1-12.