

VideoOrbits on Eye Tap devices for deliberately Diminished Reality or altering the visual perception of rigid planar patches of a real world scene

Steve Mann

James Fung

University of Toronto, 10 King's College Road, Toronto, Canada

Tel.=(416) 978-5036 Fax.=(416) 971-2326

E-mail: mann@eecg.toronto.edu, fungja@eecg.toronto.edu

Abstract

Diminished reality is as important as Augmented reality, and both are possible with a new device called a Reality Mediator. The Reality Mediator allows the wearer's visual perception of reality to be altered in such a way that the user can delete or diminish undesirable visual detritus from their perceived environment. The algorithm and apparatus which are used to achieve scene tracking and image registration are described and a practical example of diminished reality is presented. The Reality Mediator, a portable device that quantifies and resynthesizes light that would otherwise pass through one or both lenses of (an) eye(s) of a user is described. The device diverts at least a portion of eyeward bound light into a measurement system that measures how much light would have entered the eye in the absence of the device. The device has at least one mode of operation in which it reconstructs these rays of light, under the control of a portable computational system. By applying the VideoOrbits algorithm, the device can alter the light from a particular portion of the scene so a user perceives a computationally mediated version of the scene, giving rise to the possibility of computer controlled selectively diminished reality, allowing for additional information to be inserted without causing the user to experience information overload.

Keywords: Geometrical registration, mediated, diminished reality.

1 Eye Tap devices for mediating reality

Eye Tap devices have three main parts:

- a measurement system typically consisting of a camera system, or sensor array with appropriate optics;
- a diverter system, for diverting eyeward bound light into the measurement system and therefore causing the eye of the user of the device to behave, in effect, as if it were a camera;

- an aremac for reconstructing at least some of the diverted rays of eyeward bound light. Thus the aremac does the opposite of what the camera does, and is, in many ways, a camera in reverse. The etymology of the word "aremac" itself, arises from spelling the word "camera" backwards.

There are two embodiments of the aremac: (1) one in which a focuser (such as an electronically focusable lens) tracks the focus of the camera, to reconstruct rays of diverted light in the same depth plane as imaged by the camera; and (2) another in which the aremac has extended or infinite depth of focus so that the eye itself can focus on different objects in a scene viewed through the apparatus.

1.1 Focus tracking Eye Tap systems

This paper describes only the focus tracking embodiment of the Eye Tap system. The aremac has focus linked to the measurement system (e.g. "camera") focus, so that objects seen depicted on the aremac of the device appear to be at the same distance from the user of the device as the real objects so depicted. In manual focus systems the user of the device is given a focus control that simultaneously adjusts both the aremac focus and the "camera" focus. In automatic focus embodiments, the camera focus also controls the aremac focus. Such a linked focus gives rise to a more natural viewfinder experience, as well as reduced eyestrain. Reduced eyestrain is important because the device is intended to be worn continually.

The operation of the depth tracking aremac is shown in Fig 1. Because the eye's own lens L_3 experiences what it would have experienced in the absence of the apparatus, the apparatus, in effect, taps into and out of the eye, causing the eye to become both the camera and the viewfinder (display). Therefore the device is called an Eye Tap device.

Often, lens L_1 is a varifocal lens, or otherwise has a variable field of view (e.g. "zoom" functionality). In this case, it is desired that the aremac also have a variable field of view. In particular, field of view control mechanisms (whether mechanical, electronic, or hybrid) are linked in

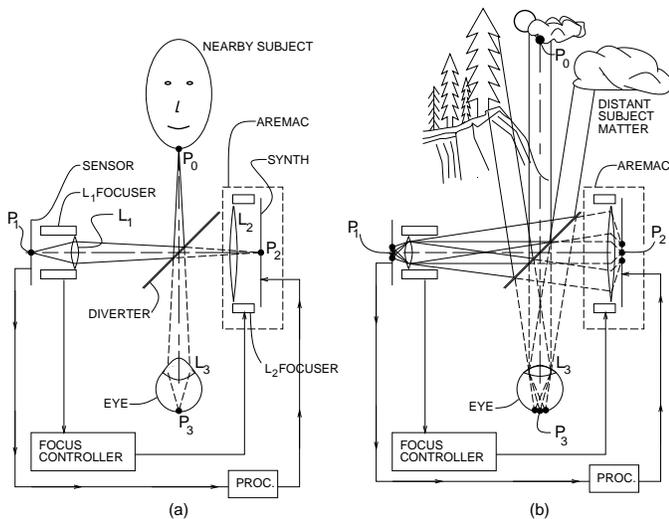


Figure 1 Focus tracking aremac: (a) with a NEARBY SUBJECT, a point P_0 that would otherwise be imaged at P_3 in the EYE of a user of the device is instead imaged to point P_1 on the image SENSOR, because the DIVERTER diverts EYEward bound light to lens L_1 . When subject matter is nearby, the L_1 FOCUSER moves objective lens L_1 out away from the SENSOR automatically, as an automatic focus camera would. A signal from the L_1 FOCUSER directs the L_2 FOCUSER, by way of the FOCUS CONTROLLER, to move lens L_2 outward away from the light SYNTHesizer. At the same time, an image from the SENSOR is directed through an image PROCessor, into the light SYNTHesizer. Point P_2 of the display element is responsive to point P_1 of the SENSOR. Likewise other points on the light SYNTHesizer are each responsive to corresponding points on the SENSOR, so that the SYNTHesizer produces a complete image for viewing through lens L_2 by the EYE, after reflection off of the back side of the DIVERTER. The position of L_2 is such that the EYE's own lens L_3 will focus to the same distance as it would have focused in the absence of the entire device. (b) With DISTANT SUBJECT MATTER, rays of parallel light are diverted toward the SENSOR where lens L_1 automatically retracts to focus these rays at point P_1 . When lens L_1 retracts, so does lens L_2 , and the light SYNTHesizer ends up generating parallel rays of light that bounce off the backside of the DIVERTER. These parallel rays of light enter the EYE and cause its own lens L_3 to relax to infinity, as it would have in the absence of the entire device.

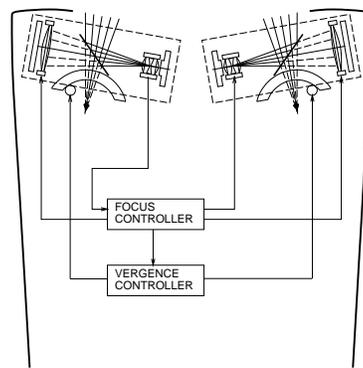


Figure 2 Focus of right camera and both aremacs (as well as vergence) controlled by autofocus camera on left side. In a two eyed system, it is preferable that both cameras and both aremacs focus to the same distance. Therefore, one of the cameras is a focus master, and the other camera is a focus slave. Alternatively, a focus combiner is used to average the focus distance of both cameras and then make the two cameras focus at equal distance. The two aremacs, as well as the vergence of both systems also track this same depth plane as defined by camera autofocus.

such a way that the aremac image magnification is reduced as the camera magnification is increased. Through this appropriate linkage, any increase in magnification by the camera is negated exactly by decreasing the apparent size of the viewfinder image.

The operation of the aremac focus and zoom tracking is shown in Fig 2. Stereo effects are well known in Virtual Reality systems [1] where two information channels are often found to create a better sense of realism. Likewise, in stereo versions of the proposed device, there are two cameras or measurement systems and two aremacs that each regenerate the respective outputs of the camera or measurement systems.

The apparatus is usually concealed in dark sunglasses that obstruct vision except for what the apparatus allows to pass through.

2 Video Orbits Image Registration for Eye Tap Reality Mediation

Because the device absorbs, quantifies, processes, and reconstructs light passing through it, there are extensive applications in mediated reality. Mediated Reality differs from Virtual Reality in the sense that Mediated Reality allows the visual perception of reality to be augmented, deliberately diminished, or, more generally computation-

ally altered.

Mediated Reality is created when virtual, or computer generated, information is mixed with what the user would otherwise normally see. The virtual information or light as seen through the display must be properly registered and aligned within the user’s field of view. To achieve this, a method of camera-based head-tracking is now described.

2.1 Why camera-based head-tracking?

A goal of personal imaging is to facilitate the use of Personal Imaging [2] systems in ordinary everyday situations, not just on a factory assembly line “workcell”, or other restricted space. Thus it is desired that the apparatus have a head-tracker that need not rely on any special apparatus being installed in the environment.

Therefore, a new method of head-tracking based on the use of the camera capability of the apparatus is needed [2], and is based on the VideoOrbits algorithm [3]. The VideoOrbits algorithm performs head-tracking, visually, based on a natural environment, and works without the need for object recognition. Instead it is based on algebraic projective geometry, and a featureless means of estimating the change in spatial coordinates arising from movement of the wearer’s head, as illustrated in Figure 3.

2.2 Algebraic projective geometry

Direct featureless methods are used for estimating the 8 parameters of an “exact” projective (homographic) coordinate transformation to register pairs of images or scene content. The approach is “exact” for two cases of static scenes: (1) images taken from the same location of an arbitrary 3-D scene, with a camera that is free to pan, tilt, rotate about its optical axis, and zoom or (2) images of a flat scene taken from arbitrary locations. The featureless projective approach generalizes inter-frame camera motion estimation methods which have previously used an *affine* model (which lacks the degrees of freedom to “exactly” characterize such phenomena as camera pan and tilt) and/or which have relied upon finding points of correspondence between the image frames. The featureless projective approach, which operates directly on the image pixels, is shown to be superior in accuracy and ability to enhance resolution. The proposed methods work well on image data collected from both good-quality and poor-quality video under a wide variety of conditions (sunny, cloudy, day, night). These fully-automatic methods are also shown to be robust to deviations from the assumptions of static scene and no parallax, although the primary application is in filtering out or modifying subject matter appearing on flat surfaces within a scene (e.g. rigid planar patches such as advertising billboards).

The most common assumption (especially in motion estimation for coding, and optical flow for computer vi-

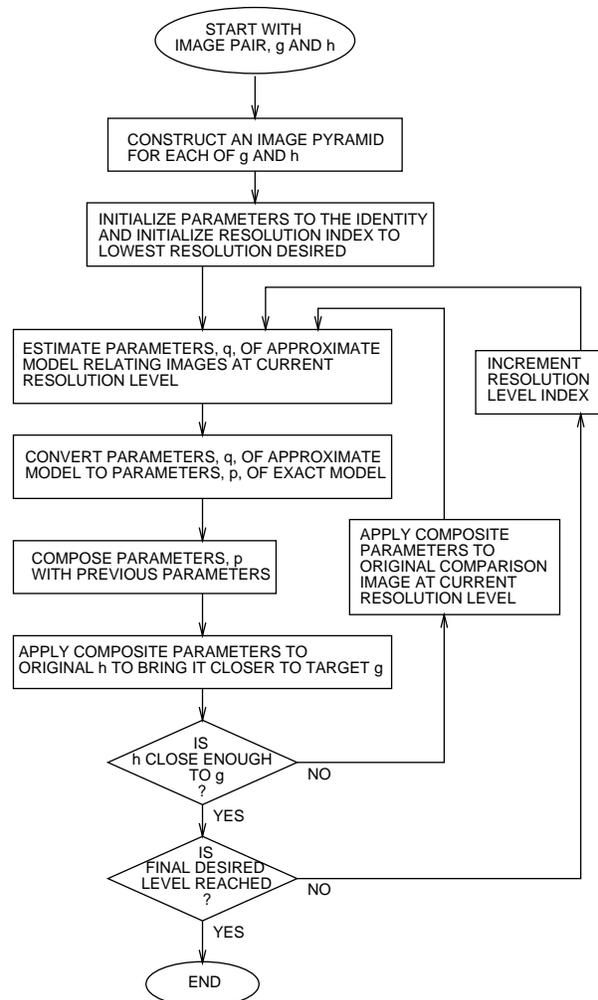


Figure 3 The ‘VideoOrbits’ head-tracking algorithm: The new head-tracking algorithm requires no special devices installed in the environment. The camera in the Personal Imaging system simply tracks itself based on its view of objects in the environment. The algorithm is based on algebraic projective geometry, and provides an estimate of the true projective coordinate transformation, which, for successive image pairs is composed using the projective group [3]. Successive pairs of images may be estimated in the neighbourhood of the identity coordinate transformation of the group, while absolute head tracking is done using the exact group by relating the approximate parameters q to the exact parameters p in the innermost loop of the process. The algorithm typically runs at 5–10 frames per second on a general-purpose computer but the simple structure of the algorithm makes it easy to implement in hardware for the higher frame-rates needed for full-motion video.

| Model | Coordinate transformation from \mathbf{x} to \mathbf{x}' | Parameters |
|---------------------|--|---|
| Translation | $\mathbf{x}' = \mathbf{x} + \mathbf{b}$ | $\mathbf{b} \in \mathbf{R}^2$ |
| Affine | $\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$ | $\mathbf{A} \in \mathbf{R}^{2 \times 2}, \mathbf{b} \in \mathbf{R}^2$ |
| Bilinear | $x' = q_{x'xy}xy + q_{x'xx}x + q_{x'y}y + q_{x'}$ $y' = q_{y'xy}xy + q_{y'xx}x + q_{y'y}y + q_{y'}$ | $bfq_* \in \mathbf{R}$ |
| Projective | $\mathbf{x}' = \frac{\mathbf{A}\mathbf{x} + \mathbf{b}}{c^T\mathbf{x} + 1}$ | $\mathbf{A} \in \mathbf{R}^{2 \times 2}, \mathbf{b}, \mathbf{c} \in \mathbf{R}^2$ |
| Relative-projective | $\mathbf{x}' = \frac{\mathbf{A}\mathbf{x} + \mathbf{b}}{c^T\mathbf{x} + 1} + \mathbf{x}$ | $\mathbf{A} \in \mathbf{R}^{2 \times 2}, \mathbf{b}, \mathbf{c} \in \mathbf{R}^2$ |
| Pseudo-perspective | $x' = q_{x'xx}x + q_{x'y}y + q_{x'} + q_{\alpha}x^2 + q_{\beta}xy$ $y' = q_{y'xx}x + q_{y'y}y + q_{y'} + q_{\alpha}xy + q_{\beta}y^2$ | $\mathbf{q}_* \in \mathbf{R}$ |
| Biquadratic | $x' = q_{x'x^2}x^2 + q_{x'xy}xy + q_{x'y^2}y^2 + q_{x'x}x + q_{x'y}y + q_{x'}$ $y' = q_{y'x^2}x^2 + q_{y'xy}xy + q_{y'y^2}y^2 + q_{y'x}x + q_{y'y}y + q_{y'}$ | $bfq_* \in \mathbf{R}$ |

Table 1 Image coordinate transformations discussed in this paper

sion) is that the coordinate transformation between frames is translation. Tekalp, Ozkan, and Sezan [4] have applied this assumption to high-resolution image reconstruction. Although translation is the least constraining and simplest to implement of the seven coordinate transformations in Table 1, it is poor at handling large changes due to camera zoom, rotation, pan and tilt.

Zheng and Chellappa [5] considered the image registration problem using a subset of the affine model — translation, rotation and scale. Other researchers [6][7] have assumed affine motion (six parameters) between frames. For the assumptions of static scene and no parallax, the affine model exactly describes rotation about the optical axis of the camera, zoom of the camera, and pure shear, which the camera does not do, except in the limit as the lens focal length approaches infinity. The affine model cannot capture camera pan and tilt, and therefore cannot properly express the “keystoning” and “chirping” we see in the real world. “Chirping” refers to the effect of increasing or decreasing spatial frequency with respect to spatial location, as illustrated in Fig 4. This chirping phenomenon is implicit in the proposed system, whether or not there is periodicity in the subject matter. The only requirement is that there be some distinct texture upon a flat surface in the scene.

2.3 Video orbits

Tsai and Huang [8] pointed out that the elements of the projective *group* give the true camera motions with respect to a planar surface. They explored the group structure associated with images of a 3-D rigid planar patch, as well as the associated *Lie algebra*, although they assume that the correspondence problem has been solved. The solution presented in this paper (which does not require prior solution of correspondence) also relies on projective group theory.

2.3.1 ‘Projective flow’: A new technique for tracking a rigid planar patch

A method for tracking a rigid planar patch is now presented. Consider first one dimensional systems, since they are easier to explain and understand. For a 1-D affine coordinate transformation, the graph of the range coordinate as a function of the domain coordinate is a straight line; for the projective coordinate transformation, the graph of the range coordinate as a function of the domain coordinate is a rectangular hyperbola (Fig 4(d)).

Whether or not there is periodicity in the scene, the method still works, in the sense that it is based on the projective flow across the texture or pattern, at all various spatial frequency components of a rigid planar patch.

The method is called ‘projective-flow’ (‘p-flow’), and arises from substitution of $u_m = \frac{ax+b}{cx+1} - x$ into the Horn and Schunk Brightness Change Constraint Equation[9].

Differentiating, setting the derivative to zero, and summing, with an additional weighting by $(cx + 1)$ gives:

$$\varepsilon_w = \sum (axE_x + bE_x + c(xE_t - x^2E_x) + E_t - xE_x)^2 \quad (1)$$

(the subscript w denotes weighting has taken place) resulting in a linear system of equations for the parameters:

$$\left(\sum \phi_w \phi_w^T \right) [a, b, c]^T = \sum (xE_x - E_t) \phi_w \quad (2)$$

where the *regressor* is $\phi_w = [xE_x, E_x, xE_t - x^2E_x]^T$.

The notation and derivations used in this paper are as described in [10], page 2139. The reader is invited to refer to [10] for a more in-depth treatment of the matter.

2.3.2 The unweighted projectivity estimator

If we do not wish to apply the ad-hoc weighting scheme, we may still estimate the parameters of projectivity in a simple manner, still based on solving a linear system of equations. To do this, we write the Taylor series of u_m :

$$u_m + x = b + (a - bc)x + (bc - a)cx^2 + (a - bc)c^2x^3 + \dots \quad (3)$$

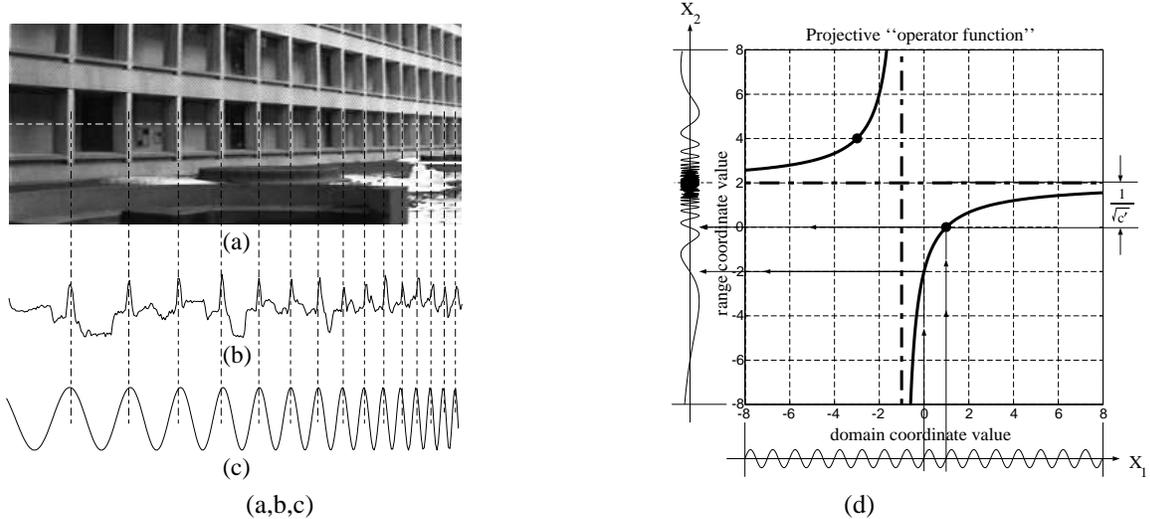


Figure 4 The ‘projective chirping’ phenomenon. (a) A real-world object that exhibits periodicity generates a projection (image) with “chirping” — ‘periodicity-in-perspective’. (b) Center raster of image. (c) Best-fit projective chirp of form $\sin(2\pi((ax + b)/(cx + 1)))$. (d) Graphical depiction of exemplar 1-D projective coordinate transformation of $\sin(2\pi x_1)$ into a ‘projective chirp’ function, $\sin(2\pi x_2) = \sin(2\pi((2x_1 - 2)/(x_1 + 1)))$. The range coordinate as a function of the domain coordinate forms a rectangular hyperbola with asymptotes shifted to center at the vanishing point $x_1 = -1/c = -1$ and ‘exploding point’, $x_2 = a/c = 2$, and with ‘chirpiness’ $c' = c^2/(bc - a) = -1/4$.

and use the first 3 terms, obtaining enough degrees of freedom to account for the 3 parameters being estimated. Letting $\varepsilon = \sum(-h.o.t.)^2 = \sum((b + (a - bc - 1)x + (bc - a)cx^2)E_x + E_t)^2$, $\mathbf{q}_2 = (bc - a)c$, $\mathbf{q}_1 = a - bc - 1$, and $\mathbf{q}_0 = b$, and differentiating with respect to each of the 3 parameters of \mathbf{q} , setting the derivatives equal to zero, and verifying with the second derivatives, gives the linear system of equations for ‘unweighted projective flow’:

$$\begin{bmatrix} \sum x^4 E_x^2 & \sum x^3 E_x^2 & \sum x^2 E_x^2 \\ \sum x^3 E_x^2 & \sum x^2 E_x^2 & \sum x E_x^2 \\ \sum x^2 E_x^2 & \sum x E_x^2 & \sum E_x^2 \end{bmatrix} \begin{bmatrix} q_2 \\ q_1 \\ q_0 \end{bmatrix} = - \begin{bmatrix} \sum x^2 E_x E_t \\ \sum x E_x E_t \\ \sum E_x E_t \end{bmatrix} \quad (4)$$

3 Planetracker in 2-D

The brightness constancy constraint equation for 2-D images [9] which gives the flow velocity components in the x and y directions, is:

$$\mathbf{u}_f^T \mathbf{E}_x + E_t \approx 0 \quad (5)$$

As is well-known [9] the optical flow field in 2-D is underconstrained¹. The model of *pure translation* at every point has two parameters, but there is only one it is common practice to compute the optical flow over some neighborhood, which must be at least two pixels, but is

¹Optical flow in 1-D did not suffer from this problem.

generally taken over a small block, 3×3 , 5×5 , or sometimes larger (e.g. the entire patch of subject matter to be filtered out, such as a billboard or sign).

Our task is not to deal with the 2-D translational flow, but with the 2-D projective flow, estimating the eight parameters in the coordinate transformation:

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{\mathbf{A}[x, y]^T + \mathbf{b}}{\mathbf{c}^T[x, y]^T + 1} = \frac{\mathbf{A}\mathbf{x} + \mathbf{b}}{\mathbf{c}^T\mathbf{x} + 1} \quad (6)$$

The desired eight scalar parameters are denoted by $\mathbf{p} = [\mathbf{A}, \mathbf{b}, \mathbf{c}, 1]$, $\mathbf{A} \in \mathbf{R}^{2 \times 2}$, $\mathbf{b} \in \mathbf{R}^{2 \times 1}$, and $\mathbf{c} \in \mathbf{R}^{2 \times 1}$.

We have, in the 2-D case:

$$\varepsilon_{flow} = \sum (\mathbf{u}_m^T \mathbf{E}_x + E_t)^2 = \sum \left(\left(\frac{\mathbf{A}\mathbf{x} + \mathbf{b}}{\mathbf{c}^T\mathbf{x} + 1} - \mathbf{x} \right)^T \mathbf{E}_x + E_t \right)^2 \quad (7)$$

Where the sum can be weighted as it was in the 1-D case:

$$\varepsilon_w = \sum \left((\mathbf{A}\mathbf{x} + \mathbf{b} - (\mathbf{c}^T\mathbf{x} + 1)\mathbf{x})^T \mathbf{E}_x + (\mathbf{c}^T\mathbf{x} + 1)E_t \right)^2 \quad (8)$$

Differentiating with respect to the free parameters \mathbf{A} , \mathbf{b} , and \mathbf{c} , and setting the result to zero gives a linear solution:

$$\left(\sum \phi \phi^T \right) [a_{11}, a_{12}, b_1, a_{21}, a_{22}, b_2, c_1, c_2]^T = \sum (\mathbf{x}^T \mathbf{E}_x - E_t) \phi \quad (9)$$

where $\phi^T = [E_x(x, y, 1), E_y(x, y, 1), xE_t - x^2E_x - xyE_y, yE_t - xyE_x - y^2E_y]$

For a more in depth treatment of projective flow, the reader is invited to refer to [10].

3.1 ‘Unweighted projective flow’

As with the 1-D images, we make similar assumptions in expanding (6) in its own Taylor series, analogous to (3). By appropriately constraining the twelve parameters of the biquadratic model we obtain a variety of 8-parameter approximate models. In estimating the ‘exact unweighted’ projective group parameters, one of these approximate models is used in an intermediate step.²

The Taylor series for the bilinear case gives:

$$\begin{aligned} u_m + x &= q_{x'xy}xy + (q_{x'x} + 1)x + q_{x'y}y + q_{x'} \\ v_m + y &= q_{y'xy}xy + q_{y'x}x + (q_{y'y} + 1)y + q_{y'} \end{aligned} \quad (10)$$

Incorporating these into the flow criteria yields a simple set of eight linear equations in eight unknowns:

$$\left(\sum_{x,y} (\phi(x, y)\phi^T(x, y)) \right) \mathbf{q} = - \sum_{x,y} E_t \phi(x, y) \quad (11)$$

where $\phi^T = [E_x(xy, x, y, 1), E_y(xy, x, y, 1)]$.

For the relative-projective model, ϕ is given by

$$\phi^T = [E_x(x, y, 1), E_y(x, y, 1), E_t(x, y)] \quad (12)$$

and for the pseudo-perspective model, ϕ is given by

$$\begin{aligned} \phi^T &= [E_x(x, y, 1), E_y(x, y, 1), \\ &(x^2E_x + xyE_y, xyE_x + y^2E_y)] \end{aligned} \quad (13)$$

3.1.1 ‘Four point method’ for relating approximate model to exact model

Any of the approximations above, after being related to the exact projective model, tend to behave well in the neighborhood of the identity, $\mathbf{A} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$, $\mathbf{c} = \mathbf{0}$. In 1-D, the model Taylor series about the identity was explicitly expanded; here, although this is not done explicitly, it is assumed that the terms of the Taylor series of the model correspond to those taken about the identity. In the 1-D case we solve the 3 linear equations in 3 unknowns to estimate the parameters of the approximate motion model, and then relate the terms in this Taylor series to the exact parameters, a , b , and c (which involves solving another set of 3 equations in 3 unknowns, the second set being nonlinear, although very easy to solve).

In the extension to 2-D, the estimate step is straightforward, but the relate step is more difficult, because we now have eight nonlinear equations in eight unknowns,

²Use of an approximate model that doesn’t capture chirping or preserve straight lines can still lead to the true projective parameters as long as the model captures at least eight meaningful degrees of freedom.

relating the terms in the Taylor series of the approximate model to the desired exact model parameters. Instead of solving these equations directly, a simple procedure is used for relating the parameters of the approximate model to those of the exact model, which is called the ‘four point method’:

1. Select four ordered pairs (e.g. the four corners of the bounding box containing the region under analysis, or the four corners of the image if the whole image is under analysis). Here suppose, for simplicity, that these points are the corners of the unit square: $\mathbf{s} = [s_1, s_2, s_3, s_4] = [(0, 0)^T, (0, 1)^T, (1, 0)^T, (1, 1)^T]$.
2. Apply the coordinate transformation using the Taylor series for the approximate model (e.g. (10)) to these points: $\mathbf{r} = \mathbf{u}_m(\mathbf{s})$.
3. Finally, the correspondences between \mathbf{r} and \mathbf{s} are treated just like features. This results in four easy to solve linear equations:

$$\begin{aligned} \begin{bmatrix} x'_k \\ y'_k \end{bmatrix} &= \begin{bmatrix} x_k, y_k, 1, 0, 0, 0, -x_k x'_k, -y_k y'_k \\ 0, 0, 0, x_k, y_k, 1, -x_k y'_k, -y_k y'_k \end{bmatrix} \\ &\begin{bmatrix} a_{x'x}, a_{x'y}, b_{x'}, a_{y'x}, a_{y'y}, b_{y'}, c_x, c_y \end{bmatrix}^T \end{aligned} \quad (14)$$

where $1 \leq k \leq 4$. This results in the exact eight parameters, \mathbf{p} .

We remind the reader that the four corners are **not** feature correspondences as used in the feature-based methods, but, rather, are used so that the two featureless models (approximate and exact) can be related to one another.

It is important to realize the full benefit of finding the exact parameters. While the ‘‘approximate model’’ is sufficient for small deviations from the identity, it is not adequate to describe large changes in perspective. However, if we use it to track small changes incrementally, and each time relate these small changes to the exact model (6), then we can accumulate these small changes using the *law of composition* afforded by the group structure. This is an especially favorable contribution of the group framework. For example, with a video sequence, we can accommodate very large accumulated changes in perspective in this manner. The problems with cumulative error can be eliminated, for the most part, by constantly propagating forward the true values, computing the residual using the approximate model, and each time relating this to the exact model to obtain a goodness-of-fit estimate.

3.1.2 Algorithm for ‘unweighted projective flow’: overview

Frames from an image sequence are compared pairwise to test whether or not they lie in the same orbit:

1. A Gaussian pyramid of three or four levels is constructed for each frame in the sequence.

2. The parameters \mathbf{p} are estimated at the top of the pyramid, between the two lowest-resolution images of a frame pair, g and h , using the repetitive method depicted in Fig. 3.
3. The estimated \mathbf{p} is applied to the next higher-resolution (finer) image in the pyramid, $\mathbf{p} \circ g$, to make the two images at that level of the pyramid nearly congruent before estimating the \mathbf{p} between them.
4. The process continues down the pyramid until the highest-resolution image in the pyramid is reached.

3.2 Mediated reality as a form of communication

The mathematical framework for mediated reality arose through the process of marking a reference frame[3] with text or simple graphics, where it was noted that by calculating and matching homographies of the plane, an illusory rigid planar patch appeared to hover upon objects in the real-world, giving rise to a form of computer-mediated collaboration[2].

This collaborative capability was suggested as an application of HI to the visually challenged, or those with a visual memory disability[11]. In this application, a computer program, or remote expert (be it human or machine) may assist in way finding, or by providing a photographic/videographic memory, such as the ability to never forget a face. (See Fig 5.)

Figure 6 show images taken with an Eye Tap system and depict the use of the reality mediator as a form of communication. Figure 6(a) is an unmediated image of a roadside advertisement. Figures 6(b),(c),(d) show images from the same sequence as seen through the reality mediator. The motion of the planar patch of the advertisement has been tracked by the Video Orbits algorithm and the advertisement is replaced by a message intended to help guide the user to their destination.

References

- [1] Stephen R. Ellis, Urs J. Bucher, and Brian M. Menges. The relationship of binocular convergence and errors in judged distance to virtual objects. *Proceedings of the International Federation of Automatic Control*, June 27–29 1995.
- [2] Steve Mann. Wearable computing: A first step toward personal imaging. *IEEE Computer*; <http://wearingcam.org/ieeecomputer.htm>, 30(2):25–32, Feb 1997.
- [3] S. Mann and R. W. Picard. Video orbits of the projective group; a simple approach to featureless estimation of parameters. TR 338, Massachusetts Institute of Technology, Cambridge, Massachusetts, See <http://hi.eecg.toronto.edu/tip.html> 1995. Also appears in *IEEE Trans. Image Proc.*, Sept 1997, Vol. 6 No. 9, p. 1281–1295.
- [4] 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.
- [5] Qinfen Zheng and Rama Chellappa. A Computational Vision Approach to Image Registration. *IEEE Transactions Image Processing*, 2(3):311–325, 1993.
- [6] M. Irani and S. Peleg. Improving Resolution by Image Registration. *CVGIP*, 53:231–239, May 1991.
- [7] L. Teodosio and W. Bender. Salient video stills: Content and context preserved. *Proc. ACM Multimedia Conf.*, pages 39–46, August 1993.
- [8] R. Y. Tsai and T. S. Huang. Estimating Three-Dimensional Motion Parameters of a Rigid Planar Patch I. *IEEE Trans. Acoust., Speech, and Sig. Proc.*, ASSP(29):1147–1152, December 1981.
- [9] B. Horn and B. Schunk. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.
- [10] Steve Mann. Humanistic intelligence/humanistic computing: ‘wearcomp’ as a new framework for intelligent signal processing. *Proceedings of the IEEE*, 86(11):2123–2151+cover, Nov 1998. <http://wearingcam.org/procieee.htm>.
- [11] Steve Mann. **Wearable, tetherless computer-mediated reality:** WearCam as a wearable face-recognizer, and other applications for the disabled. TR 361, M.I.T. Media Lab Perceptual Computing Section; Also appears in **AAAI Fall Symposium on Developing Assistive Technology for People with Disabilities**, 9-11 November 1996, MIT; <http://wearingcam.org/vmp.htm>, Cambridge, Massachusetts, February 2 1996.



(a)

(b)

(c)

Figure 5 Mediated reality as a photographic/videographic memory prosthesis: (a) Wearable face-recognizer with virtual “name tag” (and grocery list) appears to stay attached to the cashier (b), even when the cashier is no longer within the field of view of the tapped eye and transmitter (c).



(a)

(b)

(c)

(d)

Figure 6 (a) Billboards, advertising, and other visual detritus form annoying, and sometimes dangerous clutter at the sides of busy roadways and highways. This advertisement, made in the shape of an octagon, and painted red, and placed at the side of a busy road, is the visual equivalent of yelling “fire” in a crowded theatre in order to get everyone’s attention to tell them you have something for sale. (b),(c),(d) Successive frames of video processed by the Eye Tap system using the VideoOrbits planetracker. The advertisement is filtered out, to reduce visual clutter in the scene. In its place is a useful message that can help the user of the Eye Tap system keep their attention on the road, and on not getting lost. When the rigid planar patch is not sufficiently within the visual field of view, approximate tracking still works based on other planar patches present in the scene.