Robotics and Autonomous Systems 58 (2010) 276–286

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Towards a unified visual framework in a binocular active robot vision system

Gerardo Aragon-Camarasa [a,*], Haitham Fattah [b], J. Paul Siebert [a,1]

[a] Computer Vision and Graphics Group, Department of Computing Science, University of Glasgow, 17 Lilybank Gardens, Glasgow G12 8QQ. Scotland, UK
[b] Institute for System Level Integration, The Alba Centre, Alba Campus, EH54 7EG, UK

ARTICLE INFO

Article history:
Available online 28 November 2009

Keywords:
Robot head
Gaze control
Active vision
Object recognition

ABSTRACT

This paper presents the results of an investigation and pilot study into an active binocular vision system that combines binocular vergence, object recognition and attention control in a unified framework. The prototype developed is capable of identifying, targeting, verging on and recognising objects in a cluttered scene without the need for calibration or other knowledge of the camera geometry. This is achieved by implementing all image analysis in a symbolic space without creating explicit pixel-space maps. The system structure is based on the 'searchlight metaphor' of biological systems. We present results of an investigation that yield a maximum vergence error of ~6.5 pixels, while ~85% of known objects were recognised in five different cluttered scenes. Finally a 'stepping-stone' visual search strategy was demonstrated, taking a total of 40 saccades to find two known objects in the workspace, neither of which appeared simultaneously within the field of view resulting from any individual saccade.

1. Introduction

The recent maturation of digital imaging hardware and the continual advancement of image processing and analysis techniques have vastly improved the potential for the application of computer vision in real-world robotics systems. Furthermore, *binocular* robotic vision has an advantage over *monocular* vision in potentially being able to compute *range maps* (i.e. distance fields to visible surfaces) by decoding the local parallaxes between captured stereo-pairs. Binocular imaging can also be used in object recognition to provide more information and therefore generate stronger object presence/identity hypotheses than would be possible with monocular vision alone. The development of an active vision control mechanism for a binocular camera system featuring object recognition and automated visual field exploration has potential applications in autonomous roving vehicles, automatic surveillance systems and military or clinical telepresence.

In this paper we present a system that integrates visual attention, vergence, gaze control and object recognition based on point matches extracted by means of the Scale Invariant Feature Transform [1] (SIFT). The system as devised provides an efficient means for controlling an active binocular robot head by integrating low-level and high-level visual components in a uncomplicated and unified framework. Our vision system performs the key robotics task of detecting, classifying and locating known 3D objects that may be partially occluded, within a cluttered scene comprising both known and unknown objects. The essential structure of the binocular active robot vision system we have developed is sufficiently general to allow it to be readily adapted within different robotics contexts.

This paper is organised as follows. In Section 2, we describe related work and the motivation that led us to design this particular system. We then describe the design of the vergence, object recognition and gaze control systems, in Sections 3–5, respectively. Finally, Section 6 contains a summary of the system validation, its results and contributions to the field of active vision research.

2. Related work and motivation

Several binocular robot heads have been developed in recent decades. For example, the "Richard the First" head [2] and the KTH robot head [3] were capable of mimicking human head motion. More recent robot heads include the LIRA head [4], where acoustic and visual stimuli are exploited to drive the head gaze; the Yorick head [5] or the Medusa head [6] where high-accuracy calibration, gaze control, control of vergence or real-time tracking with log-polar images were successfully demonstrated.

Despite advances in binocular robot heads, few systems are reported in the literature that integrate vergence and object recognition into a complete system capable of autonomously exploring a cluttered visual field. Therefore, our motivation is to investigate the potential for state-of-the-art image processing techniques to allow binocular robotic vision systems to operate in unstructured environments.

* Corresponding author. Tel.: +44 0 141 330 1621; fax: +44 0 141 330 3119.
E-mail addresses: gerardo@dcs.gla.ac.uk (G. Aragon-Camarasa), haitham.fattah@gmail.com (H. Fattah), psiebert@dcs.gla.ac.uk (J. Paul Siebert).
[1] Tel.: +44 0 141 330 3124.

0921-8890/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2009.08.005

Vergence, in a biological context, is the act of adjusting the relative angles of a pair of eyes to centre a real-world region of interest in the fovea of both eyes such that the dynamic range of parallaxes induced is minimised. In turn, this process maximises the visual information that can be extracted and perceived by the observer. There are many different possible models for implementing vergence in the context of a robotic binocular system: for example, by means of saliency detection or stereo-matching techniques such as cepstral filtering [7], area-based matching [5] and feature-based matching [8].

In this work, feature-based matching offers advantages over area-based techniques, such as in [9], at depth discontinuities when imaged surfaces are jagged or "spiky" or give rise to occlusions which are difficult to match reliably using area-based correlation.

There are many different possible models for implementing vergence based on *point matches*. These different models are concerned with *selective* versus *non-selective* point matching, *image independent* versus *image dependent/inferred* selective vergence and *attended* versus *non-attended* vergence.

The above different models could be viewed as a *Behavioural Hierarchy* [10] that defines how the system should behave in given circumstances. The concept of modes of behaviour in gaze control is discussed in Section 3.

In the context of autonomous robot vision systems, the ability to identify and categorise objects imaged within the environment is fundamental. Accordingly, a system that can reliably identify objects in its field of view could be utilised in a broad range of applications. With regard to techniques which currently exist, however, generally applicable and robust methods are scarce. Approaches include shape-based methods such as Belongie's [11], which identifies correspondences between points on a shape and uses them to estimate an aligning transform, and Gevers' [12], which combines colour and shape information into a high-dimensional descriptor of the object for recognition purposes.

It has already been stated, however, that the integrated framework designed in this paper makes use of SIFT descriptors generated by the vergence system for the purposes of object recognition. SIFT-based object recognition has been implemented in several systems, for example as reported by Eklundh using the Yorick head, which could localise, attend and recognise objects [5], and as reported by Kragic in a domestic application of robot vision [13]. However, the above systems have been based on the assembly of ad hoc collections of vision mechanisms, including SIFT. In the work reported here, the SIFT algorithm provides the fundamental visual representation (via keypoint descriptors) manipulated by our system, as well as serving as the basis for a reasonably general purpose object recognition system. In addition, SIFT can be readily adapted for point matching based on other sensing modalities [14]. Our laboratory has now developed a version of SIFT adapted to operate on range images [15], offering the potential to extend the developed system in the future to take full advantage of its binocular imaging ability. Adopting SIFT has also considerably facilitated the development of our binocular vision system, as standard SIFT implementations are freely available.

The process of using SIFT for object recognition is described concisely by Eklundh in [13]. Assuming a set of 'known' objects and a database that contains images of a number of poses of each, SIFT features are extracted for every image in the database. Since our objective is to achieve gaze control within a binocular robot camera system, how the object recognition component is integrated with visual search processes within the system is critical to achieving the desired visual behaviour.

Human visual attention is often described as being governed by the '*searchlight metaphor*' [16]. This suggests that human visual attention is separated into two modalities of analysis running simultaneously, with the output of one feeding into the other. These modalities are categorised into pre-attentive and attentive forms (discussed further in Section 5).

In machine vision, the above paradigm was adopted by Westelius [17] to drive the attention of his hierarchical gaze control implementation. Earlier attempts at modelling attention in a computer vision context include the use of multiple feature maps [18] and [19]. More recently reported developments in visual attention include the systems developed in [20] and [21] that integrate bottom-up feature maps (colour, intensity and orientations) in order to learn features form fixation points while recognising object classes in a top-down manner; similarly, [13] uses depth recovery to segment the scene by distance as part of an object-search strategy. At first sight several of the above systems appear to share the same basic principles underlying visual attention as we report here: bottom-up, low-level cuing designed to invoke the application of a more computationally costly investigation of specific visual locations and top-down biasing designed to tune the cuing process according to the needs of the high-level visual task. A critical difference in perspective of the work reported here is that we are not attempting to model the full complexity of human attention, only that required to subserve our SIFT-based vision engine. Accordingly, SIFT keypoints themselves are sufficient to cue the search for groups of keypoints that are diagnostic of the presence of instances of specific known object classes (since no other visual representation is currently adopted in the system). That said, an explicit objective of this work is to provide a framework that can be easily adapted to different visual tasks and operating scenarios and readily extended with improved visual capabilities as they become available.

As discussed above, there are several distinct elements which drive an attention mechanism. The gaze control system adopted in this paper has been modelled on the searchlight metaphor of attention, including pre-attentive and attentive elements working in conjunction to guide the cameras.

To ensure that visual search progresses without endless backtracking, we have developed a mechanism for implementing *inhibition of return* (which also operates in a purely symbolic space) and have integrated this within the gaze control system.

The specific task that our active binocular vision system is designed to achieve in this work is automatic active exploration of a cluttered and unstructured environment in order to report the presence and locations of known objects. Our primary goal is to integrate visual mechanisms such as vergence, object recognition and gaze control in a high-level visual control architecture based on extracting symbolic visual features in order to address applications in advanced telepresence and autonomous robotics.

3. Vergence

The requirements of the vergence system specify that the cameras are driven such that they target the same real-world position. There are several different modalities of vergence conceivable, including those operating on the following contexts: when the system does not know a priori the contents of the scene; it is verging on a specific object, or salient item; the content of the scene is known a priori and one camera already targets the desired location.

Thus, the behaviour of the system is contextually defined and task-motivated. We have attempted to structure the vergence system as a hierarchy of behaviours, related to Brooks' Subsumption Architecture [22]. The two modalities considered are *Global, non-selective vergence* and *Attended, selective vergence*.

The *selective vergence* case was developed as an adaptation during the design of the gaze control system as a special case of the *non-selective vergence* case (Section 5). The remainder of this

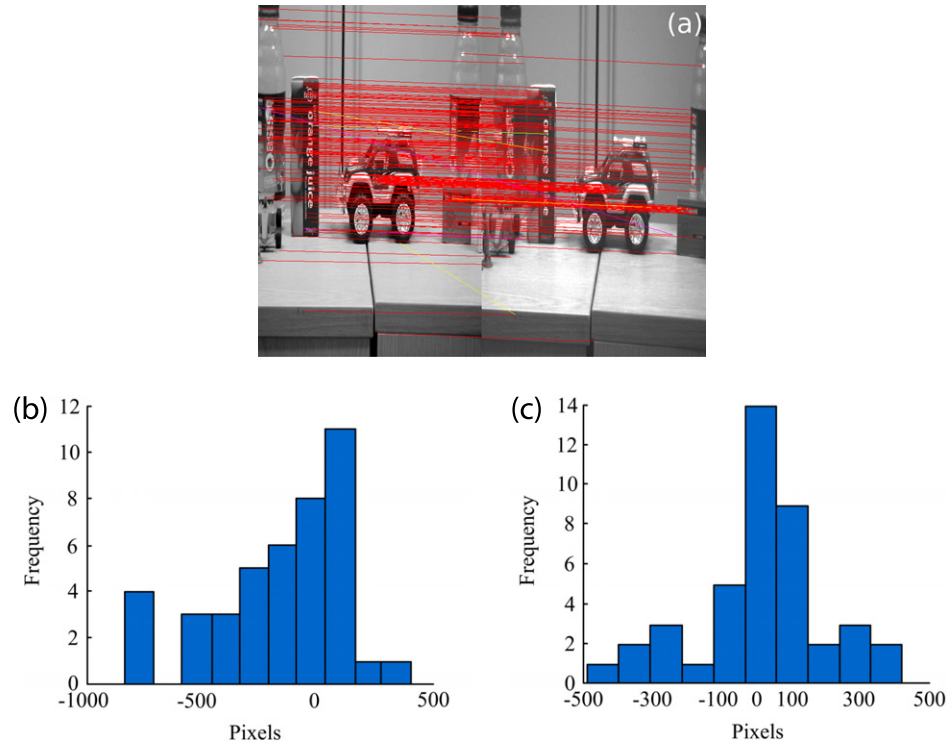


Fig. 1. (a) The stereo-pair view of a clutter scene. (b) Horizontal disparity histogram (left) and vertical disparity histogram (right).

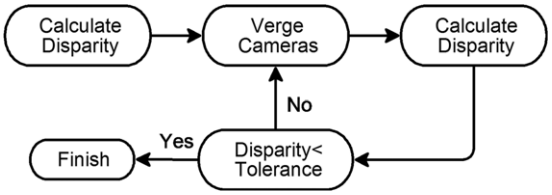


Fig. 2. The structure of the closed-loop verge control algorithm.

section, therefore, refers mainly to the development of the design of *non-selective vergence*.

The working hypothesis during the design of the vergence system was that it is possible to cause the cameras to verge by considering a global set of SIFT keypoint matches between the two camera images, i.e. keypoint correspondences between the images of the stereo-pair. For each pair of corresponding (i.e. matched) keypoints identified, the x-axis positions of these keypoints in each image are compared to produce a single-point disparity. For any given stereo-pair of images, there is likely to be a large number of such matches. An example of a stereo-pair captured by the robot head is shown in Fig. 1(a). Matched keypoints are joined by lines. The algorithmic design is summarised in Fig. 2. To facilitate closed loop vergence, the horizontal disparity is measured again after the first iteration. If the modulus of the post-verge disparity is reported to be larger than a tolerance value, another iteration is initiated.

In a scene that does not contain depth (that is to say, one in which all viewable information exists in one plane, parallel to the camera baseline), all correctly identified keypoint matches will exhibit the same disparity value. However, we know this condition is not usually true for almost all non-trivial cases. Image keypoints that correspond to real-world locations at a range of distances from the cameras will exhibit a range of disparities.

The solution developed is to use the raw disparity data to infer information about the structure of the scene by identifying clusters, or peaks, of disparities. Since each disparity corresponds to a point somewhere on a surface at a specific distance from the cameras, we hypothesise that large numbers of roughly similar disparities cue the presence of a potentially *interesting* object (i.e. an object comprising visual structure). An implicit assumption of this approach is that any object which is spatially compact in depth will form a disparity cluster around some mean distance to the cameras. Where several objects are present, the object with the most structure represented by keypoints will give rise to the largest cluster, and this can be identified (in each image axis) by detecting the highest peak in simple histograms of the keypoint horizontal and vertical disparity values. Fig. 1(b) depicts such histograms (for a bin width of 10 pixels).

An examination of the vertical disparity histogram shows a clear peak around zero. This is expected, as the cameras have a horizontal stereo baseline and are maintained in vertical alignment. Therefore, all correctly matched keypoints will exhibit near-zero vertical disparity. This assumption holds when the cameras are in a fronto-parallel position; however, as the cameras rotate away from this position, epipolar tilt induces non-zero vertical disparity between corresponding keypoints.

In order to mitigate false SIFT feature matches caused by epipolar tilt, we created two constraints associated with each SIFT keypoint matched: the *rotation* and *scale constraints*, defined as follows:

$$\left|\theta_{left} - \theta_{right}\right| \leq 20^{o} \tag{1}$$

$$\sigma_{left} \leq 0.45\sigma_{right} \tag{2}$$

where θ denotes the in-plane rotation value of the stereo-pair images of the keypoint matches in left and right cameras and σ is the scale of the keypoint matches of both camera images.

Filtering keypoint matches using (1) and (2) prior to constructing the disparity histogram enables the above vergence algorithm to operate robustly while operating with cluttered scenes.

4. Object recognition

The design of the object recognition system is a direct adaptation of the SIFT-based object recognition first described by

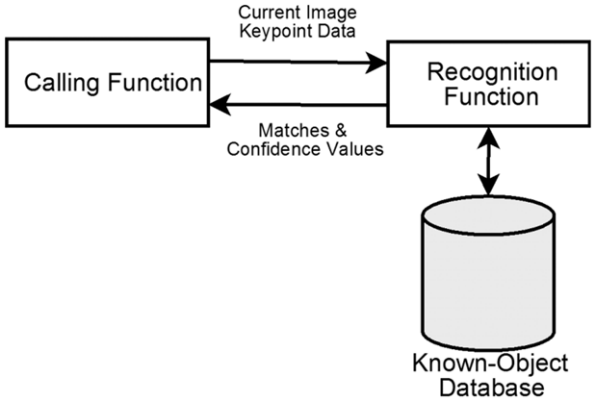


Fig. 3. The method of interfacing the recognition function. Note the use of passing keypoint data instead of image data.

Lowe [1]. The relevance of the design to this project is found in the means of integrating the object recognition system in the overall framework. For completeness, a brief overview of the design is given below.

The basic function of the object recognition procedure is to compare each input image captured by the binocular camera-pair to all pre-stored object examples held in a database (in the form of sets of keypoints rather than images). Having applied the SIFT algorithm to all training images, the generated keypoints are stored in a data structure to facilitate searching during the subsequent object recognition phase.

The object recognition system takes the keypoints extracted from the current camera images, matches these to all keypoints in the database and then applies the Generalised Hough Transform (GHT) [23]. There are typically several images of each object class in the database. Each database keypoint must, therefore, remain logically associated with an object class. When a keypoint match is found, it is registered as one vote for that object class. The integration of the recognition function into the overall framework is summarised in Fig. 3.

In an object recognition context, the GHT as Lowe described in [1] is used to strengthen a recognition hypothesis by establishing a measure of geometric consistency between test object and reference object comparisons. This is performed by assigning votes into Hough-space bins for each matched SIFT feature. When a peak or cluster of votes is detected in Hough space, it indicates a consistent interpretation for a number of features which has a much higher probability of being true than a single feature match.

When affine pose estimation is applied to a winning cluster of keypoint votes in the GHT, described in [1], this can provide a precise location of the centre, rotation and scale of the hypothesised object.

To obtain the position of any point of a database object in the scene, the affine pose estimator is used as follows:

$$y = Ax \cdot PS_{ratio} \tag{3}$$

where A is the affine transformation described above, x is the centre of the image, PS_{ratio} corresponds to the number of motor steps required per pixel of translation in the image and y is the spatial location of the object in the scene. The actuators can then be driven to target, i.e. fixate, the cameras as required using (3).

The interface to the object recognition function is intentionally low level to provide the maximum level of flexibility in its use. Notably, the recognition module does not return a set of recognised objects, but a set of all objects in the database, with the associated number of matches to each database object.

The confidence for any given object is defined as the value of the highest peak in Hough space for the recognised object. This value is used in Section 5 to saccade the cameras towards the object with the highest confidence.

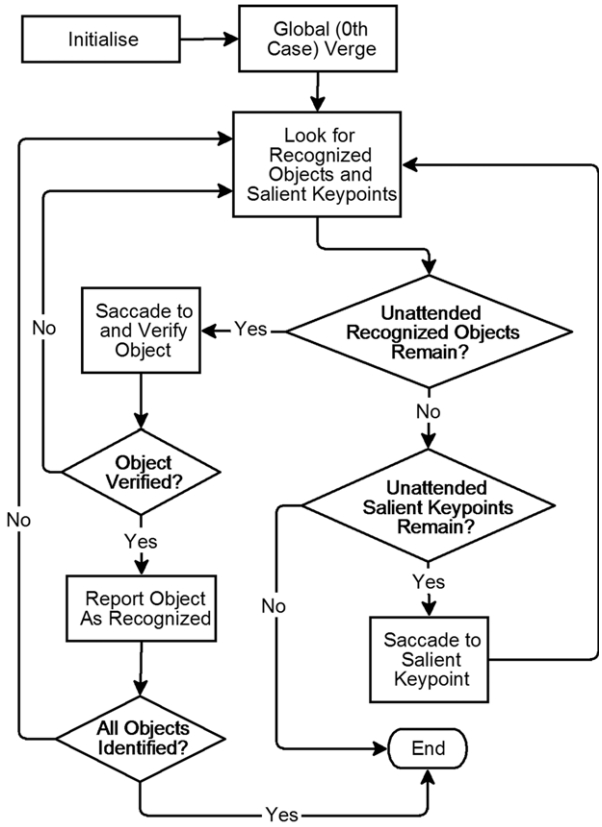


Fig. 4. A flow chart showing a high-level view of the behaviour of the gaze control system.

5. Gaze control

The design of the behavioural system aims at achieving gaze control driven by the vergence and object recognition functions (described in Sections 3 and 4) in order to undertake scene exploration. We have developed an attention system that operates purely in symbolic space represented by SIFT keypoints. This allows a single set of image features to be used for the entire heterogeneous set of tasks required.

A flow chart of the behaviour of the system can be seen in Fig. 4. The pre-attentive and the attentive functions operate in a quasi-parallel manner, with the output of the former feeding to the input of the latter.

The pre-attentive function is concerned with analysing the current field of view to detect salient visual features, i.e. putative objects and unmatched keypoints. This phase does not make recognition decisions; it is solely responsible for detecting features in the current field of view that may be of interest to the search strategy. These image features then cue the attentional 'searchlight' and are passed to the attentive function as a structured list comprising the matching object index in the database, the hypothesised object centre in camera coordinate space and the corresponding confidence value.

The attentive phase uses the information provided by the pre-attentive function to target the cameras and make recognition decisions. This phase selects which item to visit (attend) next, and directs the cameras to the reported location.

As consequence of the pre-attentive phase, the gaze control system will '*notice*' objects and keypoints only when they appear in the view of the dominant eye (the left camera). Since the cameras are only driven to look at objects and keypoints, salient items will only be registered if they appear in the field of view when the cameras are fixating on another object or salient item. This system,

therefore, follows a '*stepping-stone*' search pattern, due to the way that the system will notice a second object when saccading to target the first. An object will only reach the attention of the system if it appears close enough to a fixated object or can be reached by a 'bridge' of other objects and salient keypoints. When there are no (unattended) putative objects in the current field of view to drive fixation, salient keypoints are used to determine where to saccade next. Salient keypoints are those keypoints in the left camera image that are found both not to match any database image and also exhibit a saliency score above a threshold. The saliency score (S_{score}) for each keypoint is proportional to feature scale and spatial eccentricity:

$$S_{score} = x_{offset} \times y_{offset} \times \sigma \times 10^{-3}. \quad (4)$$

Note that x_{offset} and y_{offset} denote the horizontal and vertical distance from the left image centre, respectively, and σ is the scale value of the keypoints matched.

The mean and standard deviation of the saliency scores is computed from all unrecognised keypoints in each fixation. Only those keypoints having saliency scores that exceed three standard deviations of the currently fixated keypoint population are retained within a *working memory* pool. This strategy biases the system to discover large salient regions in the periphery of the observed fixation. In selecting which keypoint to target, the attentive function selects the unvisited keypoint with the highest saliency score from working memory.

To inhibit return to salient image features that have already been attended, a list is maintained of those salient items that have been attended (verged on) by the system (the list includes the saliency score, the SIFT descriptor of the salient point, the camera coordinate space location and if it has been attended). When the pre-attentive phase finds a new set of salient keypoints in the current image, each keypoint is compared to all recognised keypoints and salient items in the attended list using Lowe's matching algorithm [1]. If an input keypoint is found to match a keypoint that has already been recorded in the attended list, the input keypoint will be discarded.

This mechanism leads to the inclusion of the unrecognised salient keypoints as a target of the pre-attentive system. It is hypothesised that, by allowing the cameras to follow unrecognised image structures, it provides a method to guide exploration of parts of the scene that would not be reached had the cameras only fixated on recognised objects.

When saccading to a recognised object, it is necessary to verge the camera-pair such that the object of interest is centred in the field of view of each camera (Attended, selective vergence case in Section 3). The approximate location of the object is known at saccade time, as its coordinates are passed from the pre-attentive phase where they are calculated by means of the affine pose estimator (Section 4).

To ensure that the vergence operation targets only the desired object, the coordinate frame is translated to actuator units to calculate the required actuator movement to centre the object in view. Subsequently, only those database keypoints that match the target object are used in the disparity calculation (algorithm of Fig. 2), and hence only that object will be verged on and fixated.

6. Experimental design and results

6.1. Binocular camera robot head configuration

The physical robot head [24] used in this work comprises the following: one colour SONY digital camera DFW-X700 and a black and white SONY digital camera; XCD700, (each of 1024 × 768 pixels resolution) fitted with IEEE Firewire interfaces and four high-accuracy stepper motors and motor controllers (Physik Instrumente GmbH & Co.).

The hardware was interfaced to a Pentium 4 computer with a CPU clock speed of 2 GHz, with 2 Gb RAM running under Windows XP in the MATLAB programming environment.

6.2. Vergence system validation

As previously explained, by applying different modes of operation based on different visual search conditions, the vergence mechanism can be viewed as implementing a behavioural hierarchy. The *non-selective* and *selective vergence* levels of this hierarchy have been implemented in this system. As described in previous sections, the *selective vergence* case is implemented as a special case of the *non-selective* case. Therefore, validation of the vergence system is aimed primarily at the *non-selective* case. The correct function of the *selective vergence* case is validated as part of the gaze control system.

The objective of the non-selective vergence case is to minimise the total horizontal disparity between a global set of uniquely corresponding locations identified in the current camera images, when no target has been identified in the current field of view. The statistical accuracy and reliability of the vergence system is measured by observing the system behaviour when presented with a number of scenarios: (a) when all keypoints appear in a single depth plane; (b) when a disparity step (resulting from two juxtaposed planes at different distances to the cameras) is present in the field of view; and (c) in a realistic situation in which keypoints are located at a continuous range of possible depths.

To create a scene in which all identifiable detail occurs on a single plane, a printed image was mounted onto a board, which was then mounted on a bench at known distance from the camera baseline. The vergence algorithm in Fig. 2 was initiated and allowed to execute until it settled with a tolerance value of ±4 pixels. This process was repeated six times at different depth locations. In every case it took two iterations for the vergence to settle. We define the RMS vergence error to be the residual disparity measured on completion of the vergence cycle. It can be seen from Fig. 5 that there appears to be no correlation between the distance of the target from the cameras and the accuracy of the vergence. The worst single vergence error observed in all 36 verges was an error of ∼5.3 pixels from optimal. The average overall accuracy is ∼1.4 pixels of error. Both values are objectively small and therefore acceptable for the robotics applications we envisage for the system.

Likewise, to create a scene that contains identifiable detail in two separate depth planes, a second but different printed image was mounted to a board set adjacent to the previously mentioned printed image. The first image was kept at the same distance to the camera baseline, while the distance of the second image was varied. As in the first experiment, six vergence trials were performed and the vergence error was measured in the same manner. The results of this validation experiment are shown in Fig. 6.

It is notable that a lower overall accuracy is observed when comparing the results between experiments (Figs. 5 and 6). The overall mean vergence error is over twice that of the first experiment. Objectively, the mean vergence error is still sufficiently small to allow dense disparity fields to be recovered through stereo-matching. The worst single vergence result observed was ∼6.5 pixels of error.

In a real-world scenario, vergence accuracy is harder to measure quantitatively. A precise value of the vergence error could be calculated in the previous experiments as there is a clearly definable 'optimal' verge point. A sample of the images produced during this experiment is shown in Fig. 7(a). The most notable

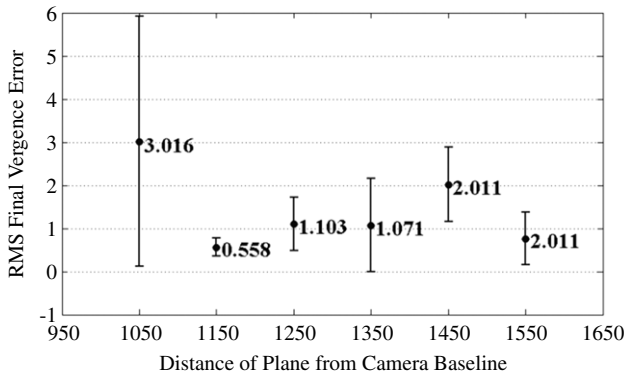


Fig. 5. The vergence errors on a single plane over six iterations at each of six distances trialled. The RMS error is given in pixels, the distance in millimetres and the error bars show ±1 standard deviation for each trial.

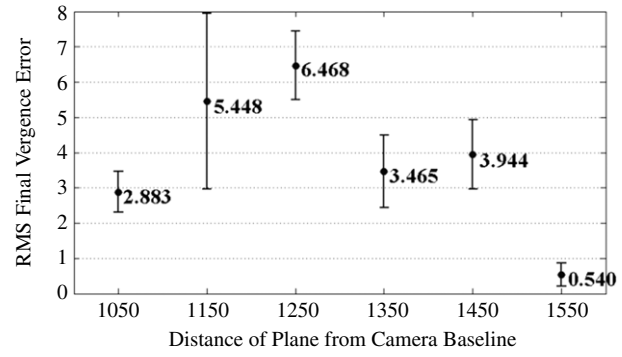


Fig. 6. The vergence errors for two separate depth planes over six iterations at each of six distances trialled. The RMS error is given in pixels, the distance in millimetres and the error bars show ±1 standard deviation for each trial.

of these images is the anaglyph showing the camera views after the verge (Fig. 7(b)). The object that exhibits fewest matches, in this case the Lion toy, is disregarded by the vergence algorithm. The resulting alignment of the skull shows the left eye to be precisely verged, whereas regions of the skull that are more distant are naturally less verged. This level of registration of the skull is, therefore, probably as good as can be expected. However, the remaining vergence residual error is still satisfactory for 3D reconstruction purposes. The average execution time required to verge the cameras was 73.8 s.

6.3. Gaze control system

The gaze control system developed combines our SIFT-based vergence technique with SIFT-based attention and recognition. To test the system, it is first necessary to isolate the different functions and operational modalities of the gaze control system, and these are listed below as three individual units that must be verified:

1. The system should detect the presence of a recognised object when it is in the field of view of the dominant camera, recording its position in the actuator space. When the system is aware of one or more possibly recognised objects, it should saccade to and verge both cameras on the object with the highest confidence of recognition (*selective* case of vergence).
2. The system will use a combination of recognised objects and salient keypoints in a "*stepping-stone*" process to explore the scene, reporting all objects recognised therein.
3. If an object has been previously recognised, no attempt should be made to return attention to that object again. When the system has not seen any potentially interesting objects, it should therefore attend the most salient previously seen keypoint (highest saliency score as described in Eq. (4).

To verify the correct operation of each function, as detailed above, we constructed five different challenging scenes containing known and unknown objects in cluttered arrangements. Fig. 8 shows the six models used in these experiments; SIFT features were extracted from each image in the database depicting different object classes and object poses. The above scenarios allowed us to produce a characterisation of the system's performance. We generated five different random initial fixations for each scene (a total of 25 visual searches were performed). Similarly, we allowed the system to perform either 45 saccades as maximum, or run until a halt condition was met, i.e. no new known objects could be detected.

Fig. 9 shows the ability of the system to detect and classify correctly objects in the field of view (first gaze function listed). As described in the previous section, the pre-attentive cycle analyses the current field of view for interesting possible known objects (Fig. 9(a)) and salient regions. The bounding boxes in Fig. 9(a) denote localised objects which should be attended in order to accept or reject an object hypothesis. The confidence value, defined in Section 4, was used to discriminate which object to attend; in this case, the appearance of the *Orange juice* object was closer to that of the stored examples in the database, and therefore its confidence value is higher than those of the other two observed objects. Fig. 9(b) represents the views of both cameras after the saccade to verify an object and the *selective* vergence cycle is performed. It can be seen that the *Orange juice* object is correctly positioned near the centre of both images. This validates the requirement of the system to be able to identify correctly the actuator-space location of an object in the field of view. The correct identification of the Orange juice demonstrates correct operation of the first gaze control function listed; in order to further characterise this function, the fixation points were recorded and analysed. These results are explained in the analysis of the second listed function.

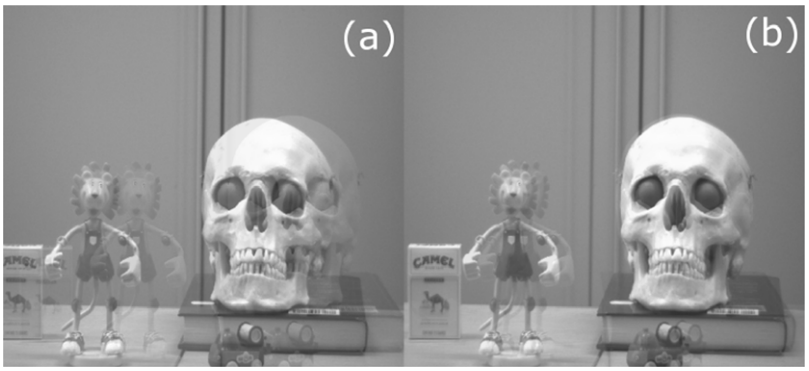


Fig. 7. (a) An anaglyph of the left and right camera images before verging. (b) An anaglyph showing the camera images after vergence has settled.

Fig. 8. (From left to right) The Iceman, Bear, Car, Cigarette box, Skull and Orange juice objects used in the trials.

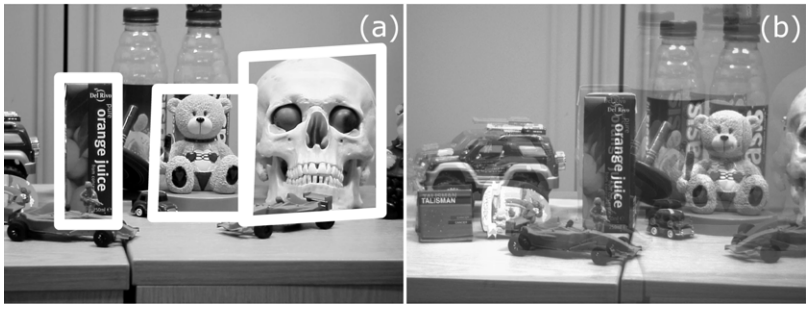


Fig. 9. (a) Initial field of view of one camera and object localised. (b) Anaglyph of the camera images after the saccade to the position of the Orange juice and before the vergence cycle.

The five scenes used in the experiments are shown in Fig. 10 for a single trial. As described, ten objects were arranged in a cluttered scene in order to investigate the ability of the system to explore a scene (second gaze function). The stepping-stone search pattern can be observed by inspection of the camera traces overlaid for each of the five trials, as shown in Fig. 10(a)–(e).

It can be seen that the object being targeted had appeared in the field of view of the previous fixation (the target/identification process is represented in Fig. 10 with a black circle and salient items with a black filled square; the initial and final fixation points are denoted as black upward- and white downward-pointing triangles). For example, in the initial fixation in Fig. 10(b) no object hypothesis was discovered; consequently, the first saccade is towards a salient item. Following this, the second and third saccades fixate on hypothesised objects; each targeted object must therefore have been detected in previous pre-attentive cycles.

It should be noted that it is not necessary that the next object attended must be selected from the current camera image; the next selected object is the most highly matched object of all unattended candidates detected. As new matches are gathered during the visual search process, isolated erroneous matches can be outvoted by correct matches accumulated during each new fixation.

In the results presented there are several examples of a saccade to an object that was identified in a fixation several cycles earlier. This behaviour can be explained by the pre-attentive cycle tending to detect first and then attend salient points near to the scene centre, where objects situated directly in front of the cameras most closely resemble their trained examples. Objects situated more peripherally in the scene less resemble their trained examples, and consequently have a lower match confidence, and therefore are detected later in the search process. Therefore the system is able to verify a region several times until the gaze control completely inhibits the region's salient information. Moreover, this type of visual behaviour has been seen in high-order vertebrates in different visual attention studies and reviews (e.g. [16,25,26]).

While it is therefore expected that the behaviour of the system will conform to the functionality described, it does not necessarily follow that 100% identification of objects in the field of view will be achieved. Table 1 summarises the visual search identification failures the system incurred. We divided the different failures in three categories; *False localisation* shows that the system was not able to centre the object in its field of view; *Not found* shows that the system did not notice the object in the visual search task; and *False hypothesis* shows that an object hypothesis was rejected during the attentive cycle.

Table 1
Table of visual failures.

Object	False localisation	Not found	False hypothesis	Total no. of failures
Skull	3	2	0	5
Iceman	3	2	0	5
Cigarette box	1	1	1	3
Car	0	3	0	3
Bear	3	0	1	4
Orange Juice	1	0	0	1
Total	11	8	2	21

Note that the *Skull*, *Iceman* and *Bear* objects are the most difficult items to locate since their 3D structure causes their 2D appearance to change significantly at each fixation or their texture detail is overly fine for SIFT features to be extracted. Likewise, the *Skull* and *Iceman* appeared on the *Not found* objects lists during the visual search trials; this supports the hypothesis that the 3D structure of these objects produced excessive 2D appearance changes that confounds SIFT matching. The *Car* object scored more *Not found* failures due to its SIFT descriptors matching with unknown objects to produce outliers and, in consequence, these outlier matches were not geometrically consistent with the reference object centre (Section 4). It must be pointed out that the 45 saccade halt condition was activated eight times, which corresponds to the eight objects that were not noticed by the system (Table 1: *Not found* column). Finally, the system maintained consistent behaviour for false hypothesis rejection; only two misclassifications were registered (scenes of Fig. 10(d) and (e)) over all of the trials, primarily due to the structure of the scene and in-plane rotation of the objects.

Therefore, the ability of the system to explore the scene identifying objects and salient regions yields ∼85.5% of correct identifi-

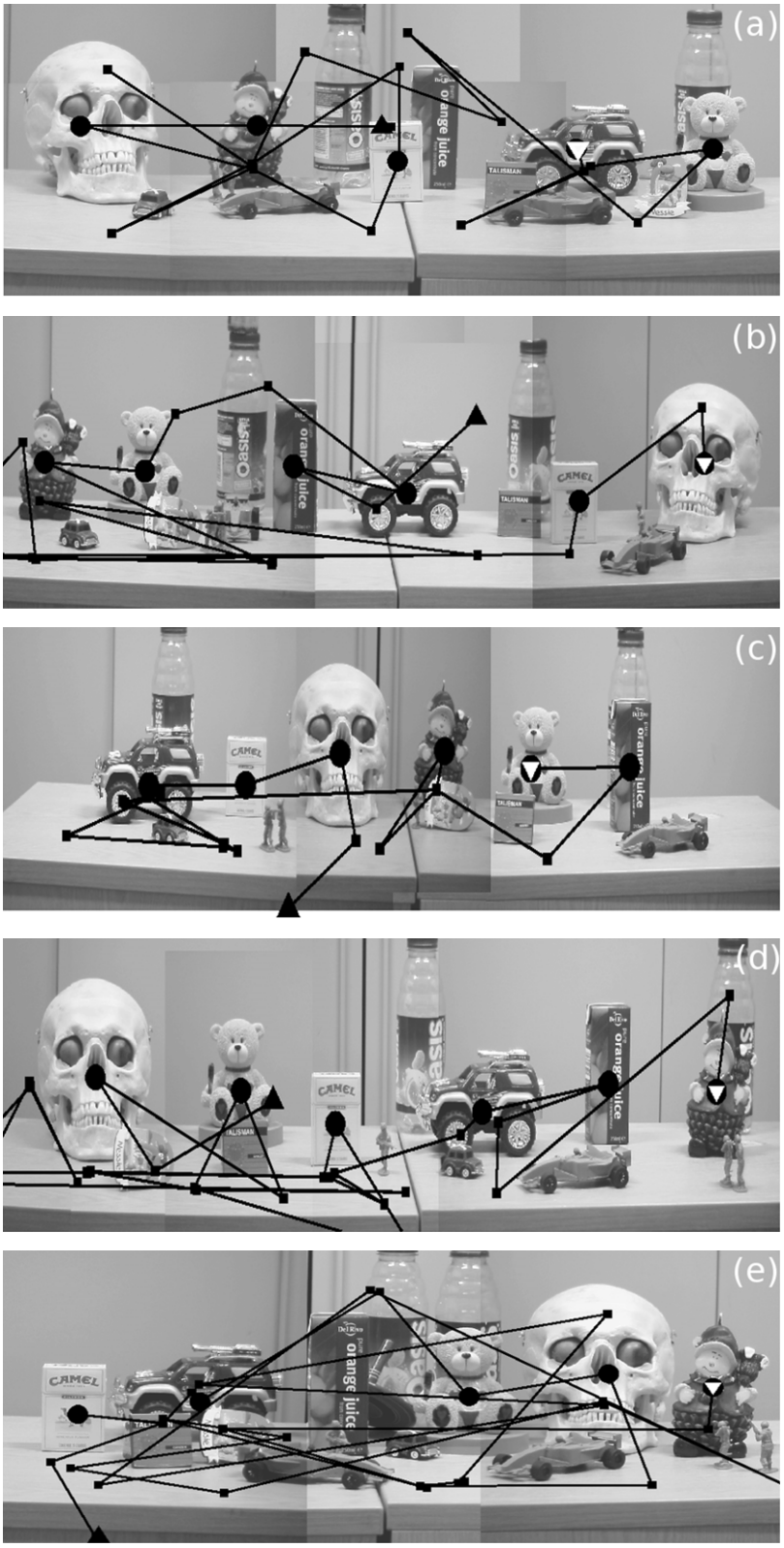


Fig. 10. (a)(b)(c)(d)(e) The five different scenes with overlaid traces of the predominant camera (image-space) used to characterise the performance.

cations over a total of 145 object observations in the 25 presented visual search trials.

For each fixation point where an object was identified, we recorded the camera coordinate-space location of the cameras in order to measure the repeatability of the system to centre the hypothesised object in the field of view (as mentioned before in the first gaze function example). Figs. 11 and 12 show the RMS fixation errors of the projected camera coordinate-space in pixels of the x and y coordinates, respectively.

The Skull object presents the poorest fixation results on the x-axis of $\sim$4.5 pixels from optimal with statistically significant deviations between measures; however, this error is unlikely to

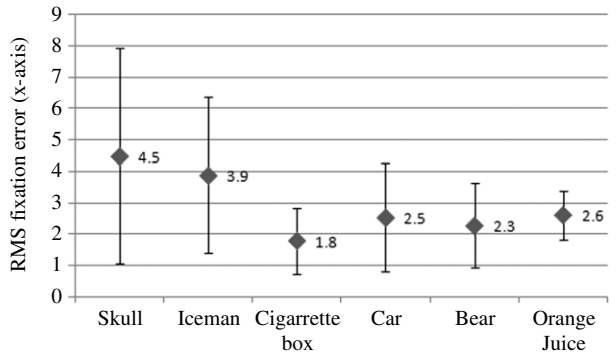


Fig. 11. The *x*-axis fixation errors for each of the six objects over 25 visual search trials. The RMS error is given in pixels (the error bars show ±1 standard deviation for each trial).

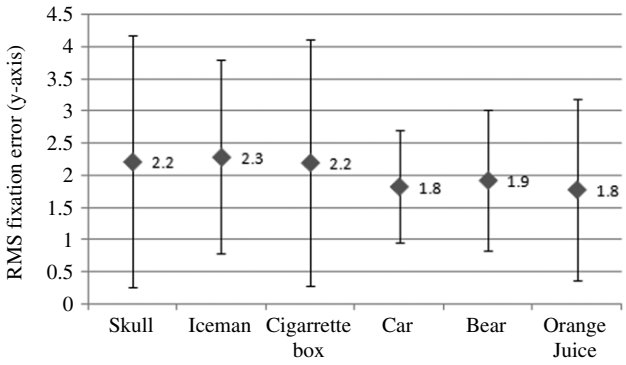


Fig. 12. The *y*-axis fixation errors for each of the six objects over 25 visual search trials. The RMS error is given in pixels (the error bars show ±1 standard deviation for each trial).

be problematic since the Skull object occupies 500 by 200 pixels in the stereo-pair images, corresponding to a viewing error of ∼1%; the vergence errors (discussed in Section 6.2) have also to be considered in order to measure the overall repeatability of the system. A more consistent average fixation error of ∼2 pixels from optimal is observed on the *y*-axis. Since the *y* displacement between cameras remains constant during visual search, vergence errors do not affect the *y* fixation accuracy, even when the epipolar tilt produces vertical disparities between the image pair. Therefore, the resultant errors are within reasonable limits for many robotics vision-based applications.

The average execution time required to explore the different scenes and to recognise the objects was ∼31.7 min. Likewise, the system performed an average of 37 saccades per trial and, as expected, it started recognising objects from the second saccade. Fig. 13 shows the ability of the system to identify objects as a function of the number of saccades performed. After 17 saccades the recall rate was exactly 80% of objects successfully identified. It can also be seen that after 37 saccades almost all the visual search trials have identified the six objects in the scene.

To verify the ability of the system to exploit unrecognised scene elements to guide exploration (third gaze function), a special cluttered scene was constructed, as shown in Fig. 14, that contained only two known but widely separated objects (the Skull and Lion toy, highlighted in boxes). The arrangement of the two known objects was such that, when the cameras were directed at one known object, the other was not present in the field of view; a gap of 300 mm between them was present.

The visual search strategy was invoked and allowed to run until both objects were found correctly. The system performed 40

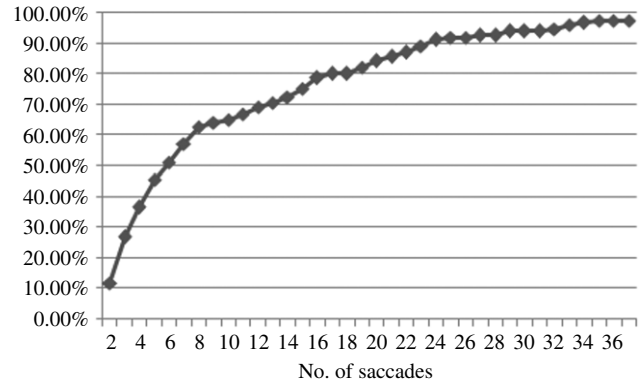


Fig. 13. Cumulative frequency of identified objects for each saccade.

Fig. 14. The scene used to verify the "*stepping-stone*" visual search and the attended objects (black boxes: Skull and Lion toy).

saccades to find both objects in the scene. These results corroborate the visual search results presented in Fig. 10, where, in order to explore this scenario, the system had to attend both unknown and salient elements to identify all known objects. In Fig. 10(a)–(e) we can observe where the gaze control behaviour drove the cameras to visit areas with salient keypoint information before attending and identifying the known objects.

The above results produced by this study suggest that we have demonstrated the ability of the system to form useful visual search patterns and recognise objects with reasonable accuracy, by testing the gaze control system against complex cluttered scenes in different arrangements and defined tasks.

7. Conclusions

The objective of the work reported here is to develop a binocular robot vision system capable of autonomous scene exploration, with the specific task of identifying and localising objects of known classes while maintaining binocular vergence. We have presented a system that demonstrates the application of several novel design principles in a functional integrated framework that essentially achieves the objectives defined in Section 2.

Adopting SIFT features as the underlying visual representation for our active gaze control system has allowed a single mechanism to combine elegantly the key functions of binocular vergence, object recognition and saccade selection.

The approach of computing a vergence signal that drives the binocular camera pair, based on finding the highest feature density peak within a SIFT derived disparity histogram, has also been found to be robust and effective. The maximum vergence error observed of ∼6.5 pixels remains within viable limits for any subsequent depth recovery task based on stereo-matching. We anticipate that, by couching the vergence mechanism as a behavioural hierarchy, it will be possible to structure this algorithm efficiently to meet the needs of different operational contexts.

Fig. 15. The scene and the robot head used in the experiment illustrated in Fig. 14.

Saccade selection drives our gaze control system (Section 5) and likewise adopts SIFT keypoints as the basis of attention and inhibition of return mechanisms. SIFT keypoints also provide the basis for a standard object recognition module which has been embedded in our system in a conceptually uncomplicated manner. In the results presented, we demonstrate that our active binocular robot vision system is capable of accurately centring its gaze on a hypothesised object in each fixation of the camera-pair in cluttered scenes, with a classification success rate of ∼85%. In addition, we demonstrated that the implemented '*searchlight-metaphor*' of visual attention in conjunction with the *stepping-stone visual search* could serve to navigate automatically amongst separated known objects embedded within unknown clutter using SIFT features as visual cues.

We argue that the current system compares favourably with current state-of-the-art binocular vision systems ([5,27]) since similar recognition rates are achieved in this work without using further sensory information or visual rectification or segmentation in the input images in order to attend and identify objects within a maximum margin error of ∼8 pixels from optimal. Similarly, the average number of search steps required by the system to locate and verify six known objects was in the interval of 2 to 37 saccades.

Our principal claim is that it is now possible for a binocular robotic vision system to direct its gaze on a scene such that it maintains binocular vergence, detects salient image features, directs its gaze to investigate these features, verifies the identification of objects and continues to investigate the workspace for recognised objects based on visual cues. All of these characteristics have been combined in a computationally parsimonious manner using SIFT descriptors. Fig. 15 shows an example of our binocular robot camera head inspecting a scene.

It should also be noted that the current system implementation is not intended for real-time operation; however, we believe that this can be achieved by means of GPU acceleration of both the SIFT algorithm [28] and critical sections of the vergence and saccade selection mechanisms.

Our current work now focuses on automatic clustering in a *continuous* Hough space to allow multiple same-class object instances to be localised accurately. In the future we propose to investigate range-map recovery [29] and use of 2.5D SIFT [15] features in conjunction with 2D SIFT features to improve object identification and 3D pose recovery. Similarly, adopting colour SIFT descriptors [14] would improve object recognition and hypothesis generation, the gaze control strategy and, in consequence, the definition of complex visual search tasks (i.e. identification of objects based on properties such as colour and spatial relations).

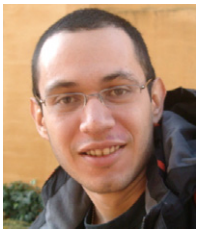
Acknowledgements

G. Aragon-Camarasa is grateful for the support in this research by the Programme AlSSan, the European Union Programme of High Level Scholarships for Latin America, scholarship no. E07D400872MX and CONACYT-Mexico.

References

[1] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[2] P. Mowforth, J.P. Siebert, Z. Jin, C. Urquhart, A head called Richard, in: Proceedings of the British Machine Vision Conference, 1990, pp. 361–366.
[3] D. Betsis, J. Lavest, Kinematic calibration of the *k*th head–eye system, Tech. Rep. S-100 44, Computational Vision and Active Perception Laboratory, Department of Numerical Analysis and Computing Science, Royal Institute of Technology (KTH), Stockman, Sweden, 1994.
[4] L. Natale, G. Metta, G. Sandini, Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head, Robotics and Autonomous Systems 33 (2002) 87–106.
[5] M. Björkman, J.O. Eklundh, Attending, foveating and recognising objects in real world scenes, in: Proceedings of British Machine Vision Conference, 2004, pp. xx–yy.
[6] A. Bernardino, J. Santos-Victor, Binocular tracking: Integrating perception and control, in: IEEE Transactions on Robotics & Automation, vol. 15, 1999, pp. 1080–1094.
[7] Y. Yeshurun, E.L. Schwartz, Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 759–767.
[8] T.A. Boyling, J.P. Siebert, A fast foveated stereo matcher, in: Conference on Imaging Science Systems and Technology, CISST 2000, AAAI Press, Las Vegas, USA, 2000, pp. 417–423.
[9] T.A. Boyling, Active vision for autonomous 3d scene reconstruction, Ph.D. thesis, University of Glasgow, 2002.
[10] L. Balasuriya, J. Siebert, An architecture for object-based saccade generation using a biologically inspired self-organised retina, in: Proceedings of the International Joint Conference on Neural Networks, Vancouver', IEEE, 2006, pp. 4255–4261.
[11] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[12] T. Gevers, A. Smeulders, Pictoseek: Combining color and shape invariant features for image retrieval, IEEE Transactions on Image Processing 9 (1) (2000) 102–119.
[13] D. Kragic, M. Björkman, H.I. Christensen, J.O. Eklundh, Vision for robotic object manipulation in domestic settings, Robotics and Autonomous Systems 52 (1) (2005) 85–100.
[14] G.J. Burghouts, J.M. Geusebroek, Performance evaluation of local colour invariants, Computer Vision and Image Understanding 113 (1) (2009) 48–62.
[15] T.R. Lo, J.P. Siebert, Sift keypoint descriptors for range image analysis, Annals of the BMVA X (2009) 1–18.
[16] E.A. Styles, Attention, Perception, and Memory: An Integrated Introduction, first ed., Psychology Press, 2005.
[17] C.-J. Westelius, Preattentive gaze control for robot vision, Ph.D. thesis, Linkping University, 1992.
[18] R. Milanese, Detection of salient features for focus of attention, in: Proceedings 3rd SGAICO Meeting, Swiss Group for Artificial Intelligence and Cognitive Science, Biel-Bienne, Switzerland. pp. 87–101, October 1992. Published as: University of Bern, Institute for Comp. Science and Mathematics, Technical Report IAM-91–004, H. Kaiser, R. Bach and H. Bunke, Eds., March 1992.
[19] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.
[20] D. Walther, U. Rutishauser, C. Koch, P. Perona, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Computer Vision and Image Understanding 100 (1–2) (2005) 41–63.
[21] V. Navalpakkam, L. Itti, Sharing resources: Buy attention, get recognition, in: Proc. International Workshop on Attention and Performance in Computer Vision, WAPCV'03, Graz, Austria, July 2003.
[22] R.A. Brooks, How to build complete creatures rather than isolated cognitive simulators, in: Architectures for Intelligence, Erlbaum, 1991, pp. 225–239.

[23] D.H. Ballard, Generalizing the Hough transform to detect arbitrary shapes, Pattern Recognition 13 (1981) 111–122.
[24] A.A. McDougall, Interfacing a robot head in MATLAB, Master's thesis, University of Glasgow, 2004.
[25] A. Yarbus, Eye Movements and Vision, vol. 100, Plenum Press, New York, 1967.
[26] M.M. Chun, J.M. Wolfe, Visual attention, Blackwell's Handbook of Perception (2001) 272–310 (Chapter 9).
[27] M. Björkman, J.O. Eklundh, Recognition of objects in the real world from a systems perspective, Kuenstliche Intelligenz 19 (2) (2005) 12–17.
[28] C. Wu, 2007. SiftGPU: A GPU implementation of scale invariant feature transform, SIFT, http://cs.un.edu/-wu/siftgpu.
[29] J. Siebert, S. Marshall, Human body 3d imaging by speckle texture projection photogrammetry, Sensor Review 20 (3) (2000) 218–226.



Gerado Aragon-Camarasa received his B.Sc. in Industrial Robotics Engineering at the National Polytechnic Institute (ESIME-IPN, Mexico City) in 2006. From 2004 to 2007 he was with the Professional Development in Automation Program at the Universidad Autonoma Metropolitana (Mexico), where he was involved in the control of processes, thermodynamics and geometric algebras. He is currently a second-year Ph.D. student in the Department of Computing Science at the University of Glasgow, supervised by Dr. J. Paul Siebert. His current research interests embrace robot vision, object recognition, computational models of human vision and geometric algebras.

Haitham Fattah is currently a research engineer and doctoral student for Codeplay Software Ltd. He graduated in 2007 from the University of Glasgow with an M.Sc. in computing science. During his undergraduate degree he specialised in computer vision, active vision systems and digital imaging, under the supervision of Dr. J. Paul Siebert. He undertook projects involving the development of a computerised test for colour vision deficiency and a SIFT-based binocular robotic vision control system.

J. Paul Siebert received his B.Sc. and Ph.D. degrees from the Department of Electronics and Electrical Engineering at the University of Glasgow, in 1979 and 1985, respectively. He is currently a Reader in the Department of Computing Science, University of Glasgow and the Computer Vision & Graphics group leader. From 1991–1997 he was with the Turing Institute, Glasgow, developing photogrammetry-based 3D imaging systems for clinical applications, and he served as Chief Executive from 1994. Prior to this he held the post of Scientist at BBN Laboratories, Edinburgh, from 1988–1991. His research interests include 3D imaging systems and tools for human and animal surface anatomy assessment, and also robot vision systems based on biologically motivated principles. He has co-authored more than 90 international journal and conference papers in these areas.