Towards Binocular Active Vision in a Robot Head

Haitham Fattah, Gerardo Aragon-Camarasa and J. Paul Siebert

***Abstract*—This paper presents the first 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 recognizing objects in a highly-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 a first pilot investigation that yield a maximum vergence error of ~6.5 pixels, while seven of nine known objects were recognized in a high-cluttered environment. 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.**

I. Introduction

THE recent maturation of digital imaging hardware and the continual advancement of image processing and analysis techniques has vastly improved the potential for uses of computer vision in real-world scenarios. 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 such as: autonomous roving vehicles, automatic surveillance, telepresence systems and military applications. 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 a binocular robot head system and a unified framework for binocular camera control.

This paper is organized as follows: in section II, 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 III, IV and V, respectively. Finally, section VI contains a summary of the system validation, its results and contributions to the field of active vision research.

II. 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 speed 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 (operating on highly cluttered images) into a complete system capable of autonomously exploring its visual field. Therefore, our motivation is to investigate the potential for state-of-the art image processing techniques to enhance the performance of binocular robotic vision systems.

Vergence, in a biological context, is the act of adjusting 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 [9]. For example, these advantages are evident when the surfaces are jagged or "spiked" or the local disparity gradient is near to an occlusion.

There are many different possible models for implementing vergence based on point matches. In the

Manuscript received May 15, 2008. (G. Aragon-Camarasa was supported by the Programme Alβan, the European Union Programme of High Level Scholarships for Latin America, scholarship no E07D400872MX).

H. Fattah is with Institute for System Level Integration, The Alba Centre, Alba Campus, EH54 7EG, UK; e-mail: fattah@dcs.gla.ac.uk).

G. Aragon-Camarasa is with the Computer Vision and Graphics group in the Department of Computing Science at the University of Glasgow, G12 8QQ, UK (corresponding author, phone: +44(0) 0141 330 1621; e-mail: gerardo@dcs.gla.ac.uk).

J. P. Siebert leads the Computer Vision and Graphics group in the Department of Computing Science at the University of Glasgow, G12 8QQ, UK (e-mail: psiebert@ dcs.gla.ac.uk, webpage: www.dcs.gla.ac.uk/~psiebert).

context of vergence, these different models are concerned with: selective versus non-selective point matching, image independent vs. image dependent/inferred selective vergence and attended vs. 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 III.

In the context of autonomous robot vision systems, the ability to identify and categorise objects imaged within the environment is essential. Accordingly, a system that can reliably identify objects in its field of view would find use 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 the SIFT data generated by the vergence system for the purposes of object recognition. SIFT-based object recognition has been implemented in several systems such as Eklundh on the Yorick head, which could localize, attend and recognize objects [5] and the work by Kragic on robotic vision in a domestic context [13]. The SIFT algorithm was adopted in this work since implementations of the SIFT algorithm are readily available and SIFT can provide the basis for a reasonably general purpose object recognition system. In addition, SIFT can serve as a framework for point matching based on other sensing modalities. Our laboratory has now developed a version of SIFT adapted to operate on range images [14], offering the potential to extend the developed system in the future to take full advantage of its binocular imaging ability.

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. The integration of object recognition is included in this system in order to demonstrate its utility and the means of doing so in a structured and computationally parsimonious manner.

On the other hand, human visual attention is often described as being governed by the searchlight metaphor [15]. 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 known as 'pre-attentive' and 'attentive' (further discussed in section V).

In machine vision, the above paradigm was adopted by Westelius [16] to drive the attention of his hierarchical gaze control. Earlier attempts of modelling attention in a computer vision context include Milanese's [17] use of multiple feature maps. More recently reported developments in gaze control include the systems implemented in [18] that perform automatic saccadic gaze control in a mobile robot unit with active binocular cameras based on keypoint features; or in [13], which uses depth recovery to segment the scene by distance as part of an object-search strategy.

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, a mechanism for implementing *inhibition of return* (which also operates in a purely symbolic space) has been developed and integrated within the gaze control system.

III. 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 is verging on a specific object or part of the scene, 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 [19]. 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 V). The remainder of this section, therefore, refers mainly to the development of the design of the *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 disparity is measured again after the first iteration. If the post-verge disparity is

reported as larger than a tolerance value, another iteration is initiated.

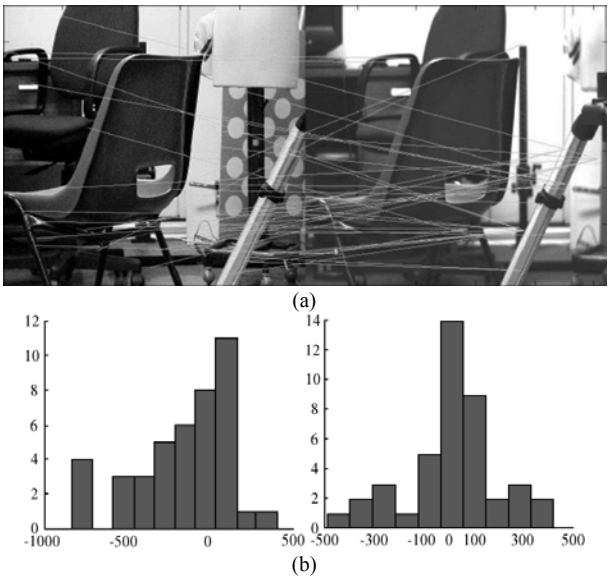

Fig. 1. (a) The stereo-pair view of a highly cluttered scene. (b) x-coordinate (left) and y-coordinate histogram of disparities.

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 number of roughly similar disparities sample the surface of a potentially *interesting* object (i.e. an object comprising visual structure). An implicit assumption of this approach is that any object that 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 by means of a simple histogram of the keypoint disparity values. An example of such a histogram (with bin width of 10 pixels), can be seen in Fig. 1(b).

An examination of the y-coordinate disparity histogram shows a clear peak around zero. This is expected, as the cameras should always be in vertical alignment, and therefore all correctly-matched keypoints will exhibit a 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 a non-zero vertical disparity between corresponding keypoints.

Therefore, we created two constraints associated with each SIFT keypoint matched, these comprise the *rotation* and *scale constraints* defined as;

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

$$\frac{\sigma_{left}}{\sigma_{right}} \leq 45\% \tag{2}$$

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

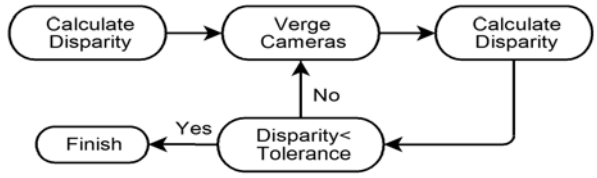

Fig. 2. The structure of the closed-loop verge control algorithm.

In order to mitigate false SIFT feature matches, the disparity histogram proposed in combination with (1) and (2) made this algorithm robust with high cluttered scenes.

IV. Object Recognition

The design of the object recognition system is a direct adaptation of the SIFT-based object recognition first described by 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 will take the keypoints extracted from the current camera images and then match these to all keypoints in the database and then apply the Generalized Hough transform (GHT) [20]. 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 geometrical consistency between the test object and a reference object. 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, this indicates a consistent

interpretation for a number of features which has a much higher probability of being true than a single feature match.

When the 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 of the hypothesized object.

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

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

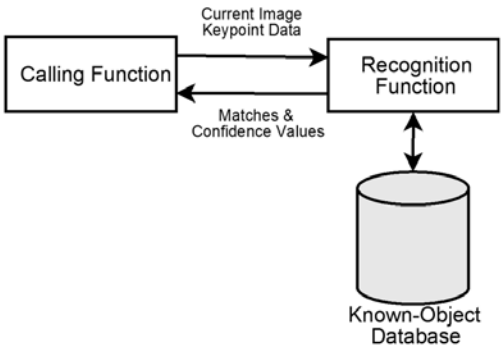

Fig. 3: The method of interfacing the recognition function. Note the use of passing keypoint data instead of image data.

where, A is the affine transformation, described above, x the centre points of the image, PS_{ratio} a pixel-space ratio which corresponds to the total number of motor steps 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 the cameras at that point.

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 recognized 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 confidence of the highest-peak in Hough space for the recognized object which is used in section V to saccade the cameras towards the object with highest confidence value.

V. Gaze Control

The design of the behavioural system aims at achieving a gaze control strategy for search that uses the vergence and object recognition functions (described in sections III and IV) to explore the scene that it is presented with. We have developed an attentional system that operates purely in symbolic space by use of the 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 work in a quasi-parallel manner, with the output of one feeding to the input of the other.

The pre-attentive function is concerned with analyzing the current field of view to detect salient features. This phase does not make recognition decisions; it is solely responsible for detecting areas of the field of view that may be of interest to the search strategy. These highlighted image features are passed to the attentive function (as in 3), which then guides the 'searchlight' to examine these areas in detail.

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

As a 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 targeting another 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 targeted object or can be reach by a 'bridge' of other objects and keypoints. Salient keypoints are those keypoints in the left camera image that are found not to match to a database image and exhibit a saliency score above a threshold. The saliency score ($S_{Saliency}$) for each keypoint is then computed as:

$$S_{Saliency} = x_{offset} \times y_{offset} \times \sigma \times 10^{-3} \tag{4}$$

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

The mean and standard deviation is computed over all saliency scores from unrecognized keypoints in each fixation. These scores are then filtered to keep only those that exceed three standard deviations of the currently input keypoint population. 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, unrecognized keypoints that have been attended (verged on) by the system (6). When the pre-attentive phase is analyzing the current image for new salient keypoints, each keypoint is compared to all keypoints in this list using the Lowe's matching algorithm. If an input keypoint is found to match to a keypoint in the visited list, it will be discarded.

$$\begin{aligned} Objects = \{ & S_{Saliency}, Descriptor, \\ & Location\ in\ the\ image, Attented \} \end{aligned} \tag{5}$$

This principle leads to the inclusion of the unrecognized, salient keypoints as a target of the pre-attentive system. It is hypothesised that by allowing the cameras to follow

unrecognized image structures, it provides a semi-guided method for exploring parts of the scene that would not be reached if the cameras only followed recognized objects.

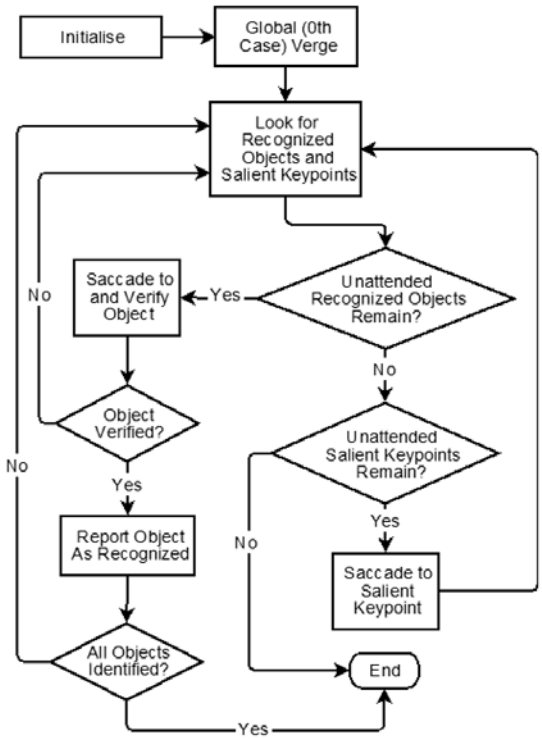

Fig. 4. A flow chart showing a high-level view of the behaviour of the gaze control system.

When saccading to a recognized 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 (3rd order case in section III). 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 IV).

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 to the target object are used in the disparity calculation (algorithm of Fig. 2), and hence only that object will be considered.

VI. Experimental Design and Results

A. Binocular camera robot head configuration

The physical robot head [21] used in this work comprises the following: two SONY cameras XCD700 (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 in RAM running under Windows XP and MATLAB.

B. Vergence system validation

As previously explained, the vergence mechanism, by applying different modes of operation based on different visual search conditions, can be viewed as implementing a behavioural hierarchy. The *non-selective* and the *selective vergence* levels of this hierarchy have been implemented in this system. As is described in previous sections, the *selective vergence* case is implemented as a special case of the non-selective case. Therefore, the experimentation on 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 situation when all keypoints appear in a single depth plane; a situation in which a disparity step (resulting from two juxtaposed planes at different distances to the cameras) is present in the field of view; and a realistic situation in which keypoints come from 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 to a board, which was 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 2 iterations for the vergence to settle. 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 most applications.

Likewise, to create a scene that contained identifiable detail in two separate depth planes, a second different printed image was mounted to a board mounted 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. The same number of iterations as in the first experiment was performed. The vergence error was measured in the same manner and the resulting data is shown in Fig. 6.

It is notable that a lower overall accuracy is observed when comparing the results in Fig. 6 with Fig. 5. 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, accuracy of vergence 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 of these images is the anaglyph showing the camera views after the verge (Fig. 7(b)). The object that exhibits fewer matches, in this case the lion toy is disregarded in the vergence. The resulting alignment of the skull shows the left eye to be precisely verged, whereas regions of the skull that are further away are, naturally, less verged. This level of overlap of the skull is, therefore, probably as good as could be expected. However, the vergence actually achieved, it is satisfactory for the purposes of 3D reconstruction. The average execution time required to verge the cameras was 73.8 seconds.

C. Gaze Control system

The gaze control system was developed to demonstrate how the SIFT-based vergence technique could be combined with SIFT-based attention and recognition in a search strategy. To test this system, it is first necessary to isolate the various functions and operational modalities of gaze control. The functions of the system are listed below as three individual units of functionality that can each be verified.

1. The system should detect the presence of a recognized 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 recognized objects, it will 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 recognized objects and salient keypoints in a `stepping stone' process to explore the scene, reporting all objects recognized therein.
3. If an object has been previously recognized, no attempt will be made to return attention to that object again. When the system has not seen any possible but unverified objects, it will saccade to the most salient, previously seen keypoint.

To verify the correct operation of the first function, as detailed above, we present the system with a highly-cluttered scene, as shown in Fig. 8(a). In the pre-attentive cycle, the skull and the car were identified, the confidence value, evaluated in section IV, was used to discriminate which object to attend (in this particular case, the car was attended).

The correct identification of the car object corroborates the ability of the system to detect and classify correctly objects in the field of view. Fig. 8(b) represents the views of both cameras after the saccade to the object is performed. Note that this image was captured before the vergence cycle; hence the poor amount of overlap. It can be seen that the car is correctly positioned near the centre of both images. This validates the requirement of the system to be able to correctly identify the actuator-space location of an object in the field of view.

To verify the second function, the system was presented with a scene containing all known objects and allowed to run until a halt condition was reached, i.e. no new objects could be detected. 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. Fig. 9 (a) and (b) shows the actuator-space motion trace of both cameras. The stepping stone search pattern can be seen by further examination of the image in Fig. 10.

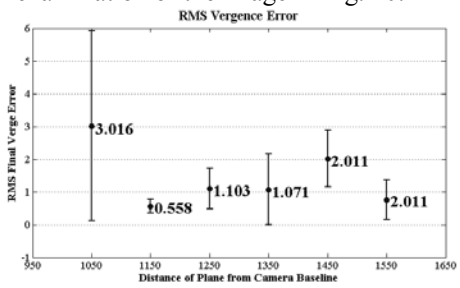

Fig. 5. The verge errors on a single plane over six iterations at each six distances. The RMS error is given in pixels, the distance in millimetres.

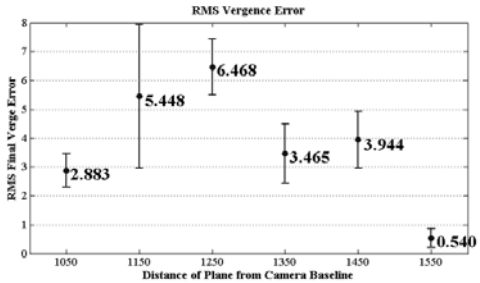

Fig. 6. The verge errors in two separate depth planes over six iterations at each six distances. The RMS error is given in pixels, the distance in millimetres.

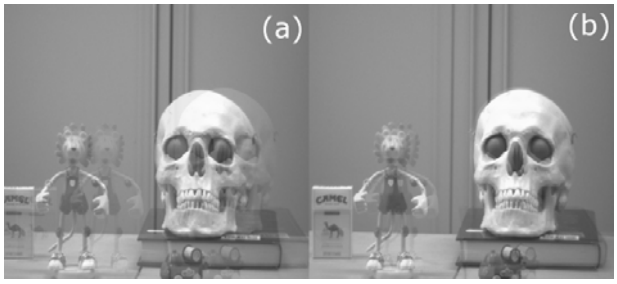

Fig. 7. (a) An anaglyph of the left and right camera images before verging. (b) An anaglyph showing the camera images after vergence had settled.

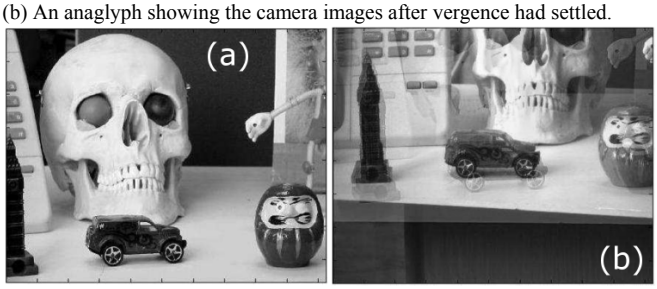

Fig. 8. (a) Initial field of view of the camera, (b) Anaglyph of the camera images after the saccade to the position of the car and before the vergence cycle.

It can be seen that, the object being targeted appeared in

the field of view in the previous fixation (the target process is represented with capital letters in Fig. 10). The system failed to identify one object in the scene, a *bike toy*, which does not have a registered fixation. The saccade denoted as C in Figure 10, which corresponds to the *mouse*, does not centre the object as expected, thus it is considered a false positive match. It is not necessary that the next object to be saccaded to is selected from the current camera image; it is the most highly matched object of all unattended candidate objects that is selected. In the results presented there are no examples of a saccade to an object that was identified several cycles prior, however this is not due to design, but simply a property of the way in which the scene is structured in this particular run. The execution time required to explore the scene and to recognize the objects shown in Fig. 10 was 31.5 minutes.

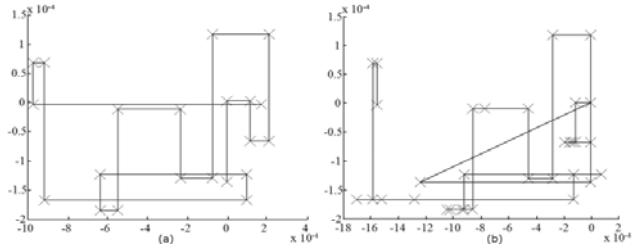

Fig. 9. The left (a) and right (b) camera position traces. Each trace is plotted in the actuator coordinate-space of that camera. Note the different horizontal axes values. Both traces begin at (0, 0).

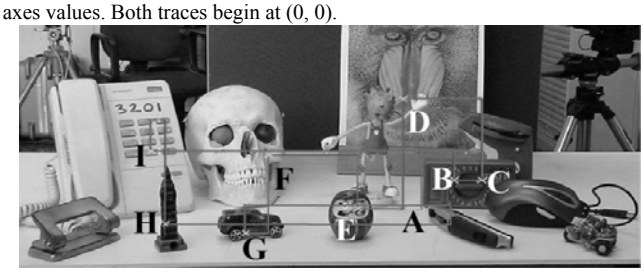

Fig. 10. The trace of one camera positing overlaid onto a photograph of the scene.

Fig. 11. The scene for the last function of the gaze control experiments.

To verify the ability of the system to exploit unrecognised scene elements to guide exploration, a scene was constructed that contained two known objects, widely separated from each other while positioned in a highly cluttered scene shown in Fig. 11. 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 300mm between them was used.

The search mode of the system was invoked and allowed to run until the objects were correctly found. The system performed 40 saccades to find both objects (Fig.12(a) and (b)) in the scene. Fig 13(a) and (b) shows the actuator-space motion trace of both cameras.

The above results produced by this first pilot study suggest that we have demonstrated the ability of the system to form a useful search pattern and recognize objects with reasonable accuracy by testing the gaze control system against a complex, cluttered scene.

VII. Conclusions

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

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

The approach of computing the vergence signal that drives the binocular camera pair, based on finding the highest feature density peak within a SIFT derived disparity histogram, proved 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. 12. Anaglyph of recognized verged objects (3rd case of vergence): (a) A lion-toy, the first object recognized (b) a skull, the second object.

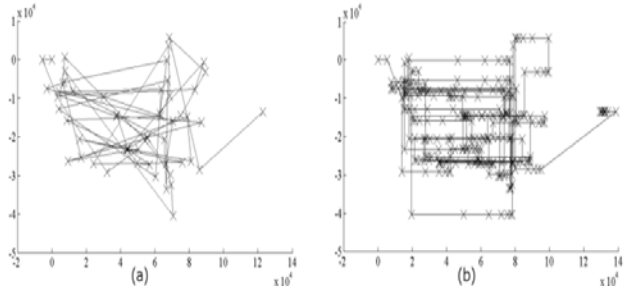

Fig. 13. The left (a) and right (b) camera position traces of the last experiment. Each one is plotted in coordinate-space of the cameras.

Saccade selection drives our gaze control system (section V) and likewise adopts SIFT keypoints as the basis of

attention and inhibition of return mechanisms.

Additionally, a functional implementation of a standard SIFT object recognition module has been embedded in a conceptually uncomplicated manner. In the preliminary results presented, seven of nine objects were recognized in a highly cluttered scene. Accordingly, the system was able to demonstrate centring its gaze on each detected object of interest in each fixation of the camera-pair, as expected. In addition, we demonstrated that the implemented 'searchlight-metaphor' of visual attention could navigate between two widely separated known objects embedded within unknown clutter.

Fig. 14. The scene and the robot head used in the experiment illustrated in Fig. 11.

Hence, 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 recognized objects based on visual cues. All these characteristics are combined in a computationally parsimonious manner using SIFT descriptors. Fig. 14 shows the experimental configuration of the robot head cameras and the scene.

It must be emphasised that we have presented the first preliminary results of validating the active binocular vision system reported here. The next objective is to perform a more complete validation involving a wider range of scenes and randomised initial fixation points in order to generate sufficient statistics to characterise reliably the performance of the system.

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 [22] 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 [23] and use of 2.5D SIFT [14] features in conjunction with 2D SIFT features to improve object identification and 3D pose recovery.

REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60(2), pp. 91–110, 2004.

[2] P. Mowforth, J. Siebert, Z. Jin, and C. Urquhart, "A head called Richard," in Proceedings of the British Machine Vision Conference 1990, Oxford, UK, p.p. 361-366, 1990., pp. 361–366, N/A, 1990.

[3] D. Betsis and J. Lavest, "Kinematic calibration of the KTH head-eye system," in ISRN KTH, 1994.

[4] L. Natale, G. Metta, and G. Sandini "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," Robotics and Autonomous Systems (39), 2002, pp. 87-106.

[5] J.-O. Eklundh and M. Björkman, "Recognition of objects in the real world from a systems perspective," Kuenstliche Intelligenz, vol. 19, no. 2, pp. 12–17, 2005.

[6] A. Bernardino and J. SantosVictor, "Binocular tracking: integrating perception and control," In IEEE Transactions on Robotics & Automation, vol. 15, no. 6, pp. 1080–94, 1999.

[7] Y. Yeshurun and E. L. Schwartz, "Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pp. 759–767, 1989.

[8] T. A. Boyling, "Active Vision for Autonomous 3D Scene Reconstruction". PhD thesis, University of Glasgow, 2002.

[9] T. A. Boyling, and J. P. Siebert, "A Fast Foveated Stereo Matcher", Conference on Imaging Science Systems and Technology, 2000, Las Vegas, pp. 417-423

[10] L.S. Balasuriya and J. P. 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, N/A, 2006.

[11] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 4, pp. 509–522, 2002.

[12] T. Gevers and A. Smeulders, "Pictoseek: Combining color and shape invariant features for image retrieval," Image Processing, IEEE Transactions on, vol. 9, pp. 102–119, January 2000.

[13] D. Kragic, M. Bjorkman, H. I. Christensen, and J.-O. Eklundh, "Vision for robotic object manipulation in domestic settings," Robotics and Autonomous Systems, vol. 52, pp. 85–100, July 2005.

[14] T. W. R. Lo, J. P. Siebert and A. F. Ayoub, "An Implementation of the Scale Invariant Feature Transform in the 2.5D Domain", 10th International Conference on Medical Image Computing and Computer Assisted Intervention, Brisbane, Australia, October 29th, 2007.

[15] M. Bjorkman and J. Eklundh, "Attending, foveating and recognizing objects in real world scenes," in Proceedings of British Machine Vision Conference, 2004.

[16] C. Westelius, "Preattentive Gaze Control for Robot Vision". PhD thesis, Linköping University, 1992.

[17] R. Milanese, "Detection of salient features for focus of attention," in Proceedings of the 3rd meeting of the Swiss group for artificial intelligence and cognitive science, 1991.

[18] Forssen, P. "Learning Saccadic Gaze Control via Motion Prediction," CRV (00), 2007, pp. 44-54.

[19] R. Brooks, "How to build complete creatures rather than isolated cognitive simulators" Architectures for Intelligence, vol. Kurt VanLehn, pp. 225–240, 1991.

[20] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," Pattern Recognition, vol. 13, pp. 111–122, 1981.

[21] A. A. McDougall, "Interfacing a Robot Head in MATLAB". MSc Thesis, University of Glasgow, 2004.

[22] Sinha, S., Frahm, J. and Pollefeys, M. "GPU-based Video Feature Tracking and Matching"(TR06-012), Technical report, University of North Carolina at Chapel Hill, 2006.

[23] J.P. Siebert, J.P. and S.J. Marshall, "Human Body 3D imaging by speckle texture projection photogrammetry", Sensor Review, Volume 20, No 3, pp 218-226, 2000.