Hybrid Stereo Camera: An IBR Approach for Synthesis of Very High Resolution Stereoscopic Image Sequences

Harpreet S. Sawhney Yanlin Guo Keith Hanna Rakesh Kumar
Vision Technologies Lab., Sarnoff Corp.*

Sean Adkins Samuel Zhou
IMAX Corp.†

Abstract

This paper introduces a novel application of IBR technology for efficient rendering of high quality CG and live action stereoscopic sequences. Traditionally, IBR has been applied to render novel views using image and depth based representations of the plenoptic functions. In this work, we present a restricted form of IBR in which lower resolution images for the views to be generated at a very high resolution are assumed to be available. Specifically, the paper addresses the problem of synthesizing stereo IMAX(R)[1] 3D motion picture images at a standard resolution of 4–6K. At such high resolutions, producing CG content is extremely time consuming and capturing live action requires bulky cameras. We propose a *Hybrid Stereo Camera* concept in which one view is rendered at the target high resolution but the other is rendered at a much lower resolution. Methods for synthesizing the second view sequence at the target resolution using image analysis and IBR techniques are the focus of this work. The high quality results from the techniques presented in this paper have been visually evaluated in the IMAX 3D large screen projection environment. The paper also highlights generalizations and extensions of the hybrid stereo camera concept.

Keywords: Image-based Rendering, Stereo Sequence Synthesis, Image Analysis

1 Introduction

This paper presents a ground-breaking method for efficiently rendering very high-resolution stereoscopic CG and real image sequences. The method has been developed for IMAX 3D motion picture projection environment which provides an immersive $3D$ experience to viewers through the projection of large format films onto large screens. IMAX 3D projection system uses dual 15 perforation 70mm film format referred to as $15/70$ format. Content for the IMAX 3D format is generated using 3D CG animation (e.g. *Cyberworld*), or using precisely manufactured binocular stereo cameras to film live action (e.g. *Galapagos*, *Into the Deep*), or a combination of both live action and CG animation through digital compositing (e.g. *T. Rex–Back to Cretaceous*). A faithful digital representation of live action content for this format requires at least $6K$ horizontal pixels of spatial resolution and 42 linear bits per pixel for color (RGB). For CG animation content, the required resolution may be less demanding, and good visual quality can be achieved with $4K$ horizontal pixels. In either case, the very high resolution format required for IMAX 3D projection limits the variety and quantity of $3D$ content that can be generated within the time and cost constraints of the industry. To create a CG animated scene of reasonable complexity, typical rendering time for each eye can be as high as 6 hours per one $4K$ resolution stereo frame on a latest 1 GHz Pentium CPU. A 45-minute IMAX 3D film requires a 100-CPU rendering farm full-time for about a year, just for rendering! For live action content, the IMAX stereo camera system, although capable of excellent quality capture, is bulky (about the size of a small refrigerator). This makes filming in some locations difficult, and limits the types of films that can be made in $3D$.

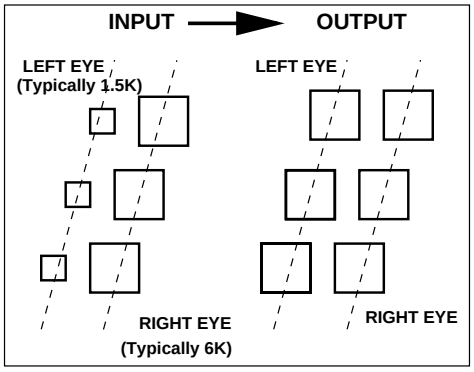


Figure 1: A schematic depicting the hybrid resolution stereo input and the full resolution output.

In order to meet the challenge of creating IMAX 3D content efficiently, we have developed image analysis and IBR methods to reduce rendering complexity while maintaining high quality and resolution. Towards this goal, we have developed the concept of a *Hybrid Stereo Camera*. Fig. 1 shows a schematic of the input and output relationship for the stereo synthesis problem underlying the hybrid stereo concept. The concept can be applied to the creation of both CG and live action content. In the context of CG rendering, the key idea is that instead of rendering both the left-right image pairs at the full resolution (typically $4K$), traditional CG rendering is used to render only one eye at the full resolution and the other eye at a much lower resolution, typically $1/4th$ of the full resolution in each dimension. Subsequently, the lower resolution eye is enhanced to match the quality and resolution of the full-resolution eye using IBR and image analysis techniques described in this paper. Since the computational complexity of image-based rendering is independent of scene complexity and depends only on the number of rendered pixels, the overall rendering time for stereo sequence rendering can be considerably reduced. Our experiments

*{hsawhney,yguo,khanna,rkumar@sarnoff.com}, CN 5300, Princeton, NJ 08543

†{szhou,sadkins@imax.com}, 2525 Speakman Dr., Mississauga, Ontario, CANADA L5K 1B1

[1]IMAX(R) is a registered trademark of IMAX Corporation.

ACM SIGGRAPH 2001, 12-17 August 2001, Los Angeles, CA, USA

have shown that even with unoptimized research prototype code, the rendering time for enhancing the resolution of the lower resolution image typically can be reduced to 20-30 minutes. We expect this time to come down to around 5 minutes with productization of the code. As a result, savings in rendering cost can be as high as 45% per stereo frame, which may result in increased production of mainstream content for IMAX 3D format.

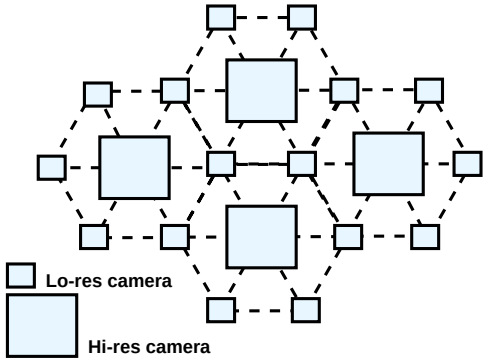


Figure 2: An example configuration of high- and low-resolution cameras that systematically cover a region of space. Any view from the viewpoint of any of the low-resolution cameras can be created at the higher resolution. Also novel views within the coverage provided by the camera configuration can be generated at high resolution.

A potential alternative method for efficient stereoscopic CG content rendering is to use the depth information available from rendered 3D image files to create a second eye through image warping. This process is called 2D-to-3D conversion [12, 5]. However, the dependence on depth information makes the 2D-to-3D conversion process vulnerable in handling certain CG effects, such as transparent or translucent objects, volumetric lighting, reflection and refraction created by ray tracing, and scenes modeled with particle systems, due to lack of clear depth information from the rendered files. On the other hand, image analysis and IBR techniques are ideally suitable for those tasks since there exists strong correspondence between left and right view image sequences.

The hybrid stereo camera concept also applies to live action content creation. In this case, a stereoscopic camera is designed such that one eye is captured at the full resolution while the second eye at a lower resolution. The lenses used for both eyes should maintain the same field of view. At least one of the views will be captured with a digital sensor (typically the lower resolution view), while the other view may remain in film format to provide high resolution required by the IMAX 3D format. Eliminating at least one of the two film cameras results in a smaller camera and a significant savings in film costs. However, in order to project a finished film we need two full resolution images for each frame. Again, the IBR method described in this paper is used to create the second high resolution image sequence.

The high resolution stereo synthesis method based on hybrid resolution inputs is a novel application of IBR since in traditional IBR, novel views are rendered "blindly" in the sense that no data is available from the viewpoint that is rendered. In our application, since a low resolution version of the high resolution image to be rendered is available, it is possible to detect discrepancies between the synthesized high resolution image and the corresponding lower resolution image. These discrepancies can be used to improve the quality of high resolution rendering.

We wish to emphasize that the hybrid resolution synthesis concept is not limited in its application to stereo only. For virtualized reality applications involving IBR, typically multiple cameras are used to tessellate a volume of space to capture live action [14]. Viewers can navigate virtually through the volume of space to see the live action from viewpoints that are different from the real cameras. If the goal is to be able to provide very high resolution views to the viewers, then using traditional technologies, one will have to use real high resolution cameras at each location. This will in general be prohibitively expensive. However, using the hybrid resolution synthesis concept, we can tessellate space using an optimal collection of high resolution and low resolution real cameras. The synthesis methods described in this paper can then be used to create the full high resolution views from the viewpoint of all the cameras. A schematic of one possible optimal configuration of high and low resolution cameras is shown in Fig. 2. For each of the low resolution cameras in the figure, a corresponding high resolution view can be generated. We will focus on the stereo sequence synthesis application in the current paper.

2 Overview & Related Work

A flow-chart of our approach is shown in Fig. 3. This section presents details at the level of blocks shown in red in the figure. Section 3 is devoted to the details of each step shown as boxes in black in the figure.

The key idea behind this work is that high resolution pixels from the sequence for one eye can be warped [23] into the coordinate system of the sequence for the other eye using an image based *analyze-test-synthesize* framework. Steps (1), (2) and (3) in Fig. 3 represent this framework. The analysis step involves generalized stereo/motion alignment between frames since frames are spatially (binocular/multi-ocular) and temporally separated. The alignment quality is then computed in the test step. Finally, the aligned images and the quality measures are used to combine multiple images to synthesize the high-resolution sequence.

The analysis step establishes correspondences between the left and right sequences at the lower resolution. Correspondences can be established at the lower resolution, since the two stereo cameras have identical fields of view but resolutions differing typically by factors of $1/4 \times 1/4$. To this end, the higher resolution frames are reduced to the lower resolution using anti-aliased filtering and down sampling with a Gaussian pyramid [3]. Correspondence map between a pair of frames may be thought of as a $2D$ flow field (vector field) where each vector represents the sub-pixel displacement that a pixel undergoes between the reference frame and the other frame within the pair.

The process of establishing correspondence has to deal with various imaging and scene conditions. In Fig. 1, ideally all the pixels in the synthesized left sequence should come from the corresponding high resolution right sequence. However, within a single stereo pair, pixels seen in one eye may not be visible in the other eye. Such occluded pixels cannot be obtained from the corresponding other eye. However, due to camera motion, these pixels may be unoccluded in frames at neighboring time instants. Therefore, a single hi-res left frame may be synthesized using not just the corresponding hi-res right frame but also the neighboring hi-res right frames.

Fig. 4(a) depicts a typical processing window. In order to synthesize a given left frame, a window of frames consisting of the lo-res left frame and a set of right hi-res frames (typically within a window of $+/-2$ frames) is used. Correspondences between a left-right stereo pair are established by exploiting the rigidity constraint that relates the two cameras [4]. The epipolar constraint between the two images can be exploited to constrain the solution of the stereo correspondence maps. Correspondences across time within the window are established using general motion, *optical flow* [6], since this allows both camera and independent object motions.

Both the stereo and motion correspondence maps are represented as $2D$ vector fields in the coordinate system of the reference lo-res

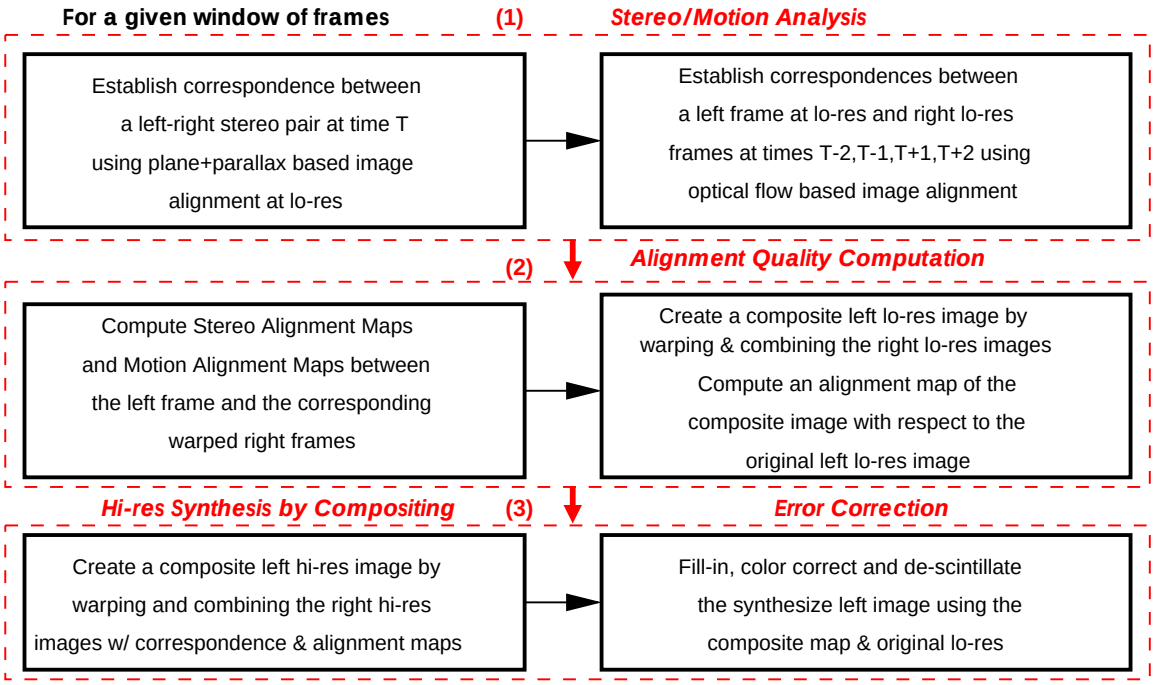


Figure 3: A flow chart of the lo-res to hi-res frame synthesis method.

left frame within the current window. Using the stereo and motion correspondence maps, each of the respective right images is warped to produce the corresponding warped images in the coordinate system of the left reference image. Note that all these images are still at the lower resolution.

The correspondence maps will in general be flawed due to the presence of unmatched regions, illumination mismatches between the left and right sequences, and other effects that may not be modeled by the stereo and motion algorithms. Therefore, a measure of the quality of correspondence is computed. The quality measure represents the similarity of color between the original reference image and a warped image at every pixel. The quality measure for a pair of images is referred to as an *alignment quality map*.

Using the warped right images, and their corresponding alignment quality maps, a composite lo-res image is synthesized by a weighted combination of the warped images with the quality measures as the weights. The synthesized image represents the best low resolution synthesis achieved using alignment. An alignment quality map for the composite image is also computed since such a map will be used to guide the synthesis at the higher resolution.

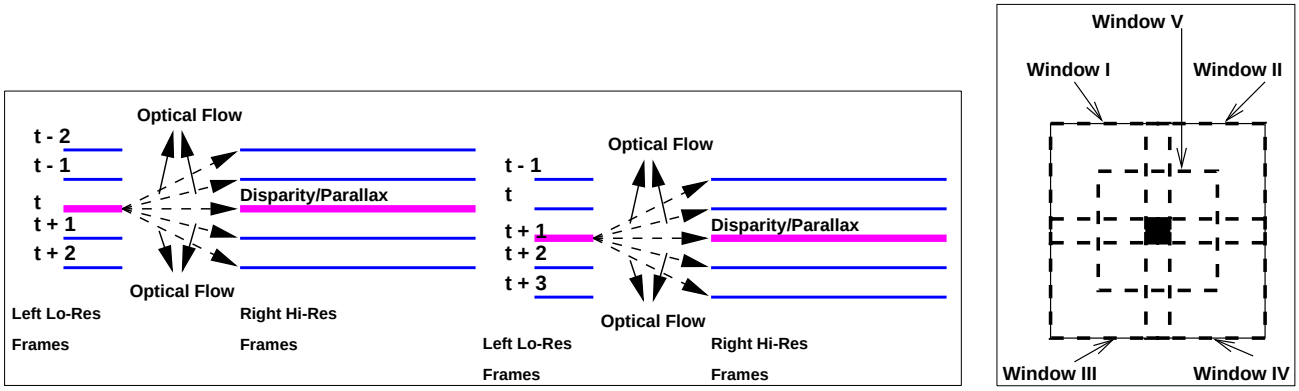
At this point, the computation moves to the target high resolution. The stereo and motion correspondence maps, their respective alignment quality maps, and the composite alignment map is projected to the higher resolution using a Gaussian pyramid [3]. A composite high resolution image is synthesized by a weighted combination of the warped images at the high resolution, where the warping is done using the projected correspondence maps. The weights at the high resolution are the projected low resolution alignment quality maps.

As mentioned before, the synthesized composite image at the high resolution may still contain artifacts from stereo mismatch. Therefore, the high resolution projected version of the low resolution composite alignment map is used to label pixels with mismatch artifacts. Pixels with low alignment measure values are labeled as mismatched. These pixels are copied from an up-sampled version of the low resolution image. In general, the regions of mismatch will be small and isolated. Therefore, filling-in these regions with the up-sampled image will not result in unpleasing stereo artifacts.

Finally, in spite of using a sliding temporal window for frame synthesis, there will still be some temporal scintillation artifacts between successive synthesized frames since the filled-in regions may be uncorrelated over time. We employ a process of de-scintillation to minimize such temporal artifacts. It is to be emphasized that even small regions of temporal scintillation can lead to irritating viewing and binocular stress. Since the temporal scintillation is primarily due to the inconsistency of filled-in low resolution regions over time, if the correlation between these regions can be increased, then the scintillation can be reduced.

A number of researchers in IBR and computer vision have used depth/parallax/stereo disparity maps and optical flow to create warped and rendered images. McMillan and Bishop [13] proposed a view based representation in which wide angle depth maps (cylindrical/spherical) and reference images are used to create new images by forward warping. Mark et al. [11] used CG rendered reference images and depth maps to efficiently render new frames using IBR for high quality rapid rendering. However, the idea of computing depth and motion fields at a lower resolution to render and composite images at a considerably higher resolution has not been exploited previously. Multiple resolution coarse-to-fine image alignment and analysis have been in the mainstream of computer vision for a number of years now [2, 21]. However, its use in high quality and high resolution frame synthesis has not been demonstrated. To the best of our knowledge, the stereo synthesis application presented in this paper is a unique application of a number of recently established image analysis and IBR techniques.

A unique aspect of our approach to IBR is the use of alignment quality measures in the process of compositing and synthesis. Image alignment based quality measures have been proposed in the past [7, 20] but according to our knowledge, they have not been used to combine multiple images to create enhanced images, especially when the multiple images may be related through non-parametric transformations such as stereo depth/disparity maps and optical flow.



(a) (b)

Figure 4: **(a):** A schematic showing the sliding temporal window used for processing hybrid resolution stereo sequences. Typically, for a left frame to be synthesized at time t, the corresponding hi-res right frame at t, and hi-res frames at $t-2$, $t-1$, $t+1$ and $t+2$ are used, as shown in the left part of the figure. The processing window slides forward one time instant for the next frame as shown on the right. **(b)**: Multiple off- and on-centered windows used in correspondence map estimation with optical flow or rigidity constraint based alignment.

3 Algorithms

There are four major components of our approach as shown in Fig. 3: stereo/motion analysis, alignment quality computation, high resolution compositing, and error correction for color mismatches and temporal scintillation.

3.1 Stereo/Motion analysis

The key analysis step is to establish dense correspondences with sub-pixel accuracy between pairs of image frames. Stereo analysis exploits the global rigidity constraint to generate correspondences between a pair of images taken at the same time instant while motion analysis exploits local smoothness of optical flow fields. It is to be emphasized that in our work on stereo analysis, knowledge of internal camera calibration parameters or external stereo camera calibration is not assumed. Specifically, our method is able to produce correspondences between stereo frames whose image planes may not lie on the same world plane, thus it handles the case of generalized stereo and not just rectified stereo only.

Iterative image alignment based image matching techniques have recently been proven to be robust for creating dense correspondence maps under a variety of 3D and motion scenarios. Image alignment based correspondence algorithms have three main features: pre-processing and representation of images, model of motion of pixels between images, and the particular matching method.

3.1.1 Representation

The images are represented by Laplacian and Gaussian multi-resolution pyramids [3]. A matching process that must accommodate a wide range of displacements can be very expensive to compute and is prone to false matches. Using a pyramid, large displacements can be computed using low spatial frequencies and high spatial frequencies can then be used to improve the accuracy of displacement estimation by incrementally estimating small displacements. Another advantage of using image pyramids is the reduction of false matches. This is because of aliasing of high spatial frequency components undergoing large motion. Aliasing is the source of false matches in correspondence solutions. Matching in a multi-resolution framework helps to eliminate problems of this type, since the large displacements are computed using images of lower spatial frequency.

3.1.2 Image motion models

The alignment algorithms used in this work have been derived from the published computer vision literature. A description of key aspects of the algorithms is included here for completeness. Please refer to the papers cited below for more details.

The alignment of image pairs needs to be performed for two different cases: stereo and motion. Both the stereo and motion correspondence maps are represented as $2D$ vector fields in the coordinate system of the reference lo-res left frame within the current window (see Fig. 4(a)). A vector field, $\mathbf{u}(\mathbf{p})$, the reference image, $I_l(\mathbf{p})$, and the second image, $I_r(\mathbf{p})$, satisfy:

$$I_l(\mathbf{p}) = I_r(\mathbf{p} - \mathbf{u}(\mathbf{p})) \tag{1}$$

where $\mathbf{p} = (x, y)$ represents pixel coordinates. Therefore, each of the right images can be warped to the coordinate system of the corresponding left image using the above equation to create:

$$I_r^w(\mathbf{p}) = I_r(\mathbf{p} - \mathbf{u}(\mathbf{p})) \tag{2}$$

where $I_r^w(\mathbf{p})$ is the warped right image that should look very similar to the left image.

Stereo case

In the stereo case, we have a pair of image frames taken at the same time instant but from different viewpoints. The motion of pixels between the two frames can be modeled by a set of global motion parameters (e.g. 3D rotation and translation) and the depth of each scene point observed in the image frames. In the method described here, we use an alternative parameterization, called *plane-plus-parallax*, where the motion of image pixels is modeled as due to the image motion of a $3D$ plane and the residual parallax field [9, 17, 18]. Plane-plus-parallax parameterization allows image matching under more general conditions, e.g. uncalibrated cameras, which is useful for real cameras. The total motion vector, $\mathbf{u}(\mathbf{p})$, with this parameterization can be written as the sum of a motion vector due to a planar surface, $\mathbf{u}_{pln}(\mathbf{p})$, and a residual parallax motion vector, $\mathbf{u}_{par}(\mathbf{p})$.

$$\mathbf{u}(\mathbf{p}) \quad = \quad \mathbf{u}_{pln}(\mathbf{p}) + \mathbf{u}_{par}(\mathbf{p}) \tag{3}$$

The motion of pixels belonging to a $3D$ plane can be approximated by a 2D quadratic transformation (8 parameters) for stereo

viewing conditions [1]. $\mathbf{u}_{pln} = (u_{pln}, v_{pln})$ is given by,

$$\begin{aligned} u_{pln}(\mathbf{p}) &= a_1 + a_2 x + a_3 y + a_7 x^2 + a_8 xy \\ v_{pln}(\mathbf{p}) &= a_4 + a_5 x + a_6 y + a_7 xy + a_8 y^2 \end{aligned} \quad (4)$$

In the above equation (u_{pln}, v_{pln}) represent the motion of pixels along the x and y axis of the image plane. The coefficients $a_1 \ldots a_8$ are the global motion parameters. Note that for discrete views with significant camera rotation, a full 8 parameter projective transformation is required to align a planar surface. However, for closely related views such as those obtained from a stereo pair the above quadratic transformation is a good approximation and is more stable to compute.

The parallax field is modeled by an epipolar direction (2 parameters with an arbitrary scale) and the parallax at every point in the image. The parallax motion vector, $\mathbf{u}_{par} = (u_{par}, v_{par})$ can be represented as:

$$\begin{aligned} u_{par}(\mathbf{p}) &= \gamma(fT_x - xT_z) \\ v_{par}(\mathbf{p}) &= \gamma(fT_y - yT_z) \end{aligned} \quad (5)$$

where (T_x, T_y, T_z) is the location of the origin of the reference camera in the coordinate system of the second camera; $\gamma = H/P_z T_\perp$, is the per-pixel parallax; H is the perpendicular distance of the 3D point from the plane; and P_z is its depth. $T_\perp$ is the perpendicular distance from the center of the *first* camera to the plane, and f is the focal length. At each point in the image, γ varies directly with the height of the corresponding 3D point from the reference $3D$ plane and inversely with the depth of the point [9, 17, 18].

Motion case

In the motion case, the two frames are taken at two different time instants. As a result, the pixel motion between frames is due to both the relative orientation between the two camera frames and also the motion of scene points moving independently. The motion of pixels is therefore unconstrained (non-parametric). In this case, we model the motion as a smoothly varying flow field (optical flow) [6].

3.1.3 Matching method

In order to align two images (an *inspection* image and a *reference* image), we use pyramid-based hierarchical image alignment techniques with different image motion models [2, 4, 9, 17]. This technique first constructs a Laplacian pyramid from each of the two input images, and then estimates the motion parameters in a coarse-to-fine manner. Within each level the sum of squared differences (SSD) measure integrated over regions of interest (which is *initially* the entire image region) is used as a match measure:

$$E(\{\mathbf{u}(\mathbf{p}\}) = \sum_{\mathbf{p}} \left(I_l(\mathbf{p}) - I_r(\mathbf{p} - \mathbf{u}(\mathbf{p}))\right)^2 \quad (6)$$

where I is the Laplacian filtered image intensity. The sum is computed over all the points within the region and $\{\mathbf{u}\}$ is used to denote the entire motion field within that region. The motion field $\{\mathbf{u}\}$ can be modeled by a set of global (e.g. plane parameters) and local parameters (e.g. flow/parallax) as described above.

Levenberg-Marquardt minimization [22] is applied to the objective function described in Eq. (6) in order to estimate the unknown motion parameters and the resulting motion field $\{\mathbf{u}\}$. Starting with some initial values (typically zero), the hierarchical estimation algorithm iteratively refines the parameters in order to minimize the SSD error from coarse to fine resolutions. After each step, the current set of parameters are used to warp the inspection image (Eq. 2), in order to reduce the residual displacement between the images.

The matching method is employed in an hierarchical manner. We first compute the motion of an average plane in the scene by using Eq. (4). The image motion expressions of Eq. (4) are substituted into Eq. (6) to obtain the complete objective function. This function is minimized using the direct hierarchical registration technique to estimate the quadratic image motion parameters $(a_1, \ldots, a_8)$. We refer to this algorithm as the *Quad* algorithm.

For stereo alignment using plane-plus-parallax, *Quad* alignment is used as a starting point to do alignment based on a joint planar and parallax estimation. The expression for $\mathbf{u}(\mathbf{p})$ from Eq. (3) is substituted into Eq. (6) to obtain the complete objective function. The resulting function is then minimized using the Levenberg Marquardt algorithm to solve for the planar motion parameters $(a_1, \ldots, a_8)$, direction of translation $\mathbf{T}$ and the parallax vector field γ.

In the case of optic flow estimation also, we use the estimation from the *Quad* registration algorithm as an initial estimate. We minimize the error function in Eq. (6) to estimate a $2D$ flow vector at each pixel using the hierarchical coarse-to-fine matching scheme [10, 2].

Both optic flow and parallax estimation are examples of local non-parametric estimation. We assume the parallax (or flow) at each pixel to be locally constant in a small window around that pixel. The parallax (or flow) for a pixel is estimated by using all the pixels in its window. This process is repeated for each pixel and results in a smoothly varying parallax (or flow) field.

We estimate the dense motion using five windows, on and off-centered around each pixel as illustrated in Fig. 4(b) . Local computation of parallax and flow is done for each window. The estimate that leads to the minimum local error is used as the estimate for the pixel under consideration. Away from occlusion boundaries, the multiple windows provide equally good estimates. However, at or close to occlusion boundaries the window with the best estimate will correspond only to the occluding surface. The non-optimal window estimates will come from the mixed estimate corresponding to the boundary between the occluded and occluding surfaces. Choosing the optimal estimate leads to crisp correspondence maps that are sub-pixel accurate at occluding boundaries.

3.2 Alignment Quality Measure & Compositing

The correspondence maps computed at the lower resolution need to be assessed for their quality. Note that both the left-right stereo pair and motion pairs are available at the lower resolution so the alignment quality can be computed between a warped and the original image. Since our goal in this work is image synthesis, we are not concerned with the accuracy of the maps with respect to ground truth maps. We are however concerned with the accuracy of prediction of the appearance of image pixels using the maps. We adopt a quality measure based on correlation of the original image with the warped image.

The correlation measure uses a variant of normalized correlation [16] to assign a value in $[-1, 1]$ to the pixel locations of a pair of aligned images. Two variations on the standard correlation measure are incorporated. First, the variance within each window is computed both for the original image values, as well as for mean normalized image values. This enables us to handle correlations in the three different color bands while reducing the effect of variable camera sensitivity to the three spectral bands. Second, when the normalized and original variances are below certain thresholds, the correlation is set to zero if the difference in means is less than a threshold, otherwise it is set to one. This handles regions where there is not enough variation in the intensities to reliably use a normalized correlation measure. The quality measure $Cval$ is computed as follows,

$$if \qquad (\sigma_1^2 \leq \sigma_{th}^2 \; AND \; \sigma_2^2 \leq \sigma_{th}^2) \qquad OR$$

$$
\begin{aligned}
& & (\sigma_{N1}^2 \leq \sigma_{Nth}^2 \; AND \sigma_{N2}^2 \leq \sigma_{Nth}^2) \\
then & & if \; \delta\mu^2 \leq k\sigma_{th}^2 \\
& then\, Cval = 1.0; & else\, Cval = 0.0; \\
else & & Cval = \frac{\sum_{\mathbf{p}} (I_1(\mathbf{p}) - \bar{I_1})(I_2(\mathbf{p}) - \bar{I_2})}{N\sigma_1\sigma_2}
\end{aligned}
$$

where σ_1^2, σ_2^2 are the respective image variances within the correlation window; $\sigma_N^2 = \sigma^2/(\mu^2 + c)$ is the mean normalized variance with μ being the mean and c a stabilizing constant to handle close to zero mean values; σ_{th}, σ_{Nth} and k are parameters, and N is the number of pixels in the correlation window.

In order to capture the total alignment quality, the geometric mean of the quality measure for the three color (RGB) spectral bands is computed. The quality measure for each frame pair and the associated correspondence maps are projected to the target higher resolution for subsequent frame synthesis. The projection to high resolution uses bi-cubic interpolation applied to the correspondence maps as well as the alignment measures.

Given the high resolution frames, correspondence maps and the quality measures, the target high resolution frame is created by a weighted combination given by:

$$
I_{h0}(\mathbf{p}) = \frac{\sum_i w_{ti} w_c(\mathbf{p_i}) I_i(\mathbf{p_i})}{\sum_i w_{ti} w_c(\mathbf{p_i})} \tag{7}
$$

where I_{h0} is the synthesized high resolution image, w_{ti} is a weight for each of the source frames i, and $w_c(\mathbf{p_i})$ is the weight given by the quality measure. Weights w_{ti} can be chosen to be inversely proportional to the temporal distance between the ith source frame and the target frame. $w_c(\mathbf{p_i})$ is equal to the quality measure if the measure is above a threshold otherwise zero. The result is that the images are composited using pixels from well aligned images.

3.3 Error correction

In the previous sub-section we described how to synthesize the high-resolution frame by warping and combining other high-resolution frames. An alignment quality measure was used to control the combination process. However, the synthesized high resolution frame may still suffer from artifacts because of remaining areas of occlusion where no high-resolution frame is able to provide the missing information, undetected mis-alignments and other effects. This may be remedied by filling in from the available low resolution frame.

We create an alignment quality map between the synthesized high-resolution frame and the original low resolution frame. This alignment quality map is denoted as the *synthesized quality map*. Using the *synthesized quality map*, misaligned pixels are copied from the original low-resolution frame where the alignment quality is low. This process is called *filling-in*. The low resolution frame can be up-sampled to the higher resolution before *filling in*.

Two problems may occur because of the *filling-in*. First there may be color mis-matches between the filled-in pixels and the neighboring high resolution pixels. We have developed a color correction procedure to solve this problem. Since the frame synthesis process ensures that misaligned pixels will be present in only a few regions, we create regions that are large enough to contain a few misaligned pixels and many aligned pixels. Using the *aligned pixels only*, we use the original image and the synthesized image to solve for a $3D$ affine color model that maps the original pixels to the synthesized pixels. This color model is subsequently applied to the original pixels in the misaligned regions. Essentially aligned pixels provide a color correction model that is applied to the mismatched pixels in each large enough region that contains both.

The second problem that occurs is temporal scintillation between frames. The temporal scintillation occurs since *filling-in* and other artifacts of synthesis are un-correlated over time. To solve this problem, we need to correlate the *filling-in* process across time. The *filling-in* at each pixel is done based on the value of the *synthesis quality map* at that pixel. We therefore need to temporally smooth the *synthesis quality map* at each time instant. We do this by tracking regions of *filling-in*. The tracking is done by computing optic flow between the *synthesized quality maps* over time. Using the optic flow, we warp neighboring maps to the coordinate system of the current frame and then average the *warped synthesis quality maps* with the original *synthesis quality map* for that frame. The smooth maps are then used to create filled in pixels and do color correction.

4 Results

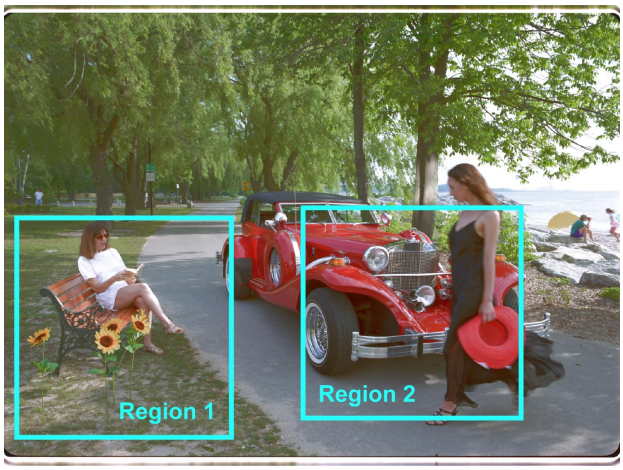


Figure 5: One full resolution image from the *Redcar* stereo sequence. The two regions for which detailed results are presented in the paper have been marked as *Region 1* and Region 2. ( )

The hybrid stereo camera concept is demonstrated with one live action sequence, *RedCar*, and one CG stereo sequence, *JoeFly*. Note that the arrangement of low and high-res frames for *RedCar* is as shown in Fig. 1 but it is reversed for *JoeFly*. The *RedCar* sequence was captured using the current $IMAX3D$ camera that uses two identical large format film cameras. Images from film were scanned at $6K$ resolution. A hybrid camera was simulated by down sampling the left sequence by $1/4 \times 1/4$ in x and y. One full frame of the sequence is shown in Fig. 5. The sequence was captured using a panning camera, with the flower stems swaying in the wind, the lady on the bench moving her head, and the lady in black moving from right to left. The range of stereo disparities between pairs at the lower resolution is as wide as 1-30 pixels (4-120 pixels at the full resolution), and the range of motion displacement between farthest pairs within the processing window is as wide as 11-50 pixels at the lower resolution (44-200 pixels at the full resolution). The top half ((A1), (A2), (A3)) of Fig. 6 shows a synthesis result as a $2K$ cutout (*Region 2* in Fig. 5) from the $6K$ synthesized frame. The middle left (A3) of the figure shows the left lo-res cutout (512×521) of the $2K$ frame used as input. The corresponding right original frame used as input and the synthesized left frame output by the algorithm are shown at the top ((A1), (A2)) of the figure. In spite of the complexity of highlights, detailed structures like the car grill and Mercedes logo, and occlusions/dis-occlusions between the lady, and the background, the synthesized frame is quite reasonable. Some thin structures like the logo are

(A1) Input: Right Original Full-res ($2K \times 2K$)

(A2) Output: Left Synthesized Full-res ($2K \times 2K$)

(A3) Input: Left Low-resolution (512×512)

(B1) Input: Right Low-resolution ($1K \times 1K$)

(B2) Input: Left Original Full-res ($4K \times 4K$)

(B3) Output: Right Synthesized Full-res ($4K \times 4K$)

Figure 6: **(A1), (A2), (A3)**: The $2K$ original right, the $2K$ synthesized left, and the $1/4$th original left frames from the live action *RedCar* sequence, respectively. **(A1)** and **(A3)** are the inputs and **(A2)** is the output. **(B1), (B2), (B3)**: The $1/4$th resolution input for the right eye, the left original $4K$ frame *JoeFly* with the *fly*, and the corresponding synthesized $4K$ right frame, respectively. **(B1)** and **(B2)** are the inputs and **(B3)** is the output. (**(A1),(A2),(A3)**: **(B1),(B2),(B3)**:)

blurred since these regions seemed to have been filled in using the up-sampled frame.

In order to qualitatively show the sharpness of disparity maps computed by the alignment algorithm, Fig. 7 shows a disparity map computed at the lower resolution for *Region1* (Fig. 5) of the *RedCar* sequence, the region containing the lady on the bench and the flowers. It is evident that the multi-window parallax based algorithm described above produces sharp boundaries at depth discontinuities. Therefore, warping with these disparity maps results in crisp boundaries in the rendered results.

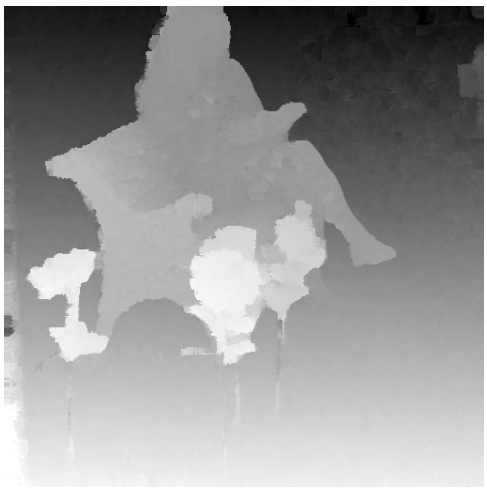


Figure 7: Stereo disparity map computed using the multiple window based parallax for a pair of frames for *Region 1* in the *Redcar* sequence.

In order to show how successive steps of processing improve the quality of synthesis, we synthesized full resolution frames at three intermediate stages along with the final synthesis using the algorithm described above. High resolution frames were synthesized (i) with one stereo pair using plane+parallax only (pp), (ii) with a window of stereo/motion frames (Fig. 4(a)) using plane+parallax and optical flow ($pp + f$), (iii) with filling-in for misaligned pixels ($pp + f + fillin$), and (iv) with additional color correction ($pp + f + fillin + CC$) for the final result. For each of the synthesized frames, we computed the alignment quality map with respect to the original and counted the number of misaligned pixels. These were pixels where the quality was below a threshold (0.8). Fig. 8 shows the progressive increase in quality of synthesis for one frame of the $2K \times 2K$ *Region1* of *RedCar* by plotting the number of misaligned (mismatched) pixels against the method used for synthesis. The number of misaligned pixels between the real frame and the synthesized hi-res frames drops from a high of 472763 after pp, to 365360 after $pp + f$, to 135252 after $pp + f + fillin$, to finally a low count of 55675 after $pp + f + fillin + CC$. The total number of pixels is $4M$.

Fig. 9 shows 960×960 cutouts of *Region1* synthesized at $2K$ for a visual comparison of intermediate results that were quantitatively shown in Fig. 8. The cutout contains the flower stems for which flow and parallax based alignment may be particularly problematic. The left frame in the figure shows that plane-plus-parallax between a stereo pair resulted in severe misalignment of the stems in the synthesized frame. However, by combining a stereo pair with optical flow based alignment of other neighboring frames, the quality of synthesis was improved as shown in the middle frame. Although three of the flower stems are correctly rendered now, the fourth one on the left is still partially missing. The rightmost frame shows the final synthesis after post-processing including filling-in and color correction, and is almost perfect. The intermediate visual result after $pp + f + fillin$ has been omitted to conserve space.

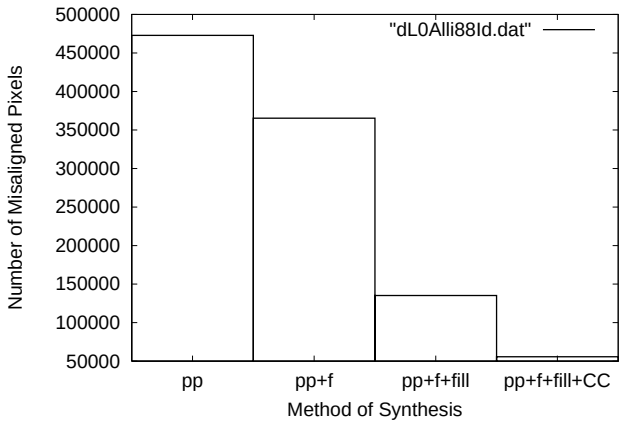


Figure 8: Bar chart showing number of mis-aligned pixels after hires frame synthesis for Region 1 (*Please see text for explanation*).

Fig. 10 shows a visual comparison, for a 960×960 cutout of *Region 2*, between the result using our algorithm (top), and an output (bottom) of simply up-sampling the low resolution images using bicubic interpolation (with some sharpening). Even though the stark difference in quality apparent on a high quality screen is lost in the print medium, closer inspection reveals that the up-sampled version lacks sharpness everywhere, especially in the regions occupied by the grill and the right headlight, and all the regions with highlights.

The deeper issue of determination of a visual threshold for acceptable loss of detail in stereo viewing of resolution mismatched inputs is related to the topic of asymmetric stereoscopic perception. Earlier studies in stereograms by Julesz [8] showed that binocular fusion can occur if either the low or high frequency spectrum are identical. Those frequency components that are not identical will cause binocular rivalry. For resolution mismatched binocular inputs, the high-resolution view will dominate the binocular percept [8, 15]. More recent studies in stereoscopic *video* ([19], for instance) have reported minimal loss of perceived quality with resolution mismatched inputs under an ITUR-601 stereo viewing environment. However, there are several factors that affect stereoscopic image quality, including the level of and methods used for resolution reduction, and image contents [19]. We emphasize that, in the IMAX 3D viewing environment, the visual threshold for asymmetrical resolution reduction is quite different from a stereo video viewing environment. For instance, the typical viewing distance in an IMAX 3D theater is less than one screen height whereas that in the cited study was four times the screen height. When we presented viewers with resolution mismatched stereo inputs, the viewers reported lack of sharpness and complained of eye stress probably due to binocular rivalry. More psychophysical work is needed to explore the issue of stereo perception in IMAX-like immersive environments with mismatched stereo inputs.

We now present synthesis results for a CG stereo sequence, *JoeFly*. *JoeFly*, created by Spans and Partner GMBH in Germany, is a CG animation movie rendered at $4K \times 4K$ resolution for the recent *CyberWorld 3D* film. Since, this sequence was used to initially test the hybrid stereo synthesis on CG content, both left and right sequences were rendered at the full resolution. For the test, the hybrid camera was simulated by down-sampling the right sequence to $1K$ resolution, and the hybrid synthesis method was applied. The bottom half ((B1), (B2), (B3)) of Fig. 6 shows one synthesized

Figure 9: **Left to right**: Cutout of a synthesized image from *Region 1* of *RedCar* showing improvement in quality of synthesis with parallax only (pp), parallax+flow only ($pp + f$), and parallax+flow and post-processing ($pp + f + fillin + CC$). (©IMAX Corporation, 1995. All rights reserved.)

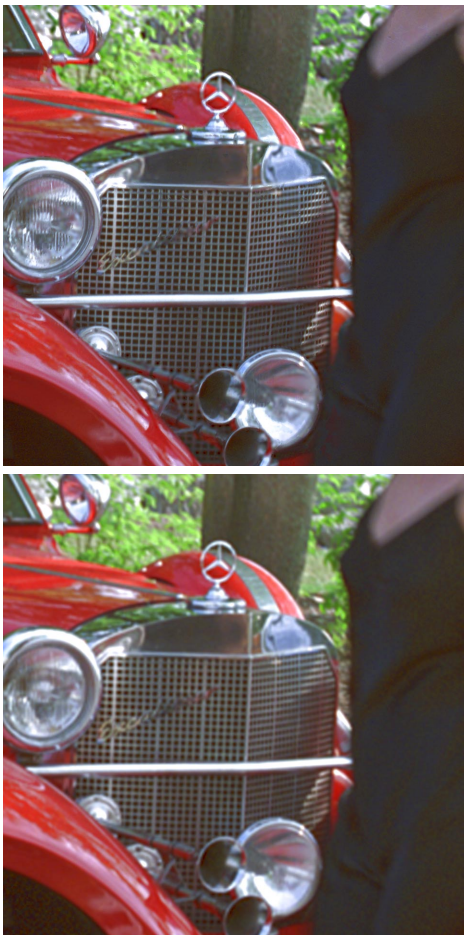


Figure 10: Cutout of *Region2*: (**top**) synthesized hi-res result, and (**bottom**) up-sampled result using bicubic interpolation for comparison. The up-sampled version lacks sharpness and detail everywhere, especially on the car grill, the horizontal bar and the headlight. (©IMAX Corporation, 1995. All rights reserved.)

frame from *JoeFly*. The original full-resolution left frame, the original low-resolution right frame and the synthesized right frame are shown. The frames show the details of a fly, flower petals and regions of transparency. There is no noticeable blemish in the synthesized frame.

As in the case of live action synthesis, Fig. 11 shows the bar graph of the intermediate and final quantitative results for one frame of *JoeFly*. Again, the bar graph shows the quality of alignment measured as the number of misaligned pixels between the original ground truth frame and the synthesized frame. The comparison was done with full resolution frames synthesized at two intermediate stages along with the final synthesis ($pp + f + fillin$). Intermediate frames were created after stereo (plane-plus-parallax, pp) alignment only, and after stereo and motion (optic flow, $pp + f$) alignment. The number of mismatched pixels drops from a high of about 180563 after pp (plane-plus-parallax) only, to about 94521 after $pp + f$ (plane-plus-parallax and flow), to a low of about 14909 after error correction for the final synthesis. This shows that for CG material too, successive steps of processing tremendously improve the quality of synthesis.

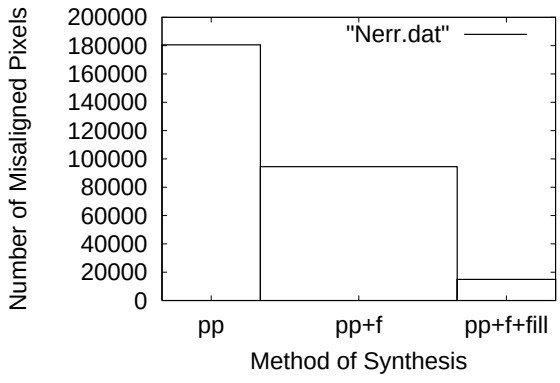


Figure 11: Bar chart showing number of mis-matched pixels after hi-res frame synthesis for one frame of *JoeFly*. (*Please see text for explanation*).

Temporal de-scintillation is performed to reduce flickering of

synthesized pixels when the frames are played as a movie. The importance of this cannot be overemphasized since even minor temporal artifacts are noticeable by the human eye especially when the hi-res frames are shown on high quality computer monitors or on the big IMAX screen. Unfortunately, it is not possible to show the results of de-scintillation on paper. We have done extensive visual evaluations of the CG and live action sequences. The original stereo sequences and the synthesized sequences, both for the *JoeFly* and *RedCar* data sets, were transferred to $15/70$ format $IMAX$ film at $4K$ and $6K$ resolutions, respectively. A blind test was done with a continuously looping film with an IMAX 3D projector. Each segment contained 10 seconds of the original left-right sequence, and the original left (*JoeFly*)/right (*RedCar*) and the synthesized right/left frame sequences, respectively. We are glad to report that the differences were not evident even to experienced viewers. We believe that the IBR based hybrid stereo camera technology presented in this paper will prove viable for the quality and efficiency requirements of IMAX films and other high quality entertainment applications.

(The results shown in the paper as well as a few more have been included as high quality images on the CD-ROM proceedings.)

5 Discussion & Extensions

The hybrid stereo camera concept can be generalized and extended both for CG and live action applications. In the case of CG stereo sequences, it is not necessary that all the frames for one view be high resolution and those for the other view be lower resolution. For example, typically the image displacements between the stereo pairs may be more than those between the motion pairs. In the alternative configuration, a frame can be synthesized by using high resolution frames from the same eye across time, thus taking advantage of smaller image displacements. Similarly, the hybrid concept can be used to advantage in traditional $3D$ warping based IBR as proposed by Mark et al. in [11]. In order for complex CG scenes to be rendered at near real-time rates, Mark et al. proposed that key frames may be rendered using traditional rendering. Subsequently, near real time IBR can be used to render a denser collection of frames. The concept of a hybrid camera suggests that the key frames need not be rendered at the highest resolution required, but a mixture of high and low resolution key frames can be used. High resolution hybrid IBR can be used to efficiently create a dense collection of output frames.

In Fig. 2 a spatial configuration of mixed resolution cameras for capturing and synthesizing high resolution sequence of live action was shown. With the increase in the power of video processing platforms, and a steady decline in the cost and size of digital cameras (e.g. CMOS cameras), capture and manipulation of live events for entertainment and surveillance applications will steadily be on the rise. Therefore, mixed resolution configurations and the hybrid camera processing may provide cost effective solutions for real time high quality rendering. Furthermore, the hybrid concept is not limited to generation of views from the position of real cameras only. It can be combined with depth and view based IBR to create interpolated and extrapolated new views.

Acknowledgments

We are grateful for the extremely valuable help in data handling and testing provided by Ed Lepieszo and Carol Harrison of IMAX Corp., and the software and system support provided by Vince Paragano and Doug Corliss of Sarnoff Corp. Furthermore, we would like to thank Spans and Partner Inc. and IMAX Corp. for the use of the images.

References

[1] G. Adiv. Determining 3D motion and structure from optical flows generated by several moving objects. *IEEE TPAMI*, 7(4):384–401, 1985.

[2] J. R. Bergen et al. Hierarchical model–based motion estimation. In *Euro. Conf. on Comp. Vision*, pages 237–252, 1992.

[3] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE TCOMM*, 31(4):532–540, 1983.

[4] K. J. Hanna and Neil E. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *Proc. Intl. Conf. on Computer Vision*, pages 357–365, 1993.

[5] Phil Harman. Home based 3d entertainment - an overview. In *IEEE Intl. Conf. on Image Processing*, pages 1–4, 2000.

[6] B.K.P. Horn. *Robot Vision*. MIT Press, 1986.

[7] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12:5–16, 1994.

[8] B. Julesz. *Foundations of Cyclopean Perception*. The University of Chicago Press, 1971.

[9] R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: A parallax based approach. In *ICPR*, pages 685–688, 1994.

[10] B. D. Lucas and T. Kanade. An iterative image registration technique with an application in stereo vision. In *IJCAI*, 1981.

[11] W. R. Mark, L. McMillan, and G. Bishop. Post-rendering 3d warping. In *Symp. on Interactive 3D Graphics (Providence, RI)*, pages 7–16, 1997.

[12] Y. Matsumoto et al. Conversion system of monocular image sequence to stereo using motion parallax. In *SPIE Stereo. Disp. and VR Sys.(Vol 3012)*, 1997.

[13] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *SIGGRAPH*, pages 39–46, 1995.

[14] P.J. Narayanan, P.W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *ICCV*, pages 3–10, 1998.

[15] Michael G. Perkins. Data compression of stereo pairs. *IEEE Trans. on Comm.*, 40(4):684–696, 1992.

[16] W. K. Pratt. *Digital Image Processing (Second Edition)*. Wiley, 1991.

[17] H. S. Sawhney. 3D geometry from planar parallax. In *CVPR*, pages 929–934, 1994.

[18] A. Shashua and N. Navab. Relative affine structure: Theory and application to 3D reconstruction from perspective views. In *CVPR*, pages 483–489, 1994.

[19] L. Stelmach et al. Human perception of mismatched stereoscopic 3d stereo inputs. In *ICIP*, 2000.

[20] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proc. Intl. Conf. on Computer Vision*, 1999.

[21] R. Szeliski and J. Coughlan. Spline-based image registration. *IJCV*, 22(3):199–218, 1997.

[22] W.H.Press, B.P.Flannery, S.A.Teukolsky, and W.T.Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1986.

[23] G. Wolberg. *Digital Image Warping*. IEEE Computer Society Press, CA, 1990.