What do we perceive from motion pictures? A computational account

Loong-Fah Cheong and Xu Xiang

Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Received June 19, 2006; revised December 20, 2006; accepted December 22, 2006; posted January 17, 2007 (Doc. ID 72021); published May 9, 2007

Cinema viewed from a location other than a canonical viewing point (CVP) presents distortions to the viewer in both its static and its dynamic aspects. Past works have investigated mainly the static aspect of this problem and attempted to explain why viewers still seem to perceive the scene very well. The dynamic aspect of depth perception, which is known as structure from motion, and its possible distortion, have not been well investigated. We derive the dynamic depth cues perceived by the viewer and use the so-called isodistortion framework to understand its distortion. The result is that viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems. Despite this shift, the key properties of the perceived depths remain largely the same, being determined in the main by the accuracy to which extrinsic motion parameters can be recovered. For a viewer seated at a noncentral position and watching the movie screen at a slant angle, the view is related to the view at the CVP by a homography, resulting in various aberrations such as noncentral projection.

OCIS codes: 330.5510, 330.7310.

1. INTRODUCTION

Three projections underlie the creating and viewing of motion pictures, namely, (a) the projection from the 3-D real scene to the film of the camera, (b) the back projection from the film onto the viewing screen, and (c) the projection from the screen to the human retina. These projections are assumed to be perspective in this paper.

Mathematically, only the audience located at a certain viewing position sees a "veridical" version of the scene, that is, as if he or she is seeing through the director's eyes and making the same movement. We call this position the canonical viewing position (CVP). All other positions receive visual stimuli different from the veridical version; the differences include dynamic visual cues such as optical flow, as well as depth information arising from such dynamic cues. Paradoxically, picture viewing is apparently not limited to the location at the CVP. A remarkably large number of positions in front of the projector can serve as reasonable viewpoints allowing layout within the motion picture to appear relatively normal. It is fortunate that the human visual system has this ability, for without it, the design of cinema theater and home entertainment systems would be severely constrained.

The paradox of the unnoticed distortions has been studied by researchers for about two decades. Cutting[1] argues that the slant at which pictures are viewed is usually small, and consequently the distortions of the retinal image are too small to be noticed. Perkins[2] claims that such invariance is a byproduct of the viewer's expectations with known shapes. For example, if the retinal image is similar to the image that would be created by a cube, prior expectations force the percept to that of a cube. The invariance thus comes from the viewer's experience with objects whose shapes are familiar or usually follow certain rules (right angles, parallel sides, symmetry). A third explanation claims that the invariance is the consequence of altering or re-interpreting the retinal image by recovering the position of the screens surface. For example, it is known[3] that the vanishing points of three mutually orthogonal lines are sufficient to recover the principal point. Banks *et al.*[4] argue that a local slant mechanism is used to estimate the foreshortening due to viewing obliqueness and then adjust the retinal image to undo the foreshortening.

All these hypotheses mainly attempt to deal with the static aspect of the paradox. Yet cinema is very much an art of camera motion, as attested by the original names of "kinetoscope" and "moving pictures." For Metz,[5] indeed, movement is the principal reason for the effect of reality within film. Motion dynamically changes the viewing perspectives of the spectators both in space and in time to give the unique reality effect, allowing the viewers to inhabit the visual space of the person(s) producing the film narrative. The depth information carried by motion cues is particularly relevant as cinema is typically viewed from a distance of 20 m or more, a condition under which accommodation, convergence, and stereoscopic depth perception are inactive. Last but not least, it is often through motion that the content or the meaning in a shot is expressed and the attention of the viewers captured or shifted, allowing the film's intentions to be communicated. Thus motion cue and depth perception arising from it should be the principal object of investigation in cinematic perception.

In theory, the optical flow present in the motion pictures and the dynamic depth cues arising therefrom should also experience distortion but have received very little attention. In fact, it is not even clear what sort of distortion is experienced by the viewer as far as the dynamic aspect is concerned. This neglect is partly due to

the distortion of depths arising from errors in the motion cues being an analytically complex problem; geometrical analysis that sheds light on this problem has only recently been formulated.[6,7] Our work focuses on this dynamic aspect of cinematic perception and investigates computationally the distortion of both the camera motion parameters and the depth recovered from such distorted motion cues.

To recover the spatial structure using optical flow present on the picture screen amounts to the classical structure-from-motion (SFM) problem with a slight twist. We will introduce this modified SFM model in Section 3. The typical SFM problem has been the central problem of computer vision since the 1980s. It recovers the structure of 3-D scene and the 3-D relative motion between the scene and the observer from the projection of the 3-D relative motion onto a 2-D surface. If the 3-D motion parameters can be estimated perfectly, depth recovery can be achieved accurately; in other words, one can perceive the spatial arrangement of objects. However, this veridical space recovery from SFM is difficult to achieve, as has been shown both computationally and experimentally. Either the 3-D motion estimates contain errors with the result that depths are distorted, or the intrinsic parameters of the camera are unknown, in which case one can recover only the so-called projective depth,[8] which is related to the true depth by a projective transformation.

Since errors in motion estimates are highly likely, there have been various error analyses in the past[9–14] in terms of the local minima and ambiguities of the SFM algorithms. However, there has been much less analysis of the behavior of depth distortion given some errors in the motion estimates. Cheong *et al.*[6] developed a geometric account of the depth error behavior via the so-called isodistortion framework. It showed that even with known intrinsic parameters but with errors in the 3-D motion estimates, the distortion transformation from physical to perceived space is already highly complex, in fact more complicated than that of the projective transformation. It is a space Cremona transformation, which is a rational transformation between two projective spaces.[15]

Given such potentially complex distortion behavior, Cheong and Xiang[7] were then motivated by the importance of special generic motions favored by biological visual systems. One such motion is lateral motion, which consists of lateral translation plus rotation. Such motion will, despite errors in the estimates, yield a special type of Cremona transformation that preserves depth order. We say that such transformation exhibits ordinal depth invariance. Another generic motion type is forward motion (forward translation plus rotation), which gives rise to conditions conductive for 3-D motion recovery but not for depth recovery. The idea here is that different motion types are suited for specific tasks; this is important since there is no general motion algorithm that can work well under all motion-scene configurations.

The SFM process is further complicated by the presence of intrinsic parameters such as the focal length and the principal point coordinates. Cheong and Xiang[7] further showed that as long as the focal length is not dynamically varying (i.e., the camera is not performing a zoom operation) and the error in the principal point estimation is small enough, the aforementioned properties of spatial perception under different generic motions are still preserved.

Whether visual systems in nature have precise knowledge of the eyes' intrinsic parameters when processing visual tasks is still unknown. Nevertheless psychophysics researchers studying the perception of scene structure from dynamic cues[16–18] tend to assume that the brain uses a calibrated visual system and neglect the problem of calibration altogether. This is mainly due to the elaborate model needed to model the complex intrinsic parameters of human eyes, making it very difficult to incorporate them into computational analysis. In this paper, we consider only the typical intrinsic parameters used for modeling a pinhole camera.[19] The extent to which these intrinsic parameters are calibrated determines the type of space that can be perceived from motion cues. A host of models proposed for the geometric structure of this perceived visual space, varying from Euclidean geometry[20] to hyperbolic[21] to affine[22] and others.[23,24] Recently Droulez and Cornilleau-Pérès's anamorphosis glasses[25] showed that the visual system is able to recalibrate a Riemannian metric adapted to the glasses' deformation, and a Euclidean geometry can be perceived after the plastic adaptation. Another experimental result[26] that supports the assumption that brain cognition is more "Euclidean than affine or projective" is that when perceiving the orientation of a surface drawn using curves, subjects preferentially consider the orthogonality cue rather than parallelism. Viéville *et al.*[27] report that the human visual system is able to take intrinsic parameter variations into account during perceptual tasks. It has also been argued[28] that the more recently evolved vision-for-perception system is quite different from those of the more ancient vision-for-action system, and the latter is based on Euclidean object metrics.

In spite of these results, there is psychophysical evidence that suggest human vision is not Euclidean under all conditions,[29,30] especially in the impoverished scenarios typically encountered in psychophysical experiments (e.g., random dots in motion). For instance, Cheong *et al.*[29] reported that the recovery of curvatures under lateral translation is subject to varying degrees of uncertainty depending on the motion-scene configuration. In particular, the theory proposed therein explained why the reconstructed second-order shape tends to be more distorted in the direction parallel to the translational motion than that in the orthogonal direction. This orientational anisotropy has also been reported in many psychophysics papers. Refs. 31–35 studied the perception of second-order shapes under active vision, and it was found that some types of shapes can be perceived quite accurately, whereas others are more difficult to be distinguished. Thus, on the whole, it seems that human's visual space deriving from motion cues is quite plastic and grades from being nearly Euclidean to nonmetrical depending on tasks and conditions.

In this paper we seek to use the isodistortion framework to analyze the nature of the depths recovered from dynamic cues under the cinema configuration. We show that, with respect to SFM, viewing a movie in a cinema from a general position differs from viewing a 3-D real

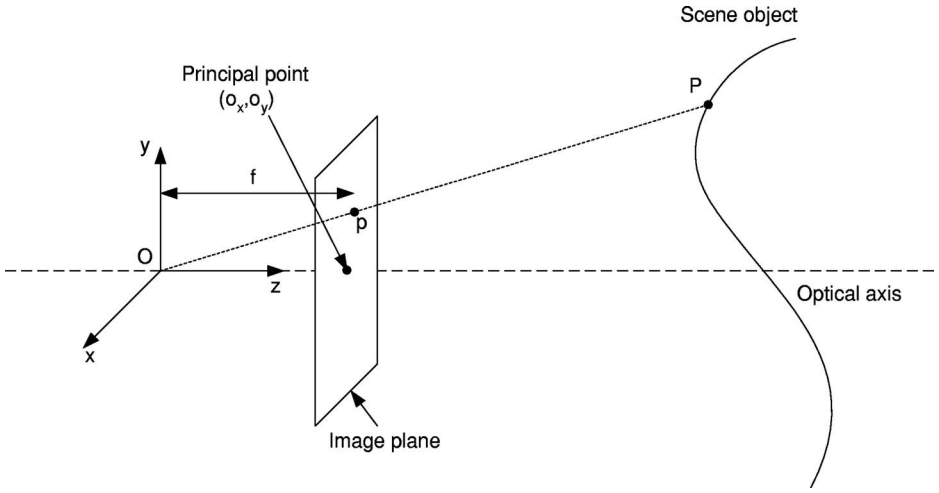


Fig. 1. Image formation model: O is the optical center. The optical axis is aligned with the Z axis and the horizontal and vertical image axes are aligned with the X and the Y axis, respectively. A world point $\mathbf{P}=(X,Y,Z)^T$ is projected to its image pixel coordinate (x,y). The focal length is denoted by f and the principal point by (o_x,o_y).

scene primarily in that the visual system experiences an altered optical flow resulting from changed intrinsic parameters. To be exact, this statement is only correct for a viewer seated at a reasonably central position. The impacts of these altered optical flows on depth perception from different seating positions are elucidated and compared with SFM under normal condition. Results show that even with the shift in the intrinsic parameters, the key properties of the recovered depth-from-motion remain largely the same despite some differences from the case of normal uncalibrated SFM discussed by Cheong and Xiang,[7] and these key properties are determined primarily by the degree of accuracy with which the extrinsic parameters can be recovered and by the types of motions being executed. In sum, the main contribution of this paper is to show the geometric laws governing distortions in the perceived space arising from motion cues and to make explicit those situations that lead to different types of distortions. The implications of these results for SFM under cinematic viewing and uncalibrated vision in general will be further discussed below, but these speculative possibilities have to be further investigated by comprehensive psychophysical tests in the light of the types of distortion and their motion-scene dependence that are unraveled here.

2. MODEL AND PREREQUISITE

Figure 1 introduces the notation associated with the general projection process. The reference frame is attached to the optical center at O. A world point $\mathbf{P}=(X,Y,Z)^T$ is projected to its image pixel coordinate (x,y) by the following well-known transformation[19]:

$$\mathbf{p}=\begin{pmatrix}x\\y\\1\end{pmatrix}=\mathbf{K}\Pi_0\mathbf{P}=\begin{bmatrix}f & s_\theta & o_x\\0 & f & o_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}\begin{pmatrix}X\\Y\\Z\\1\end{pmatrix}, \quad (1)$$

where we have expressed $\mathbf{p}$ and $\mathbf{P}$ in homogeneous coordinates, with slight abuse of notation in using $\mathbf{p}$ and $\mathbf{P}$ for both homogeneous coordinates and inhomogeneous coordinates, since it is usually clear from the context which one we are referring to. The constant 3×4 matrix Π_0 represents the perspective projection, and the upper triangular 3×3 matrix $\mathbf{K}$ is the intrinsic parameter matrix used to model a pinhole camera, with the focal length denoted by f, the x and y coordinates of the principal point by (o_x,o_y), respectively, and the skew factor by s_θ.

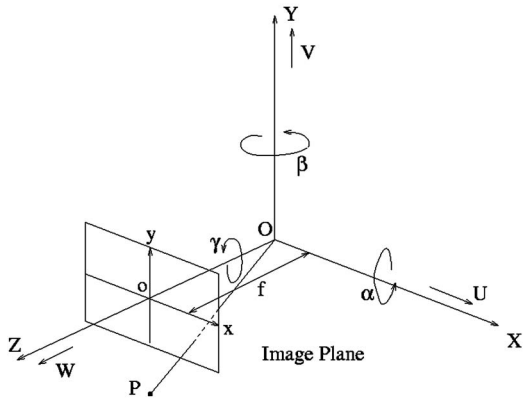


Fig. 2. Three-D camera motion with a translational velocity $\mathbf{v}=(U,V,W)^T$ and a rotational velocity $\mathbf{w}=(\alpha,\beta,\gamma)^T$. The motion induces a relative motion between the static scene point $\mathbf{P}$ and the camera.

We now present the notation associated with the conventional SFM problem, ignoring all the intrinsic parameters except f. This is equivalent to saying one has perfect estimates of other intrinsic parameters so that one can appropriately transform the image coordinates to obtain $s_\theta=o_x=o_y=0$. If the camera undergoes a motion with a translational velocity $\mathbf{v}=(U,V,W)^T$ and a rotational velocity $\mathbf{w}=(\alpha,\beta,\gamma)^T$ (see Fig. 2), the motion induces a relative motion between the static scene point $\mathbf{P}$ and the camera. The relative 3-D velocity of $\mathbf{P}$ (with respect to the camera) can be written as

$$\dot{\mathbf{P}}=-\mathbf{v}-\mathbf{w}\times\mathbf{p}, \quad (2)$$

from which the well-known 2-D motion field equations[36] can be derived:

$$u = \frac{W}{Z}x - f\frac{U}{Z} + \frac{xy}{f}\alpha - f\left(1 + \frac{x^2}{f^2}\right)\beta + \gamma y, \tag{3}$$

$$v = \frac{W}{Z}y - f\frac{V}{Z} - \frac{xy}{f}\beta + f\left(1 + \frac{y^2}{f^2}\right)\alpha - \gamma x, \tag{4}$$

where (u,v) is the optical flow at the feature point (x,y) on the image plane.

It is well known that the depth can be recovered only up to an arbitrary scale factor; thus we can set $W=1$ without loss of generality, except for the case where $W=0$. Henceforth Z represents the scaled depth Z/W with W set to 1 unless explicitly stated otherwise.

If the 3-D motions have been estimated, Z can in turn be obtained from Eq. (3) or (4). Usually a direction $\mathbf{n}$ is chosen according to some criteria to recover Z. Thus Z can be obtained as

$$Z = \frac{(x - x_0, y - y_0)\cdot\mathbf{n}}{(u - u_{rot}, v - v_{rot})\cdot\mathbf{n}}. \tag{5}$$

In Eq. (5), $(x_0,y_0)=(f(U/W),f(V/W))$ denotes the focus of expansion (FOE), (u_{rot},v_{rot}) are the rotational components of Eqs. (3) and (4), respectively, and $\mathbf{n}$ is the unit vector in the direction chosen to recover Z. As an example, $\mathbf{n}$ can be along the normal flow direction because the flow along this direction can be most reliably estimated. In the case where optical flow can be recovered well, other considerations might lead to the choice of recovering depth along the direction emanating from the estimated FOE (x_0,y_0) based on the intuition that this direction (also known as the epipolar direction) contains the strongest translational flow and thus provides the best estimate of depth.

It follows that if there are some errors in the estimation of the extrinsic parameters, Z will be estimated with errors, that is, a distorted version of the space will be perceived. The detailed analysis of this depth distortion will be deferred to Section 4, after we have introduced the modified form of the SFM problem under the cinema viewing configuration.

3. STRUCTURE FROM MOTION UNDER CINEMA VIEWING CONFIGURATION

A. Optical Axes of Viewer and Projector Parallel

We first consider the case in which the viewer's optical axis is parallel to the projector's optical axis, and the screen is oriented in a frontoparallel manner to the projector and the viewer. This is applicable to most cinema viewers who are seated not near the side or right at the front (see Fig. 3). As the seats are designed to face forward, the viewers will do so unless they are positioned so far to the side or front that they are obliged to tilt their viewing axis toward the central area of the screen. We assume the cinema images captured by the director have been transferred to film for optical projection and we call this film the projector film. We also assume monocular viewing to focus just on motion cues. We use subscripts p, v, s to represent quantities associated with projector, actual viewer, and screen, respectively. The distances (along the Z axis) from the screen to the projector and to the viewer are D_p and D_v, respectively. The focal length of the projector and of the viewer's visual system are f_p and f_v, respectively.

Consider the simplest case where the viewer's optical axis is not only parallel to but also coincident with the projector's optical axis. Then clearly the feature points $\vec{x}_p$, $\vec{x}_s$, and $\vec{x}_v$ [see Fig. 3(a)] are related by

$$\vec{x}_p = \frac{f_p\vec{x}_s}{D_p}, \tag{6}$$

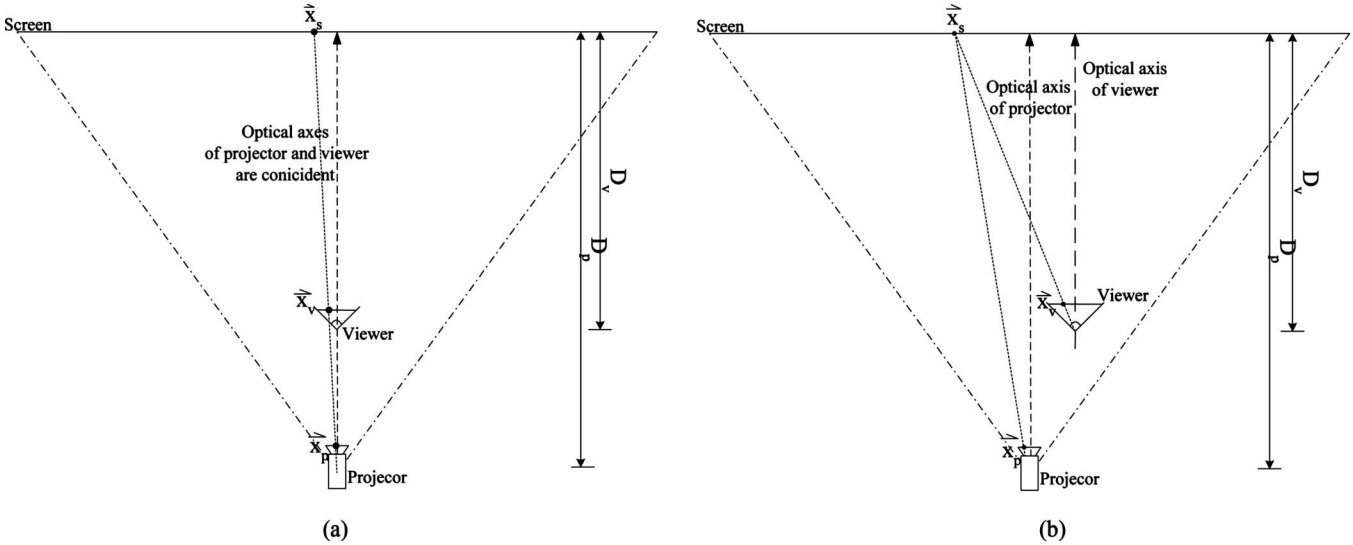


Fig. 3. Simple cinema viewing configuration with the optical axes of the viewer and the projector parallel, a situation applicable to most viewers who are seated not near the side or right at the front. $\vec{x}_p,\vec{x}_s,\vec{x}_v$ represent, respectively, the feature points on the projector film, screen, and viewer's retina corresponding to the same world point. The distances (along the Z axis) from the screen to the projector and to the viewer are D_p and D_v, respectively. (a) Optical axes of viewer and projector are coincident. (b) Optical axes of viewer and projector are not coincident but are parallel to each other.

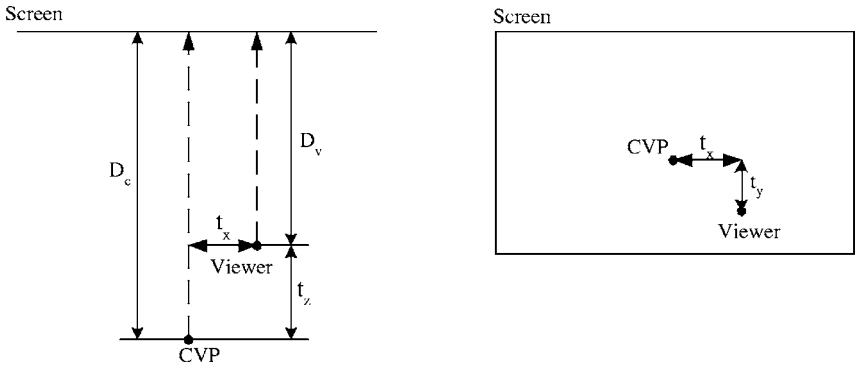


Fig. 4. Configuration in which the viewer's and projector's optical axes are parallel but not coincident. The viewer is displaced from the CVP by (t_x,t_y,t_z).

$$\vec{x}_v = \frac{f_v \vec{x}_s}{D_v} = \frac{\vec{x}_p}{k}, \tag{7}$$

where $\vec{x}_s$ is given in metric units, $\vec{x}_p$ and $\vec{x}_v$ are in pixel units, and

$$k = \frac{D_v f_p}{D_p f_v}.$$

Assume there is a 2-D motion flow $\vec{u}_p=(u_p,v_p)$ on the projector film and the corresponding flows on the screen and the viewer's retina are denoted by $\vec{u}_s$ and $\vec{u}_v$, respectively. From Eqs. (6) and (7), we have

$$\vec{u}_p = \frac{f_p \vec{u}_s}{D_p}, \tag{8}$$

$$\vec{u}_v = \frac{f_v \vec{u}_s}{D_v} = \frac{f_v D_p \vec{u}_p}{D_v f_p} = \frac{\vec{u}_p}{k}. \tag{9}$$

Equation (9) suggests that the flow $\vec{u}_v$ perceived on the retina is scaled by a factor k compared with the corresponding flow $\vec{u}_p$ on the projector film. The flow is given by

$$u_p = \frac{W}{Z}\left(x_p - f_c\frac{U}{W}\right) + \alpha\frac{x_p y_p}{f_c} - \beta\left(\frac{x_p^2}{f_c} + f_c\right) + \gamma y,$$

$$v_p = \frac{W}{Z}\left(y_p - f_c\frac{V}{W}\right) - \beta\frac{x_p y_p}{f_c} + \alpha\left(\frac{y_p^2}{f_c} + f_c\right) - \gamma y, \tag{10}$$

where the 3-D motion parameters (U,V,W) and (α,β,γ) represent the motion experienced by the director's camera, and f_c is the focal length of the director's camera. Expanding the horizontal component of $\vec{u}_v$ in Eq. (9) and bringing in equation (10), we obtain

$$u_v = \frac{W}{Z}\left(x_v - \frac{f_c U}{kW}\right) + \alpha\frac{x_v y_v}{f_c/k} - \beta\left(\frac{x_v^2}{f_c/k} + \frac{f_c}{k}\right) + \gamma y_v. \tag{11}$$

A similar expression can be written for the vertical component of the flow v_v. From Eq. (11), we see that the flow field experienced by the viewer indirectly through the screen is one that arises from the same external motion and depths experienced by the director's camera, i.e., $U,V,W,\alpha,\beta,\gamma,Z$, but with a modified focal length f_v' $=f_c/k=f_v D_p f_c/D_v f_p$. Thus only when the viewer is seated at the CVP $[D_v=(D_p f_p)/f_c]$ is the motion field $\vec{u}_v$ undistorted (i.e., the same as that experienced by a viewer making the 3-D motion himself/herself).

Clearly, if the viewer is able to revise the estimate of the intrinsic parameter from f_v to f_v', he or she is then no worse off than in the case of having to solve the SFM problem when experiencing an undistorted 2-D motion flow. Even when the viewer is not able to estimate the new focal length, we shall show later that the effect of this focal length error is benign as far as scene structure recovery is concerned.

The above analysis can now be extended to the case where the viewer's and the projector's optical axes are parallel but not coincident [Fig. 3(b)]. If the viewer is displaced from the CVP by (t_x,t_y,t_z), there will be a shift in the principal point (o_x,o_y). Since $o_x/t_x=-f_v/D_v$ and o_y/t_v $=-f_v/D_v$, we have $(o_x,o_y)=[-t_x(f_v/D_v),-t_y(f_v/D_v)]$ (see Fig. 4). Thus the optical flow can be written in the following form:

$$u_v = \frac{W}{Z}\left(x_v + t_x\frac{f_v}{D_v}\right) - f_v'\frac{U}{Z} + \frac{\left(x_v + t_x\frac{f_v}{D_v}\right)\left(y_v + t_y\frac{f_v}{D_v}\right)}{f_v'}\alpha - f_v'\left(1 + \frac{\left(x_v + t_x\frac{f_v}{D_v}\right)^2}{f_v'^2}\right)\beta + \gamma\left(y_v + t_y\frac{f_v}{D_v}\right), \tag{12}$$

$$v_v = \frac{W}{Z}\left(y_v + t_y\frac{f_v}{D_v}\right) - f_v'\frac{V}{Z} - \frac{\left(x_v + t_x\frac{f_v}{D_v}\right)\left(y_v + t_y\frac{f_v}{D_v}\right)}{f_v'}\beta + f_v'\left(1 + \frac{\left(y_v + t_y\frac{f_v}{D_v}\right)^2}{f_v'^2}\right)\alpha - \gamma\left(x_v + t_x\frac{f_v}{D_v}\right). \tag{13}$$

This is similar to the optical flow that would be obtained if the principal point of the viewer's optical system were not (0, 0), as we have assumed so far, but were given by $[t_x(f_v/D_v),t_y(f_v/D_v)]$.

In sum, for the simple scenario where the viewer's and the projector's optical axes are parallel, the motion estimation problem is no more complex than an uncalibrated

SFM problem, where in particular the focal length f'_v might be different over time because the director is using different lenses, but over most of the time f'_v would not be dynamically varying (i.e., f'_v piecewise constant) unless the director were using the zooming shot. The difference from the usual uncalibrated SFM problem is that the principal point offset $[t_x(f_v/D_v), t_y(f_v/D_v)]$ can be very much larger [especially $t_x(f_v/D_v)$] than one usually encounters in a computer vision problem.

In general, if the intrinsic parameters are not calibrated, the visual system cannot recover the Euclidean geometry of the scene from motion cues, but only its projective or affine geometry.[37] The question remains whether a human "calibrates" his or her visual system. Indeed, the anatomy of the eye varies from the fovea to the periphery and such parameters also change over time. Computationally, estimation of intrinsic parameters from the motion cues is possible, even for varying focal length and principal point.[38] However, it is numerically ill-conditioned and always requires higher-order constraints. It seems that such higher-order mechanisms cannot explain how the brain may estimate eye intrinsic parameters.[27] It could be that the visual system needs only to obtain a very rough estimate of the intrinsic parameters, and knowing these rough estimates, it is sufficient to obtain certain aspects of depth information.[7] If this were indeed the case, then what the cinema viewer experiences is just a more severe version of the situation with larger errors in the estimates for the intrinsic parameters.

We shall see in Section 4 how these errors in the intrinsic parameters (as well as the extrinsic parameters) will affect spatial perception from motion cues. In particular, we will compare the changes caused by the large principal point offset $[t_x(f_v/D_v), t_y(f_v/D_v)]$ in the characteristics of the depth distortion. More important, we show that despite these errors, certain key qualitative properties of the structure recovered from motion remain unchanged.

B. Optical Axes of Viewer and Projector Not Parallel

In general, a movie viewer could watch the screen with a slant angle if he/she does not sit right on the optical axis of the projector (see Fig. 5). To relate how the dynamic cues are transformed as a result of the slant, we relate the view of a viewer seated at CVP and that of a viewer seated at a general position via a homography (a 3×3 linear transformation) induced by the screen plane.[39] Here we again assume that the axis of the projector is perpendicular to the screen. In other words, no keystone distortion[40] is present in the cinema. Thus the unit vector normal to the screen plane is given by $\mathbf{N}=(0,0,-1)^T$. Then the coordinates of a feature point $\mathbf{x_v}$ of a viewer seated at CVP and $\mathbf{x'_v}$ of a viewer seated at the general position can be related by a simple homography $\mathbf{H}$,

$$\mathbf{x'}_\mathbf{v} = \mathbf{H}\mathbf{x}_\mathbf{v}, \tag{14}$$

where the homography is given by[39]

$$\mathbf{H} = \mathbf{K'}_\mathbf{v}\left(\mathbf{R} + \frac{1}{D_c}\mathbf{t}\mathbf{N}^T\right)\mathbf{K}_\mathbf{v}^{-1}. \tag{15}$$

In Eq. (15) the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ denote the rigid transformation between the viewer and the CVP. We have assumed that the human eye is modeled by a pinhole camera; thus $\mathbf{K'_v}$ and $\mathbf{K_v}$ are the intrinsic parameter matrices characterizing the eyes of the viewer seated at the CVP and the general position, re-

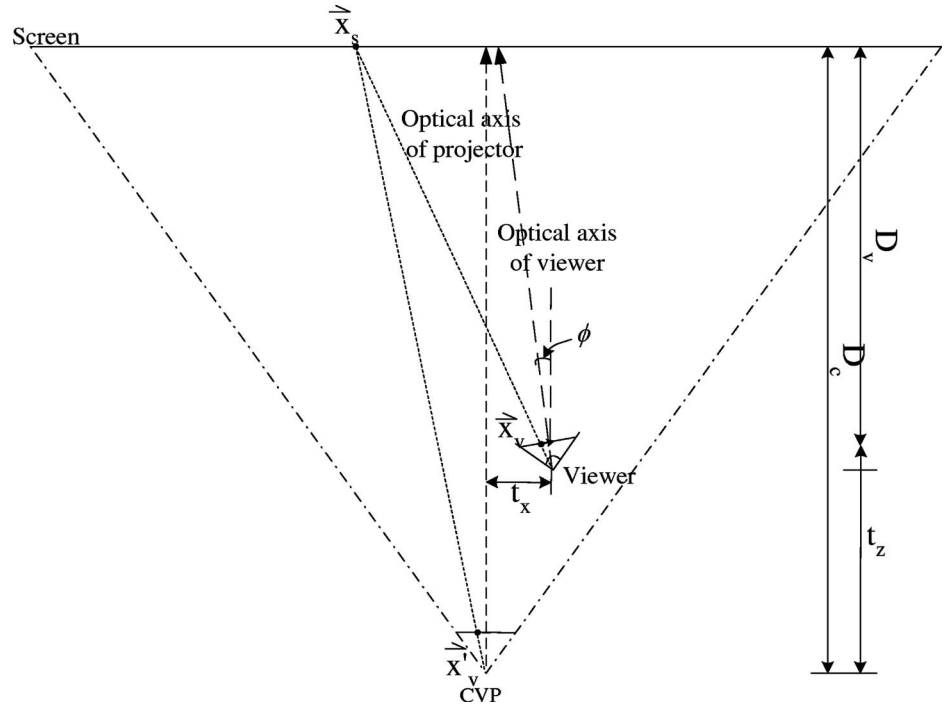


Fig. 5. General configuration with a slant angle ϕ between the viewer's optical axis and the vertical axis. The view of a viewer seated at CVP (with feature $\vec{x}_v$) is related to that of a viewer seated at a general position (with feature $\vec{x}'_v$) via a homography induced by the screen plane.

spectively, with the form of $\mathbf{K}_\mathbf{v}'$ and $\mathbf{K}_\mathbf{v}$ given by that of $\mathbf{K}$ in Eq. (1). We assume the s_θ, o_{xv}, o_{yv} for the eyes at both positions to be zero. We also assume the focal lengths f_v for the eyes at both positions to be identical, because given the typical distance of the screen, both focal lengths will correspond to the eyes at the most relaxed state. Thus

$$\mathbf{K}_\mathbf{v} = \begin{bmatrix} f_v & 0 & 0 \\ 0 & f_v & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{16}$$

We assume that the viewer is gazing at a region near the central part of the screen, and we first derive the simple case where the viewer is at the same vertical level as the projector. There is an angle of ϕ between the viewer's optical axis and the vertical axis but there is no rotation around the X axis. Conceptually, one can also reduce any rotations around both X and Y axes to a single rotation about the Y axis by a suitable in-plane rotation of the X and Y coordinate axes. Referring to Fig. 5, the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ can thus be written, respectively, as

$$\mathbf{R} = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix}, \tag{17}$$

$$\mathbf{t} = [t_x \; 0 \; t_z]^T. \tag{18}$$

The homography $\mathbf{H}$ is then readily obtained as

$$\mathbf{H} = \mathbf{K}_\mathbf{v} \begin{bmatrix} \cos\phi & 0 & -\sin\phi - \dfrac{t_x}{D_c} \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi - \dfrac{t_z}{D_c} \end{bmatrix} \mathbf{K}_\mathbf{v}^{-1}. \tag{19}$$

Using the projection model of Eq. (1), the whole projection process can then be written as

$$\mathbf{x}'_\mathbf{v} = \mathbf{H}\mathbf{x}_\mathbf{v} = \mathbf{H}\mathbf{K}_\mathbf{v}[\mathbf{I}|0]\mathbf{X}, \tag{20}$$

where $\mathbf{X}$ is the 3-D point that gives rise to $\mathbf{x}_\mathbf{v}$ ($\mathbf{x}_\mathbf{v}$, $\mathbf{x}'_\mathbf{v}$, $\mathbf{X}$, all expressed in homogeneous coordinates).

Equation (20) shows that there is a new "intrinsic parameter matrix" $\mathbf{H}' = \mathbf{H}\mathbf{K}_\mathbf{v}$ underlying the image formation process of a cinema viewer seated at a general position. Unfortunately the intrinsic parameter matrix induced by the homography is not an upper-triangular matrix like that of a typical intrinsic parameter matrix:

$$\mathbf{H}' = \mathbf{H}\mathbf{K}_\mathbf{v} = \begin{bmatrix} f_v\cos\phi & 0 & -f_v\sin\phi - \dfrac{t_x}{D_c}f_v \\ 0 & f_v & 0 \\ \sin\phi & 0 & \cos\phi - \dfrac{t_z}{D_c} \end{bmatrix}. \tag{21}$$

If ϕ is sufficiently small, we can simplify $\mathbf{H}'$ such that the effect of ϕ can be regarded as a perturbation of $\mathbf{K}_\mathbf{v}$. First, the lowest right entry can be approximated by

$$\cos\phi - \frac{t_z}{D_c} \approx \frac{D_c - t_z}{D_c} = \frac{D_v}{D_c}. \tag{22}$$

Since $\mathbf{H}'$ is up to an arbitrary scale factor that is inherent in homogeneous representation, we can scale the whole matrix such that the lowest right entry is unity, so

$$\mathbf{H}' = \mathbf{H}\mathbf{K}_\mathbf{v} \begin{bmatrix} f_v\dfrac{D_c}{D_v}\cos\theta & 0 & -\dfrac{t_x}{D_v}f_v \\ 0 & f_v\dfrac{D_c}{D_v} & 0 \\ \dfrac{D_c}{D_v}\sin\theta & 0 & 1 \end{bmatrix} \tag{23}$$

Clearly, if ϕ is small enough that $\cos\phi \approx 1$ and $\sin\phi \approx 0$, then the intrinsic parameter matrix reduces to that of Subsection 3.A. However, in the general case, the matrix does not have the typical form for the intrinsic parameter matrix in view of the nonzero lowest left entry. Then x'_v, y'_v can be written, respectively, as

$$x'_v = \frac{f_v\dfrac{D_c}{D_v}X\cos\theta - \dfrac{t_x}{D_v}f_vZ}{Z + \dfrac{D_c}{D_v}X\sin\theta}, \tag{24}$$

$$y'_v = \frac{f_vY}{Z + \dfrac{D_c}{D_v}X\sin\theta}. \tag{25}$$

The projection at the general position can thus be regarded as one not only with changes in focal length and principal point offset, but now these changes also vary in magnitude from the fovea to the periphery [the denominators in Eqs. (24) and (25) change as X increases from fovea to periphery]. In other words, the projection rays do not intersect at one point, with the result that we have a noncentral projection system.[41,42]

How would this affect the viewer seated at this general position? The human visual system is itself prey to nonideal effects like spherical aberration, coma, and other asymmetries expected from a biological system. For instance, the optical surfaces may lack rotational symmetry and their nominal centers of curvature may not lie on a common axis; such meridional changes in radius of curvature lead to ocular astigmatism.[43] Though these aberrations occur in the human eyes, their visual effects are minimal. For instance, the astigmatic image falls on the peripheral retina, which has relatively poor resolving power compared with the retina at the macula. Thus peripheral spatial vision performance seems little affected, though the effect of these off-axis errors on spatial and temporal sampling in the periphery is not yet completely determined, with some recent studies being described in Refs. 44–46.

The question here is whether there is a need for the visual system to recalibrate itself (to whatever extent) with such more severe aberrations being introduced. The answer depends on the type of space recoverable—or indeed

being recovered—by the human visual system under both everyday SFM and under the cinema viewing condition. We now come to the central question of depth distortion under both situations in Section 4.

4. DEPTH DISTORTION ARISING FROM ERRONEOUS ESTIMATION OF 3-D MOTION AND INTRINSIC PARAMETERS

A. Isodistortion Framework

The isodistortion framework was first introduced by Cheong *et al.*[6] The isodistortion framework seeks to understand the geometric laws under which the recovered scene is distorted due to some errors in the estimated camera parameters. This is motivated by the fact that it is unlikely for a human visual system to recover the exact motion parameters, and hence it is important to understand how the perceived space is distorted by such errors in the motion estimates. The distortion in the perceived space recovered from motion is visualized by looking at the locus of equal distortion, known as the isodistortion surfaces. This makes explicit the systematic way in which depths are distorted and leads to its algebraic characterization by Cremona transformation.[15]

We introduce further notations for our distortion analysis. The estimated parameters are denoted by the hat symbol (ˆ) and errors in the estimated parameters by the subscript e. The error of any estimate r is defined as $r_e = r - \hat{r}$.

Referring to Eq. (5), we note that if there are errors in the estimates of the extrinsic parameters, these errors will in turn cause errors in the estimation of the scaled depth. We replace the motion terms in Eq. (5) by their various estimates. Furthermore, assuming no noise in the optical flow (since we are primarily concerned with how errors in the 3-D motion parameters affect depth reconstruction), we can express the optical flow (u, v) in terms of the true motion parameters via Eqs. (3) and (4). The distorted depth $\hat{Z}$ is obtained as follows:

$$\hat{Z} = Z\left(\frac{(x-\hat{x}_0, y-\hat{y}_0)\cdot\mathbf{n}}{(x-x_0, y-y_0)\cdot\mathbf{n} + Z(u_{rot_e}, v_{rot_e})\cdot\mathbf{n}}\right). \quad (26)$$

Equation (26) shows that errors in the motion estimates distort the recovered relative depth by a factor D, given by the terms in the bracket, which among other terms, contains the term $\mathbf{n}$. As mentioned in the discussion following Eq. (5), the value of $\mathbf{n}$ depends on the scheme we use to recover depth. In this paper, we choose to recover depth along the estimated epipolar direction, i.e., $\mathbf{n} = (x-\hat{x}_0, y-\hat{y}_0)^T / \sqrt{(x-\hat{x}_0)^2 + (y-\hat{y}_0)^2}$. Such a choice is reasonable because the estimated epipolar direction contains the strongest translational flow and hence is the most reliable direction to recover Z. Hence the distortion factor D becomes

$$D = \frac{(x-\hat{x}_0)^2 + (y-\hat{y}_0)^2}{(x-x_0, y-y_0)\cdot(x-\hat{x}_0, y-\hat{y}_0) + Z(u_{rot_e}, v_{rot_e})\cdot(x-\hat{x}_0, y-\hat{y}_0)}. \quad (27)$$

The complexity of Eq. (27) can be intuitively grasped with a graphical approach in its first analysis. For specific values of the parameters x_0, y_0, $\hat{x}_0$, $\hat{y}_0$, α_e, β_e, γ_e and for any fixed distortion factor D, Eq. (27) describes a surface $g(x, y, Z) = 0$ in the xyZ space. The entire ensemble of such surfaces, each for a different value of D, describes the distortion action of the motion errors on any points in the 3-D space. Normally, under general motion, a complicated distortion characteristic may arise. Readers are referred to Refs. 6 and 7 for a full description of the geometry of the distortion. We briefly summarize the salient features for the purposes of this paper.

Algebraically, it was shown in Ref. 6 that given such motion errors, the transformation from the physical to the perceived space belongs to the family of Cremona transformations whereby the homogeneous coordinates of a point in the perceived space $[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}, \hat{\mathcal{W}}]$ is related to the actual point $[\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}]$ by

$$[\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{\mathcal{Z}}, \hat{\mathcal{W}}] = [\phi_1, \phi_2, \phi_3, \phi_4],$$

where the quantities ϕ_i are homogeneous polynomials in $[\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}]$. Such transformation is bijective almost everywhere except on the set of what is known as fundamental elements where the correspondence between the two spaces becomes one-to-many.[15] The complex nature of this transformation makes it clear that in general it is very difficult to recover metric depth accurately. What is less clear is the feasibility of recovering some of the less metrical depth representations under specific motions. For instance, the ordinal representation of depth constitutes one such reduced representation of depth where only depth order is available. Cheong and Xiang[7] showed that although in the general case small amounts of motion errors can have significant effects on depth recovery, there exist generic motions that allow robust recovery of such partial depth information. In particular, lateral motion is better than forward motion in terms of yielding ordinal depth information and other aspects of depth recovery, in spite of the ambiguity between the camera rotation and translation being more severe in this case. On the other hand, forward motion leads to conditions more conductive to 3-D motion estimation compared with the case of lateral motion, but it is not necessarily good for depth recovery. This dichotomy between forward and lateral motion means that it is important for a biological system to choose a motion intelligently so as to accomplish tasks robustly.

In the case of uncalibrated motion with fixed intrinsic parameters and reasonably small principal point offset, the distortion factor D becomes[7]

- for lateral motion ($W=0$),

$$D = \frac{\hat{f}\hat{U}}{fU + (\beta f - \hat{\beta}\hat{f})Z}, \tag{28}$$

- for forward motion ($U=V=0$),

$$D = \frac{x^2 + y^2}{[(x - o_{xe})x + (y - o_{ye})y] + [-(\beta f - \hat{\beta}\hat{f})x + (\alpha f - \hat{\alpha}\hat{f})y]Z}. \tag{29}$$

It was shown in Ref. 7 that lateral motion is better than forward motion in terms of yielding ordinal depth information, in spite of the fact that the ambiguity between the camera rotation and translation is more severe in this case. On the other hand, forward motion leads to conditions more conducive to 3-D motion estimation compared with the case of lateral motion, but it is not necessarily good for depth recovery. The aforementioned factors regarding the dichotomy in depth and motion recovery are not affected, in spite of possible errors in the intrinsic parameters. However, if the intrinsic parameters are allowed to vary dynamically (equations for D under such case not shown here; see Ref. 7 for a fuller account), then even ordinal depth information might not be recoverable under lateral motion.

The upshot of characterizing depth distortion behavior under these generic types of forward and lateral motions are the following two aspects: (1) It shows that the reliability of a reconstructed scene has quite a different behavior from that of the motion estimates. For instance, if the motion contains dominant lateral translation, it might be very difficult to lift the ambiguity between translation and rotation. However, in spite of such motion ambiguity, certain aspects of depth information seem recoverable with robustness. Indeed, in the biological world, lateral motions are often executed to judge distance and relative ordering. On the other hand, psychophysical experiments[47] reported that under pure forward translation, human subjects were unable to recover structure unless favorable conditions such as large field of view existed. Thus it seems that not all motions are equal in terms of robust depth recovery and that there also exists a certain dichotomy between forward and lateral translation as far as motion and depth recovery are concerned. (2) Understanding the depth recovered under these two very different motion types gives us an epistemological idea about the geometry of the perceived space under general motions, in the sense that the behavior of depth reconstruction at these two opposite poles of translation spectrum delimits the type of general depth distortion behavior somewhere between the two poles. Clearly, in the absence of other depth cues, or without using additional scene knowledge, Euclidean or even affine depth recovery may not be possible in general.

B. Depth Distortion in Cinema

We now apply the isodistortion framework to examine the SFM problem under the cinema viewing configuration. Like previous isodistortion analyses, we restrict ourselves to scenes where only the camera is moving, or we assume that in scenes where there are independently moving objects, these objects have been properly segmented. We focus on the situation depicted in Fig. 3(b), which has been shown to be equivalent to an uncalibrated SFM problem for the viewer, with mostly fixed but possibly unknown focal length and potentially very large principal point offset.

One might ask to what extent the notion of generic motion employed in the previous analyses is valid or relevant in the cinema context. In cinematography, camera motions are not arbitrary, but are dictated by the need to communicate meanings and by the mechanics of filmmaking. For instance, a panning shot is often used to establish the scenes of a new shot and to track an object or person. A dolly shot [translation in depth; see Fig. 6(a)] is used to move in closer to a subject or to affect a first-person viewpoint shot as the protagonist moves forward. Shots with more complex combinations of motions are possible; for instance, translation and rotation are often coupled together in tracking shots using the setup illustrated in Fig. 6(b). Nevertheless, it is reasonable to say

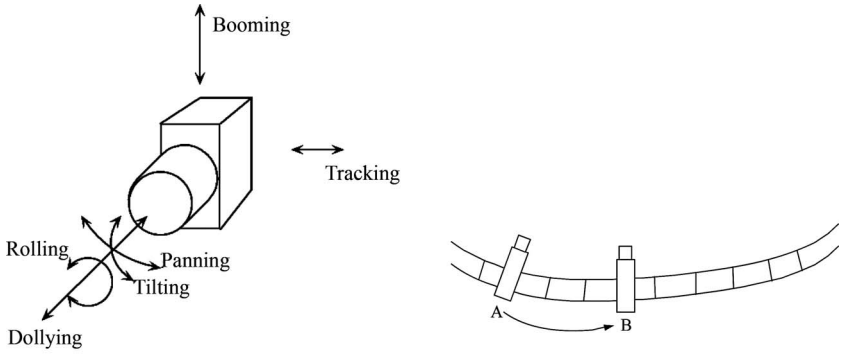


Fig. 6. Camera operations: (a) Basic terminologies for translational and rotational operations. (b) Typical camera operation on rail that results in lateral translation and rotation.

that in terms of translation, the shots exhibit either primarily forward/backward translation or primarily lateral translation. Thus, consistent with the assumption made in the previous paper,[7] we can hypothesize that the viewer is at least aware what generic type of motion is being executed by the camera. That is, the motion estimates are such that

- for lateral motion, $\hat{W}=W=0$;
- for forward motion, $\hat{U}=U=\hat{V}=V=0$.

We ignore zooming motion and its possible confusion with forward translation. Even though zoom lenses are prevalent nowadays, the experience of zooming motion is not a natural phenomenon to our eyes. Excessive zooming in or out may irritate the viewer and hence, zooming is not commonly used, except in some cases where special effects are required.[48,49] For instance, in the film *Vertigo* (1958), Hitchcock makes Scotty's illness visible and intelligible through the simultaneous combination of a forward zoom and a dollying out (backward translation), described in Ref. 50 as a "combination of approach and retreat whose complex confusions of perspective briefly induce all the sensations of nausea in the spectator." This rare use of zooming is fortunate, as it is difficult to separate the flow field induced by a zooming-in from the flow field simultaneously created by a forward translation. It also justifies our decision to ignore such motion in our analysis.

Next, we also assume that the contribution of γ_e is very small. Camera operations in cinematography usually minimize rotation about the optical axis (rolling) so as to avoid causing excessive discomfort to viewers. Last, in our first presentation of the distortion characteristics, we make an assumption that will allow us to better grasp the major geometrical features of the depth distortion: Within a limited field of view, second-order rotational terms in the image coordinates are small relative to the linear and constant terms. This is the case when the visual system focuses its attention on the fovea region under normal viewing conditions. Even if this assumption is removed, given their typical magnitudes, these terms do not qualitatively affect the nature of the depth distortion. However in the cinema viewing configuration, we will reinstate those second-order terms caused by the principal point offset $[t_x(f_v/D_v), t_y(f_v/D_v)]$ as the latter is large and no longer negligible.

C. Lateral Motion

If we assume that the viewer is aware of the type of generic motion being made, then under lateral motion all $\mathbf{n}$ will be in the same direction given by $\mathbf{n}=-(\hat{U},\hat{V})^T/\sqrt{\hat{U}^2+\hat{V}^2}$ for the epipolar reconstruction scheme of recovering depth. For notational convenience, we can rotate the X and Y axes without loss of generality so that $\mathbf{n}$ becomes $(1,0)^T$, or $(\hat{U},\hat{V})$ lies in the direction (1,0) [though (U,V) need not lie in that direction]. From Eqs. (12) and (13), the optical flow caused by lateral motion can be written as

$$u_v = -f'_v\frac{U}{Z} + \frac{(x_v-o'_x)(y_v-o'_y)}{f'_v}\alpha - f'_v\left(1+\frac{(x_v-o'_x)^2}{f'^2_v}\right)\beta + \gamma(y_v-o'_y), \tag{30}$$

$$v_v = -f'_v\frac{V}{Z} - \frac{(x_v-o'_x)(y_v-o'_y)}{f'_v}\beta + f'_v\left(1+\frac{(y_v-o'_y)^2}{f'^2_v}\right)\alpha - \gamma(x_v-o'_x), \tag{31}$$

where

$$(o'_x,o'_y)=\left(-t_x\frac{f_v}{D_v},-t_y\frac{f_v}{D_v}\right).$$

Plugging in the value of $\mathbf{n}$, the optical flow given by Eqs. (30) and (31), the estimated motions, and the estimated intrinsic parameters $\hat{o}'_x$, $\hat{o}'_y$, $\hat{f}'_v$ into Eq. (28) we obtain the distortion factor D:

- for the case of uncalibrated SFM under normal viewing conditions (second-order terms due to rotation and principal point ignored),

$$D=\frac{\hat{f}'_v\hat{U}}{f'_vU+(\beta f'_v-\hat{\beta}\hat{f}'_v)Z}, \tag{32}$$

- for the case of cinema viewing configuration (second order terms due to principal point dominant),

$$D=\frac{\hat{f}'_v\hat{U}}{f'_vU+[(\beta f'_v-\hat{\beta}\hat{f}'_v)+O^2(x_v,y_v)]Z}, \tag{33}$$

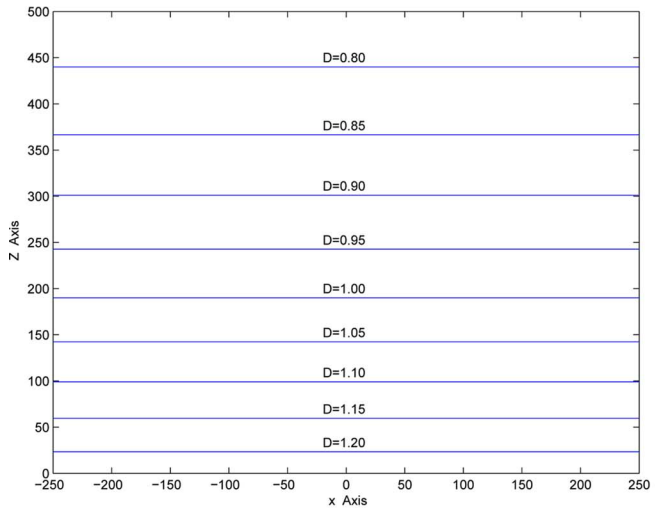
where

$$O^2(x_v,y_v)=\beta\frac{(x_v-o'_x)^2}{f'_v}-\alpha\frac{(x_v-o'_x)(y_v-o'_y)}{f'_v}-\hat{\beta}\frac{(x_v-\hat{o}'_x)^2}{\hat{f}'_v}+\hat{\alpha}\frac{(x_v-\hat{o}'_x)^2(y_v-\hat{o}'_y)}{\hat{f}'_v}. \tag{34}$$

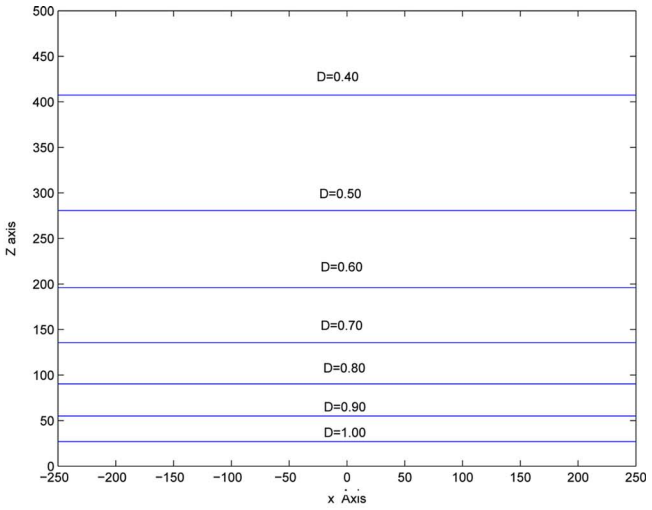
The distortion factor expressed in Eq. (32) for normal viewing conditions has the form $1/(a+bZ)$, where $a=f'_vU/\hat{f}'_v\hat{U}$ and $b=(\beta f'_v-\hat{\beta}\hat{f}'_v)/\hat{f}'_v\hat{U}$ are constants for all the scene points. It has the property that the distortion preserves the depth order of any two recovered depths $\hat{Z}_1$ and $\hat{Z}_2$ under certain conditions that are likely to hold (see Ref. 7 for details). For instance, if $Z_1>Z_2$, it can be readily shown that, given either of the following conditions, depending on the sign of a,

- $(a+bZ_1)(a+bZ_2)>0$ if $a>0$,
- $(a+bZ_1)(a+bZ_2)<0$ if $a<0$,

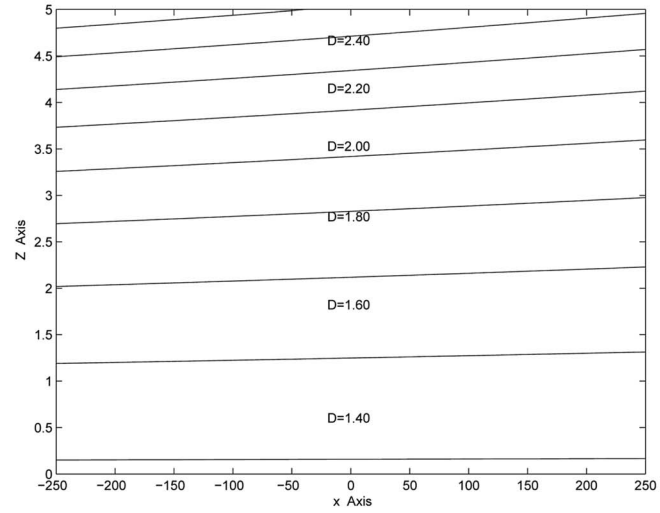
then the transformation $\hat{Z}=DZ$ preserves the depth order of the two points, that is, $\hat{Z}_1>\hat{Z}_2$. Since $a=f'_vU/\hat{f}'_v\hat{U}$, the condition $a>0$ means that f'_vU and $\hat{f}'_v\hat{U}$ have the same sign. This condition can easily be met by the human visual system; thus we can focus on just the first condition. The requirement $(a+bZ_1)(a+bZ_2)>0$ simply means that



(a)



(b)



(c)

Fig. 7. (Color online) Families of isodistortion contours for lateral motion obtained by intersecting the isodistortion surfaces with the xZ plane. $FoV=53°$, $f=f'_v=309.0$, $U=V=0.81$, $\beta=-0.002$, $\alpha=0.002$. (a) Viewer at CVP with errors only in the 3-D motion estimates: $\hat{U}=1.0$, $\hat{\beta}=-0.001$. (b) Viewer with optical axis parallel to and coincident with the projector's optical axis: $\hat{U}=1.0$, $\hat{\beta}=-0.001$, $\hat{f}'_v=303.0$. (c) Viewer with optical axis parallel to but not coincident with the projector's optical axis: $\hat{U}=1.0$, $\hat{V}=1.0$, $\hat{\alpha}=0.001$, $\hat{\beta}=-0.001$, $\hat{f}'_v=303.0$, $o'_x=o'_y=10{,}000$.

the two estimated depths should have the same sign. This condition can be easily assured: just check the sign of $\hat{Z}_1$ and $\hat{Z}_2$. If they are of the same sign, the depth order of $\hat{Z}_1$ and $\hat{Z}_2$ is correct; otherwise, just reverse the depth order. Furthermore, if the errors in the motion estimates are small enough, then this perceived ordinal depth space converges to a metric space.

Now consider Eq. (33). It is of a similar form $1/(a+bZ)$, but with b given by the nonconstant expression

$$b = \frac{(\beta f'_v - \hat{\beta}\hat{f}'_v) + O^2(x_v, y_v)}{\hat{f}'_v \hat{U}},$$

where $O^2(x_v, y_v)$ is given by Eq. (34). Clearly, ordinal depth is not preserved, since the value of b depends on (x_v, y_v). However, if the offset terms o'_x, o'_y, $\hat{o}'_x$, and $\hat{o}'_y$ in $O^2(x_v, y_v)$ dominate (x_v, y_v), then b remains largely the same over a local region, and to the extent that b is constant, the ordinality of depths recovered within this local region is likely to be preserved. See Fig. 7 for the values of D in the $x-Z$ plane under various viewing positions. We make the following observations:

- The sign of b decides whether the perceived space is compressed or expanded [compare Figs. 7(a) and 7(b) with 7(c)], with the depth order preserved irrespective of the sign of b.
- There is no qualitative difference between the distortion in Fig. 7(a) and that in Fig. 7(b) despite the addition of second order-rotational terms (which results in the bending of the contour) and the error in the focal length. This echoes the result of our previous paper[7] that calibration is not the determining factor in the quality of the perceived space.
- With the large principal point offset error found in the cinema viewing condition, the bending is made more pronounced by the second-order terms arising from this offset and is further aggravated by the shift in the origin. This results in difficulty in deciding depth orders across large visual angle, which seems to be consistent with our experience of sitting in an extreme off-center position in the cinema.

D. Forward Motion

For the case of forward motion and adopting the same epipolar reconstruction scheme, $\mathbf{n}$ can be expressed as $(x,y)^T/\sqrt{x^2+y^2}$. The distortion factor D can then be expressed as

- for the case of normal uncalibrated SFM (with all second-order terms ignored)

$$D = \frac{(x_v - \hat{o}'_x)^2 + (y_v - \hat{o}'_y)^2}{(x_v - o'_x)(x_v - \hat{o}'_x) + (y_v - \hat{o}'_y)(y_v - \hat{o}'_y) + O'Z}, \quad (35)$$

where

$$O' = -(\beta f'_v - \hat{\beta}\hat{f}'_v)(x_v - \hat{o}'_x) + (\alpha f'_v - \hat{\alpha}\hat{f}'_v)(y_v - \hat{o}'_y); \quad (36)$$

- for the case of cinema viewing configuration (second-order terms due to principal point dominant)

$$D = \frac{(x_v - \hat{o}'_x)^2 + (y_v - \hat{o}'_y)^2}{(x_v - o'_x)(x_v - \hat{o}'_x) + (y_v - o'_y)(y_v - \hat{o}'_y) + O''Z}, \quad (37)$$

where

$$O'' = -\hat{\beta}_f(x_v - \hat{o}'_x) + \hat{\alpha}_f(y_v - \hat{o}'_y),$$

$$\hat{\beta}_f = \beta f'_v - \hat{\beta}\hat{f}'_v + O_x^2(x_v, y_v),$$

$$\hat{\alpha}_f = \alpha f'_v - \hat{\alpha}\hat{f}'_v + O_y^2(x_v, y_v);$$

$$O_x^2(x_v,y_v) = -\beta\frac{(x_v - o'_x)^2}{f'_v} + \alpha\frac{(x_v - o'_x)(y_v - o'_y)}{f'_v} + \hat{\beta}\frac{(x_v - \hat{o}'_x)^2}{\hat{f}'_v} - \hat{\alpha}\frac{(x_v - \hat{o}'_x)(y_v - \hat{o}'_y)}{\hat{f}'_v},$$

$$O_y^2(x_v,y_v) = -\alpha\frac{(y_v - o'_y)^2}{f'_v} + \beta\frac{(x_v - o'_x)(y_v - o'_y)}{f'_v} + \hat{\alpha}\frac{(y_v - \hat{o}'_y)^2}{\hat{f}'_v} - \hat{\beta}\frac{(x_v - \hat{o}'_x)(y_v - \hat{o}'_y)}{\hat{f}'_v}.$$

From both Eqs. (35) and (37), we can see that D cannot be expressed in the form of $1/(a+bZ)$ with constant a and b. Indeed, for a particular value of D, the corresponding isodistortion surface is a cone. It has also been shown[51] that all D surfaces in the 3-D space intersect on a common line. As can be seen the distortion factor varies rapidly in a small neighborhood [see Figs. 8(a) and 8(b)] around the forward direction, and thus depth reconstruction is much more difficult than that in the case of lateral motion. While the presence of the second-order terms may change the shape of the isodistortion contours toward the periphery, the key properties discussed above regarding depth distortion are still true. In particular, ordinal depths are no longer recoverable. On the contrary, it has been shown[52] that forward motion leads to conditions favorable for motion recovery.

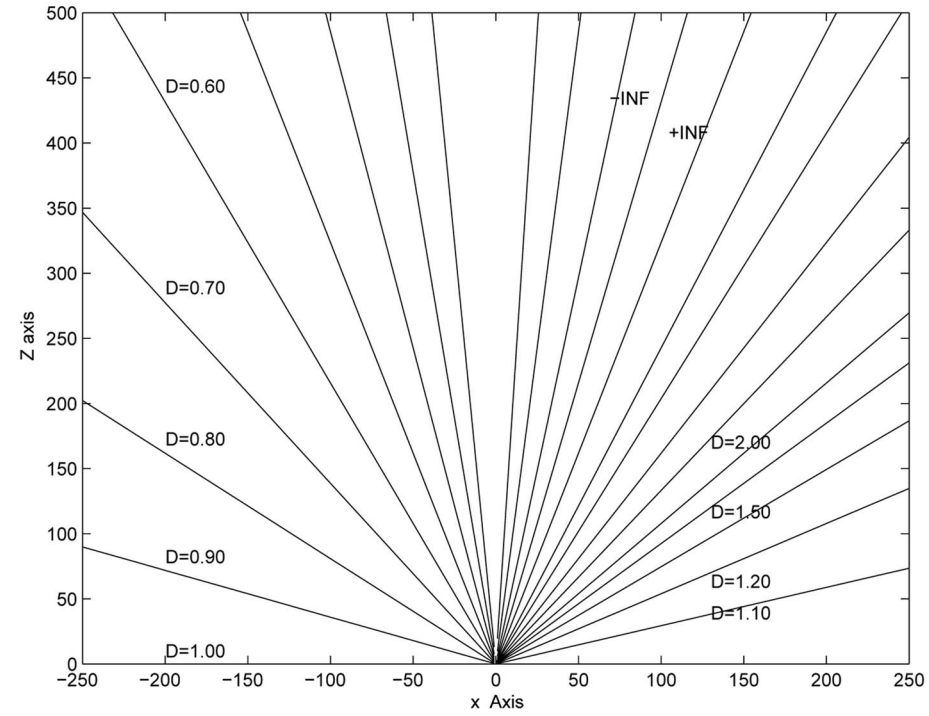
In sum the discussion so far in this section has shown that while the multiple projection processes in a cinema viewing configuration (with optical axis of viewer and projector being parallel) may vary the isodistortion equations, they do not alter the essential properties of depth distortion arising from both lateral and forward motion. In other words, since in the first place it is difficult even under normal viewing condition to obtain exact motion estimates from motion cues, the key properties of the perceived depths are already laid down, with changes in the intrinsic parameters (brought about by the cinema viewing conditions) contributing only to quantitative but not qualitative change.

5. DISCUSSION

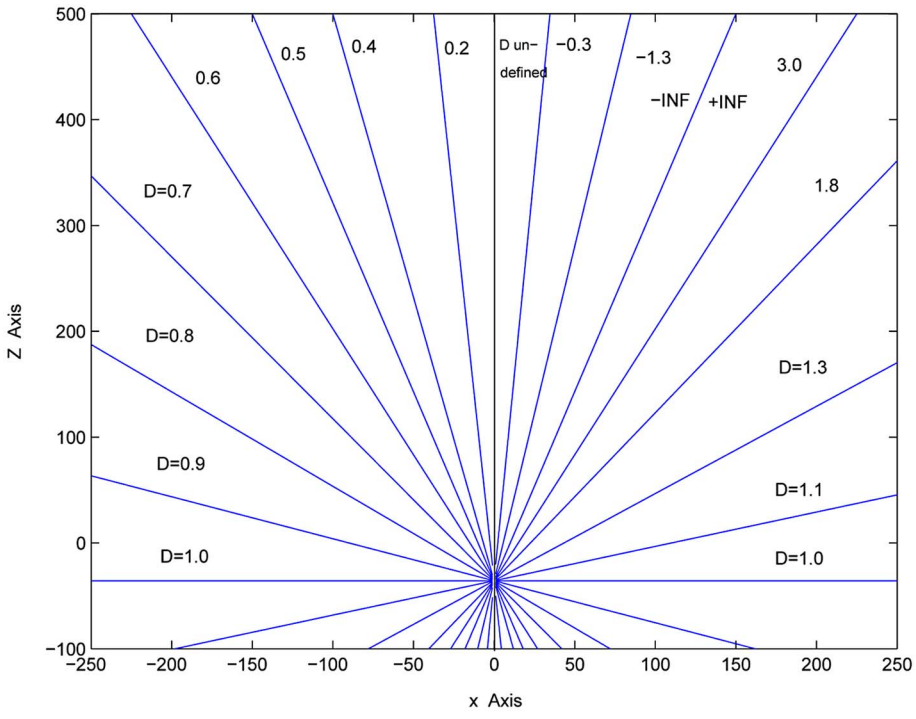
Various psychophysical experiments showed that we cannot recover the Euclidean space from two views even in our everyday activities. This is manifest in various psychophysical phenomena such as apparent frontal parallel plane, apparent distance bisection, and foreshortening of visual space at increasing distance under stereo vision[53] (note that human stereopsis is mathematically equivalent to a lateral monocular translation along the interocular distance, followed by an eye rotation equal to the convergence angle). AFPP has also been reported for the case of motion.[29] This inability to recover the veridical space is also mirrored by the computational difficulties encountered in depth reconstruction algorithms from motion and stereo cues. In particular, the resultant distortion in the recovered depth is modeled by the distribution of the isodistortion surfaces presented in this paper. For instance, Figs. 7(a) and 7(b) explain the compression of stereoscopic space noted by various researchers.[54–56] The surprising thing is that we function remarkably well in everyday life, and this seeming paradox parallels that happening in the cinema.

The results of this paper showed the link between everyday SFM and that occurring in the cinema. In particular, viewers seated at a reasonably central position experience a shift in the intrinsic parameters of their visual systems. What are the implications of these results? Is there a need to calibrate these changes in the intrinsic parameters? It is an open question whether the human visual system does this. As far as SFM is concerned, there is no need to calibrate if in the first place, we are not even able to estimate extrinsic motion parameters accurately under everyday SFM conditions. Such errors in the motion parameters render Euclidean space recovery impossible and in fact already determine all the important properties of space distortion. Changes or errors in the intrinsic parameters introduce further changes in the perceived shape but the qualitative nature of distortion remains the same.

Clearly, without a comprehensive psychophysical investigation, we cannot comment conclusively on the nature of space representation used by the human visual system. The psychophysics literature has proposed a number of models explaining different aspects of depth perception using well-defined geometries, such as similarity, conformal, affine, or projective transformations mapping physical space into perceived space. The epistemological considerations (what can and what cannot be recovered) raised by this computational inquiry do constrain the likely forms of space recovered from motion cues over two



(a)



(b)

Fig. 8. (Color online) Families of isodistortion contours for forward motion. (a) Viewer seated at CVP, $f_v=309.0$, $\beta_e=0.001$, $\alpha_e=0.001$. (b) Viewer seated on the optical axis of the projector with $D_v<D_c$, $f'_v=309.0$, $\hat{f}'_v=303.0$, $\beta=-0.002$, $\hat{\beta}=-0.001$, $\alpha=0.002$, $\hat{\alpha}=0.001$. INF stands for infinity.

views. It seems that that recovery of metrical depth information from motion is in general very difficult, and the transformation between physical and perceptual space is more complicated than previously thought. For the case of

motion or stereo it is rational and belongs to the family of Cremona transformations in general, although under specific motion-scene configurations, it reduces to simpler forms such as the projective transformation.

Doumen *et al.*[57] indicate that even under the ecologically valid conditions where there are redundant sources of depth information, the concept of a single visual space is problematic, let alone a Euclidean one. Visual space remains recalcitrant to the determination that a single geometry seeks to impose upon it: Results showed that the visual space is deformed differently over various tasks. Thus, tasks—considered as the web of interactions between the organisms and the physical world—need to be studied as the partitions of the physical world that are cognitively relevant for such organisms. Although our work does not explicitly address how tasks are interwoven with perception, the distortion framework put forth does indirectly support the notion of task dependency of perception. Specifically, under general motion-scene configuration, the perceived space is distorted by the complex Cremona transformation. However, under specific motion carried out by a task such as visual tracking with lateral motion, the distortion might reduce to a simpler form. In the case of the aforementioned lateral motion, the form of the distortion is a special projective transformation with nice properties.[7] Thus, the perceptually relevant geometry under the action of visual tracking is the attendant (distorted) visual space arising from the specific motor action of tracking. In general, different tasks with cooccurring movements would give rise to distinct space distortions described by the various subclasses of the Cremona transformation.

Let us explore the cinematographic implications even we do suffer from depth distortion arising from motion cues (not considering the role played by other cues). The director's camera motion is not arbitrary; in order to make cinematic communication possible, there is an intimate relationship between camera movements and the types of motions that human observers use. Some camera movements are routinely accomplished by the cameraman or the producer, including the introduction to a scene, the unfolding of an event, the tracking of an object, and the directing of attention. From our discussion in Section 4 about the nature of generic motions, we know that the establishment shots favored by directors to introduce scenes will yield reliable ordinal depth information because of the lateral motions employed in these shots. The same holds true for tracking shots. The preceding is true irrespective of whether there is calibration of the intrinsic parameters or not, and as long as the seat position is not too far off to the side. Such qualitative appreciation of the scene depth might be sufficient to render cinematic communication between the director and the audience possible.

On the other hand, shots with primarily forward motion present conditions favorable for motion recovery but not for depth recovery, regardless of whether the intrinsic parameters are calibrated or not.[52] Such shots are mainly used in closing in toward a subject or to affect a first-person view as he or she navigates through some environment. In the latter scenarios, the ability to recover the direction of motion well is obviously important for the appreciation of the meaning of the shot. Aspects of structural information might also be important in order for the viewer to "inhabit" the space of the protagonist, although its recovery from motion cues might not be feasible. Which particular structural aspect needs to be recovered is task-dependent; for instance, the ability to estimate the time-to-collision is important for shots depicting chases, say, through tight corridors. Fortunately, such information can be recovered directly from the optical flow without going through the step of full structural recovery.[58–60]

Finally, it must be added that even though the distortion may seem severe for two-frame SFM, the viewing conditions experienced by human beings are typically not so impoverished in depth cues, be it in everyday life or in the cinema. For instance, merely extend the SFM problem to multiple views, and the recovered structure has to obey the constraint of rigidity. Other cues such as the static perspective cue play an important role too. The studies by Stevens and Brookes,[61] Sparrow and Stine,[62] and Cornilleau-Pèrés *et al.*[63] have shown that static cues can dominate stereopsis or motion cues for the perception of plane orientation. Cutting[1] showed that the nonrigidity predicted by a motion cue for a viewer not seated at the CVP is not perceived, and one explanation is that the static cues overrule the motion cue. Indeed, static cues might also be used to recover H, the homography that relates the view of a person seated at the CVP to that of one seated at a general position. For instance, the orthogonality assumption among the detected vanishing directions enables partial self-calibration of the principal point from just a single view.

Hillis *et al.*[64] and Knill and Saunders[65] presented evidence of depth cue combination in an optimal manner; such optimality has been argued in visual perception or in motor actions.[66,67] However, it is far from certain that optimal depth cue combination can explain away the cinematic paradox. For instance, Doumen *et al.*[57] showed that even under an ecologically valid scenario where there are redundant sources of depth information such as shading, texture, or motion, systematic deviations from veridical settings may still arise. Irrespective of whether human observers are able to use the redundant information effectively or not, it is important to understand the intrinsic distortion arising from a particular depth cue. Equipped with a more in-depth understanding of such intrinsic distortion, we can then investigate how human observers can (or cannot) use the additional information effectively.

6. CONCLUSION

This paper offers an analytic account of several properties of the perceived visual space arising from motion cues when viewing the cinema from a location other than the CVP. In Section 3, we prove that as far as SFM is concerned, the perception of pictures viewed from this location with optical axis parallel to the projector's axis can be treated as one where the viewer experiences a change in the intrinsic parameters. Such changes remain within the framework of uncalibrated SFM proposed for machine vision, and thus the viewer can use an algorithm similar to the various self-calibration algorithms proposed in the

computational vision community, if such algorithm exists at all in the human brain. If the viewing axis and the projector axis are not parallel, then such viewing configuration not only changes the intrinsic parameters, but the amounts of the changes themselves are a function of the eye's eccentricity, a situation not dissimilar to the complex geometry of a foveated eye.

Even if the viewer does not or is not able to calibrate these intrinsic parameters, we show that the situation is not as serious as it seems. We investigate the properties of the SFM recovery and find that the ability of depth perception from motion is not made significantly worse under the cinema configuration. In other words, the estimation errors of the intrinsic parameters will not change the essential properties of depth recovered from motion. Lateral motion still leads to robust ordinal depth recovery, whereas for forward motion, the chief factor contributing to severe distortion in depth recovery is the difficulty in estimating the extrinsic parameters well enough in the first place.

Loong-Fah Cheong and Xu Xiang's e-mail addresses are, respectively, eleclf@nus.edu.sg and cvexx @nus.edu.sg.

REFERENCES

1. J. E. Cutting, "Rigidity of cinema seen from the front row, side aisle," J. Exp. Psychol. Hum. Percept. Perform. **13**, 323–334 (1987).
2. D. N. Perkins, "Compensating for distortion in viewing pictures obliquely," Percept. Psychophys. **14**, 13–18 (1973).
3. B. Caprile and V. Torre, "Using vanishing points for camera calibration," Int. J. Comput. Vis. **4**, 127–140 (1990).
4. M. S. Banks, H. F. Rose, D. Vishwanath, and A. R. Girshick, "Where should you sit to watch a movie?" Proc. SPIE **5666**, 316–325 (2005).
5. C. Metz, *Film Language: A Semiotics of the Cinema* (U. Chicago Press, 1991).
6. L. F. Cheong, C. Fermuller, and Y. Aloimonos, "Effects of errors in the viewing geometry on shape estimation," Comput. Vis. Image Underst. **71**, 356–372 (1998).
7. L. F. Cheong and T. Xiang, "Characterizing depth distortion under different generic motions," Int. J. Comput. Vis. **44**, 199–217 (2001).
8. T. Viéville, O. D. Faugeras, and Q. T. Luong, "Motion of points and lines in the uncalibrated case," Int. J. Comput. Vis. **17**, 7–41 (1996).
9. G. Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 477–489 (1989).
10. K. Daniilidis and M. E. Spetsakis, "Understanding noise sensitivity in structure from motion," in *Vision Navigation*, Y. Aloimonos, ed. (Hillsdale, 1993), pp. 61–88.
11. E. Grossmann and J. Santos-Victor, "Uncertainty analysis of 3-D reconstruction from uncalibrated views," Image Vis. Comput. **18**, 685–696 (2000).
12. J. Oliensis, "A new structure-from-motion ambiguity," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 685–700 (2000).
13. R. Szeliski and S. B. Kang, "Shape ambiguities in structure-from-motion," IEEE Trans. Pattern Anal. Mach. Intell. **19**, 506–512 (1997).
14. G. S. Young and R. Chellappa, "Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field," IEEE Trans. Pattern Anal. Mach. Intell. **14**, 995–1013 (1992).
15. H. P. Hudson, *Cremona Transformations in Plane and Space* (Cambridge, 1927).
16. V. Cornilleau-Pérès and J. Droulez, "The visual perception of 3D shape from self-motion and object-motion," Vision Res. **34**, 2331–2336 (1994).
17. J. T. Todd and V. J. Perotti, "The visual perception of surface orientation from optical flow," Percept. Psychophys. **61**, 1577–1589 (1999).
18. F. Domini, C. Caudek, and S. Richmann, "Distortions of depth-order relations and parallelism in structure from motion," Percept. Psychophys. **60**, 1164–1174 (1998).
19. O. D. Faugeras, *Three-Dimensional Computer Vision* (MIT, 1993).
20. J. J. Gibson, *The Perception of the Visual World* (Houghton Mifflin, 1950).
21. R. K. Luneburg, *Mathematical Analysis of Binocular Vision* (Princeton U. Press, 1947).
22. J. T. Todd and P. Bressan, "The perception of 3-dimensional affine structure from minimal apparent motion sequences," Percept. Psychophys. **48**, 419–430 (1990).
23. W. C. Hoffman, "The lie algebra of visual perception," J. Math. Psychol. **3**, 65–98 (1966).
24. J. J. Koenderink and A. J. van Doorn, "Relief: pictorial and otherwise," Image Vis. Comput. **13**, 321–334 (1995).
25. J. Droulez and V. Cornilleau-Pérès, "Visual perception of surface curvature, the spin variation and its physiological implications," Biol. Cybern. **62**, 211–224 (1990).
26. T. S. Meese and M. G. Harris, "Computation of surface slant from optic flow: orthogonal components of speed gradient can be combined," Vision Res. **37**, 2369–2379 (1997).
27. T. Viéville, J. Droulez, C. H. Peh, and A. Negri, "How do we perceive the eye's intrinsic parameters?" Tech. Rep. RR 4030 (INRIA, 2000).
28. M. A. Goodale and D. A. Westwood, "An evolving view of duplex vision: separate but interacting cortical pathways for perception and action," Curr. Opin. Neurobiol. **14**, 203–211 (2004).
29. L. F. Cheong, T. Xiang, V. Cornilleau-Pérès, and L. C. Tai, "Not all motions are equivalent in terms of depth recovery," in *Computer Vision and Robotics*, J. X. Liu, ed. (Nova, 2005), pp. 99–134.
30. J. S. Tittle, J. T. Todd, V. J. Perotti, and J. F. Norman, "Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis," J. Exp. Psychol. Hum. Percept. Perform. **21**, 663–678 (1995).
31. V. Cornilleau-Pérès and J. Droulez, "Visual perception of surface curvature: Psychophysics of curvature detection induced by motion parallax," Percept. Psychophys. **46**, 351–364 (1989).
32. J. F. Norman and J. S. Lappin, "The detection of surface curvatures defined by optical motion," Percept. Psychophys. **51**, 386–396 (1992).
33. B. J. Rogers and M. E. Graham, "Anisotropies in the perception of three-dimensional surfaces," Science **221**, 1409–1411 (1983).
34. W. J. M. Damme and W. A. van de Grind, "Active vision and the identification of 3D shape," Vision Res. **11**, 1581–1587 (1993).
35. W. J. M. Damme, F. H. Oosterhoff, and W. A. van de Grind, "Discrimination of 3-D shape and 3-D curvature from motion in active vision," Percept. Psychophys. **55**, 340–349 (1994).
36. H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," Nature (London) **293**, 133–135 (1981).
37. O. D. Faugeras, "Stratification of 3D vision: projective, affine, and metric representations," J. Opt. Soc. Am. A **12**, 465–484 (1995).
38. A. Heyden and K. Åström, "Euclidean reconstruction from image sequences with varying and unknown focal length and principal point," In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 1997), pp. 438–443.

39. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. (Cambridge U., 2004).
40. R. Petrozzo and S. W. Singer, "Cinema projection distortion," in *Proceedings of the 141st Society of Motion Picture and Television Engineers Technical Conference and Exhibition* (SMPTE, 1999).
41. P. F. Sturm and S. Ramalingam, "A generic concept for camera calibration," in *Proceedings of European Conference on Computer Vision*, T. Pajdla and J. Matas, eds. (Springer, 2004), pp. 1–13.
42. P. F. Sturm, "Multi-view geometry for general camera models," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2005) pp. I: 206–212.
43. W. N. Charman, *Visual Optics and Instrumentation* (Macmillan, 1991).
44. S. J. Galvin, D. R. Williams, and N. J. Coletta, "The spatial grain of motion perception in human peripheral vision," Vision Res. **36**, 2283–2296 (1996).
45. P. Artal, A. Derrington, and E. Colombo, "Refraction, aliasing, and the absence of motion reversals in peripheral vision," Vision Res. **35**, 939–947 (1995).
46. A. Guirao and P. Artal, "Off-axis monochromatic aberrations estimated from double pass measurements in the human eye," Vision Res. **39**, 207–217 (1999).
47. S. Ullman, *The Interpretation of Visual Motion* (MIT, 1979).
48. M. Martha, *Producing Videos: A Complete Guide* (Australian Film, Television, and Radio School, 1997).
49. B. Leeuw, *Digital Cinematography: Lighting and Photographing Computer-Generated Animation* (Morgan Kaufmann, 1997).
50. D. Raymond, *The Strange Case of Alfred Hitchcock* (MIT, 1978).
51. L. F. Cheong and C. H. Peh, "Depth distortion under calibration uncertainty," Comput. Vis. Image Underst. **93**, 221–244 (2004).
52. L. F. Cheong and X. Xiang, "Error characteristics of sfm with unknown focal length," in *Proceedings of Asian Conference on Computer Vision*, P. J. Narayanan, S. K. Nayar and H. Y. Shum, eds. (Springer, 2006).
53. K. N. Ogle, *Researches in Binocular Vision* (Hafner, 1964).
54. W. C. Gogel, "The analysis of perceived space," in *Foundations of Perceptual Theory*, S. C. Masin, ed. (Elsevier, 1993), pp. 113–182.
55. J. M. Loomis and J. W. Philbeck, "Is the anisotropy of perceived 3-d shape invariant across scale?" Percept. Psychophys. **61**, 397–402 (1999).
56. M. Wagner, "The metric of visual space," Percept. Psychophys. **38**, 483–495 (1985).
57. M. J. A. Doumen, A. M. L. Kappers, and J. J. Koenderink, "Visual space under free viewing conditions," Percept. Psychophys. **67**, 1177–1189 (2005).
58. D. N. Lee, "The optic flow field: The foundation of vision," Philos. Trans. R. Soc. London, Ser. B **290**, 169–178 (1980).
59. R. C. Nelson and J. Aloimonos, "Obstacle avoidance using flow field divergence," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 1102–1106 (1989).
60. M. Subbarao, *Interpretation of Visual Motion: A Computational Study* (Pitman, 1988).
61. K. A. Stevens and A. Brookes, "Integrating stereopsis with monocular interpretations of planar surfaces," Vision Res. **28**, 371–386 (1998).
62. J. E. Sparrow and W. M. Stine, "The perceived rigidity of rotating eight-vertex geometric forms; extracting nonrigid structure from rigid motion," Vision Res. **38**, 541–556 (1998).
63. V. Cornilleau-Pèrés, M. Wexler, J. Droulez, E. Marin, C. Miège, and B. Bourdoncle, "Visual perception of planar orientation: dominance of static depth cues over motion cues," Vision Res. **42**, 1403–1412 (2002).
64. J. M. Hillis, S. J. Watt, M. S. Landy, and M. S. Banks, "Slant from texture and disparity cues: optimal cue combination," J. Vision **4**, 967–992 (2004).
65. D. C. Knill and J. Saunders, "Do humans optimally integrate stereo and texture information for judgments of surface slant?" Vision Res. **43**, 2539–2558 (2003).
66. D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference," Annu. Rev. Psychol. **55**, 271–304 (2004).
67. J. Trommershauser, S. Gepshtein, L. T. Maloney, M. S. Landy, and M. S. Banks, "Optimal compensation for changes in task-relevant movement variability," J. Neurosci. **25**, 7169–7178 (2005).