3D scene's object detection and recognition using depth layers and SIFT-based machine learning

Article *in* 3D Research · September 2011
DOI: 10.1007/3DRes.03(2011)6

2 authors:

Tsampikos Kounalakis
Aalborg University
13 PUBLICATIONS **22** CITATIONS
SEE PROFILE

George A. Triantafyllidis
Aalborg University
59 PUBLICATIONS **675** CITATIONS
SEE PROFILE

3D Res. **2,** 03(2011)6
10.1007/3DRes.03(2011)6

3DR EXPRESS

3D scene's object detection and recognition using depth layers and SIFT-based machine learning

T. Kounalakis • G. A. Triantafyllidis

Received: 06 June 2011 / Revised: 04 July 2011 / Accepted: 14 July 2011

Abstract This paper presents a novel system that is fusing efficient and state-of-the-art techniques of stereo vision and machine learning, aiming at object detection and recognition. To this goal, the system initially creates depth maps by employing the Graph-Cut technique. Then, the depth information is used for object detection by separating the objects from the whole scene. Next, the Scale-Invariant Feature Transform (SIFT) is used, providing the system with unique object's feature key-points, which are employed in training an Artificial Neural Network (ANN). The system is then able to classify and recognize the nature of these objects, creating knowledge from the real world.

Keywords 3D object detection and recognition, SIFT, Machine learning

1. Introduction

Biological vision such as human vision is a complex process of organisms to extract information of the world by using the visible light. Due to the development of computer science, we try today to replicate human vision and perception with computer vision. Object detection and recognition are some of the most desirable applications in this field.

In general, the object detection and recognition algorithms are divided in two main categories: the appearance based methods and the feature extraction methods. The appearance based methods are traditional approaches and they process the 2D image by using the structural and/or frequency fields. Examples of these methods are edge detection, Hough transform, corner detection, and Fourier transform. The problem is that these techniques perform efficiently only on a controlled environment (controlled light and object pose). The feature extraction methods are algorithms that process the 2D image and extract features to signify that image. The advantage of this technique is that the features are unique for each image and more robust on light and pose changes. Some of these algorithms are Scale-Invariant Feature Transform (SIFT)[1], Speeded Up Robust Features (SURF)[2], Gradient Location and Orientation Histogram (GLOH)[3], and Local Energy based Shape Histogram (LESH)[4]. The main disadvantage of these techniques is that the extracted features are unique to each object and cannot describe a similar one. Although these two approaches are not similar, both are referring to object detection and recognition using two dimensional imaging.

However, the real world is described by the geometric model of the three dimensional (3D) space. Based on this fact, the proposed system is able to extract each object from real world scenery, by using depth maps of image stereo pairs. Then, each object is processed separately by using the feature extraction algorithm of SIFT. The SIFT features of the objects detected in the 3D space, are the keys for obtaining the knowledge, by employing an artificial neural network (ANN) and thus, succeeding object recognition.

The novelty of the proposed system resides in two main points: first, the combination of depth map and SIFT-based ANN for utilizing the scene's 3D information along with the adaptiveness of machine learning and second, and the fact that the system classifies objects which are similar and not necessarily identical to object categories. It is also important to clarify that the SIFT algorithm is not applied on the full 2D image, but only on the detected object, cropped from the whole scene, containing the shape and the 2D texture of this specific object, detected form the proposed depth map processing. This feature introduced by this algorithm actually succeeds an efficient combination of the SIFT theory with the depth maps processing for object recognition. The performance of these two stages is important for the efficiency of the system. The choice of the SIFT algorithm compared to other similar techniques is based on the fact that SIFT is more robust in our case of the detected objects (shape and texture).

T. Kounalakis[1] • G. A. Triantafyllidis[1] (✉)
[1] Applied Informatics and Multimedia Dept.,
Technological Educational Institute of Crete, Heraklion, Greece
Tel.: +30-2810-379189
Fax.: +30-2810-371994
E-mail: gt@teicrete.gr

On the other hand, the system's architecture based on the detected object's SIFT key-points analysis, imposes two basic requirements: First the accurate object detection (and cropping from the whole image) and second, the fact that it requires a specific number of SIFT features to be found in the detected object as well as during the training, since the classification performance is based on these features. But, this requirement is also present in the human visual system: if we cannot see and realize the features of an object, it is difficult to understand what this object is.

After this short introduction the rest of the paper is organized as follows: In the following chapter, we present the methodology of the proposed system, next the experimental results, the conclusions and future work.

2. Methodologies

The proposed system is consisted of three main stages (see Fig. 1): a) disparity map calculation and depth layers pre-processing, b) object detection and SIFT feature extraction, and c) neural network training and use for object recognition. The detailed analysis of each stage is depicted in Fig. 2.

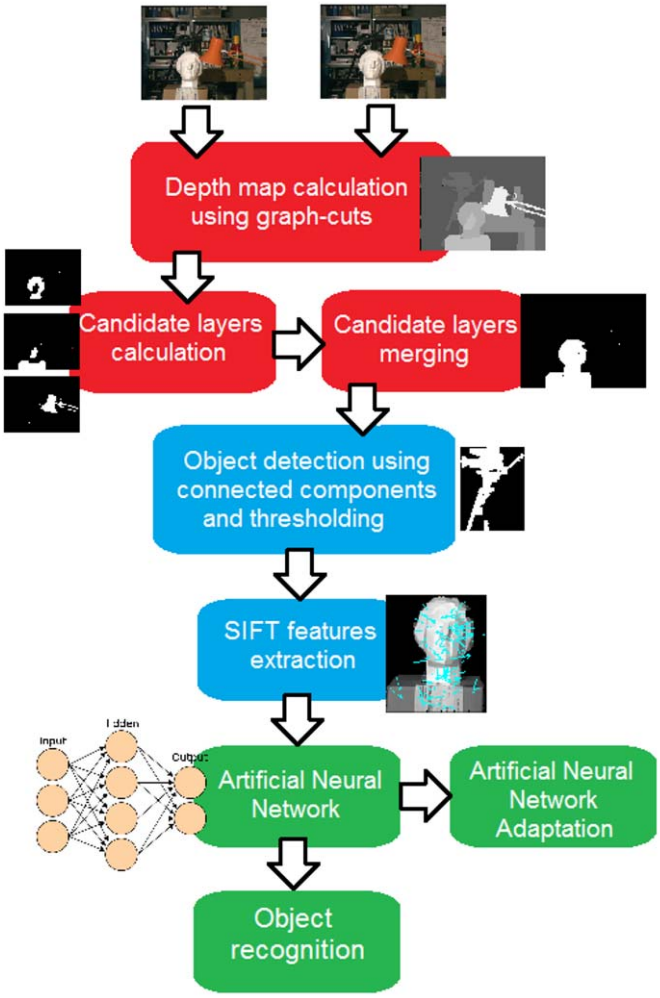

Fig. 1 System overview

2.1 Disparity map extraction and depth layer pre-processing

In a real world, scenery objects are usually found in different positions of depth. So, the system should extract each object from the scenery and process it. In this context, the first stage of the proposed system includes a calculation of visual correspondence in the stereo image pair, resulting in a depth map and then a depth layer by layer preprocessing. This process concludes to the candidate layers which are needed for the next stage of object detection.

A conventional method for depth map calculation provides an outcome with high energy on areas of pixels that do not appear in both images (occluded areas)[5]. To overcome this problem, the MATCH algorithm[6] was employed. This algorithm computes the visual correspondence in a stereo image pair by using graph cuts technique, which transforms the stereo image pair to graphs. Then max-flow min-cut theorem is used to find the cut with the smallest cost, providing a disparity (and depth) map with low energy.

The disparity map is then processed in depth layers which are slices of space according to depth value. Each depth layer L derived from the depth map D, has a number of pixels N_L:

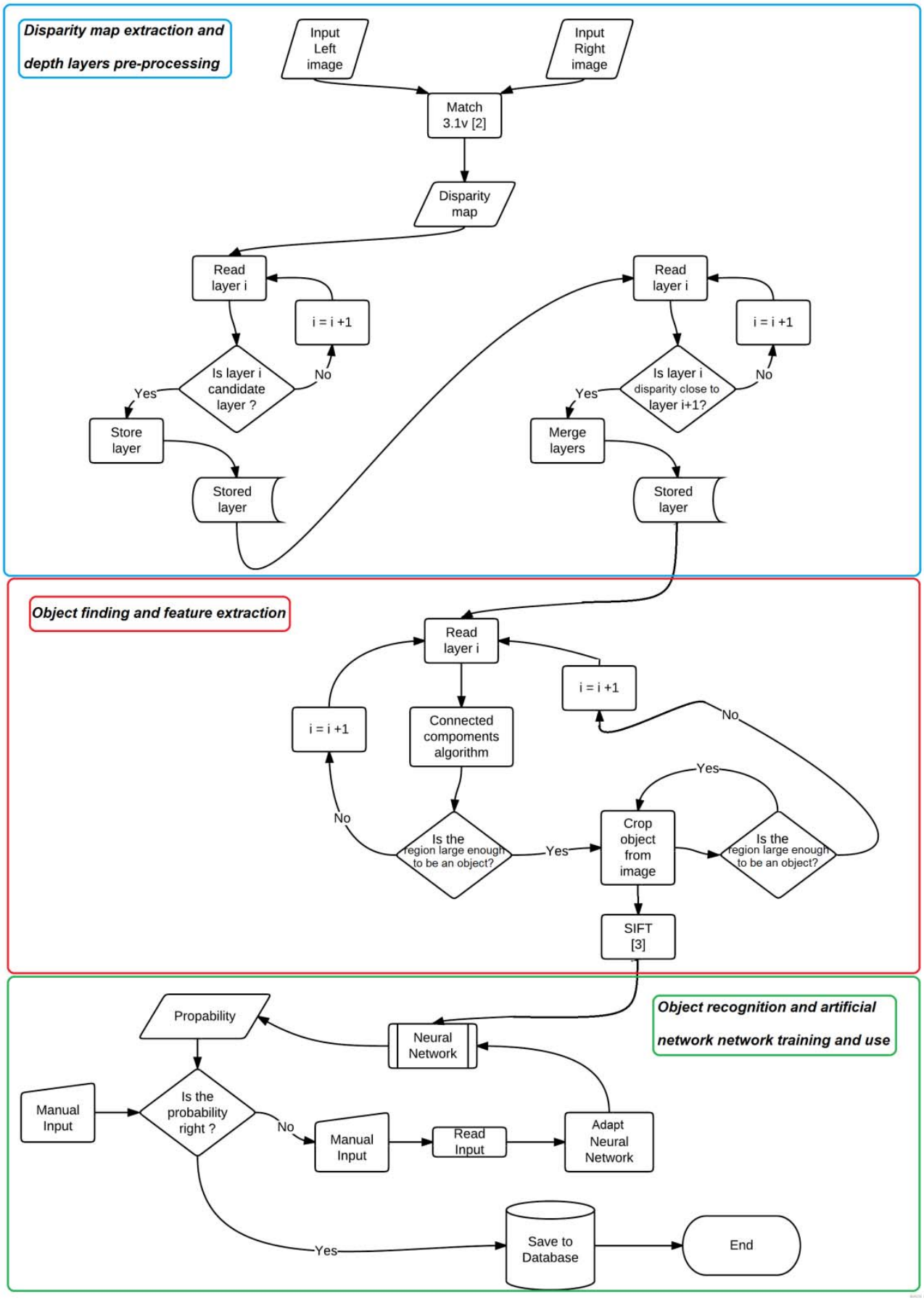

Fig. 2 System details

$$N_L = \sum_i \sum_j p(i,j)$$

where

$$p(i,j) = \begin{cases} 1 & if \quad p(i,j) \in L \\ 0 & if \quad p(i,j) \notin L \end{cases}$$

where i is the width and j is the height of the image frame. By defining a minimum number of pixels threshold T_D (needed for discarding meaningless depth layers), the layers that pass through the process $N_L > T_D$ are called candidate layers (see Fig. 3).

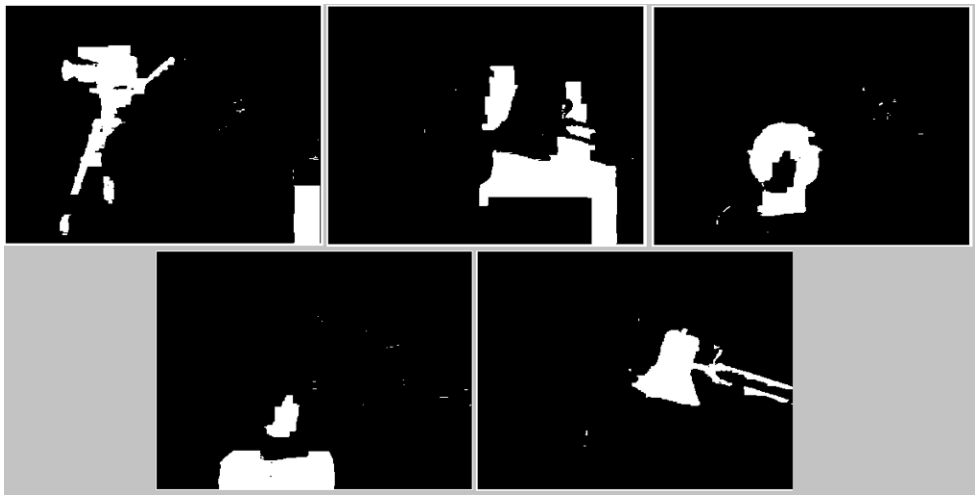

Fig. 3 Extracted candidate layers

2.2 Object detection and feature extraction

In this stage, all object-containing layers are processed, aiming at object detection and extraction of object's SIFT features. First, the algorithm of connected components labeling[7] is used to group pixels into regions for each candidate layer. The outcome of the connected components algorithm are image regions, which are called candidate objects (see Fig. 4).

Fig. 4 Candidate layer merging

The number of pixels N_O composing a candidate object is then defined as:

$$N_O = \sum_i \sum_j p(i,j)$$

where,

$$p(i,j) = \begin{cases} 1 & if \quad p(i,j) \in O \\ 0 & if \quad p(i,j) \notin O \end{cases}$$

where O is the considered candidate object. This information for each candidate object O is used by the proposed system to determine if the candidate objects are real objects by comparing N_O with a pre-defined threshold T. This threshold defines the minimum number of pixels required for a candidate object to be considered as a real object. Obviously, big values of this threshold result in considering only the dominant objects of the scene. If a candidate object is classified as a real object, its texture is then cropped and stored for further processing (see Fig. 5).

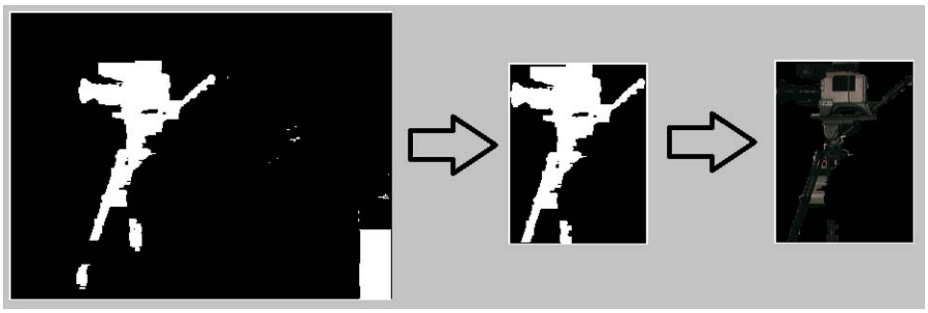

Fig. 5 Object cropping from candidate layer

Each considered object is now finally ready for its features extraction. For this process, the SIFT was chosen to be employed, which is an algorithm that detects and describes features invariant to image scale and rotation (see Fig. 6). SIFT features are highly distinctive and can be considered as key-points.

Since the detected (and cropped from the whole scene) objects containing the object's shape and texture and not the full images are employed for the SIFT calculation, the number of possible missing number of SIFT key-points is minimized. Also, by this cropping, we try to eliminate extraneous or erroneous SIFT key-points, as well as the need of feature hierarchy that is common in some 2D object detectors. This property makes the SIFT features an excellent input for machine learning systems.

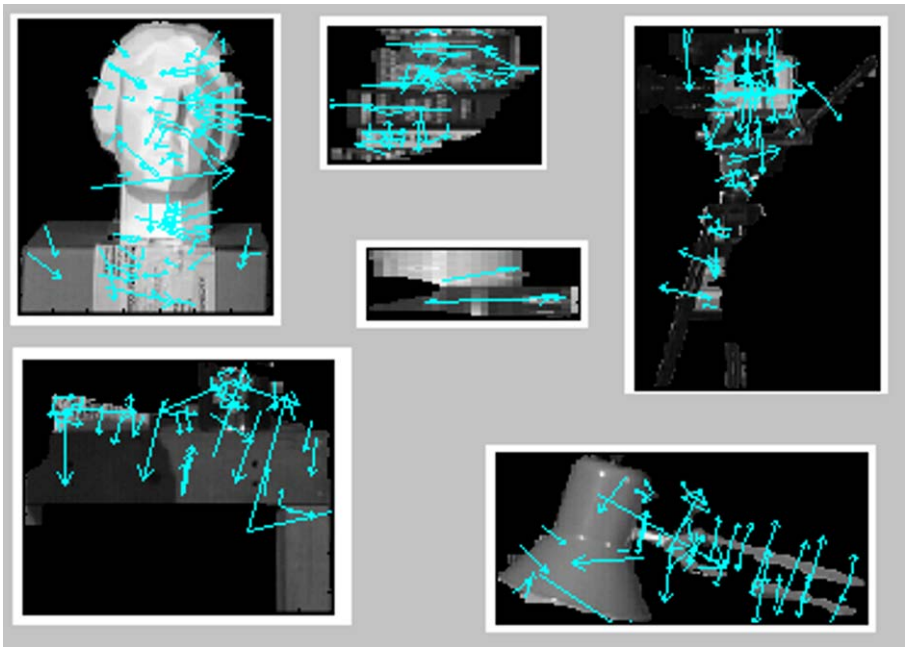

Fig. 6 SIFT key-points extracted from each object.

2.3 Object recognition via machine learning techniques

The third and final stage of the system is separated to two different parts: the machine learning (artificial neural network) and the object recognition. The training process must be executed first because it produces the ANN. A known database of object's SIFT features is used for this purpose, defined as $D:\{C_n,n=1,2,3,..\}$, where C_n are the object categories used for recognition and n is the number of the different object categories. The second part is the use of the ANN, with input the SIFT features of the detected object to be classified. The result of the system is $max(Pr(O \cap C_n)$, which is the maximum probability of classifying the object O with a known object category.

The performance of the proposed method is strongly associated with the number of the SIFT key-points extracted from the detected objects (during the training and the testing). More specifically, an object category having a small number of SIFT key-points for ANN training can result to a low probability of successful recognition of such objects. Similarly, a test object short on SIFT key-points may lead to false classification. Generally, we may state that if we have a database with object categories of sufficient key-points, the system's classification performance could be an absolute success.

But this not the case in the real world, since small objects or very uniform objects indeed provide the system with a small number of SIFT key-points. Also, failures in accurate object detection may result in false classification, due to the irrelevant SIFT key-points that affect the ANN performance. So we introduce a user supervision action over the system by asking the user if the algorithm provided the right outcome. If the user disagrees with the presented classification result, the considered object can be manually re-classified to another group of the database. Moreover, if the category in which the object belongs does not exist the user may even create it. In such way, the new SIFT features will be imported to the database and the ANN will be adapted accordingly, making the future recognitions more accurate.

3. Results

As stated before, the input of the SIFT-based ANN are the detected (and cropped from the whole scene) objects, containing the shape and the texture, and not some full 2D images. These inputs are produced by processing the depth maps derived from the stereopairs. So, we cannot use a simple 2D image library, but instead we need a stereopair library displaying objects which can be classified in some categories. In this context and for the needs of our experiments, we constructed a database of fifteen stereoscopic images (Fig. 7). This database is public available on http://users.teicrete.gr/gt/database.html.The content of these stereopairs are twenty five objects (Fig. 8), which are grouped in five different categories. The categories are: statues, tables, lamps, bottles and teddy bears.

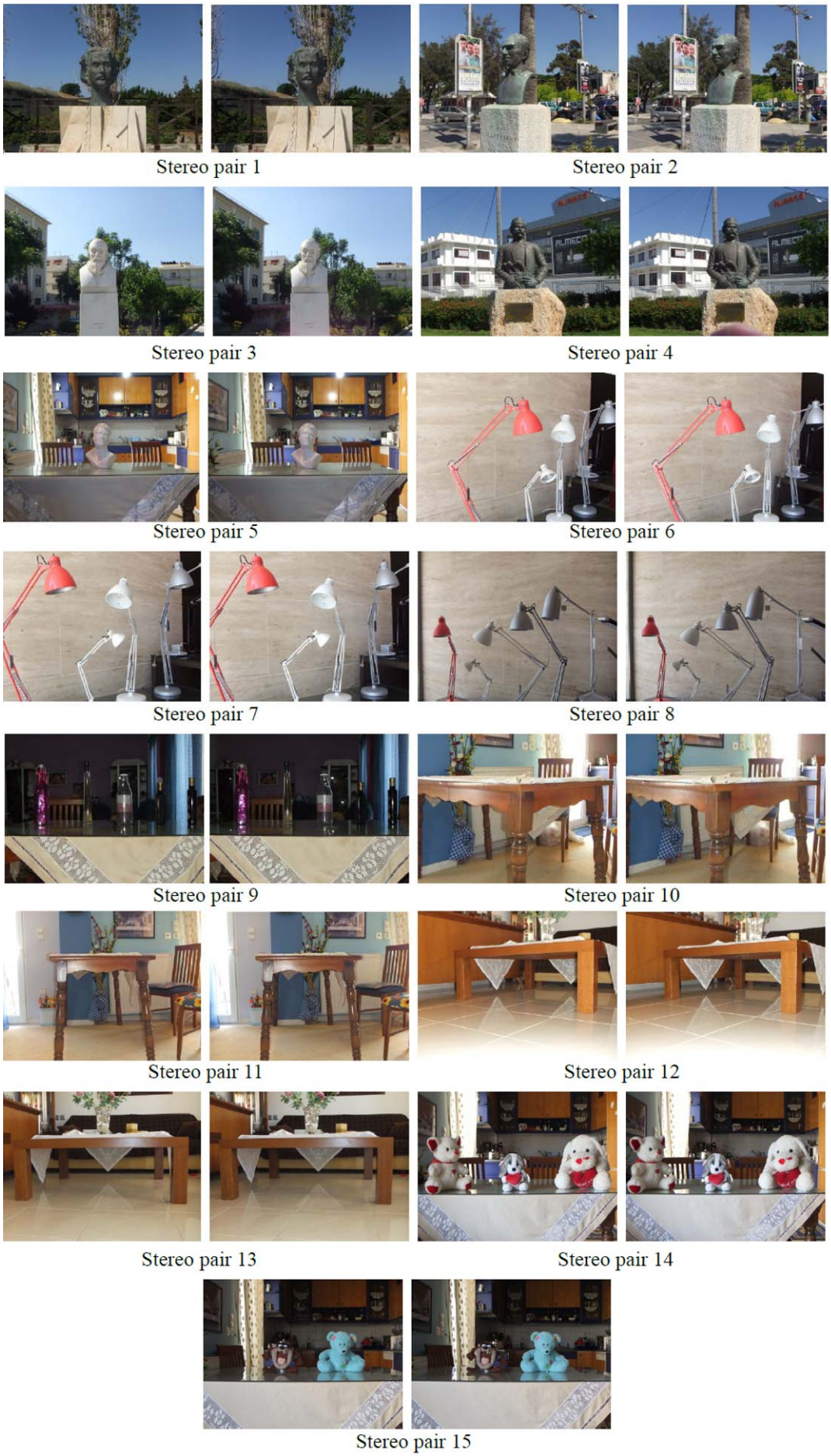

Fig.7 The database of stereoscopic images used for the ANN training

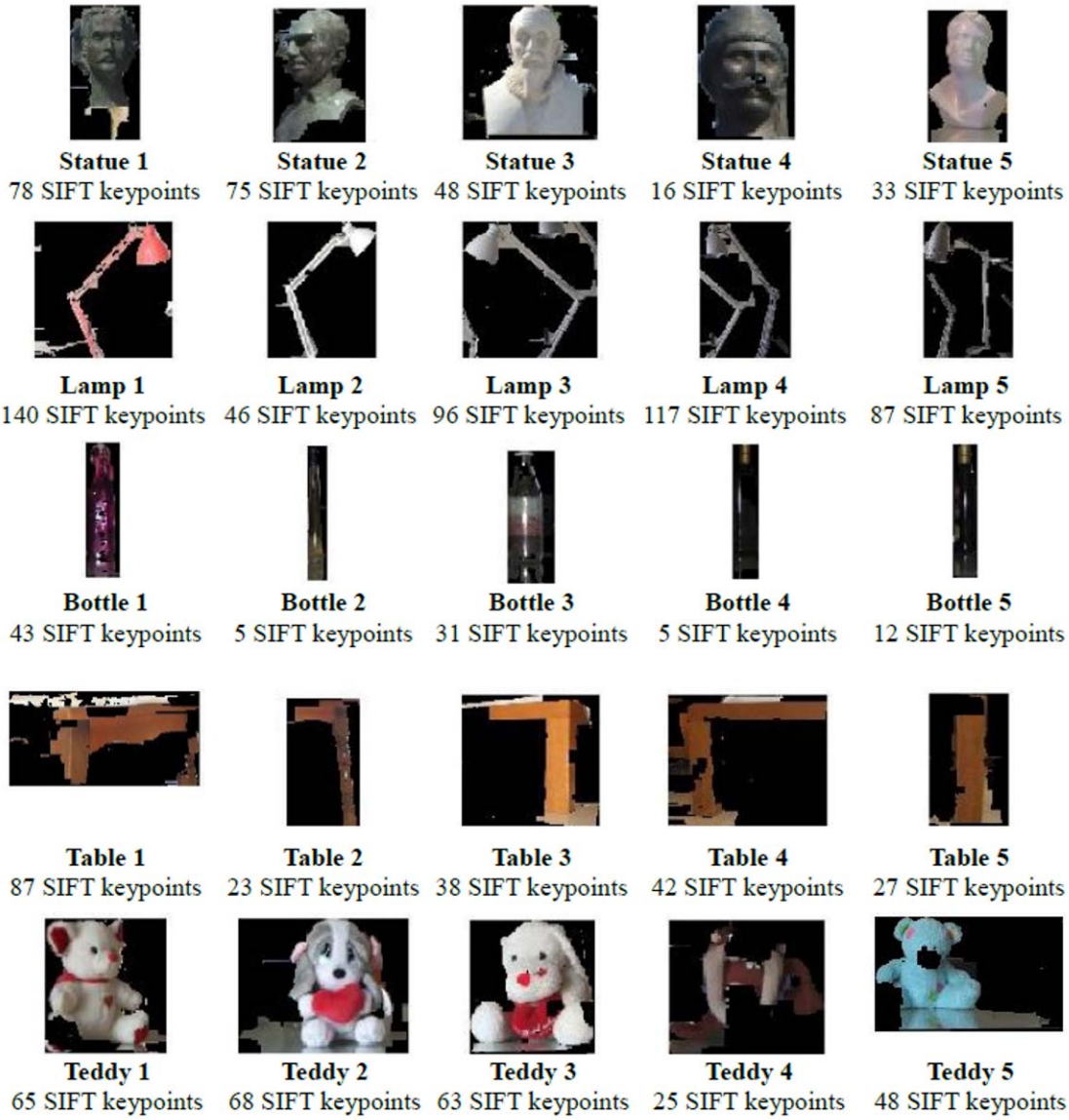

Fig. 8 Extracted objects (shape and texture) from the training database

These objects of the database are used for the ANN training of our system, while for the ANN test we used three new stereopairs with unknown objects and the stereoscopic image pair from University of Tsukuba.

We first tested the nearest neighbor algorithm that is used for the SIFT key-points matching. The object classification using only SIFT matching without the use of any ANN performed very poorly. This was an expected result since this algorithm can match only identical key-points. To overcome this problem, ANN is employed and as a result the proposed scheme succeeds in classifying the SIFT key-points of an object to the SIFT key-points of a similar (and not necessarily identical) object group.

The general ANN architecture in our experiments is a three layered SIFT-trained ANN with tan-sigmoid transfer function in every layer. The size of the input layer was 128, as the size of the SIFT key-points. The size of the second layer was 32 neurons and the output layer had three neurons. The ANN was trained with the Levenberg - Marquardt back-propagation method (see Fig. 9). This ANN architecture was selected since it proved that provides the best possible results.

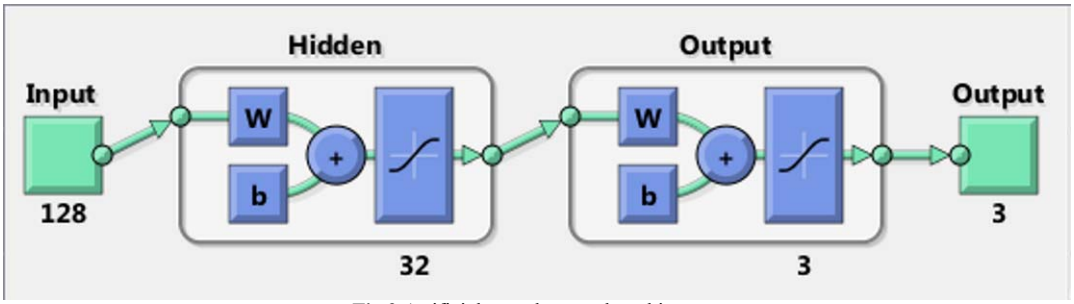

Fig.9 Artificial neural network architecture

Fig. 10 First testing stereo pair

Fig. 11 First testing stereo pair's depth map

Fig. 12 Second testing stereo pair

Fig. 13 Second testing stereo pair's depth map

Fig. 14 Third testing stereo pair

Fig. 15 Third testing stereo pair's depth map

3.1 First testing phase

First, we test the proposed system's performance using three new stereopairs. For testing these stereopairs, we need first to perform the ANN training by employing five object categories as described previously: First, the twenty-five objects of the database's stereopairs are extracted using the depth maps, then the SIFT key-points are calculated for each object and finally the ANN is trained with five object categories.

At this point, it is useful to recall the fact stated before that the performance of the proposed method is depended on the number of the SIFT key-points extracted from the detected objects (during the training and the testing). In this context, the main objective of our first test stereopair (Fig. 10) is the confirmation of our statement that small numbers of SIFT key-points in one category may cause object miss - classification.

So, at the first stereopair there is bottle, which is our test object. Generally, bottles are objects that do not produce high numbers of SIFT key-points due to their uniform texture and shape. Figure 11 shows the depth map from where the texture and the shape of the bottle object can be extracted. Then the bottle's SIFT key-points are calculated. The number of key-points provided is just two. So, the ANN produces the probability of fifty-fifty chance this unknown object to be either a lamp or a statue (see the first row of the Table 1).

In such cases of false object classification due to insufficient SIFT key-points, we introduce the user interaction procedure (by ANN adaptation) to correct the results. This procedure is described in the second testing phase (subsection 3.2).

Table 1

Objects	ANN matched SIFT key-points	Total object SIFT key-points	Matching percentage (%)	Categorized as
ANN results testing three stereo pairs				
Reference bottle object (first stereopair 1)	1	2	50	Statue / Lamp
Reference statue object (first stereopair 2)	3	7	42.86	Statue
Reference statue object (first stereopair 3)	10	17	58.82	Statue

We continue the first testing phase, with a second testing stereopair displaying a statue from a distant shot, as shown in Fig.12, while Fig. 13 shows the depth map. It is easily assumed that this image will also provide little SIFT key-points, because of the small area that the statue possesses in the stereo pairs. But unlike the previous test with the bottle object, this statue has more features in its texture and shape resulting in more SIFT key-points. Indeed, the number of the SIFT key-points that the statue provides is seven.

Moreover, unlike the bottle category, the statue category is a category which produces many SIFT key-points during the training procedure. So, in the experiment, the ANN indeed classifies the statue correctly (see the second row of the Table 1).

The third stage of this testing phase employs a stereopair which shows the same statue but in a closer position to the camera as shown in Fig. 14. The respective depth map is shown in Fig. 15. The SIFT transform provides the system with seventeen key-points (ten more than the seven key-points detected when the same statue was in a more distant place). The ANN classifies the object correctly as a statue (see the third row of the Table 1) presenting a higher matching percentage than the previous classification of the distant statue.

To sum up the first testing phase, the three testing stereopairs (bottle, distant statue and close statue) shows three different cases (see also Table 1):

The first (bottle) has few SIFT key-points as test object and its category also produces few SIFT key-points during the ANN training. This SIFT key-point insufficiency results in false classification. In such cases, user interaction by ANN adaptation may correct the results.

The second (distant statue) has also few SIFT key-points as test object, but its category produces many SIFT key-points in the ANN training, resulting in correct classification.

The third (close statue) has many SIFT key-points as test object and its category produces also many SIFT key-points in the ANN training, resulting in correct classification.

These results prove our theory that the performance of the proposed method is strongly associated with the number of the SIFT key-points extracted from the detected objects (during the training and the testing).

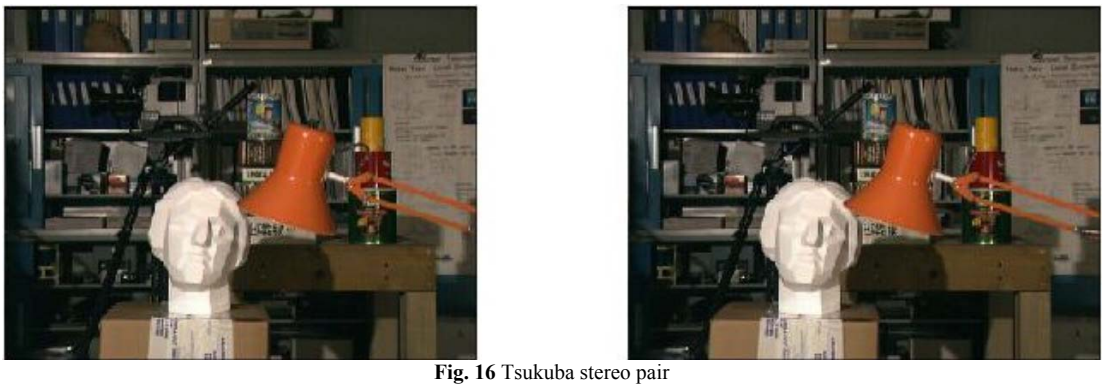

Fig. 16 Tsukuba stereo pair

Fig. 17 Tsukuba stereo pair depth map

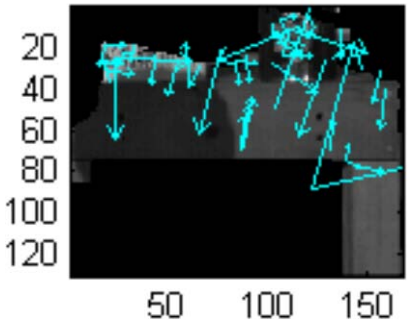

Fig. 18 Tsukuba's table object and its extracted SIFT key-points

3.2 Second testing phase

During the second testing phase, we use the Tsukuba university stereopair (Fig. 16) to test our system's performance. This test aims to examine how the system handles multi – object stereopairs and to prove the performance and the necessity of the user interaction. We start with the typical procedure of the object detection described in the proposed system and six different objects are obtained from the depth map (Fig. 17). These objects are the statue, the lamp, the camera, the table, the books on the table and the cans on the table (see Fig. 6). In order to be compatible with our database categories with which the ANN has been trained, we concentrated on recognizing the statue, the lamp and the table.

The results of the SIFT-based ANN are summarized in Table 2, which shows that the system classifies correctly the statue and the lamp, but is unable to classify correctly the table. The table object has not been recognized correctly because of the poor object detection performed after the depth map processing. The algorithm fails to detect accurately the table and considers as table also some cans and books which are located on the table. This fact results in some extra SIFT key-points that are produced by the other objects which are mistakenly considered as table (see Fig. 18). These extra SIFT key-points are used in the ANN test, but are irrelevant to the SIFT-keypoints that a table object normally produces. Therefore, the ANN is affected and performs a false classification.

Next, we help the ANN correct the erroneous results with user interaction as explained before in Chapter 2.3. The ANN is now adapted to the new user assigned data. That means that the training set of the table category has been updated also with the specific extra SIFT keypoints produced by the specific table cropping. So, after the ANN

adaptation by the user, the results of Table 3 finally provide the desired results, classifying correctly all the objects.

Table 2

Objects	ANN matched SIFT key-points	Total object SIFT key-points	Matching percentage (%)	Categorized as
ANN results testing Tsukuba stereo pair				
Reference statue object	64	88	71.91	Statue
Reference lamp object	22	48	45.83	Lamp
Reference table object	26	54	48.15	Statue

Table 3

Objects	ANN matched SIFT key-points	Total object SIFT key-points	Matching percentage (%)	Categorized as
ANN results testing Tsukuba stereo pair with user interaction				
Reference statue object	64	88	71.91	Statue
Reference lamp object	29	48	60.42	Lamp
Reference table object	33	54	61.11	Table

4. Conclusions and Future Work

It is concluded that SIFT-based ANN can successfully classify unknown objects to object categories by using a simple and efficient framework, in which the cropped objects (containing the shape and the texture) detected by processing the depth maps, are the inputs of the SIFT algorithm in order that SIFT key-points can be detected and feed an ANN which classifies these objects according to its training.

We also proved that the system's performance depends on the number of the SIFT key-points. Therefore, when the system is trained by a large SIFT feature object database, having variety of object categories, the percentage of the correct object recognition is increasing. The system also depends on the object detection result from the depth map processing. Possible such errors are dealt with user interaction which produces ANN adaptation and eventually lead to a fully autonomous system with low errors.

Future work will be routed to advanced database management. Databases which implement queries and relations between object categories can lead to a more sophisticated and interactive system. Also advance database may change knowledge representation with ontologies. Implementations of new algorithms that demand less features and provide better results can increase the performance[8]. Different object detection methods can be also examined. Finally, new and efficient 3D feature extraction algorithms that extract information directly from 3D image, are possible to improve the recognition results.

References

1. D. G. LOWE (2004) Distinctive Image Features from Scale-Invariant Key-points, *International Journal of Computer Vision*, **60**(2): 91-110.
2. H. BAY, A. ESS, T. TUYTELAARS, L. VAN GOOL (2008) SURF: Speeded Up Ro-bust Features, *Computer Vision and Image Understanding (CVIU)*, **110**(3): 346-359.,
3. K. MIKOLAJCZYK, C. SCHMID (2005) A performance evaluation of local descrip-tors, *IEEE Tr. on Pattern Analysis and Machine Intelligence*, **10**(27): 1615-1630.
4. M. S. SARFRAZ, O. HELLWICH (2008) Head Pose Estimation in Face Recognition across Pose Scenarios, *Proceedings of VISAPP 2008, Portugal*, pp. 235-242.
5. G. A. TRIANTAFYLLIDIS, D. TZOVARAS, M. G. STRINTZIS (2000) Occlusion and Visible Background and Foreground areas in Stereo: A Bayesian Approach, *IEEE Trans. on Circuits and Systems for Video Technology*, **10**(4):563-576.
6. V. KOLMOGOROV, R. ZABIH (2001) Computing Visual Correspondence with Oc-clusions using Graph Cuts, *In International Conference on Computer Vision*.
7. R. M. HARALICK, L. G. SHAPIRO (1992) Computer and Robot Vision, Volume I, *Addison- Wesley* pp. 28-48.
8. P. TURCOT, D. G. LOWE (2009) Better matching with fewer features: The selection of useful features in large database recognition problems, *ICCV Workshop on Emergent Issues in Large Amounts of Visual Data (WS-LAVD)*, Kyoto, Japan.