www.ietdl.org

Published in IET Computer Vision
Received on 28th September 2009
Revised on 3rd December 2010
doi: 10.1049/iet-cvi.2009.0094

ISSN 1751-9632

Evaluation of two-part algorithms for objects' depth estimation

R. Kouskouridas A. Gasteratos E. Badekas

Production and Management Engineering, Democritus University of Thrace, Vas. Sofias 12, 67100 Xanthi, Greece
E-mail: rkouskou@pme.duth.gr; gasteratos@ieee.org; badekas@anadelta.com

Abstract: During the last decade, a wealth of research was devoted to building integrated vision systems capable of both recognising objects and providing their spatial information. Object recognition and pose estimation are among the most popular and challenging tasks in computer vision. Towards this end, in this work the authors propose a novel algorithm for objects' depth estimation. Moreover, they comparatively study two common two-part approaches, namely the scale invariant feature transform SIFT and the speeded-up robust features algorithm, in the particular application of location assignment of an object in a scene relatively to the camera, based on the proposed algorithm. Experimental results prove the authors' claim that an accurate estimation of objects' depth in a scene can be obtained by taking into account extracted features' distribution over the target's surface.

1 Introduction

Over the past decade, significant research efforts were devoted to the building of autonomous vision systems, capable of providing location and direction competencies to robots. The procedures of obstacle avoidance or object manipulation can be accomplished by integrating vital visual information derived from pose estimation techniques. Algorithms that were recently proposed utilise either visual sensors [1, 2] or the latter combined with inertial ones [3]. Another open research topic in computer vision is the depth estimation. In [4], the discrete Fourier transform (FT) is used to estimate both local and global transformation of the spatial information. On the other hand, one of the most efficient ways to calculate a scene's depth is the adoption of a disparity estimation method. Stereo vision frameworks invoke correspondences derived from two slightly different images to extract sufficient spatial information. A recent survey of existing disparity estimation methods is presented in [5].

During the past few years, remarkable efforts were made to build new vision frameworks for robust object recognition in cluttered environments. To this end, researchers emphasised in creating recognition schemes based on appearance features with local estate [6, 7]. Algorithms of this field extract features with local extent that are invariant to possible illumination, viewpoint, rotation and scale changes. During the past decade, several techniques that enforce the essential role of local features in demanding pattern recognition tasks were presented [8, 9]. The two main sub-mechanisms of such frameworks are the detectors and descriptors of areas of interest, respectively. The efficiency of the two sub-mechanisms is investigated in [9], where detectors and descriptors are evaluated for object recognition purposes. The two most widely used object recognition frameworks based on local appearance features are the scale invariant feature transform (SIFT) [10] and speeded-up robust features (SURF) [11]. The common issue in both of them is the fact that their detector depicts significant efficiency [9, 12, 13] since it is not affected by possible image alterations. On the other hand, the most important drawback of SIFT's and SURF's descriptor, is the fact that its performance alters significantly under possible rotation, scale, viewpoint and illumination changes. Methods that consist of both a detector and a descriptor are referred in the literature as two-part approaches.

The main idea behind interest location detectors is the pursuit of points or regions in a scene containing unique information. Harris and Stephens [14] were the first to implement an interest point detector, known as Harris Corner detector. Owing to the fact that it provides significant repeatability, many recent studies [15, 16] have adopted it in demanding object recognition tasks. Furthermore, several variations of Harris Corner detector, such as Harris-Laplace [12] and Harris-Affine [13], were presented in an attempt to provide enhanced efficiency. In a following step a descriptor organises the information collected from the detector in a discriminating manner. In [17], a new invariant descriptor called Spin Image that outperforms Gabor filter was presented. The gradient location and orientation histogram (GLOH), which was proposed in [9], produces descriptor histograms that are calculated on a fine circular grid.

Matched features from 2D images are combined in order to produce the 3D model of a pre-recognised object. In [18], a method able to compute camera poses from single query images and to efficiently search for 3*D* models in a city-scale database is presented. It employs viewpoint invariant patches (VIP) that are based on the creation of

ortho-textures for the 3*D* models and on the detection of local features, for example, SIFT or SURF, on them. Time-of-flight cameras can be used for the precise 3*D* environment mapping, as it is shown in [19]. In [20], an image-based visual servo with natural landmarks is presented, as well as, a real-time method for estimating and tracking the 3*D* pose of a rigid object. On the other hand, in [21], pose estimation tasks are fulfilled, by exploiting SIFT in a different manner the contour of the 3*D* model is extracted, while new correspondences obtained from SIFT and the contour are taken into account for the final pose estimation. In addition, local features are utilised in biologically inspired vision systems capable of adequately estimate objects' pose in a scene. In [22], a real-time vision system that integrates a series of algorithms for object recognition, tracking and pose estimation tasks is presented.

In this paper we investigate the construction of a simple and easy-to-build framework for location assignment of an object in a scene. Initially, a database is built and the objects – targets are registered. Images of each object are captured at different distances from the camera and the measured depth d_O is stored. The ultimate goal is to estimate objects' distance from the camera (Z^*) by taking into account the spatial information obtained during training. Thus, considering a given features' distribution over the object's surface corresponding to a known depth, the object's distance from the camera, in cases where the distribution alters, can be computed. The basic assumption underlying the proposed method's motivation derives from the Thales' intercept theorem, as depicted in Fig. 1, where a rather simplistic case of four points on the surface of a plane is illustrated for demonstrative purposes. The further an object is positioned from the camera the denser the distribution of its features becomes, and this relation is linear. The proposed method excels in simplicity, computational cost and execution time. Furthermore, its database can be easily modified for the needs of challenging multi-object recognition tasks and, more important, combinations of detectors and descriptors can be utilised. Our method expands any two-part approach for the needs of objects' depth estimation in a scene. By exploiting vital information derived from SIFT's and SURF's detector and descriptor, we are able to estimate the distance between the camera's frame and the recognised object. In turn, one of the proposed method's drawbacks is the fact that it requires at least two images per object, one obtained during training and one captured through image sequences representing a real scene. Moreover, partial occlusions affect directly the efficiency of the algorithm since the object features for the boundary are lost around some corners. In Section 4 our claim that the distributions' similarity estimation can produce an accurate measurement about objects' depth is experimentally proved.

2 Two-part algorithms

2.1 Sscale Iinvariant Ffeature Ttransform

SIFT's detector starts with the image being convolved with the variable-scale Gaussian function for the production of the scale-space image. Afterwards, the stable key-point locations are detected by using scale-space extrema in the difference-of-Gaussian (DoG) function convolved with the image. After the efficient key-point location assignment by the detector, information around a feature point is exploited by the descriptor. Initially, a consistent orientation to each key point based on local image properties is estimated. For each image sample, the gradient magnitude and the orientation are computed using pixels' intensity values differences. The final descriptor representation is a $4 \times 4 \times 8 = 128$ element feature vector with magnitude and orientation derived from the algebraic sum of the orientation histogram contents for every key point.

2.2 Speeded-up robust features

Interest point detection is performed by adopting the basic Hessian matrix approximation and, thus, by utilising integral images, as proposed in [23]. For the needs of the efficient detection of blob-like structures, SURF's detector is based on the Hessian matrix. A key point is found where the determinant of the Hessian matrix becomes maximum. The construction process of SURF's descriptor is divided into two phases. In the first stage, and with a view to the descriptor's invariance to a possible image rotation, a

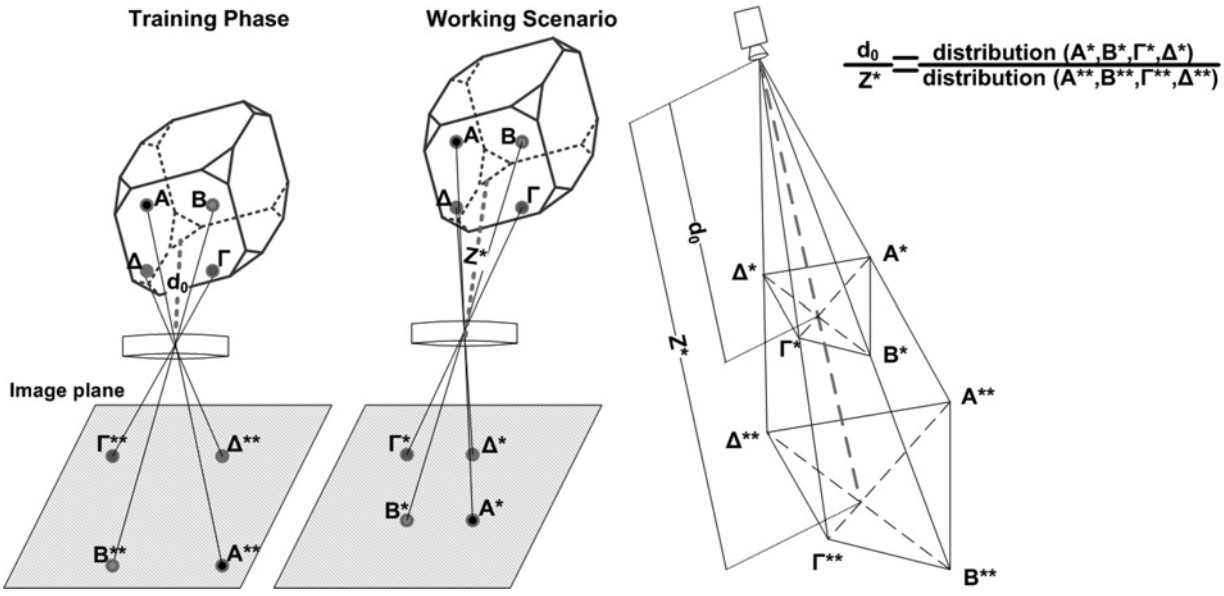

Fig. 1 *Basic assumption of the proposed method*

reproducible orientation of the interest points is estimated. For every interest point an orientation, which is estimated by calculating the sum of all Haar wavelet responses within a sliding window of size ($\pi/3$), is assigned. In the second phase, a 20 s size square region centred around the key point and oriented along the orientation extracted in the previous stage is constructed. This region is split up regularly into smaller 4×4 square sub-regions and for each of them Haar wavelet responses on 5×5 sample points are computed. The final descriptor representation is a feature vector with 64 elements that is extracted by taking into account the vectors of each sub-region.

3 Objects' depth estimation using any two-part algorithm

The proposed objects' depth estimation process is motivated by the idea that the detected features are located on given geometric positions, and thus they can be considered as the corners of a polyhedron, the centre of gravity of which is computed and it is associated to the actual centre of mass of the sought object. Once the features' centre of mass is known and the object is recognised, the distance of the object from the camera is trivial, given at least one recorded position of the object. Practically, the proposed technique tries to estimate the geometrical proportion between an image of a known object (contained in the database) and another one in a scene containing the same object in a different arrangement. By taking into account the distribution of objects' features around their centre in the image space we can estimate spatial information about the object. This could be accomplished by comparing an object's image to that of a scene's containing it and, therefore transform information from image space to the real world.

The main idea underlying our algorithm is the maintenance of any two-part approach's ability in object recognition while making an attempt to further exploit it in order to assign to the recognised object its distance from the camera. To this end, we have constructed a database containing images of several objects. With a view to database's enrichment, these objects were photographed from different viewpoints and distances from the camera. These distances were accurately measured with a laser device and recorded to evaluate the proposed position estimation technique. Each object was captured from four different viewpoints and two different distances from the camera. Moreover, by taking into account SIFT's and SURF's matching sub-procedures we have built an online scene search engine. Estimations derived from this engine are taken into account for the depth of the centre of mass of the object's feature estimation task.

The main stages of the proposed algorithm are as follows:

Stage I: Apply the detector mechanism to the scene's and object's image, in order to estimate the features position in each of them.

Stage II: Obtain the N features that match in the two images by applying the matching sub-procedure of the two-part algorithm. Define as (X_{S_i}, Y_{S_i}), $i = 1, \ldots, N$ the positions of the N features in the scene image and $(X_{\mathrm{o}_i}, Y_{\mathrm{o}_i})$, $i = 1, \ldots, N$ the positions of the N features in the object image.

Stage III: Define as (X_{S_c}, Y_{S_c}) and $(X_{\mathrm{o}_c}, Y_{\mathrm{o}_c})$ the features' centres of mass for both images. This is accomplished by estimating the mean values of the feature positions in the two images

$$X_{S_c} = \frac{1}{N}\sum_{i=1}^{N} X_{S_i} \quad \text{and} \quad Y_{S_c} = \frac{1}{N}\sum_{i=1}^{N} Y_{S_i}$$

$$X_{\mathrm{o}_c} = \frac{1}{N}\sum_{i=1}^{N} X_{\mathrm{o}_i} \quad \text{and} \quad Y_{\mathrm{o}_c} = \frac{1}{N}\sum_{i=1}^{N} Y_{\mathrm{o}_i}$$

Stage IV: Calculate the mean Euclidean distance (in pixels) of each feature from the corresponding centre of mass that is extracted in the previous stage. Set as E_S and E_0 the mean Euclidean distances in the scene and object image, respectively. The following relations are used

$$E_S = \frac{1}{N}\sum_{i=1}^{N} \sqrt{(X_{S_i} - X_{S_c})^2 + (Y_{S_i} - Y_{S_c})^2}$$

$$E_{\mathrm{o}} = \frac{1}{N}\sum_{i=1}^{N} \sqrt{(X_{\mathrm{o}_i} - X_{\mathrm{o}_c})^2 + (Y_{\mathrm{o}_i} - Y_{\mathrm{o}_c})^2}$$

Stage V: Estimate d_S which corresponds to the ratio of the two mean distances E_S and E_O. Furthermore, we introduce the pre-computed depth d_O, which is obtained during the training session and when the object is captured alone

$$d_S = \frac{E\mathrm{o}}{E_S}$$

Stage 1 is apprehended as the training session of our algorithm where the database is constructed. This is a controlled process where each object entering the database is separately captured and its distance from the camera is measured. The background used is a neutral, uniform one, while in cases of more complex scenes we segment the target objects. In this phase, for each image in the database the key-point features are extracted using the detector mechanism. Images of each object are captured at different distances from the camera and the measured depth d_O is stored. This process is performed off-line; thus, execution time is not taken into account. The results are stored for further use at the next phases. In Stage 2, the matching sub-procedure of the two-part algorithm is performed. Especially, descriptors that are common in both images (scene and object) are extracted. It is apparent that, one image representing the scene is compared with several others, representing the object from different viewpoints. Furthermore, the locations of the common features are stored for further use. In Stage 3, the object's depth estimation sub-procedure takes place till the conclusion of the algorithm. Moreover, at this phase the features' centres of mass in both images are calculated. The last ones are obtained by estimating the mean values of feature locations in both representations. In Stage 4, the distance of each key point from the centre of mass is calculated. This is measured in pixels with the use of Euclidean distance. By the end of this sub-routine, we are able to collect spatial information of an object in a scene. This is accomplished by simply estimating the distribution of trained features around their centre of mass. Finally, in Stage 5, the object's

doi: 10.1049/iet-cvi.2009.0094

distance from the camera is computed. The pre-computed depth d_O, measured during the training session of the first phase, is taken into account. The ratio d_S is used to measure the proportion of the object's features to those found in the scene.

Concluding, after the necessary training session and the database construction at the first stages of the method, an online search engine takes over. It is responsible for querying in the scene for objects contained in the trained database. Whenever an object is found, the scene is compared with the image of the object, providing the majority of common matches. Finally, features' information from both images is interpolated with a view to objects' position allocation.

4 Experimental results

The proposed object's depth estimation method is evaluated through extended tests containing several scenes. The tests were executed on a typical PC with a core2duo 2.2 GHz processor, 2 GB RAM and Windows XP operating system. The camera used (Grasshopper by Point Grey Research [24]) is able to capture images up to 1280 × 960 pixels resolution and is connected to the PC via a firewire port. Data transmission is accomplished by using IEEE 1394b transfer protocol. Initially, in Fig. 2 a scene comprising four different objects is shown. We captured five different images representing the same scene under possible viewpoint changes as it is shown in Fig. 2*a*. Furthermore, after each experiment, the images were stored in the database along with each object's distance d_O. The latter were stored in a matrix format as shown in Table 1. Afterwards, we applied the training session (Stage 1) of the proposed method, where each object is photographed separately and under altering viewpoint and illumination conditions. This stage is devoted to the extraction of interest key points using both SIFT and SURF. Thus, we have constructed a large database containing information of objects' identity along with their spatial information needed for the adequate fulfillment of the proposed object's depth estimation method.

Then, we applied the next stage (Stage 2) of our approach where matches between scene's and object's images are obtained by using the matching mechanisms of both SIFT and SURF. Furthermore, we estimated features' centre of mass for both images (scene and comparing object) as it is implemented in Stage 3. Moreover, the mean Euclidean distances and the ratio d_S are calculated by applying Stage 4 and Stage 5, respectively. Finally, each object's distance from the camera is estimated by multiplying the extracted ratio d_S by the corresponding pre-computed depth d_O. The results for the this first scene (shown in Fig. 2) are illustrated in Table 2 where one can see that the accuracy of the proposed object's depth estimation method never falls below 75% while the average stays at 92.8%, with a standard deviation (σ) of 5.81. Furthermore, SIFT outperforms significantly SURF in the object's depth estimation for the scene of Fig. 2 since its performance oscillates at around 96% ($\sigma = 2.36$), while SURFs at 88% ($\sigma = 5.66$).

However, in common environments, rigid objects such as those in Fig. 2, seldom exist. The most likely is that scenes captured in an everyday workspace may contain non-rigid objects – targets. For this reason, we assessed the proposed technique in a second scene comprising four non-rigid objects as illustrated in Fig. 3. Furthermore, each object's distance from the camera (ground truth measurements) is

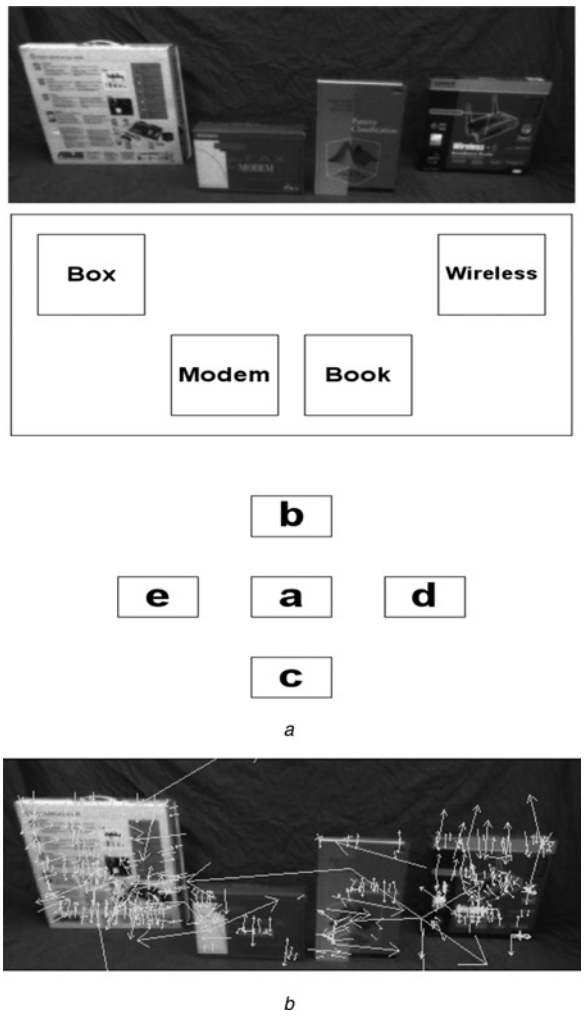

Fig. 2 *Scene that contains 4 different objects is captured from several viewpoints (a,b,c,d,e) 2a and the extracted features 2b*

a Four different objects and the corresponding viewpoints
b Same scene after the feature extraction process

Table 1 Ground truth measurements for the scene shown in Fig. 2

Viewpoints and distances, cm – ground truth					
Object	*a*	*b*	*c*	*d*	*e*
book	100	73	123	100	103
modem	102	75	125	107	101
box	125	100	145	131	118
wireless	122	97	143	117	128

stored in a matrix format and presented in Table 3. By following exactly the same procedure as the first experiment we obtained the results shown in Table 4. Once again SIFT's accuracy remains higher than SURF's (94.31% and $\sigma = 2.87$ to 87.01% and $\sigma = 3.94$), while the overall average accuracy remains at relatively high standards and at approximately 90.66% ($\sigma = 5.01$).

Finally, as occlusions are of the most common problem in computer vision and occluded objects might result to information loss, we tested the robustness of the proposed object's depth estimation technique against them. Towards this concept, we introduced the scene shown in Fig. 4, containing the same non-rigid objects but in a different

Table 2 Experimental results for SIFT's and SURF's performance under different viewpoints 2(a), 2(b), 2(c), 2(d) and 2(e) of Fig. 2's scene

Viewpoints and distances, cm					
Object	d_O	SIFT		SURF	
		Meas	acc, %	Meas	acc, %
book	100	100.4	99.5	89.9	89.99
modem	102	103.9	98.11	88.9	87.19
box	125	118.7	95.02	134.9	98.07
wireless	122	115.4	94.61	112.8	92.54
book	73	72.3	99.12	84.1	84.76
modem	75	81.9	90.77	92.1	77.08
box	100	93.1	93.16	109.2	90.77
wireless	97	96.7	99.72	106.8	89.87
book	123	115.4	93.86	132.1	92.59
modem	125	119.7	95.83	94.6	75.74
box	145	142.6	98.35	138.3	95.40
wireless	143	140	97.91	134.3	93.94
book	100	97.4	97.46	91.2	91.24
modem	107	101.5	94.93	88.2	82.46
box	131	129	98.48	144.1	89.98
wireless	117	113.9	97.38	128.7	89.97
book	103	106.8	96.22	99.9	97.03
modem	101	103.3	97.71	114.5	86.6
box	118	117.8	99.91	132.1	88.05
wireless	128	132	96.86	144.6	87.02%

Meas stands for measured and acc for accuracy

arrangement, so that the objects overlap with each other. The viewpoint angles of the camera remain the same as those of the previous scenes (e.g. Fig. 3*a*). In addition, the ground truth measurements for each object and viewpoint are presented in Table 5. The proposed method's results for this scene are illustrated in Table 6. The overall efficiency of the algorithm is around 89.6% with $\sigma = 5.18$, while SIFT (approx. 93.5% and $\sigma = 2.89$) outperform SURF (approx. 85.85% and $\sigma = 4$) in position assignment of the object in the scene illustrated in Fig. 4.

It is apparent that, the main idea behind the proposed method is the efficient exploitation of features' distribution over an object's surface. As a result, an important aspect arises in cases of partial occlusions an object's parts containing significant amounts of features. In such cases, the proposed method fails to estimate accurately the object's distance from the camera. A simple example of occlusion is shown in Fig. 5: Figs. 5*b* and *c* demonstrate the distribution of SIFT features over the object's surface in a scene with no occlusion and with a partial occlusion, respectively. In this experiment, the accuracy of the proposed method, in case of an occlusion, is around 70%. Moreover, we introduce another scene, which is depicted in Fig. 6, that contains the four previously examined objects. Needless to say that, the algorithm fails in cases where the recognition approach is ineffective. In this experiment, the box and the ball were fully visible, the bunny was partially occluded, whereas the cup is almost invisible by the camera. Box's and ball's depths were very accurately estimated (94.18 and 95.02%, respectively), but the cup was not recognised and bunny's depth, because of the partial occlusion, was mis-calculated (75.41%). We have evaluated our method through several tests comprising

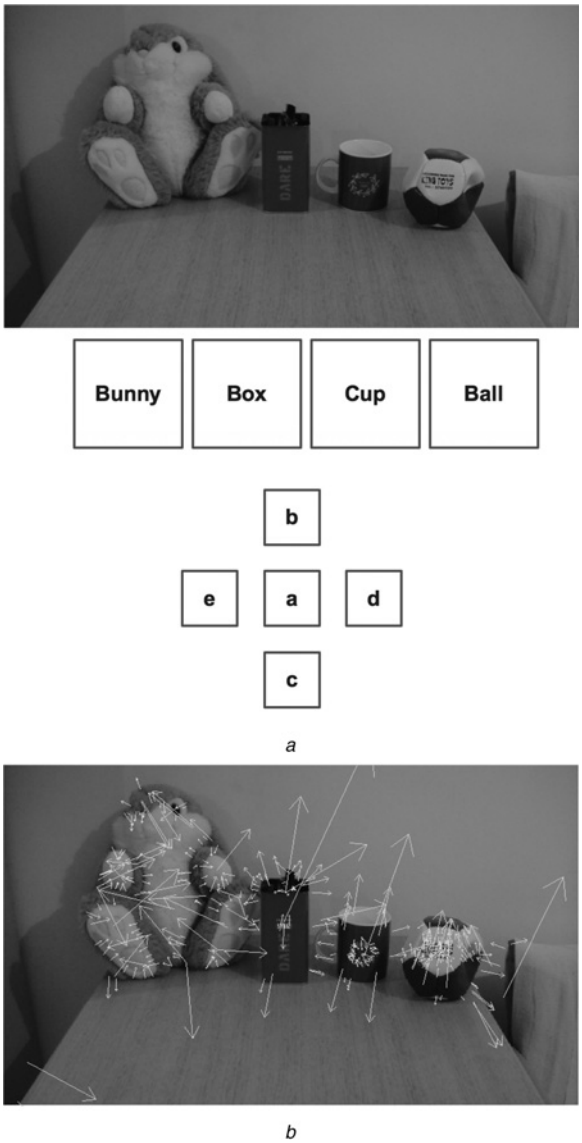

Fig. 3 *Second scene containing four additional objects captured from several viewpoints (a,b,c,d,e) 3a while the extracted features are shown in 3b*

a Four additional objects and the corresponding viewpoints
b Results of the feature extraction process

Table 3 Ground truth measurements for the scene shown in Fig. 3

Viewpoints and distances, cm – ground truth					
Object	*a*	*b*	*c*	*d*	*e*
bunny	115	100	135	120	110
box	112	98	132	115	117
cup	110	98	130	113	120
ball	118	101	131	115	125

a series of partially occluded subjects. Through the series of the experiments, the efficiency of our algorithm has a mean accuracy value 55% with $\sigma = 7.97$. Although this might drastically affect a demanding manipulation task, the whole problem can be simplified by utilising more than one camera in order to assess more integrated data. Multi-camera systems exhibit the advantage to minimise the

doi: 10.1049/iet-cvi.2009.0094

Table 4 Experimental results for SIFT's and SURF's performance under different viewpoints 4(a), 4(b), 4(c), 4(d) and 4(e) of Fig. 3's scene

Viewpoints and distances, cm					
Object	d_O	SIFT		SURF	
		meas	acc, %	meas	acc, %
bunny	115	111.3	96.85	105.9	92.11
box	112	108.3	96.71	121.3	91.68
cup	110	114.9	95.53	125.0	86.33
ball	118	122.3	96.35	105.1	89.07
bunny	100	95.2	95.23	111.7	88.28
box	98	101.6	96.24	82.1	83.78
cup	98	106.2	91.55	109.9	87.84
ball	101	108.1	92.94	118.9	82.25
bunny	135	121.1	89.71	122.0	90.40
box	132	140.2	93.72	119.9	90.87
cup	130	127.0	97.71	115.1	88.61
ball	131	130.1	99.38	155.0	81.67
bunny	120	111.3	92.78	104.1	86.75
box	115	99.1	86.2	126.8	89.72
cup	113	118.2	95.39	130.7	84.31
ball	115	122.1	93.82	96.6	84.06
bunny	110	104.1	94.68	127.2	84.36
box	117	109.5	93.66	122.0	95.69
cup	120	128.9	92.56	142.5	81.18
ball	125	130.9	95.27	148.2	81.44

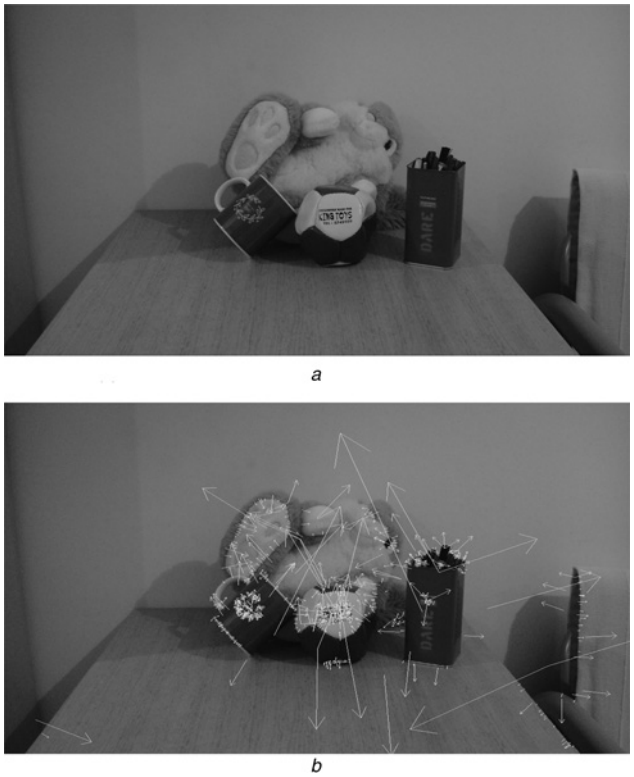

Fig. 4 *Scene that contains the same fours objects in different geometrical constraints that produce occlusions*

a Alternative placement of the four objects introduced in 3
b Extracted features

Table 5 Ground truth measurements for the scene shown in Fig. 4

Viewpoints and distances, cm – ground truth					
Object	*a*	*b*	*c*	*d*	*e*
bunny	112	92	137	118	108
box	119	100	125	113	125
cup	102	84	126	110	106
ball	118	96	122	128	116

Table 6 Experimental results for SIFT's and SURF's performance under different viewpoints 6(a), 6(b), 6(c), 6(d) and 6(e) of Fig. 4's scene

Viewpoints and distances, cm					
Object	d_O	SIFT		SURF	
		meas	acc, %	meas	acc, %
bunny	112	120.5	92.40	96	85.73
box	119	113.1	95.12	108.1	90.86
cup	102	114.1	88.10	114.5	87.73
ball	118	120.1	98.20	112.4	95.26
bunny	92	101.1	90.01	78.1	84.93
box	100	104.9	95.00	111	88.99
cup	84	82.3	97.99	100.1	80.75
ball	96	102.9	92.78	114.1	81.14
bunny	137	130.1	95.00	124.1	90.61
box	125	132.1	94.27	109.1	87.31
cup	126	119.2	94.64	112.4	89.22
ball	122	133.9	90.24	107.1	87.79
bunny	118	122.4	96.19	102.6	87.02
box	113	121.4	92.56	129	85.80
cup	110	101.2	92.04	131.6	80.31
ball	128	138.2	91.96	104.1	81.37
bunny	108	116.7	91.87	129	80.50
box	125	119.9	95.94	144.2	84.64
cup	106	117.2	89.40	120.7	86.05
ball	116	105.1	90.61	137.9	81.08

possibility of partially occlusions of objects in a scene. Yet, in single-camera systems, a solution would be to detect whether the distribution of the features is uniform around the centre of mass and decide on the reliability of the measurement.

We conclude by presenting comparative results of the proposed integrated object recognition and depth estimation algorithm with a commercially available framework: The Visual Pattern Recognition system (ViPR version 3.0), manufactured by Evolution Robotics [25], which offers adequate database construction processes that enables its usage in demanding pattern recognition tasks. In Table 7, the attributes of both the proposed method and the ViPR are presented. ViPR is a framework for real-time object detection and tracking, yet it is also prepared with a depth estimation routine. As it can be seen, the common issue in both of them constitutes the fact that they are able to recognise multiple objects per frame, which in turn could be apprehended as an important advantage in modern pattern recognition frameworks. We have exhaustively tested both the proposed method's and ViPR's efficiency throughout several experimental sets comprising the objects belonging in the database that is available at [26]. In Table 8, the accumulated

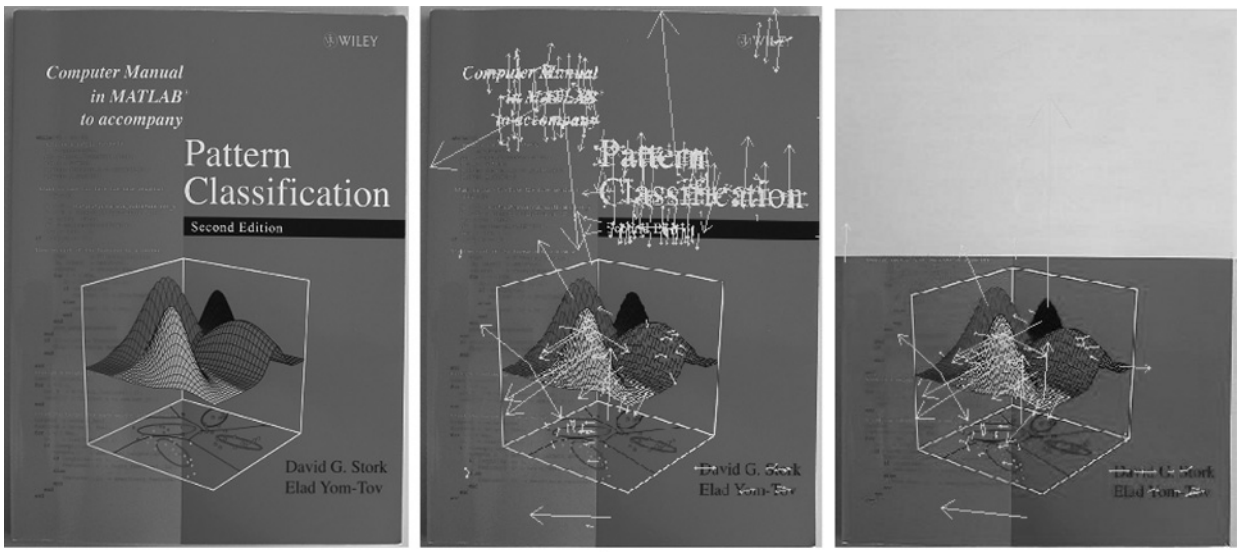

Fig. 5 *Features' distribution over the surface of an non-occluded 5b and a partially occluded 5c object 5a*

a A book
b Extracted SIFT features of the non-occluded book
c Extracted SIFT features of the partially occluded book

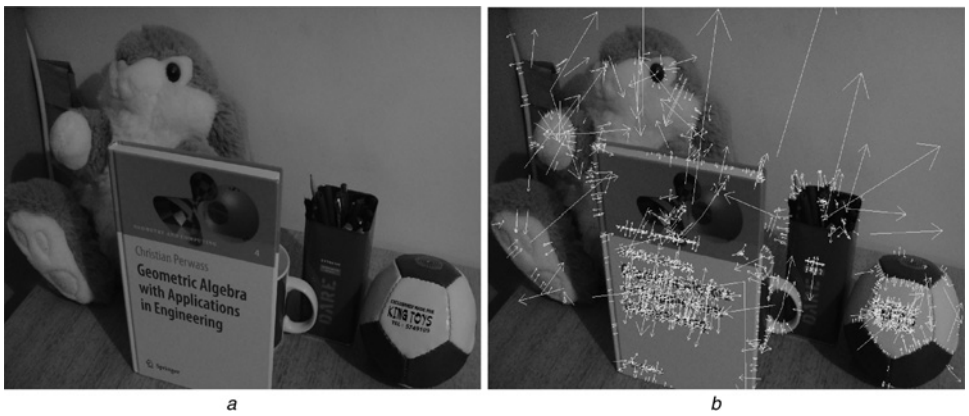

Fig. 6 *System in cases of total objects' occlusions (e.g. cup) fails to estimate their depth*

a A scene containing the four previously introduced objects
b Extracted SIFT features of the non-occluded objects and those that are partially occluded
In instances of partial occlusion (e.g. bunny) the proposed method calculates targets' distance from the camera with reduced efficacy

Table 7 Comparison of the proposed method and ViPR

	Proposed method	ViPR
cameras used	firewire, USB	mainly USB
maximum resolution	1280 × 960	320 × 240
objects' depth estimation error	≤15%	≤30%
execution time	2 s/object	real time

comparison results for the 70 different objects are illustrated. The proposed method, which adopts either SIFT or SURF, outperforms ViPR in objects' position estimation tasks. With more details, SIFT was once again proven to be the most effective solution providing 90.11% mean efficiency, when SURF's and ViPR's remained around 87.04 and 80.21%, respectively. Generally, the advantages of our method is its ability to utilise both firewire and USB (Universal Serial Bus) cameras with resolution up to 1280 × 960 pixels and overall depth estimation error below 15%. On the contrary, ViPR that excels in object recognition and tracking processes mainly uses USB cameras with 320 × 240 maximum resolution, while working real time with overall estimation error below 30%.

To sum up, we have proposed a novel technique for the estimation of the distance between the camera and the centre of mass of the object's features. Its main advantages are its simplicity, the limited computational cost and the execution time. Furthermore, its database can be easily modified for the purposes of multi-object recognition. On the other hand, the fact that it requires a pre-constructed database containing spatial information of the trained objects, is its main drawback. As a result, it cannot be utilised in operations that entail online object recognition. In turn, the proposed method could be adopted in applications where the need of targets' position assessment is a prerequisite. As a result, demanding pick and place tasks, such as moving plants in a green house, could be adequately accomplished. Furthermore, after the construction of an adequate database and considering a manipulator, several tasks including the clearing of a

doi: 10.1049/iet-cvi.2009.0094

Table 8 Evaluation of our method using (i) SIFT and (ii) SURF against ViPR in distance from camera estimation tasks for the objects contained in the database available at [26]

Object #	Ground truth	SIFT		SURF		ViPR	
		Mean error, cm	Std error	Mean error, cm	Std error	Mean error, cm	Std error
1	85	4.14	7.09	11.48	0.99	20.25	12.78
2	80	2.74	9.99	8.45	3.01	17.55	9.03
3	83	9.48	0.16	6.52	1.43	27.76	16.71
4	78	8.84	0.03	11.56	4.53	14.15	0.99
5	71	11.12	1.75	12.24	5.25	22.56	16.77
6	73	10.82	1.77	12.81	1.93	20.78	14.37
7	81	10.11	3.08	11.76	1.54	19.85	7.38
8	85	9.21	2.47	10.85	3.07	15.81	8.81
9	60	4.66	0.26	10.18	1.05	22.14	7.89
10	58	7.06	5.31	10.17	2.49	8.17	20.35
11	57	9.56	2.31	10.98	3.9	7.69	23.14
12	61	9.27	1.45	10.46	0.61	16.58	9.75
13	58	8.62	0.71	11.48	3.65	15.47	8.51
14	105	7.76	1.38	7.11	1.04	16.52	2.16
15	95	7.07	1	9.47	2.08	23.94	16.24
16	100	11.61	2.29	12.51	5.42	15.31	11.23
17	99	12.48	0.08	12.52	2.36	19.05	5.92
18	70	10.18	2.12	14.91	5.33	17.33	6.58
19	68	4.15	6.73	7.89	1.18	3.33	11.67
20	72	7.18	1.04	9.81	0.24	11.86	7.09
21	102	7.11	4.7	9.34	0.55	13.93	4.27
22	99	9.45	0.57	10.12	2.74	7.75	2.44
23	101	4.83	3.05	11.53	0.32	19.85	3.86
24	98	5.57	2.29	6.58	1.88	5.52	1.52
25	110	2.41	8.42	9.5	0.45	10.89	6.02
26	108	11.54	1.33	11.52	3.59	23.89	11.01
27	114	7.36	3.97	6.82	6.15	19.06	5.08
28	80	3.01	5.01	3.39	3.58	13.11	4.68
29	78	10.44	2.67	13.48	5.4	5.9	3.5
30	92	4.57	8.87	6.25	6.04	14.41	10.14
31	99	10.24	2.67	10.85	2.65	19.53	12.35
32	55	7.21	0.58	11.58	3.82	14.87	6.28
33	66	6.76	3.28	7.83	3.91	10.02	0.7
34	68	9.31	0.9	11.48	1.11	9.44	0.81
35	57	6.59	3.62	8.66	0.6	12.62	1.86
36	64	7.28	2.57	9.23	2.67	12.23	6.65
37	48	7.17	0.26	9.62	1.86	14.33	8.46
38	68	9.85	3.25	11.71	1.11	12.89	3.32
39	86	8.31	3.66	11.26	1.61	16.83	3.43
40	79	12.53	4.89	16.34	5.23	15.52	10.08
41	92	9.2	0.86	7.83	0.82	13.11	8.08
42	96	13.49	11.83	15.18	3.78	27.07	16.21
43	68	6.03	6.83	7.59	3.19	19.03	9.73
44	82	9.08	4.88	13.47	4.66	20.92	14.3
45	76	6.17	0.64	8.69	0.8	23.87	13.3
46	63	11.28	2.54	7.83	2.81	18.34	8.38
47	124	8.04	0.02	14.25	3.92	25.58	19.11
48	129	5.59	7.37	7.86	2.99	19.53	5
49	134	5.66	1.19	11.51	1.81	19.85	4.04
50	119	6.73	2.95	7.89	2.85	17.99	3.45
51	115	7.41	0.52	9.88	0.63	23.68	10.47
52	121	9.57	3.61	7.66	4.67	19.74	3.89
53	111	8.06	4.39	13.01	0.53	24.34	17.34
54	108	8.65	0.03	11.47	2.71	23.43	19.02
55	116	7.53	4.52	9.67	0.91	25.43	19.55
56	68	9.37	5.31	13.49	0.42	12.17	12.64
57	62	8.06	5.14	12.53	0.14	16.78	8.75
58	59	3.83	3.87	7.48	2.44	11.89	2.87
59	55	5.19	9.01	10.18	9.76	13.44	1.48

Continued

Table 8 *Continued*

Object #	Ground truth	SIFT		SURF		ViPR	
		Mean error, cm	Std error	Mean error, cm	Std error	Mean error, cm	Std error
60	70	2.17	4.25	4.52	4.03	17.83	6.89
61	85	5.88	2.11	9.78	1.34	18.83	11.83
62	82	5.78	6.77	4.52	5.6	2.83	9.25
63	82	6.34	3.19	8.06	4.61	11.91	4.17
64	78	8.49	2	7.94	0.68	13.83	0.78
65	76	9.17	0.78	10.67	0.82	8.08	5.25
66	91	9.33	0.97	11.54	0.55	10.82	4.69
67	86	11.82	3.9	11.33	3.73	15.28	11.14
68	78	9.51	2.11	11.99	1.9	15.83	9.34
69	81	8.49	4.28	18.51	6.45	18.08	11.35
70	68	8.93	2.38	13.34	4.25	16.49	11.97

conference centre or the tidying of a child's bedroom after a party etc., could be achieved by adopting the proposed method.

5 Discussion

In this paper we proposed a novel computer vision technique for objects' depth estimation suitable for adoption to any two-part algorithm. It is based on the observation that the features extracted from any two-part algorithm correspond to spots on the object's surface and their centre of mass is related to the one of the objects. Thus, by extracting these features at known positions of the sought object, one can estimate its distance form the camera. The proposed technique was tested on the two most common two-part algorithms, namely the SIFT and the SURF, and was found to outperform with the first one. However, it is easily adopted for any other two-part algorithm that can be found in the literature. Moreover, it is compositionally inexpressive and its training part is done off-line. Furthermore, its efficiency depends on the distribution of object's features over its surface and as a result, the algorithm fails to estimate object's distance from the camera in case all the object's features detected are located on an occluded part of it. Target applications includes robotics or any other one, where apart from the object recognition, an estimation of the position of the recognised object is essential.

6 Acknowledgments

This work is supported by the E.C. under the FP6 research project for Autonomous Collaborative Robots to Swing and Work in Everyday EnviRonment ACROBOTER, FP6-IST-2006-045530. http://www.acroboter-project.org

7 References

1 Schweighofer, G.: 'Robust pose estimation from a planar target', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (12), pp. 2024–2030
2 Chandraker, M.K., Stock, C., Pinz, A.: 'Real-time camera pose in a room', *Lect. Notes Comput. Sci.*, 2003, **2626**, pp. 98–110
3 D, Xu, Li, Y.F.: 'A new pose estimation method based on inertial and visual sensors for autonomous robots'. IEEE Int. Conf. on Robot. Biomimetics. Sanya, China, December 2007, pp. 405–410
4 Torralba, A., Oliva, A.: 'Depth estimation from image structure', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, pp. 1226–1238
5 Naplapntidis, L., Sirakoulis, G.C., Gasteratos, A.: 'Review of stereo vision algorithms: from software to hardware', *Int. J. Optomechatronics*, 2008, **2**, (4), pp. 435–462
6 Nister, D., Stewenius, H.: 'Scalable recognition with a vocabulary tree'. Proc. IEEE Comput. Society Conf. on Computer Vision and Pattern Recognition, New York, USA, June 2006, pp. 2161–2168
7 Sivic, J., Zisserman, A.: 'Video google: a text retrieval approach to object matching in videos'. Proc. Ninth IEEE Int. Conf. on Computer Vision, Nice, France, October 2003, pp. 1470–1477
8 Liao, M.Z.W., Ling, W., Chen, W.F.: 'A novel affine invariant feature extraction for optical recognition'. Int. Conf. on Mach. Learn. Cybern., Hong Kong, China, August 2007, vol. 3, pp. 1769–1773
9 Mikolajczyk, K., Schmid, C.: 'A performance evaluation of local descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630
10 Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis..*, 2004, **60**, (2), pp. 91–110
11 Bay, H., Tuytelaars, T., Van Gool, L.: 'Surf: speeded up robust features'. Proc. Ninth Eur. Conf. on Comput. Vis., Graz, Austria, May 2006, vol. 3951, p. 404
12 Mikolajczyk, K., Schmid, C.: 'Scale & affine invariant interest point detectors', *Int. J. Comput. Vis..*, 2004, **60**, (1), pp. 63–86
13 Mikolajczyk, K., Tuytelaars, T., Schmid, C., *et al.*: 'A comparison of affine region detectors', *Int. J. Comput. Vis.*, 2005, **65**, (1–2), pp. 43–72
14 Harris, C., Stephens, M.: 'A combined corner and edge detection'. Proc. Fourth Alvey Vis. Conf., Manchester, UK, August 1988, pp. 147–151
15 Rothganger, F., Lazebnik, S., C., S., Ponce, J.: '3D object modeling and recognition from photographs and image sequences', *Toward Category-Level Object Recognition*, 2006, **4170**, pp. 105–126
16 Schmid, C., Mohr, R.: 'Local grayvalue invariants for image retrieval', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (5), pp. 530–535
17 Lazebnik, S.C.S., Ponce, J.: 'A sparse texture representation using local affine regions', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (8), pp. 1265–1278
18 Wu, C., Fraundorfer, F., Frahm, J.M., Pollefeys, M.: '3D model search and pose estimation from single images using VIP features'. IEEE Comput. Soc. Conf. on Comput. Vis. Pattern Recognition Workshops, Anchorage, Alaska, June 2008, pp. 1–8
19 May, S., Droeschel, D., Holz, D., *et al.*: '3D pose estimation and mapping with time-of-flight cameras'. IEEE Int. Conf. on Intelligent Robots and Systems, Workshop in 3D-Mapping, Nice, France, 2008, pp. 1–8
20 Choi, C., Baek, S.M., Lee, S.: 'Real-time 3D object pose estimation and tracking for natural landmark based visual servo'. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Nice, France, 2008, pp. 3983–3989
21 Gall, R.B.J., Seidel, H.P.: 'Robust pose estimation with 3d textured models', *Lect. Notes Comput. Sci..*, 2006, **4319**, p. 84
22 Bjiörkman, M., Kragic, D.: 'Combination of foveal and peripheral vision for object recognition and pose estimation'. IEEE Int. Conf. on Robotics and Automation, April 2004, pp. 5135–5140
23 Viola, P., Jones, M.: 'Rapid object detection using a boosted cascade of simple features'. Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, December 2001, vol. 1, pp. 511–518
24 http://www.ptgrey.com/products/grasshopper/index.asp
25 http://www.evolution.com/core/ViPR/
26 http://robotics.pme.duth.gr/rigas/Objects.rar

doi: 10.1049/iet-cvi.2009.0094