FURTHER DEVELOPMENTS ON 'HEADCAM': JOINT ESTIMATION OF CAMERA ROTATION+GAIN GROUP OF TRANSFORMATIONS FOR WEARABLE BI-FOVEATED CAMERAS

Steve Mann

MIT Bldg. E15-389, 20 Ames Street, Cambridge, MA02139
steve@media.mit.edu http://wearcam.org (http://18.85.20.100/pencigraphy)

ABSTRACT

An eyeglass-mounted camera system with wearable multimedia computer ('WearCam') was recently proposed. In particular, 'WearCam' contains two miniature cameras: one (wide-angle in *landscape* orientation) provides the overall contextual information from the wearer's perspective, while the other (telephoto, in *portrait* orientation) provides close-up details, such as faces. This 'bi-foveated' scheme was found to work well within the context of a recently proposed model of image motion characterized by a projective (homographic) coordinate transformation together with a gain transformation. Applications of 'WearCam' include personal safety (crime prevention and personal documentary), perceptual intelligence/situational awareness (in the context of personal wearable multimedia), and homographic modeling (wearable, tetherless computer-mediated reality). A pencigraphic image representation is presented where the photometric response function of the camera is determined to within a constant, and the registered images are assembled into a photometric environment map, yielding an estimate, to within a constant scale factor, of the number of photons of light coming from each angle, toward the wearer.

1. INTRODUCTION

The author's 'wearable computer/personal imaging' invention (Fig 1) provides imagery from the perspective of the wearer. The current realization of the apparatus comprises two cameras built into a pair of ordinary eyeglasses (Fig 1). Various realizations of the have been built (some alternate embodiments appear in Fig 1).

'Personal imaging' is a field that is rich in signal processing applications. Signals available from the author's current apparatus comprise video from two cameras, wearable radar (operating at 24.36GHz), biosensors (ECG, EMG, respiration, skin resistance, signals pertaining to footsteps, etc.). As an example signal-processing application for the personal safety device (PSD), consider the gunman who asks for your wallet. Your PSD senses that your footsteps have stopped, yet your heart rate has suddenly increased. Since the heart rate jumped up without any apparent athletic rationale, the DSP in your PSD concludes that the scene (in this case your assailant) is of interest, and devotes maximum bandwidth toward video capture, assuming that the foveal portion of your view would probably contain something that you would desire to remember later.

THIS WORK SPONSORED, IN PART, BY HEWLETT PACKARD RESEARCH LABS.

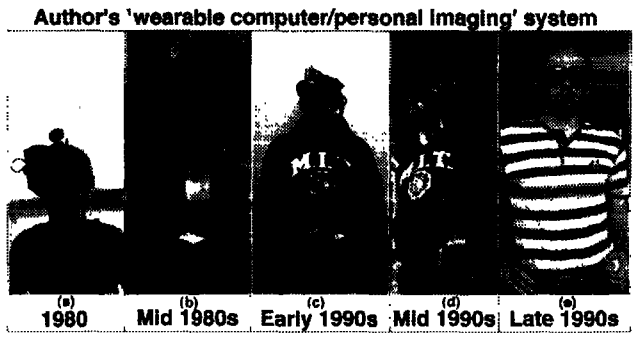


Figure 1: Evolution of author's wearable computer/personal imaging invention. Large head-mounted CRT and separate inbound and outbound communications antennas of the late 70s were awkward. Author's waist-mounted television of the mid 80s was somewhat more comfortable but not constantly visible. Small viewfinders from consumer video cameras of the late 80s made possible an eyeglass-based system which later evolved toward author's current embodiment built into ordinary eyeglasses.

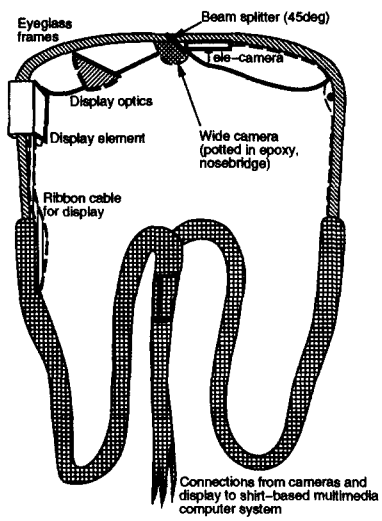


Figure 2: Current rig has two miniature cameras and display built into ordinary eyeglasses. This bi-foveated scheme was found to be useful in a host of applications ranging from crime-reduction (personal safety/documentar), to situational awareness and shared visual memory.

Figure 3: Author (at left) sights multiple picture of self and others, through imaging apparatus in front of group, to illustrate various embodiments of online living/imaging.

A contribution of this paper is a simple method of "scanning" out a scene, from a fixed point in space, by panning, tilting, or rotating a camera, whose gain (iris, AGC, or the like) is also allowed to change of its own accord (e.g. arbitrarily). Photometric self calibration together with gain estimates allows for an estimate, $kq(x,y)$, which is linearly proportional to the number of photons per second, $q(x,y)$, of light received from each incoming ray. The unknown constant of proportionality, k, is constant over all incoming light rays (e.g. a single scalar constant for the whole composite image). Thus the emphasis here is on processing video signals from WearCam.

Hartley demonstrated the simplicity and utility in estimating the calibration parameters of a camera by means of rotating (yaw, pitch, roll but no translation) it about its center of projection[1]. The assumption of zero parallax may arise out of a deliberate control over camera motion (as in Hartley's procedure). Alternatively, the zero-parallax assumption is useful and important as a first step in the more difficult problem of estimating depth and structure from a scene (first modeling the motion as a projective coordinate transformation, and then estimating the residual epipolar structure or the like[2][3][4][5]). In our application, that of generating an environment map, zero-parallax is assumed.

The problem of assembling multiple pictures of the same scene into a single image commonly arises in mapmaking (with the use of aerial photography) and photogrammetry[6], where zero-parallax is also generally assumed. Fully automatic featureless methods of combining multiple pictures have also been previously proposed[7][8]. The emphasis of this work was on subpixel image shifts; the underlying assumptions and models (affine, and pure translation, respectively) were not capable of accurately describing more macroscopic image motion (arising from large changes in camera orientation, etc.).

In motion-estimation work, the commonly-used affine model fails to capture the essence of the chirping and keystoning effects encountered in a panning or tilting camera (especially important to 'WearCam' because of its use of an extremely wide-angle contextual camera). hence the featureless method of estimating the parameters of a projective group of coordinate transformations was first proposed in [9], and later in [10][11]. The earlier method of [9] [10] differs from [11] in that, in the former, a simple and direct method is provided that does not require a nonlinear optimization strategy.

Fully automatic methods of seamlessly combining multiple pictures of the same scene, where differently exposed pictures are combined to extend dynamic range have been proposed[9][12].

The output of a typical camera, f, is not linear with respect to the incoming quantity of light, q. A common model for the nonlinearity, f, is the classic response curve[13]:

$$f(q) = \alpha + \beta q^{\gamma} \quad (1)$$

Methods to estimate the unknown response curve from pictures that differ only in exposure, have been proposed[14]. These methods are based on computing the joint histogram between differently exposed pictures, and then estimating the function $g(f)$, defined by

$$g(f(q(x,y))) = f(kq(x,y)) \quad (2)$$

where $q(x,y)$ is the quantity of light received in a first exposure, and $kq(x,y)$, the quantity of light received in a second exposure, is k times that of the first exposure.

1.1. Automatic Gain Control (AGC)

If what is desired is a picture of increased spatial extent or spatial resolution, the nonlinearity is not a problem, so long as it is not image dependent. However, most low-cost cameras (especially miniature devices suitable for mounting in eyeglasses) have a built in automatic gain control (AGC), electronic level control, auto iris, or some other form of automatic exposure[1] which cannot be disabled. This means that the unknown response function, $f(q)$, is image dependent, and will therefore change over time, as the camera framing changes to include brighter or darker objects.

Recently a joint estimation of the projective coordinate transformation and gain was proposed [12]. Using this projective+gain estimator, it turns out that AGC, rather than being an impediment, becomes an advantage, providing additional information about the scene (extended dymanic range measurement capablility) as well as the camera. This is especially important for WearCam contributing to the hands-free nature of the apparatus, so that one need not make any adjustments when, for example, entering a dimly lit room from a brightly lit exterior.

If we extend the concept of "motion estimation" to include both 'domain motion' (motion in the traditional sense) as well as 'range motion' (Fig 4), we may think of the projectivity (pan, tilt) as contributing to the former, while the gain (e.g. effects of AGC) to the latter.

2. BACKGROUND: JOINT ESTIMATION OF DOMAIN MOTION AND 'RANGE MOTION'

As in[10], we consider one dimensional "images" for purposes of illustration, with the understanding that the actual operations are performed on 2-D images. The 1-D projective+gain group is defined in terms of the "group[2]" of projective coordinate transformations, taken together with the

[1] I refer to all of these methods of automatic exposure control as AGC, whether or not they are actually implemented using gain.

[2] To be strictly mathematically correct, the projective group is written $(ax+b)/(cx+d)$, but $d \neq 0$ in practical engineering problems (physically "reasonable" camera motion), so we may divide by d.

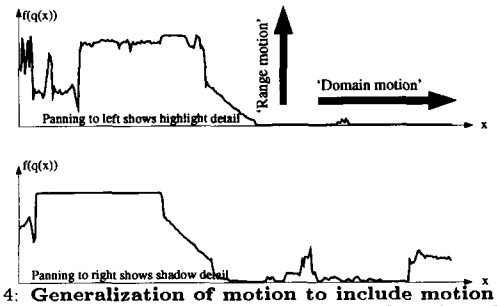


Figure 4: **Generalization of motion to include motion in the range as well as the domain of the image function:** 'Domain motion' is motion in the traditional sense, in this case, arising from the camera panning to the right. 'Range motion' refers to a tone-scale adjustment, in this case arising as a consequence of the fact that the camera is panning to point more and more into the darkness of an open doorway, causing the automatic gain control (AGC) to increase the exposure. Thus there is some upwards 'motion' of the image function as well as leftwards motion. Just as panning the camera across causes information to leave the frame at the left, and new information to enter at the right, increased exposure causes highlight detail to leave from the top and new shadow detail to enter from the bottom.

one-parameter group of gain (image darkening/lightening) operations:

$$p_{a,b,c,k} \circ f(q(x)) = g_k(f(q(\frac{ax+b}{cx+1}))) = f(kq(\frac{ax+b}{cx+1})) \quad (3)$$

where g_k characterizes the gain operation.

The law of composition is defined as: $(p_{abc}, p_k) \circ (p_{def}, p_l) = (p_{abc} \circ p_{def}, p_k \circ p_l)$ where the first law of composition on the right hand side is the usual one for the projective group (a subgroup of the projective+gain group), and the second one is that of the one-parameter gain (image lightening/darkening) subgroup.

Two successive frames of a video sequence are related through a group-action that is near the identity of the group, thus one may think of the Lie algebra of the group as providing the structure locally. As in previous work[10] an approximate model which matches the 'exact' model in the neighbourhood of the identity is used.

For the 'gain group' (which is a one parameter group isomorphic to addition over the reals, or multiplication over the positive reals), the approximate model may be taken from Eq 1, by noting that

$$g(f(q)) = f(kq) = \alpha + \beta(kq)^\gamma = k^\gamma f + 1 - \alpha k^\gamma \quad (4)$$

This equation suggests that linear regression on the joint histogram between two images will provide an estimate of α and γ, while leaving β unknown, which is consistent the fact that the response curve may only be determined up to a constant scale factor[12].

From (4), using the (generalized) brightness change constraint equation[12] and minimizing the sum of squared errors yields a linear solution in substituted variables (that are easily related to the variables of the approximate model):

$$\begin{bmatrix} \sum x^4F_x^2 & \sum x^3F_x^2 & \sum x^2F_x^2 & -\sum x^2FF_x & -\sum x^2F_x \\ \sum x^3F_x^2 & \sum x^2F_x^2 & \sum xF_x^2 & -\sum xFF_x & -\sum xF_x \\ \sum x^2F_x^2 & \sum xF_x^2 & \sum F_x^2 & -\sum FF_x & -\sum F_x \\ \sum x^2FF_x & \sum xFF_x & \sum FF_x & -\sum F^2 & -\sum F \\ \sum x^2F_x & \sum xF_x & \sum F_x & -\sum F & -\sum 1 \end{bmatrix} \begin{bmatrix} (bc-a)c \\ a-bc \\ b \\ k^\gamma \\ 1-\alpha k^\gamma \end{bmatrix} = -\left[\sum x^2F_x(F+F_t), \sum xF_x(F+F_t), \sum F_x(F+F_t), \sum F(F+F_t), \sum (F+F_t)\right]^T$$

where $F(x,t) = f(q(x))$ at time t, $F_x(x,t) = (df/dq)(dq(x)/dx)$, at time t, and $F_t(x,t)$ is the frame difference of adjacent frames.

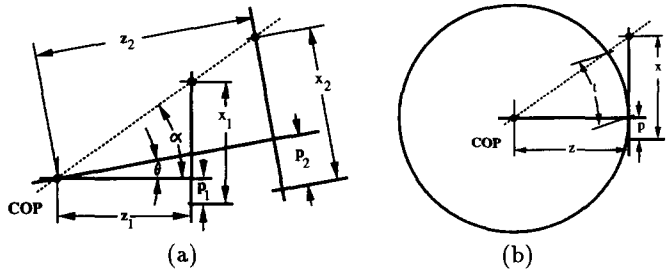


Figure 5: **A rotating uncalibrated camera** (a) depicted as two separate cameras. Calibration constants, z_1 and z_2 are equal but distinguished for clarity. Likewise, $p_1 = p_2$. (b) Appropriate coordinate transformation simplifies problem to joint estimation of pure translation and gain change.

3. THE HARTLEY CONSTRAINT

For 2-D images in 3-D, there are nine degrees of freedom, so the approximate model (using the feedback loop of [10] to obtain solution of the 'exact' motion model) must have at least 9 parameters. Constraining the parameters of the projective group through estimation of the calibration matrix[1], to the three rotation parameters and one gain parameter, will reduce sensitivity to noise, while at the same time, better fitting the true underlying phenomena (camera rotation and AGC) that gave rise to the image motion. For 1-D images, a representation for the rotation+gain group is:

$$[e^\theta, 0; 0, k] \quad (5)$$

An *uncalibrated* camera (Fig 5(a)) is assumed.

Proposition 1 *The set of transformations governed by pure rotation (Fig5 (a)) and gain transformations forms a group (5). Furthermore, any such transformation may be expressed as a conjugation with both a calibration matrix and gain adjustment operation.*

Proof: From Fig5 (a), $\frac{x_2}{z_2} = \tan(\arctan(\frac{x_1-p_1}{z_1}) - \theta) + \frac{p_2}{z_2}$. Thus,

$$x_2 = \frac{\frac{z\cos\theta + p\sin\theta}{z\cos\theta - p\sin\theta}x_1 - \frac{(z^2+p^2)\sin\theta}{z\cos\theta - p\sin\theta}}{\frac{\sin\theta}{z\cos\theta - p\sin\theta}x_1 + 1} \quad (6)$$

The same coordinate transformation results if we conjugate a rotation operator, represented by $[\cos\theta, -\sin\theta; \sin\theta, \cos\theta]$ with calibration matrix $[k, p; 0, 1]$ (multiplying the three matrices). The complete transformation is obtained by following this coordindate conjugation with conjugation by a gain adjustment operation. □

The problem of estimating the parameters of this (5) group may be greatly simplified by considering a coordinate transformation operator, T, defined by $t = T(x) = z\arctan((x-p)/z)$. The geometric intuition for T comes from Fig 5(b), where t is the distance along a circle of radius equal to the principal distance, z.

Proposition 2 *The transformation T reduces the group parameter estimation problem to one of joint estimation of pure translation and gain change.*

Proof: consider two images $g(f(q((a_2x+b_2)/(c_2x+1)))) = f(k_2q(x_2))$ and $h(f(q((a_3x+b+3)/(c_3x+1)))) = f(k_3q(x_3))$. Substituting into (6) yields translation $t = t_3 - t_2 = z\theta$, and the gain change $k = k_3/k_2$. □ This means that a simple Fourier-based approach may be used to estimate θ[15] and k[14], which is equivalent to using a family of projective chirps, rather than ordinary sines and cosines, as analysis primitives.

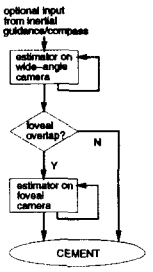
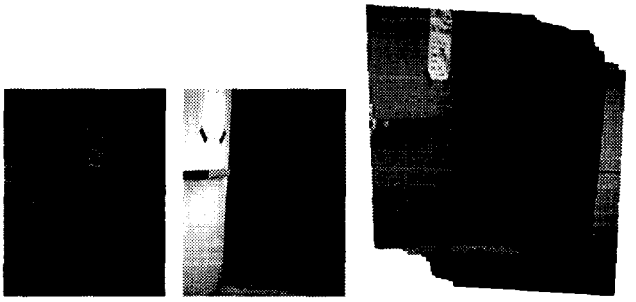


Figure 6: Signal processing approach for bi-foveated 'WearCam'. Note also that the spatial coordinates are propagated according to the projective group's law of composition while the gain parameters between the wide-camera and foveal-camera are not directly coupled.



(a) (b) (c)

Figure 7: Image compositing to extend both dynamic range and spatial extent. (a) Exterior door to building is brightly illuminated: camera's AGC results in a dark exposure. (b) As camera pans to right, pointing into dark interior, AGC causes gain to increase, revealing details inside. (c) Image composite from a sequence of 40 frames. Because the result is a floating-point image of much greater dynamic range, locally non-monotonic processing must be performed to adequately print the image. Note that we can see clearly both the building interior and the white sign on the exterior door.

4. APPLICATION TO REAL IMAGES

To construct a single floating-point image of increased spatial extent and increased dynamic range, each pixel of the output image is constructed from a weighted sum of the images whose coordinate-transformed bounding boxes fall within that pixel. The weights in the weighted sum are the so-called 'certainty functions', which are found by evaluating the derivative of the corresponding 'effective response function' at the pixel value in question[14].

In order to print a picture of such dynamic range it is often necessary to relax the monotonicity constraint, and perform some local tone-scale adjustments[16]. (See Fig 7(c)).

5. BI-FOVEATION

Signal processing with respect to bi-foveated cameras is a special consideration. In particular, since the geometry of one camera is fixed (in epoxy or the like) with respect to the other, there exists a fixed coordinate transformation that maps any image captured on the wide camera to one that was captured on the foveal camera at the same time. Thus when there is a large jump between images captured on the foveal camera — a jump too large to be considered in the neighbourhood of the identity — one may look to the wide camera for contextual reference (greater overlap), apply the estimation between the two wide images, and then relate these to the two foveal images. (The procedure is illustrated in Fig 4).

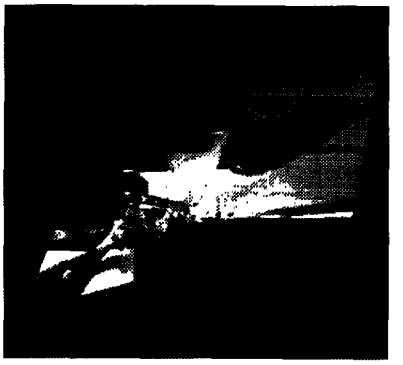


Figure 8: With a very small number of frames and little overlap, a more expressive composite (showing image boundaries more clearly) may be constructed. Here, extreme perspective is used to juxtapose the author's outstretched arm (lower left) at the deli counter with the surveillance camera (upper right).

Image composites with very little overlap may be used for their expressive/narrative qualities (as in the "ShootingBack" documentary, Fig 5

6. REFERENCES

[1] Richard I. Hartley. Self-calibration of stationary cameras, g.e. crd, schenectady, ny, 12301.

[2] O. D. Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. International Journal of Pattern Recognition and Artificial Intelligence, 2(3):485–508, 1988.

[3] G. Adiv. Determining 3D Motion and structure from optical flow generated by several moving objects. IEEE Trans. Pattern Anal. Machine Intell., pages 304–401, July 1985.

[4] Amnon Shashua and Nassir Navab. Relative Affine: Theory and Application to 3D Reconstruction From Perspective Views. Proc. IEEE Conference on Computer Vision and Pattern Recognition, June 1994. 1994.

[5] Nassir Navab and Steve Mann. Recovery of relative affine structure using the motion flow field of a rigid planar patch. Mustererkennung 1994, Tagungsband., 1994.

[6] William A. Radlinski American Society of Photogrammetry. Editor-in-chief: Morris M. Thompson. Associate editors: Robert C. Eller and Julius L. Speert. Manual of photogrammetry. Schaum's Outline Series. McGraw-Hill Book Company, Falls Church, Va., 3d ed edition, 1966. 2 v. (xx, 1199 p.) illus. 27 cm.

[7] M. Irani and S. Peleg. Improving Resolution by Image Registration. CVGIP, 53:231–239, May 1991.

[8] A.M. Tekalp, M.K. Ozkan, and M.I. Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration. In Proc. of the Int. Conf. on Acoust., Speech and Sig. Proc., pages III–169, San Francisco, CA, Mar. 23-26, 1992. IEEE.

[9] S. Mann. Compositing multiple pictures of the same scene. In Proceedings of the 46th Annual IS&T Conference, Cambridge, Massachusetts, May 9-14 1993. The Society of Imaging Science and Technology.

[10] S. Mann and R. W. Picard. Virtual bellows: constructing high-quality images from video. In Proceedings of the IEEE first international conference on image processing, Austin, Texas, Nov. 13-16 1994.

[11] R. Szeliski and J. Coughlan. Hierarchical spline-based image registration. CVPR, 1994.

[12] Steve Mann. 'pencigraphy' with AGC: Joint parameter estimation in both domain and range of functions in same orbit of the projective-Wyckoff group. Technical Report 384, MIT Media Lab, Cambridge, Massachusetts, December 1994. also appears in: IEEE International Conference on Image Processing (ICIP 96), Lausanne, Switzerland, September 1996.

[13] Charles W. Wyckoff. An experimental extended response film. S.P.I.E. NEWSLETTER, JUNE-JULY 1962.

[14] S. Mann and R.W. Picard. Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. Technical Report 323, M.I.T. Media Lab Perceptual Computing Section, Boston, Massachusetts, 1994. Also appears, IS&T's 46th annual conference, pages 422-428, May 1995.

[15] Bernd Girod and David Kuo. Direct estimation of displacement histograms. OSA Meeting on IMAGE UNDERSTANDING AND MACHINE VISION, June 1989.

[16] T. G. Stockham, Jr. Image processing in the context of a visual model. Proc. IEEE, 60(7):828–842, July 1972.