In L. Itti, G. Rees & J. Tsotsos, eds, Neurobiology of Attention, Academic Press/Elsevier, San Diego (in press).

Attentive wide-field sensing for visual telepresence and surveillance [⋆]

James H. Elder [a,*] Fadi Dornaika [b] Bob Hou [a] Ronen Goldstein [a]

[a] *Centre for Vision Research, York University, Toronto, Canada*

[b] *Heudiasyc Laboratory, CNRS, University of Technology of Compiegne, France*

Abstract

Physical and computational constraints limit the spatial resolution and field-of-view (FOV) achievable in any single sensor. In the human eye, a compromise has evolved in which resolution is high near the optical axis, falling off with eccentricity. The success of this design depends upon visual mechanisms for rapidly detecting interesting events in the periphery and motor mechanisms for accurately redirecting the fovea to peripheral targets for more detailed analysis.

Prototype artificial visual systems based conceptually on these principles are being developed in our laboratory. We partition visual sensing into two components. A *pre-attentive* component provides a large, fixed FOV at low resolution, allowing detection of events of interest over an entire visual environment. An *attentive* component provides a much smaller, shiftable FOV at high resolution, designed to recognize and interpret events detected by the pre-attentive system. These prototype sensors serve as useful testbeds for computational models of visual attention, and may lead to new telepresence and surveillance technologies.

Key words: spatial resolution, registration, motion, tracking, periphery, saccadic control, trans-saccadic integration, memory

[⋆] This work was supported by the Ontario Centre for Research in Earth and Space Technology and by the Canadian Institute for Robotics and Intelligent Systems

[*] Centre for Vision Research, York University, 4700 Keele Street, North York, Ontario, Canada M3J 1P3. jelder@yorku.ca, elderlab.yorku.ca

I Introduction

Typical machine vision systems employ a single sensor with relatively small FOV. This can be sufficient for narrow applications, e.g., assembly line inspection, where the location of the object of interest is strongly constrained. However, machine vision research is increasingly concerned with more human-like visual tasks, e.g. surveillance of a large, open environment, and this has led to increasing interest in wide FOV machine vision sensing, particularly panoramic sensing (e.g. Oh and Hall (1987), Yagi and Kawato (1990), Ishiguro, Yamamoto, and Tsuji (1992), Nayar (1997)).

The advantages of a wide FOV for surveillance and teleconferencing applications are clear, however these advantages come at the expense of resolution. Switching from the 14 deg FOV of a typical lens to the 360 deg FOV of a panoramic camera results in a 26-fold reduction in linear resolution. For a standard 640 × 480 pixel camera, horizontal resolution is reduced to roughly 0.5 deg/pixel, a factor of 60 below human foveal resolution.

The human visual system has evolved a bipartite solution to the FOV/resolution tradeoff. The FOV of the human eye is roughly 160×175 deg - nearly hemispheric. Central vision is served by roughly five million photoreceptive cones that provide high resolution, chromatic sensation over a five degree field of view, while roughly one hundred million rods provide relatively low-resolution achromatic vision over the remainder of the visual field(Wandell, 1995). The effective resolution is extended by fast gaze-shifting mechanisms and a memory system that allows a form of integration over multiple fixations (Irwin & Gordon, 1998).

Variations on this architecture are found in other species. Many insects have panoramic visual systems; the springing spider, for example, has four eyes that capture movement over the entire viewing sphere and two small-FOV high resolution eyes used in predation and mating (Moller, Lambrinos, Roggendorf, Pfeifer, & Wehner, 2001).

Exploration of such heterogeneous visual architectures for computer vision is just beginning. Early attempts have focused on the design and fabrication of space-variant (foveated) sensor chips (Spiegel et al., 1989; Ferrari, Nielsen, Questa, & Sandini, 1995; Pardo, Dierickx, & Scheffer, 1997; Wodnicki, Roberts, & Levine, 1997). However, since the density of photoreceptive elements on these sensors is no greater than for regular sensors, they do not provide a resolution advantage over traditional chips, and hence do not address the fundamental resolution/FOV tradeoff.

The problem is more squarely addressed by mosaicing systems that compose mosaics from individual overlapping high-resolution images obtained by a single camera rotated about its optical centre (Szeliski, 1994; Kumar, Anandan, Irani, Bergen, & Hanna, 1995). Such systems are useful for recording high-resolution ”still life” panoramas, but are of limited use for dynamic scenes, since the instantaneous field of view is typically small. An alternative is to compose the mosaic from images simultaneously recorded by many identical cameras with overlapping fields of view. A disadvantage of this approach is the multiplicity of hardware and independent data channels that must be integrated and maintained. For example, a standard 25mm lens provides a field-of-view of roughly 14 × 10 degrees. Allowing for 25% overlap between adjacent images to support accurate mosaicing, achieving this resolution over a hemispheric field of view would require on the order of 260 cameras!

Consideration of biological visual systems inspires a more practical solution to the FOV/resolution dilemma for machine vision. Instead of employing dozens or hundreds of identical sensors, it may be feasible to employ a small number of very different sensors with complementary properties. For example, in a human-inspired machine vision system, one sensor with low resolution but large FOV may play the role of the peripheral visual system while a second sensor, with high-resolution but small FOV, serves as the fovea.

While in the human these two subsystems share a single sensor substrate (the retina) and common optics, their machine vision analogues are more easily realized as distinct physical components. Since technological constraints limit the total number of photoreceptive elements on a reasonably-priced chip, achieving fovea-like resolution in a machine sensor depends upon telephoto optics that limit the FOV. Obtaining the large FOV required by the peripheral component requires a much shorter focal length, hence separate optics.

This physical split, while necessary, introduces significant complications. The displacement of the optical centres of the two subsystems introduces parallax, making accurate correspondence of information in the two sensors nontrivial. This correspondence problem is further complicated by the disparity in resolution between the sensors. Interestingly, this problem is exactly that faced by the human visual system in attempting to accurately land a saccade on a peripheral target. The efficacy of the human saccadic system thus provides some evidence that the proposed approach may be feasible.

As for the human eye, an artificial fovea is only useful if it can be rapidly directed to important visual events in the scene. This obviously requires a fast, accurate and reliable 2D rotational (e.g. pan/tilt) motion platform. But it also requires a set of fast visual mechanisms capable of identifying and localizing these important visual events at relatively low resolution.

In this article, we will outline recent research in our laboratory to design, build and evaluate attentive wide-field sensors based on these principles. These prototypes, while presently at an early stage of development in terms of their attention capabilities, are fully-functioning real-time active vision systems, and as such can serve as interesting testbeds for theories of attention such as those proposed in this volume.

Attentive wide-field sensing may also be useful in a number of application areas. In telepresence applications, the attentive sensor can be directed toward events or activities of interest to a remote human observer, while low-resolution data from the pre-attentive sensor continues to provide the observer with a sense of context or 'situational awareness' (Geisler & Perry, 1998). Recent work with saccade-contingent displays (Loschky & McConkie, 1999) has shown that video data viewed in the periphery of the human visual system can be substantially subsampled with negligible subjective or objective impact. While our prototype sensors are not eye-slaved, this prior work suggests that attention-contingent sampling for human-in-the-loop video is feasible and potentially useful.

Attentive wide-field sensing may also be useful in autonomous surveillance applications. Events detected in the pre-attentive sensor may generate saccade commands to allow more detailed inspection/verification at the higher resolution of the attentive sensor.

II Physical Design

Fig. 1 shows two wide-field attentive vision systems designed and built in our laboratory. Both prototypes employ two sensors. The pre-attentive sensors have large, fixed FOVs and play the role of the peripheral visual system. The attentive sensors have small FOVs but are mounted on pan/tilt platforms, allowing gaze to be rapidly redirected. In both systems, the attentive sensor is based on a 25mm lens with an FOV of roughly 13 deg, providing an angular resolution of 1.2 arcmin, roughly half the resolution of the human visual system. An example attentive image is shown in Fig. 2(c).

The principal difference between the two systems lies in the pre-attentive sensor. In Sensor 1 (Fig. 1(a)), the pre-attentive sensor is based on a parabolic catadioptric video sensor (Nayar, 1997) purchased from Cyclovision Technologies (now RemoteReality$^{\text{TM}}$). This sensor provides a panoramic FOV (Fig. 2(a)), permitting surveillance from the centre of a visual environment. In Sensor 2 (Fig. 1(b)), the pre-attentive sensor is based on a more conventional wide-angle lens with 2.1mm focal length, providing a FOV of 130 degrees (Fig. 1(b)), somewhat less than the human eye. This FOV is well-suited to surveillance of a visual environment from a corner location. The second system is also designed to be lighter, faster, cheaper and smaller.

The raw video frames from the pre-attentive sensors are geometrically distorted (Fig. 1(a-b)). These distortions are caused by the curvature of the parabolic mirror in Sensor 1 (Nayar, 1997), and by radial and tangential error of the preattentive lens in Sensor 2. In both systems, these distortions are removed online in software prior to further processing.

In both sensors, the motion platforms have been designed so that the axes of rotation intersect approximately at the optical centre of the attentive visual system, thus minimizing motion parallax. However, the physical separation of the pre-attentive and attentive sensors creates significant parallax between the two video streams. In Sensor 1, the optical baseline is 22 cm. In sensor 2, our efforts at compressing the package resulted in a baseline of only 7.5 cm, much closer to the average human interocular separation of 6 cm. This reduction is important in limiting the search region required for online, dynamic fusion of the pre-attentive and attentive visual streams.

III Fusion

In telepresence systems, continuous data from remote sensors are displayed for a human observer. In autonomous surveillance systems, these data are interpreted directly by computer vision algorithms. In either case, incorporating wide-field attentive sensing raises the problem of data fusion. For example, in a remote surveillance system monitored by a human security specialist, pre-attentive and attentive visual data could be displayed on separate monitors. But this would require the observer to shift their gaze back and forth between the displays, mentally integrating the two disparate video streams. Accurate registration of the visual data from the two streams raises the possibility of integrating the visual data in a single seamless window, lowering the workload for the human observer.

When an event of interest is detected in the pre-attentive stream, either by a human observer or by an automatic algorithm, the attentive sensor must be rapidly and accurately redirected to the corresponding location. Further automatic interpretation of a pre-attentively detected event may also depend upon an estimate of the location of the event within the attentive FOV based on pre-attentive data. These computations depend upon accurate registration of pre-attentive and attentive visual streams.

Accurate registration in turn depends on computing correspondences between the two visual streams. In a typical computer vision stereo sensor, the intrinsic parameters of the cameras are identical and their extrinsic parameters are fixed and can be estimated in advance. This greatly simplifies correspondence, facilitating feature comparison and restricting the search space to fixed epipolar lines.

Computing correspondence for wide-field attentive sensing is more complicated, since the attentive sensor is not fixed, and the intrinsic parameters of the two sensors are very different. On the other hand, the system is not completely unconstrained: the pre-attentive sensor is fixed, and the attentive sensor has only 2 rotational degrees of freedom. These constraints, coupled with non-uniform prior distributions on the distance of surfaces from the sensor, simplify the fusion problem considerably.

Most critically, visual scenes are coherent: the distance of visible surface points projecting to neighbouring sensor pixels are highly correlated. To exploit this property, we model the scene as piecewise planar, and approximate the correspondence between attentive and pre-attentive coordinate frames using a table of 2D projective mappings (homographies), indexed by the pan/tilt coordinates of the attentive sensor.

III.A Calibration

For a static scene, 8-parameter homographies can be estimated using a manual technique and a simple calibration rig. Pre-attentive/Attentive image pairs of the scene are captured at regular pan/tilt intervals of the attentive sensor. 12-16 point pairs are manually localized in each image pair, and the corresponding least-squares homographies are estimated using standard techniques. The result is a table of homographies, indexed by pan/tilt coordinates of the attentive sensor. In operation, given arbitrary pan/tilt coordinates, the corresponding homography can be estimated from the table using bilinear interpolation.

For each image pair we also store the projection of the attentive image centre into pre-attentive coordinates. This allows construction of a second table, indexed by pre-attentive coordinates, that provides the pan/tilt coordinates required to centre the attentive sensor at a specific pre-attentive location. Bilinear interpolation is used to generate a pan/tilt command given an arbitrary saccadic target in pre-attentive coordinates, generated either by human or machine attention algorithms.

This fixed system of coordinate transforms between the two sensors will be accurate only if viewing distance is large or if dynamic variations in depth are small relative to viewing distance. Since neither condition holds in general, we calibrate the system for intermediate distances and then use a dynamic fusion algorithm to obtain more precise registration during operation.

III.B Dynamic Fusion

Automatic image registration is a well-studied problem, but the problem of registering pre-attentive and attentive images presents new challenges. The main difficulty is the extreme difference in resolution between the sensors (from 10:1 to 16:1 linear resolution ratio in our prototypes): Fig. 1(c-d) shows an example of the same subject viewed in attentive and pre-attentive streams. The problem is complicated further by resolution inhomogeneities of the pre-attentive sensor (more pronounced for Sensor 1 than Sensor 2).

We address the problem in a three-stage approach. In the first stage, we use the pan/tilt coordinates of the attentive sensor to index the calibration table and determine the pre-computed homography relating the two sensor images. Fig. 2(e) shows an example of fusion based on pre-calibration. Since the subject is out of the local plane of calibration, there is significant registration error (note, for example, the mislocation of the shoulder of the subject on the right). Nevertheless, the translation and scaling parameters of the pre-calibrated homography are good enough to seed a search region within which the true homography should lie, within predefined confidence limits.

In the second stage, we compute a coarse registration within this search region using a parametric template matching technique on a multi-resolution representation of the attentive image that accommodates the inhomogeneity of the pre-attentive image. This provides an estimate of the translation and scale factors between the two streams.

In the third stage, this coarse registration is used to bootstrap a refinement process in which a full 2D projective mapping is computed. We have studied two different refinement methods. The first recovers point matches between either high-gradient pixels or interest points, and then uses a robust estimation procedure (RANSAC (Fischler & Bolles, 1981)) to estimate the complete 2D projective transformation. The second method directly estimates geometric and photometric transforms between the images by minimizing intensity discrepancies. In empirical evaluations we have found the second, direct method to be superior in both accuracy and reliability. An example of the resulting fusion is shown in Fig. 2(f). Note that visual features in the attentive and pre-attentive streams join smoothly at the boundary of the attentive image. Circular vignetting has been used to further smooth the transition between the streams. All coordinate transforms and blending are performed using standard PC graphics hardware using OpenGL, allowing the systems to operate at roughly 17 frames per second (fps).

IV Saccadic Control

In telepresence mode, the gaze of the attentive sensor can be controlled by a human observer monitoring the fused display. When an interesting event is noted in the pre-attentive field, the observer uses a point-and-click mouse interface on the pre-attentive field to drive the attentive sensor to the event location.

In automatic surveillance mode, saccades are determined by automatic attention algorithms. In human vision, one of the most powerful exogenous attention cues is visual motion (ref). A fundamental issue in motion detection is how to select the spatial scale of analysis. In our case, the purpose of detection is to drive the attentive sensor to the point of interest to resolve the change. Thus it is natural to match the scale of analysis to the FOV of the attentive sensor in pre-attentive coordinates. In this way, saccades will resolve the greatest amount of motion energy.

In our motion detection algorithm, successive pre-attentive RGB image pairs are differenced, rectified, and summed to form a primitive motion map. This map is convolved with a separable kernel that approximates the FOV of the attentive sensor in panoramic coordinates, and thresholded to prevent generation of saccades due to sensor noise and vibration. The location of the maximum of the resulting motion map determines the next fixation.

V Tracking and Smooth Pursuit

Human tracking data can be useful in evaluating the relative importance of simultaneous events of interest detected in the pre-attentive sensor. In a security application, for example, these data may be used to distinguish normal horizontal walking motion, from the motion of someone running or scaling a wall.

Tracking data is also important for accurate interception of events by the attentive sensor. If the velocity of the target is not taken into account in saccadic planning, the attentive sensor will perpetually lag the target. Tracking data from both the pre-attentive and attentive streams can also be used for smooth pursuit, i.e. to control attentive gaze to remain locked to the moving target.

Automatic visual tracking of human activity has been the subject of intensive research in recent years (Aggarwal & Cai, 1997). Traditional methods use conventional cameras and are based on extracting and/or matching features such as occluding contours, skin color and head regions.

For wide-field attentive sensing applications, we are particularly interested in techniques that can be applied to low-resolution imagery, allowing tracking of activity in the pre-attentive stream. At such resolutions, tracking based upon detailed modeling of the human body are unlikely to work, since individual body parts may not be resolved.

In our prototypes we have been studying tracking algorithms based upon relatively low-level features such as intensity, intensity gradients and colour. We have found that combining these multiple features leads to a tracking system that is much more robust to occlusions, scale changes and non-rigid motions than systems based upon a single feature. An example of human body tracking based on this multi-feature approach is shown in Fig. 3(a). The tracking algorithm operates in real time (17 fps).

VI Memory

What information the human visual system retains over a sequence of fixations is a subject of significant debate (e.g. Rensink, O'Regan, and Clark (1997)). There is no question, however, that humans have some forms of visual memory (iconic, short-term, long-term).

In any wide-field attentive sensing system, there is a tradeoff between spatial and temporal resolution. The pre-attentive sensor provides good temporal resolution (17 fps in our prototypes), but poor spatial resolution. The attentive sensor provides good spatial resolution, but poor temporal resolution, in that it may visit a particular portion of the visual scene relatively infrequently. By titrating the information from the two stream in different ways over space and time, one can adjust this tradeoff, and potentially adapt it to the dynamics of a particular visual environment. For example, in a completely static scene. A high resolution visual image can be built up over time using a sequence of attentive fixations tiling the viewing sphere. In a more dynamic scene, changing portions of the mosaic may be updated using more immediate low-resolution data from the pre-attentive sensor.

We have implemented a primitive version of this idea. In our attentive wide-field sensor prototypes, the display duration of attentive images from previous fixations is determined by a memory parameter. At one extreme, previous attentive data are immediately replaced by more recent low resolution data from the pre-attentive sensor. At the other extreme, a sequence of fixations builds up a persistent high resolution mosaic. In intermediate modes, attentive data from previous fixations gradually fade into more recent low-resolution pre-attentive data (Fig. 3(b)). In this way, resolution in space can be traded off against resolution in time, depending upon the nature of the application and the nature of the visual scene.

VII Future Directions

It seems likely that the the human attention and saccadic systems have co-evolved with the systematic falloff in photoreceptor density, resolution and cortical magnification as a function of eccentricity. Attentive wide-field sensing platforms, which can mimic both the high resolution of the human fovea and the wide FOV of the human peripheral visual system thus provide an interesting testbed on which to evaluate models of human attention.

It remains to be seen how effective and immersive an experience an attentive wide-field sensor can deliver in a telepresence application. Given the speed of eye movements and the intolerance of the visual system to lag, eye-slaved systems are impractical in many telepresence applications. We wish to investigate the degree to which intelligent attention algorithms and system memory can be used to provide an effective visual experience in situations where lags are significant and bandwidth is limited.

In the near-term applications are likely in more loosely coupled telepresence environments where the attentive sensor is not directly slaved to the eye. In a remote learning application, for example, an attentive sensor at the lecture site might track an instructor, keeping her face and gestures in clear view to the remote students. A raised hand might be detected by a pre-attentive sensor at the classroom site, directing an attentive sensor to provide the instructor with high-resolution video of a student who wishes to ask a question.

Surveillance applications are also feasible in the nearer term. Fairly simple face detection and tracking algorithms in the pre-attentive stream may be sufficient to guide the attentive sensor to deliver high resolution imagery of human activities. Even without effective face recognition or activity interpretation algorithms , such technologies could be useful in improving the quality of archival databases and reducing the workload of human security specialists.

References

Aggarwal, J., & Cai, Q. (1997). Human motion analysis: a review. In *Nonrigid and articulated motion workshop proceedings* (p. 90-102). San Juan, Puerto Rico: IEEE.

Ferrari, F., Nielsen, J., Questa, P., & Sandini, G. (1995). Space variant imaging. *Sensor Review*, *15*(2), 17-20.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381-395.

Geisler, W. S., & Perry, J. S. (1998). A real-time foveated multi-resolution system for low-bandwidth video communication. In B. Rogowitz & T. Pappas (Eds.), *Human Vision and Electronic Imaging, SPIE Proceedings* (Vol. 3299, p. 294-305). San Jose, CA.

Irwin, D. E., & Gordon, R. D. (1998). Eye movements, attention and trans-saccadic memory. *Visual Cognition*, *5*(1/2), 127-155.

Ishiguro, H., Yamamoto, M., & Tsuji, S. (1992). Omni-directional stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *14*(2), 257-262.

Kumar, R., Anandan, P., Irani, M., Bergen, J., & Hanna, K. (1995). Representation of scenes from collections of images. In *Proc. of the ieee workshop on representation of visual scenes* (p. 10-17). Los Alamitos, CA.

Loschky, L., & McConkie, G. W. (1999). Gaze contingent displays: Maximizing display bandwidth efficiency. *Army Research Laboratory Advanced Displays and Interactive Displays Federated Laboratory Third Annual Symposium.*

Moller, R., Lambrinos, D., Roggendorf, T., Pfeifer, R., & Wehner, R. (2001). Insect strategies of visual homing in mobile robots. In B. Webb & T. Consi (Eds.), *Biorobotics - methods and applications.* AAAI Press / MIT Press.

Nayar, S. (1997). Catadioptric omnidirectional camera. *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 482-488.

Oh, S. J., & Hall, E. L. (1987). Guidance of a mobile robot using an omnidirectional vision navigation system. *Proc. Soc. Photo-Optical Instrumentation Engineers (SPIE)*, *852*, 288-300.

Pardo, F., Dierickx, B., & Scheffer, D. (1997). CMOS foveated image sensor: Signal scaling and small geometry effects. *IEEE Transactions on Electron Devices*, *44*(10), 1731-1737.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368-373.

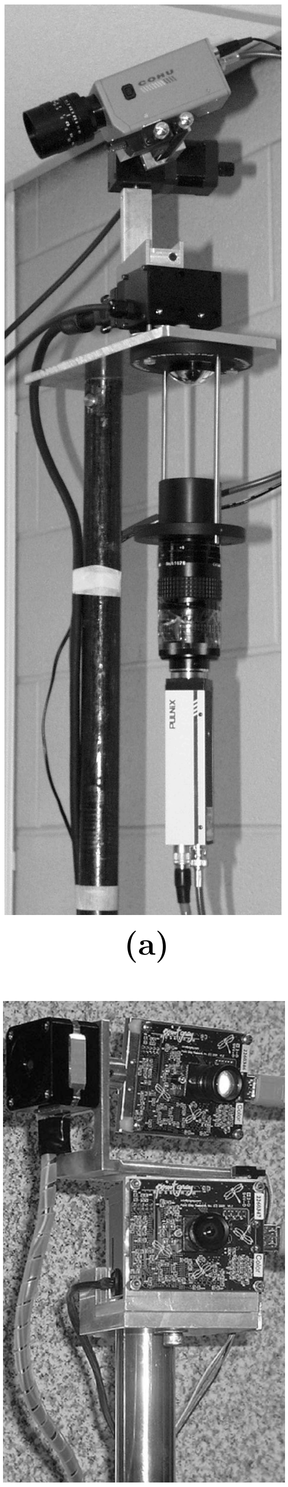
Spiegel, J. van der, Kreider, G., Claeys, C., Debusschere, I., Sandini, G., Dario, P., Fantini, F., Belluti, P., & Soncini, G. (1989). A foveated retina-like sensor using CCD technology. In C. Mead & M. Ismail (Eds.), *Analog VLSI implementation of neural systems* (p. 294-305). Boston: Kluwer.

Szeliski, R. (1994). *Image mosaicing for tele-reality applications* (Tech. Rep. No. CRL 94/2). 1 Kendall Square, Cambridge, Massachusetts: Cambridge Research Laboratory.

Wandell, B. (1995). *Foundations of vision.* Sunderland, Massachusetts: Sinauer.

Wodnicki, R., Roberts, G. W., & Levine, M. (1997). Design and evaluation of a log-polar image sensor fabricated using a standard 1. 2 um ASIC CMOS process. *IEEE Journal of Solid-State Circuits*, *32*(8), 1274-1277.

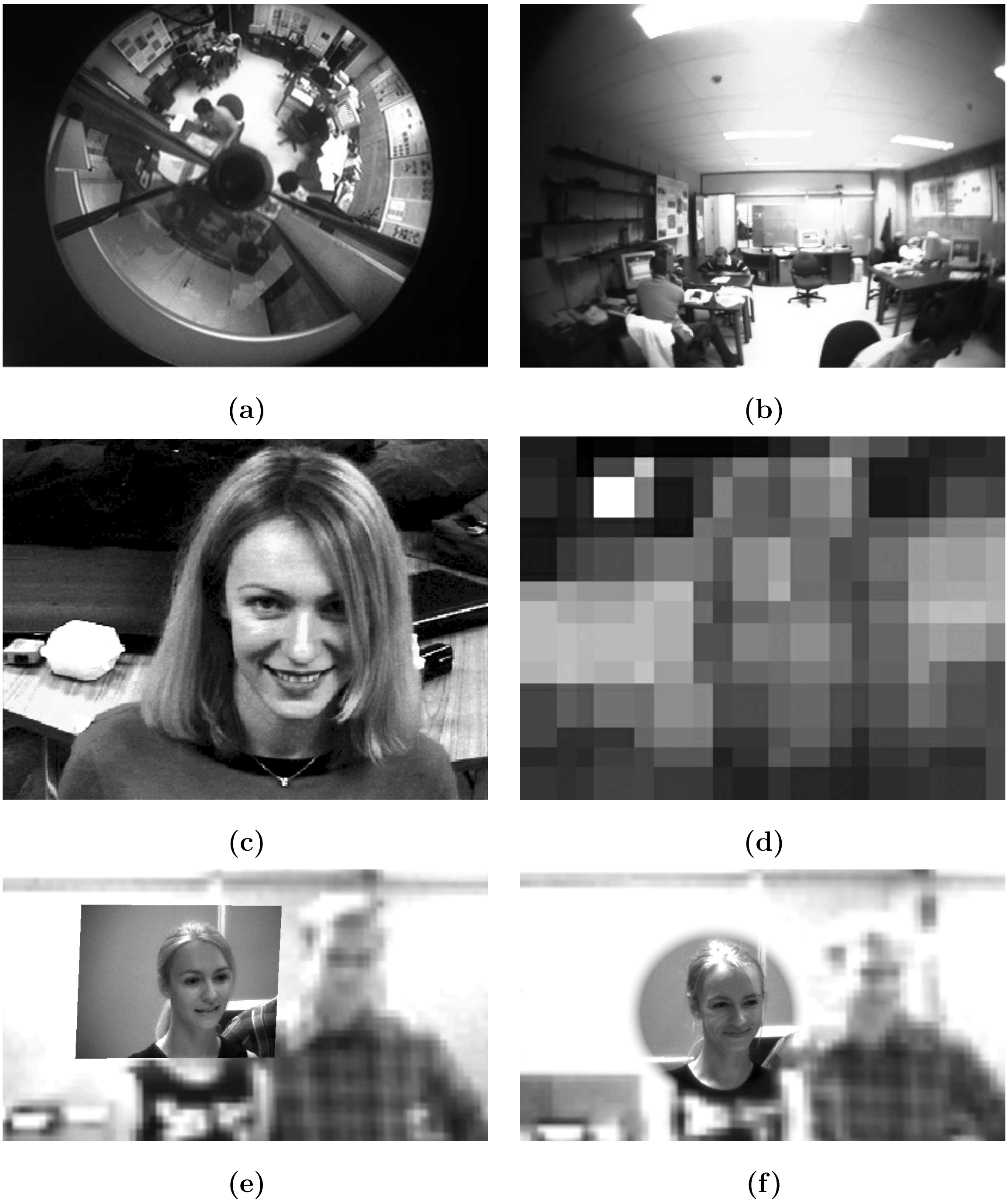
Yagi, Y., & Kawato, S. (1990). Panoramic scene analysis with conic projection. *Proc. Int. Conf. on Robots and Systems (IROS).*



(a)

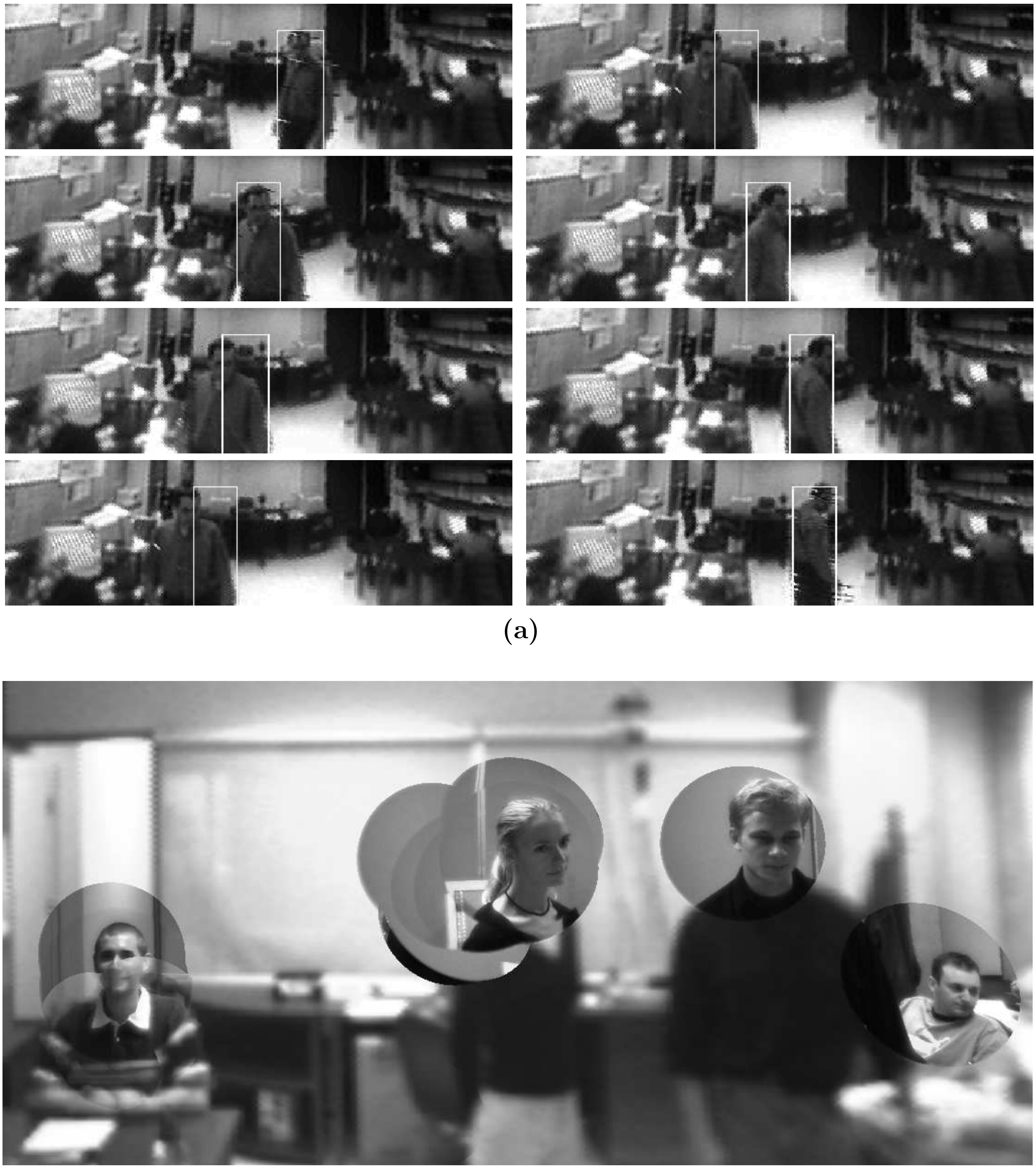
(b)

Fig. 1. Two prototype attentive wide-field sensors. **(a)** Sensor 1. Attentive camera is mounted on a gimbaled pan/tilt platform at the top of the sensor. Pre-attentive camera is mounted on the bottom, imaging panoramic view reflected from parabolic mirror. Apparatus shown measures roughly 75 cm in height. **(b)** Sensor 2. Two identical digital board cameras are mounted in close proximity on vertical axis, with he attentive sensor on top. Apparatus shown measures roughly 20 cm in height.



(a) (b) (c) (d) (e) (f)

Fig. 2. Wide-field attentive sensing imagery. **(a)** Raw, uncorrected pre-attentive image from Sensor 1. **(b)** Raw, uncorrected pre-attentive image from Sensor 2. **(c)** Example attentive image from Sensor 1. **(d)** Corresponding image from Sensor 2. Note the gross difference in resolution. **(e)** Fused imagery based on pre-calibrated homography. **(f)** Improved registration based on direct dynamic fusion algorithm. Circular vignetting has been used to smooth the transition between resolutions.



(a)

(b)

Fig. 3. **(a)** Pre-attentive tracking of human motion. Rectangle indicates estimated body position and size. **(b)** Trading off spatial and temporal resolution using a primitive form of trans-saccadic memory.