A Head Called Richard

P. Mowforth*, P. Siebert†, Z. P. Jin* & C. Urquhart‡

*The Turing Institute, George House, 36 North Hanover Street, Glasgow, Scotland, G1 2AD;
†BBN Systems and Technologies, Edinburgh, UK, EH14 4AP;
‡BBN and Dept. of EEE, Heriot Watt University, Edinburgh, UK, EH1 2HT

This paper describes two preliminary experiments concerned with the construction of a robot head. The initial design and research is aimed at producing a system with two cameras and two microphones on a system capable of operating with the same degrees of freedom and reflex times as its biological counterpart. Whilst the primary goal of the project is develop an anthropomorphic system with the sensory reflex capabilities of a human head, the system will also contain some non-anthropomorphic components. The most obvious of the non-anthropomorphic components is a spatially and temporally programmable light source. Some preliminary results are presented.

Whilst focus and aperture have been readily automated in devices such as the camcorder, the automation of vergence, tracking and the basic orientation reflexes are not well advanced. Such built-in reflexes are critical components of any sophisticated vision system. Without a vergence or slow tracking system, stereo and motion solutions are limited to a narrow operational range involving the limits of data fusion. Gaze control allows the system to direct attention, for example, to get a particular view which helps reconcile some ambiguity in an otherwise static system. Gaze may also be used to provide closed-loop tracking which may help simplify the computation of intrinsic image information.

There have been few attempts to build vision systems that include cameras which can be motor driven on some of their degrees of freedom. Probably the best two examples are those at Rochester [7] and at Munich [6]. Each of these two systems are strictly limited in terms of their functionality. Early versions of the Rochester system used a vergence control system driving the cameras in opposite directions so as to maintain coincidence between two images whilst the Munich system allows XY gaze control of a single camera directed at a target feature.

The integration of other sensory modalities into orientation reflexes for directing, for example, a vision system has also received little attention. Systems have been built which attempt to localise sound sources and use this information to direct autonomous behaviour (Brooks, 1989) but these reflexes have not been coupled with vision systems for integrated tracking or gaze control.

This paper outlines several of the ideas behind a project to build a robot head and provides some preliminary results for one of the reflex systems as well as giving some details of its integrated texture projection system. This novel projection system allows the head to project illumination onto a scene with computer control in both temporal and spatial (textural) domains.

VERGENCE CONTROL

Vergence control is necessary to bring images captured from two locations in space into approximate correspondence. This may be used directly to provide information about range as well as providing a stereo process with favourable starting conditions. Stereo matching algorithms contain fusional limits for the maximum allowable disparity; the vergence mechanism simply attempts to maximise the amount of the image pair which can then be fused. Rather than provide a globally best vergence signal from a stereo pair, vergence mechanisms typically attempt to find correspondence between small subsets of the images, i.e. central foveas [3].

Whilst one important requirement for the project is real time execution of the vergence mechanism, the requirement for precision is not as high as for the stereo process.

Based on the above understanding, we have developed a vergence control algorithm which fulfills the above

BMVC 1990 doi:10.5244/C.4.64

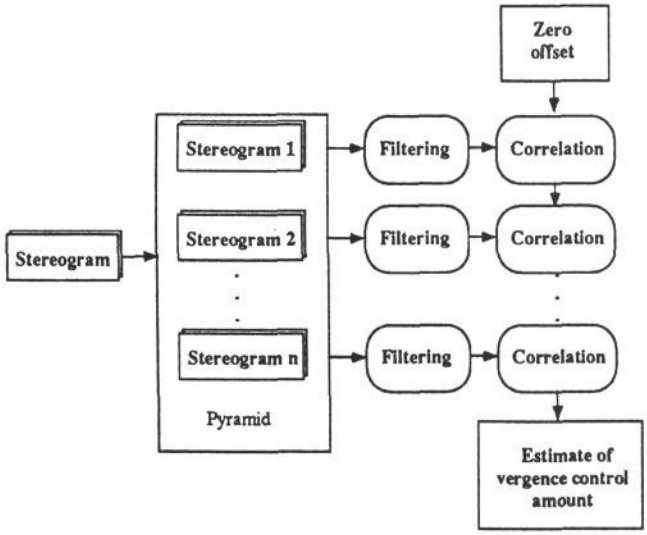


Figure 1: Software architecture of the vergence control algorithm.

mentioned requirements.

Algorithm

The algorithm design is strongly constrained by the requirement for fast execution time. To achieve this, a pyramid image architecture has been adopted. Figure 1 provides a summary of the software architecture for the vergence control algorithm.

First a pyramid is built from the input stereogram. Then, the image pair is blurred by a small sigma Gaussian function at each pyramid level. This filtering is necessary in order to achieve left-right matches at each level. Next, a small fixed size window is applied to the left and right images, again at each level. The window is able to move to the left or to the right. Correlation is now carried out between the image pair masked by these windows as a search operation. The search starts from the image pair at the top of the pyramid using a zero offset of the windows. By moving the windows either to the left or to the right, the best match is looked for and obtained which then serves as the starting offset of the windows at the next lower level of the pyramid and another round of search begins. This process continues until the best match has been found at the bottom of the pyramid. The window offset which leads to this best match is used for vergence control.

The progressive focusing of the solution through a series of octave separated levels is effectively a coarse-to-fine fovea. The algorithm is based on the Multiple Scale Signal Matching algorithm developed for matching in stereo and motion computation [5]. Like *MSSM*, the algorithm also produces a confidence value accompanying its estimate of vergence.

Figure 2: A natural corridor scene.

Performance

The algorithm is programmed in C and runs on a sun SPARCstation 1. For a 128×128 8 bits per pixel image pair, the typical time consumed by different part of the program is 20-30ms to read in the image pair, 20-30ms to build the pyramid, 30-40ms for matching and searching and 10-20ms for the other miscellaneous operations. The total time taken to produce a vergence estimate is 80-120ms. The algorithm is of pixel accuracy.

The algorithm has been tried out on a natural corridor scene and a random-dot image. From each of these two images, two, partially overlapping sub-images were extracted with known amounts of offset. A third test was performed where various amounts of Gaussian noise was added to the corridor stereogram.

Figure 2 is a natural corridor scene which served as a base image from which stereograms of different disparity were derived.

Figure 3 is a stereogram derived from the natural corridor image.

Figure 4 is the same stereogram as above but with added Gaussian random noise.

Figure 5 a is a random-dot image; b is a stereogram derived from a.

Figure 6 shows that with a natural image pair, the algorithm may properly function in the symmetric interval from -51 to 51 pixels. With added noise, the confidence

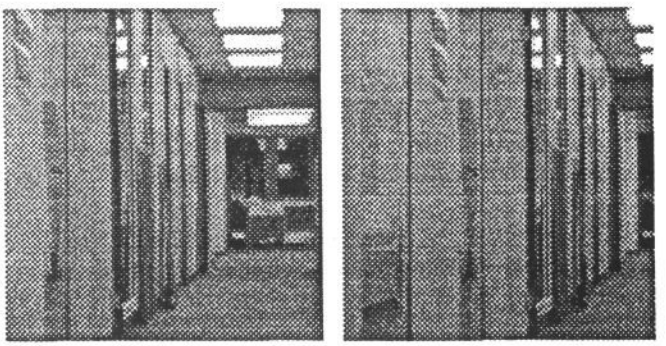


Figure 3: A stereogram derived from the natural corridor scene image.

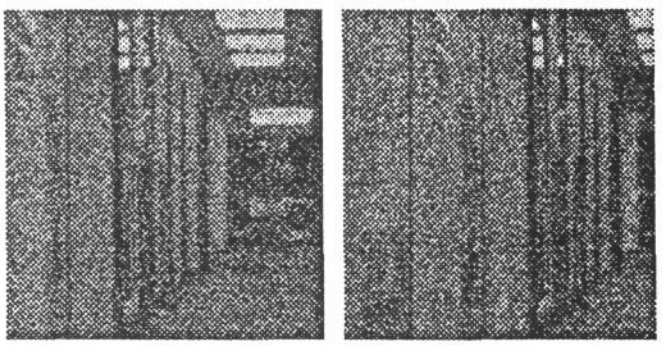
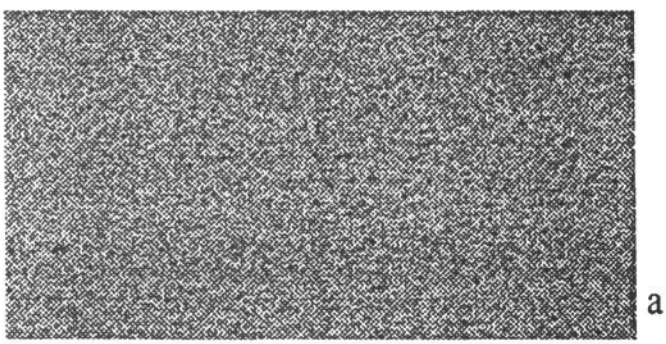
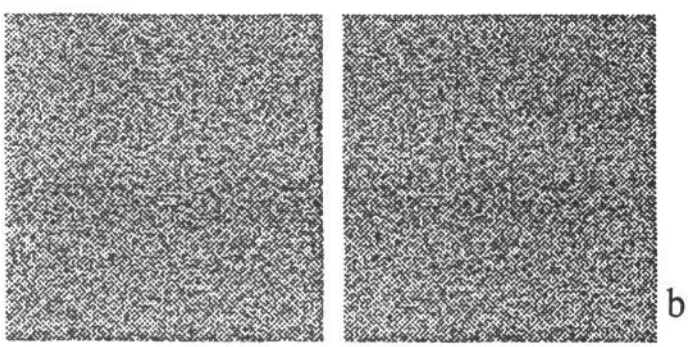


Figure 4: A noise corrupted natural corridor scene stereogram.



a



b

Figure 5: a random-dot image. b random-dot stereogram derived from a.

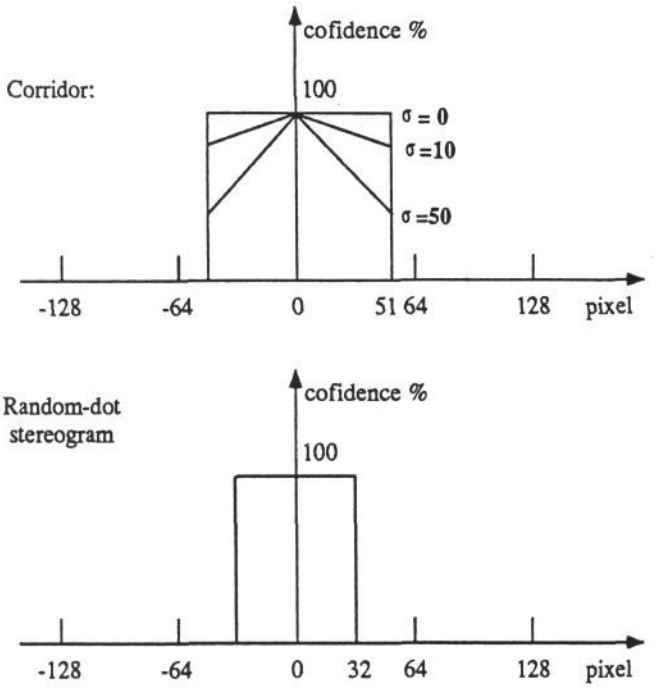


Figure 6: Algorithm performance - see text for details.

value fell according to the amount of noise present. The estimate of error was 2 pixels for a disparity of 51 pixels and noise sigma of 50.

INTEGRATED DYNAMIC PROJECTION

All vision systems that are intended to sense objects that do not intrinsically radiate their own light energy require some form of illumination source by which to sense such objects. Indeed, the majority of vision installations utilise some form of artificial illumination when operating indoors. Under such circumstances we are presented with the opportunity of utilising a controlled illumination source that may be structured to enhance the operation of any vision algorithms executed by the robot head.

Accordingly, a novel feature of the robot head design is the incorporation of a programmable projection based illumination source. This source will illuminate the scene using a steerable projector that tracks the motion of the stereo cameras to ensure that the projector field of illumination remains congruent with the fields of view of the stereo cameras [1].

Integration of Active and Passive Sensing

Active vision techniques typically employ some form of structured illumination (e.g. light striping, moire fringing etc. [2]) and rely upon interpreting contrast differences provided by this illumination to compute surface range estimates. By definition, active techniques employ interpretation algorithms that are intimately coupled to the structure of the adopted lighting source and are accordingly *ad hoc* in nature, i.e. incapable of operating in a purely passive mode and thereby utilising unstructured lighting.

We intend to effect the integration of active and passive sensing techniques by providing controlled illumination to bear on specific interpretation problems that give rise to interpretation ambiguity, e.g. when the illumination configuration details are unknown and shading analysis is attempted. Previous attempts to combine active and passive sensing modalities have been reported [10], however these simply combine the results of passive interpretation with structured illumination based ranging. Our approach has been to utilise specific modes of illumination suited to particular passive interpretation tasks. When shading analysis is being performed, a simple illumination field of known characteristic can be provided. When stereo interpretation is being performed, the projector can be switched to illuminate the scene with a Gaussian texture field. We have previously reported [8] that the performance of the scale space stereo algorithms can be dramatically improved by the projection of textured light onto the scene. Hence a projection mode would be selected according to the interpretation task at hand. Indeed interpretation algorithms that perform the tasks of stereo and shape-from-shading exhibit complementary performance characteristics depending upon the density or nature of intrinsic surface texture patterns on the surfaced of objects contained in the scene.

Several advantages can be gained by adopting this approach to integrating active and passive vision techniques. Controlled programable illumination makes possible the investigation, development and utilisation of passive interpretation algorithms under ideal sensing conditions. From the perspective of vision research this is highly expedient as individual sensing mechanisms become increasingly better understood in isolation while their combination remains difficult. From a systems perspective, the robot head will be capable of operating, albeit with degraded performance, in a passive mode using unstructured lighting, e.g. outdoors or when subject to extraneous illumination fields, and hence will be inherently robust in operation.

In short, the system will be based passive sensing techniques but can also be enhanced in operation by the use of controlled active illumination. This paradigm compliments the major project goal of utilising a dynamic approach to visual interpretation, in the sense of employing sensor platform mobility or motion to resolve interpretation ambiguity [1].

Enhancement of Stereo Using Textured Light

As stated above, we have undertaken a number of experiments that investigate enhancing a passive scale-space stereo algorithm, MSSM [5] by illuminating the imaged scene with a Gaussian noise field, i.e. textured light.

There are four main problems which can lead to match failures in the stereo matching process [4]:

- Photometric variations at a point viewed from two angles.
- Lack of texture in a region.
- Presence of repetitive texture patterns.
- Occlusion between camera views of some points in stereo pair.

The problem of photometric variations is overcome in the MSSM algorithm by utilising a covariance measure to search for disparity estimates that does not rely upon correlating absolute grey-level intensities. Random texture projection both addresses the problem of match failure where the intrinsic texture is sparse and counteracts the matching ambiguity caused by repetitive texture patterns.

Projected random texture gives rise to an extended spectral distribution that provides matching energy at all scales in scale space. The low frequency limit of the texture spectral distribution sets the maximum scale size and hence disparity that can be matched unambiguously. The high frequency limit of the spectral distribution sets the smallest scale size and hence the final degree of scale space match refinement. Ideally the texture spectrum should be constant or "white" to achieve

[1] A broad analogy might be the lamp on a miner's helmet!

Figure 7: Left image captured using random texture projection. Right image captured using white light as normal.

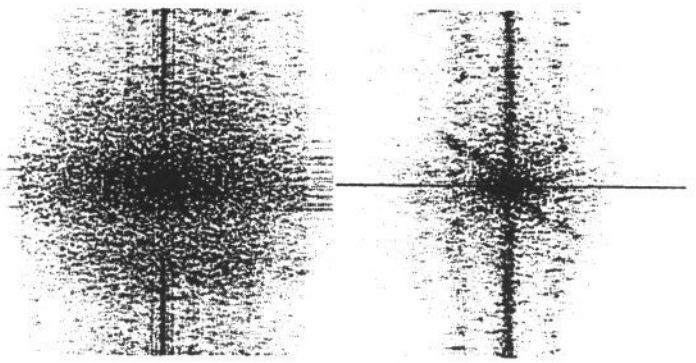


Figure 8: 2DFT power spectrum of images from figure 7 respectively.

the optimum autocorrelation characteristics. However, in practice random texture projection can provide "random dot stereograms" of the imaged scene and thereby counter both lack of texture and the effects of repetitive intrinsic surface markings.

We have investigated in detail [9] the optimum spectral characteristics of the projected textured field (for our current camera/projection apparatus) in terms of the texture size and aspect ratio that yields the greatest horizontally biased spectral energy envelope in the sensed image. Figure 7 shows the left hand images of stereo pairs with and without texture projection respectively. In this example the texture was provided by a simple overhead projector containing a laser printed texture transparency. Figure 8 shows the increased spectral energy due to texture projection, comparing the DFT power spectrum of these images. Figure 9 shows wire-frame reconstructions of the range image surface obtained by matching the texture projected stereo pair using MSSM.

Briefly, the final cause of match failure listed, occlu-

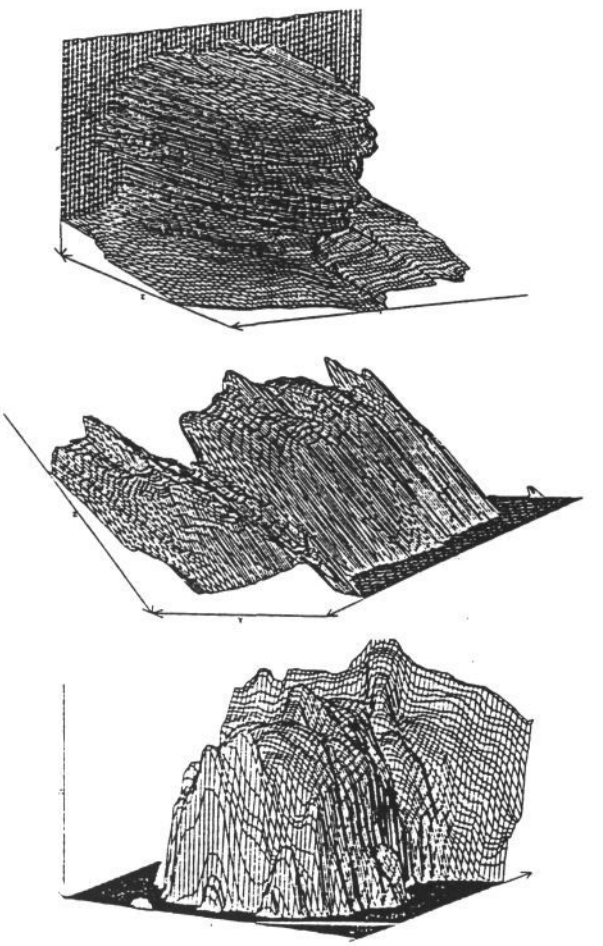


Figure 9: Wire-frame reconstructions of range information recovered from textured projected stereo pair (left image of figure 7.)

sions, can be detected when utilising active texture projection as low confidence matches. We have investigated [9] detecting occlusions using both this effect and the shape of the correlation search graph itself.

Discussion

This paper has demonstrated two applications of Multiple Scale Signal Matching as part of a single project. The work has helped outline a framework for the development of an anthropomorphic head capable of a number of integrated reflex functions. Because it is recognised that there will be occasions where the head is unable to provide good measures of its environment, the system is enhanced with the capability for active illumination designed to help facilitate its otherwise passive systems.

It is intended to extend the use of the signal matching technology to provide fast matching between microphone signals from the robot head. The purpose here will be to provide orientation information so enabling sounds to provide a direction reflex for vision.

Acknowledgements: This work is supported under a UK IED project funded in part by the Department of Trade and Industry, by the Esprit DIMUS project, The Turing Institute and BBN Systems and Technologies, UK.

References

[1] J. Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. In *DARPA image understanding workshop*, pages 552–573, Los Angeles, LA, Feb 1987.

[2] P. Besl. Active, optical range imaging sensors. *Machine vision and applications*, 1:127–152, 1988.

[3] R. H. S. Carpenter. *Movements of the eyes.* Pion, London, 1977.

[4] S. Cochran and G. Medioni. Accurate surface description from binocular stereo. In *DARPA image understanding workshop*, pages 857–869, Palo Alto, LA, May 1989.

[5] Z. Jin and P. Mowforth. A discrete approach to signal matching. Research Memo TIRM-89-036, The Turing Institute, Glasgow, UK, January 1989.

[6] B. Mysliwetz and E. D. Dickmanns. A vision system with active gaze control for real-time interpretation of well structured dynamic scenes. In L. O. Hertzberger and F. C. A. Groen, editors, *Intelligent Autonomous Systems Conference*, pages 477–483, Amsterdam, 1987. North Holland.

[7] T. J. Olson and R. D. Potter. Real-time vergence control. Research Report TR-264, University of Rochester, Computer Science Dept, Rochester, New York, 1988.

[8] P. Siebert and C. Urquhart. Active stereo: textured enhanced reconstruction. *Electronics letters*, 26(7):427–430, March 1990.

[9] C. Urquhart. An investigation into active and passive methods for improving the performance of scale-space stereo. 5th year project report, BBN Systems and Technologies, Edinburgh, UK, MArch 1990.

[10] Y. F. Wang and J. K. Agargwal. Integration of active and passive sensing techniques for representing 3d objects. *IEEE Trans. Robotics and Automation*, 5(4):460–470, 1989.