Separation of Transparent Layers using Focus

Yoav Y. Schechner* Nahum Kiryati* Ronen Basri **

*Department of Electrical Engineering,
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

**Department of Applied Mathematics
Weizmann Institute of Science
Rehovot 76100, ISRAEL

Abstract

Consider situations where the depth at each point in the scene is multi-valued, due to the presence of a virtual image semi-reflected by a transparent surface. The semi-reflected image is linearly superimposed on the image of the object that is behind the transparent surface. A novel approach is proposed for the recovery of the superimposed layers. By searching for the images in which either of the objects (layers) is focused, the transparent areas are detected and an estimate of the depth map of each layer is obtained. As a result of the focusing, an initial separation of the layers is achieved. The separation is enhanced via mutual blurring of the perturbing components in the images, based on the depths estimate and the parameters of the imaging system.

1. Introduction

The approach of depth from focus (DFF) consists of obtaining image *slices* of the scene (imaging with different focus settings) from which depth is extracted by a search for the slice maximizing a focus criterion [1-6]. DFF methods concentrated on cases in which the depth, at each point of the image, is single valued. However, the situation in which several (typically two) linearly superimposed contributions exist is often encountered in real-world scenes. For example [7], looking out of a room window, we see both the outside world (termed *real object* [8,9]), and a semi-reflection of the objects inside the room, termed *virtual objects*. The treatment of such cases is important, since the combination of several unrelated images may greatly degrade the ability to understand them and also confuses autofocusing devices. The detection of the phenomenon also indicates the presence of a transparent surface in front of the camera, at a distance closer than the imaged objects.

The term *transparent layers* is used in the context of scenes semi-reflected from transparent surfaces [7,10,11] (in the current work we do not refer to viewing through an object having a variable opacity, since there the superposition is not linear [12]). The image is decomposed into layers, each with an associated depth and intensity distribution. We adopt the common layer representation, in which within each layer the relative depth variations are small compared to the inter-layer difference. Approaches to reconstructing the layers [7,8,10-13] relied mainly on motion and stereo.

The treatment of multiple objects in the axial dimension has been considered in the field of microscopy [14-17]. The emphasis [15-17] has usually been put on the reconstruction of the continuous volume, rather than discrete layers. In [14] a method for DFF was demonstrated in a layered situation, but due to the very small depth of field used, the interfering layer was very blurred so no reconstruction process was necessary.

In this work the phenomenon of multi-valued depth is first detected and the depth-map of each of the objects is estimated by means of an extension of the DFF algorithm. We assume the depth of each layer is approximately constant over patches Then, the limited depth of field is exploited to separate and reconstruct the intensity distribution of multiple layers. We concentrate on the common case of two layers. The generalization to a larger number of layers can be easily derived.

2. Detection of transparency, and DFF

The distances to the real and virtual objects are assumed to differ greatly. This assumption holds in many practical situations. Thus, if the lens aperture is large enough, only one of the objects may be in-focus. Imaging is first done with different focus settings, so as to sample the 3D viewed world into a few slices. A focus measure, calculated in each of these slices, is searched as a function of the slice-index. A new method to find the focus is presented.

2.1. The optical system

An imaging system telecentric on the image side [5] ensures a constant magnification even if the sensor plane is out of focus (the defocused contributions will be used

later for layer reconstruction). The depth scan is performed by moving the sensor array axially, enabling the efficient coverage of long object distances, up to infinity. A model of the system is shown in Fig. 1.

The object points u_1 and u_2 are focused at points v_1 and v_2, respectively, which are two of the axial positions of the sensor of the camera. Point u_1 is defocused when the sensor is at v_2. The radius of the support [19] of the geometrical 2D blur PSF is

$$r_{2,1} = (a / F)\Delta v , \qquad (1)$$

where $\Delta v \equiv |v_2 - v_1|$. The same relation is obtained for the blur-radius of the image of u_2, when the sensor plane is at v_1. Thus, the marginal rays emanating from axial points in the object space are parallel to each other when emerging in the image space (Fig. 1). Hence, the 2D point spread function (PSF) does not depend on the position of the sensor array, but only on the distance between the focused-image plane and the sensor plane. We adopt the standard assumption, that the properties of the imaging system are invariant to transversal shift. We thus conclude that the imaging is a 3D space-invariant operation [18] at the image space $(\bar{x}, \bar{y}, v)$, where the transversal coordinates in this space are related by $(\bar{x}, \bar{y}) = [1 - u / F]^{-1}(x, y)$ to the object coordinates, and $(1 / v) = (1 / F) - (1 / u)$. Recall that for a single 2D image, different points of the scene are blurred differently - according to their depth. However, the entire 3D effect of the telecentric system is space invariant in image space, regardless of the scene.

2.2. Depth sampling

In some of the previous work [2,4,6,14], the axial movement between consecutive slices corresponded to a single step of the step-motor or was arbitrarily chosen. We suggest that, by more careful planning, the depth sampling may be sparser. Let the axial sampling positions of the sensor be at v_z, where z is the slice index. Consider an on-axis object point for which the corresponding image point is at $v_1 < v \le v_2$ (see Fig. 1). The 2D PSF over the plane at v_z has radius $r_z = a|v - v_z|/F$. Suppose that $\tilde{r}$ is the radius of the smallest blur kernel that leads to detectable defocus. By requiring that $r_2, r_1 < \tilde{r}$ and $r_z|_{z \neq 2,1} > \tilde{r}$, we obtain two slices that seem almost equally focused, while the others are blurred. This bounds the depth estimation (after the geometrical transformation) to be between v_1 and v_2. Sampling the depth more densely will give multiple sharp images of the same object points, but not tighten the bounds. Thus, we require $r_{2,1} = \tilde{r}$. The radius $\tilde{r}$ is related to the transverse (2D) sampling period, $\Delta\bar{x}$, of the sensor array. Assuming $\tilde{r} \approx \Delta\bar{x}$, and substituting it into Eq. (1), leads to the axial sampling period

$$\Delta v \approx F\Delta\bar{x} / a . \qquad (2)$$

Taking the first sample at $v=F$ to enable focusing on infinity, the axial sampling positions are

$$v_z = F + (z-1)F\Delta\bar{x} / a , \quad \text{where} \quad z = 1,2,3,\ldots K , \qquad (3)$$

and the number of slices is $K = 1 + Fa/\left[\Delta\bar{x}(u_{\min} - F)\right]$, where $u_{\min}$ is the minimal viewed depth.

A more rigorous derivation is based on 3D spatial frequency considerations, as we showed in [19]. Eq. (2) is associated with the 3D Nyquist rate, based on the characteristics [17] of the geometrical PSF, while it is four times denser than required by physical optics [19] (when the imaging system is diffraction limited).

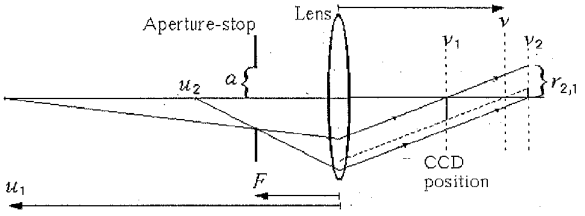

Fig. 1 The telecentric imaging system model. Only the lower marginal rays are shown. a is the radius of the stop. F is the focal length.

2.3. Detection of layers, and depth recovery

A conventional focus-measure is first calculated in each slice. Common criteria [1,3,6,14] are sensitive to 2D variations in the slice (for example, calculating the gradient response). This is done on each slice, leading to "slices of local focus-measure", $FOCUS(\bar{x}, \bar{y}, z)$, where z is the slice index. We assume for simplicity that the scene can be divided into patches in which the objects have a roughly constant depth. In the sequel we continue the analysis separately in each patch.

Naively, one might suggest to average $FOCUS(\bar{x}, \bar{y}, z)$ over the patch to obtain $FOCUS(z)$. Ideally, in the presence of several layers, each of the layers would lead to a main peak in $FOCUS(z)$. However, mutual interference may shift the peaks off their original positions, and even lead to the appearance of only a single peak in some "average" position (see Fig. 2). For this reason transparent scenes confuse conventional autofocusing devices.

Since the layers are generally unrelated, the chance that a brightness edge in one of them will appear in the

same spot as an edge of the other is small. Since edges (and other feature-dense regions) are dominant contributors to the focus criterion, it would be wise not to mix them by brute averaging over the entire patch. If point $(\bar{x},\bar{y})$ is on an edge in one layer, while on an ordinary, smooth region in the other layer, then the peak of the edge in $FOCUS(\bar{x},\bar{y},z)$ will not be greatly affected by the contribution of the other layer. So, we suggest to rely on feature-dense regions to extract depth information and associate it with the entire patch.

For a specific pixel $(\bar{x},\bar{y})$ in the slices, the focus measure is analyzed as a function of the slice index. For each pixel the local maxima of this function are found. The result is expressed as a binary vector of local maxima positions. For example, if the focus measure has local maxima at the 1st and 5th slice (out of 6), the vector is (1,0,0,0,1,0). A vote table is formed by summing the "hits" in each slice-index over all pixels in the patch. Each vote is given a weight that depends monotonically on its value $FOCUS(\bar{x},\bar{y},z)$, to enhance the contribution of high focus-measure values, such as those arising from edges, while reducing the random contribution of featureless areas. The vote table eventually is as seen in Fig. 2. The number of layers in the scene is equal to the number of significant values. Assuming a-priori that the maximum number of layers is two (as in most cases), the two highest values are used. The patches in which the transparency was detected are segmented. Via Eq. (3) the distances of the layers from the camera correspond, roughly, to the slice indices that received the highest number of votes.

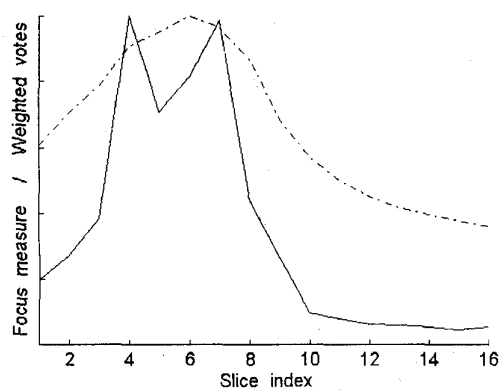

Fig. 2. [Dashed line]: The conventional focus measure of an experimental scene, as a function of the slice index. It mistakenly detects a single focused state at the 6th slice. [Solid line]: The locations histogram of detected local maxima of the focus measure (the same scene). The highest numbers of votes (positions of local maxima) are correctly accumulated at the 4th and 7th slices - where the layers would individually be focused.

3. Layer reconstruction

Following the detection of the slices in which either of the layers is in focus, we have estimates of the distance of each layer from the lens. The imaging system is under our control, and we assume that its parameters are known. We can thus calculate the blur kernel of each layer, when the camera is focused on the other one.

Let layer f_1 be superimposed on layer f_2. Consider the slices g_a and g_b, in which either layer f_1 or layer f_2, respectively, is in focus. The other layer is blurred

$$g_a = f_1 + f_2 * h_{2a} \quad , \quad g_b = f_2 + f_1 * h_{1b} \quad , \qquad (4)$$

where * denotes convolution. Due to the telecentricity, $h_{1b} = h_{2a} = h$, where h is the common blur kernel.

The reconstruction of the layers may be visualized in the frequency domain, where Eqs. (4) take the form of two linear constraints (see Fig. 3). The solution, which corresponds to their intersection, uniquely exists for $H \neq 1$. The slopes of the lines representing the constraints are reciprocal to each other. As the frequency response H approaches 1 (that is, at low frequencies), the slopes of the two lines become similar, hence the solution is more sensitive to noise in G_a and G_b. When H=1 (i.e., for the DC component), the constraints coincide into a single line implying infinite number of solutions in the noiseless case; in the presence of noise in the input images the lines become parallel (no solution). Due to energy conservation, the average gray level is not affected by defocusing. We can only limit this component to satisfy

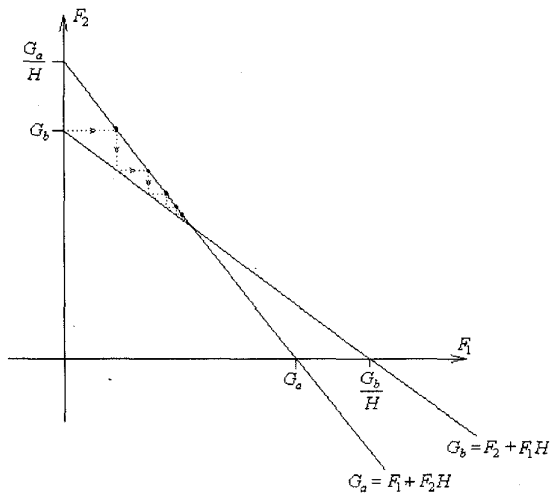

Fig. 3. Visualization of the convergence of the suggested iterative algorithm in the transversal frequency domain. For each frequency, the constraints take the form of straight lines.

$$\mathrm{DC}(\hat{f}_1) + \mathrm{DC}(\hat{f}_2) = \mathrm{DC}(g_a) = \mathrm{DC}(g_b) \quad , \qquad (5a)$$

$$\hat{f}_1 \geq 0 \quad , \quad \hat{f}_2 \geq 0 \quad , \quad \hat{f}_1 \leq 1 \quad , \quad \hat{f}_2 \leq 1 \quad . \qquad (5b)$$

where $\hat{f}_1$ and $\hat{f}_2$ are the estimations of f_1 and f_2, respectively. On the other hand, *the problem is well posed and stable at the high frequencies.*

This behavior is quite opposite to many typical reconstruction problems. However, it is not unique to this algorithm of transparency separation, but also seen in the results obtained using motion. In [10], the reconstructions of semi-reflected scenes are clearly highpass filtered versions of the superimposing components. In [11], one of the objects is "dominant". As the dominant object is faded out in the reconstruction, it leaves considerable low-frequency contamination. In regions of translational motion the spatiotemporal energy of each layer resides on a plane [7,12] in the spatiotemporal frequency domain, which passes through the origin. Any two of these frequency planes have a common frequency line passing through the origin (the DC), whose components are thus generally inseparable.

To bypass similar problems [15], an iterative approach has been used. The method suggested here, which iteratively applies the constraints of Eq. (4), is visualized as alternating vectors parallel to the axes of Fig. 3. For $H<1$ it converges to the solution from any initial hypothesis. As H decreases (roughly speaking, as the frequency increases), the lines approach perpendicularity, thus convergence is faster.

The slices g_a and g_b may be taken as the initial hypotheses for $\hat{f}_1$ and $\hat{f}_2$, respectively. With these initial conditions, we obtain (in the 2D spatial frequency domain) at the m'th iteration

$$\hat{F}_{1,m} = \hat{T}_m(G_a - G_b H) + H^{m+1} G_b \quad ,$$

$$\hat{F}_{2,m} = \hat{T}_m(G_b - G_a H) + H^{m+1} G_a \qquad (6)$$

for even m, where

$$\hat{T}_m = \sum_{l=0}^{m/2} H^{2l} \qquad (7)$$

Eq. (7) is a geometrical series, converging to the inverse filter as $m \to \infty$, for $H < 1$. According to Eq. (6), $\hat{T}_m$ has a major effect on the amplification of noise added to the raw images g_a and g_b. At high transversal spatial frequencies, $H \to 0$, so the amplification of additive noise is negligible.

The more iterations done, the unknown surroundings affect larger portions of the image. In the spatial domain, Eq. (7) turns into a convolution kernel t_m. Approximating the blur and reconstructing kernels by their discrete versions, the spatial support of t_m is $(2rm+1)$ pixels long, where r is the radius of the support of h. For an image N pixels wide, the number of iterations m is chosen to satisfy $2rm+1 \leq \varepsilon N$ where ε is a small fraction.

The above result suggests the possible existence of a basic limit to the ability to separate layers. If r is very large, only few iterations can be done, if at all, and we cannot improve the image much. However, the initial slices already show a good separation of the individual layers, since the defocused layer is very blurred and thus is hardly disturbing. On the other hand, if r is small, confusing images are initially created but a large number of iterations can be carried out.

The reconstruction kernel can be calculated a-priori to a length of about εN. Yet, this approach has a greater complexity than the iterative one, unless the blur kernel is separable. If it is, the complexity turns out to be similar. However, this approach may be very efficient, if convolution is implemented by the FFT algorithm. Nevertheless, in an iterative process, the dynamic range constraints (5b) can be conveniently imposed [16].

Reconstruction may also be achieved using a single focused slice and a pinhole image [19]. The axial positions of the system components are the same for both images, hence no geometrical distortions are present. This relaxes the telecentricity requirement, previously needed for the reconstruction stage.

In practice, the imaging PSF will be slightly different than the one used in the reconstruction. This may be due to inaccurate prior modeling or calibration of the imaging PSF. This is also a consequence of error in depth estimation. Our analysis [19] showed that the overall effect of this error and the reconstruction operation is to slightly contaminate the reconstruction with the low and middle frequency components of both layers (assuming the error in depth estimation is small).

4. Examples

4.1 Simulation

A simulated scene consists of a "bridge" layer at a depth of 10m. The other layer ("astronaut") was given a variable depth in the range of 0.8-1m. Fig. 4 is the combined scene as imaged via a pinhole camera. (the images in this example are shown in full contrast). The dimensions of the imaging system are of order of magnitude of common macroscopic systems [19].

The algorithm successfully detected the layers as described in Sec. 2. The significance of the secondary maximum of the voting counts was determined using thresholds. The depth maps of the objects were correctly

Fig. 4. The scene as seen through a pinhole camera

estimated within the depth of field of the system, except for a few marginal patches estimated at the some mid-distance between the layers Some patches were detected as single-layered, since there are large textureless parts of the "astronaut" layer that do not indicate the presence of any depth-layer. The focused slices are shown in the top of Fig. 5. Even though some layer separation is obtained, significant crosstalk is seen. The reconstructed layers, after 7 iterations, are shown in the bottom.

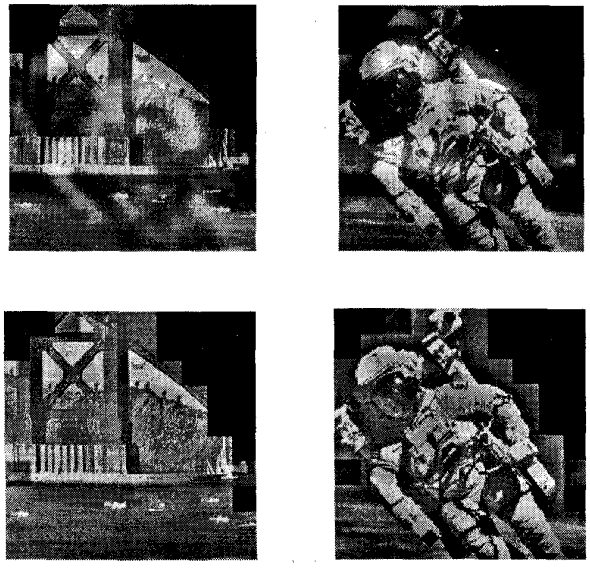

Fig. 5. Simulation results. The image patches where transparency was detected are segmented and shown. The rest are darkened. [Upper-left]: The slice in which the far layer is in focus. [Upper-right]: A composition of slice patches, in each of which a region of the close layer is in focus, creating an overall focused image of the layer. [Lower row]: Reconstructed layers. A few marginal patches were detected at the wrong distance and introduce local disturbance.

4.2 Experiment

The experimental setup consisted of a "vase" picture that was partly-reflected from the glass-cover of a "crab" picture situated at a distance of 2.8m from the lens, making a total optical distance of 5.3m from the lens. The imaging system consisted of a 75mm triplet lens, a CCD sensor array mounted on a linear step motor, and an aperture stop. Using laser beams, the system was aligned to be telecentric, as in Fig. 1. We set the diameter of the stop to be $2a$=15mm, for which our calculation showed that vignetting is well avoided within the field of view. The depth variations of these objects were negligible with respect to the depth of field. A 256×256 part of each of the images was cropped and used. An image of the scene as viewed through a pinhole camera (mimicked using a 3.75mm aperture stop) is shown in Fig. 6.

According to geometric considerations (2), the axial sampling period should be about 0.13mm. However, calibration showed that the standard deviation of the PSF is about 1 pixel at the focused state, and increases about 3 times slower than expected, as a function of defocusing (probably due to electronic and optical blur). This leads to an increase of about the same factor in the sampling period. This was consistent with our subjective sensation of minimum detectable defocus. We thus sampled the axial position of the CCD in a period of 0.338mm, corresponding to 266 steps of the motor. This demonstrates that acquiring images at minimal detectable defocus intervals, as suggested in Sec. 2.2, significantly reduces the number of images that need to be taken.

We used the focus measure $17|\nabla_{\text{row}} g|+13|\nabla_{\text{column}} g|$ on each slice g. The distances between rows and between columns of the CCD are 13μ and 17μ , respectively. Thus, these factors correct the "digital" gradient to be consistent with the physical one. We treated the image as a single patch. The results of the focus search are shown in Fig. 2. The mean of the focus measure failed to detect the layers. However, the vote table clearly succeeded to find the layers (the weight of each vote is equal to the square of the value of its corresponding focus measure). The depths deduced by this method are correct, within the uncertainty imposed by the empirical depth of field of the system. The slices in which either of the layers is focused are shown in the top of Fig. 7.

We used the PSF estimated by calibration, rather than a theoretical model, for the reconstruction of the layers. The PSF when the sensor array was in front of the plane of best focus turned out to be quite different than the PSF on the opposite side. In this example we imposed constraints (5b) within the iterations by the method used in [16]. The results after 11 iterations are shown in the bottom of Fig. 7.

Figure 6: The scene viewed through a small aperture system focused at infinity. The image is enhanced to partly compensate for the effects caused by the low light input.

To conclude, the new DFF algorithm combined with the axial sampling criterion demonstrated successful detection and depth estimation of transparent layers. The layer separation achieved via focusing was enhanced by a reconstruction algorithm. The suggested technique, though simple, provides valuable guidelines and a basis for better algorithms.

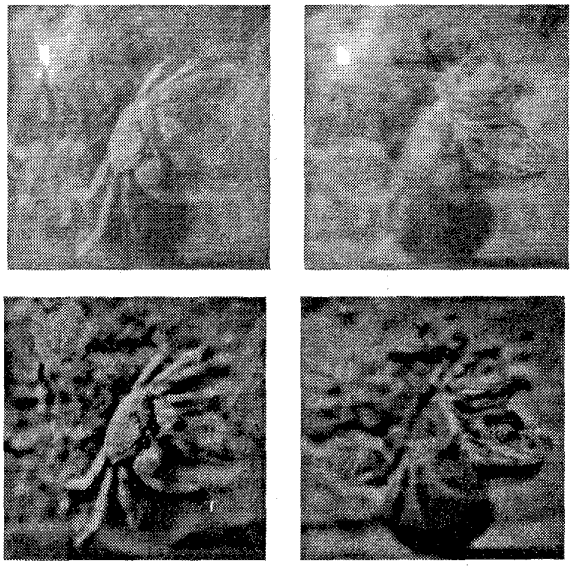

Figure 7: [Upper row]: The slices in which either of the transparent layers is focused. [Lower-left]: The reconstructed "crab" layer. The image is much improved. in particular, the dark silhouette of the "vase" was mostly filled and is less noticeable. [Lower-right]: The reconstructed "vase" layer has some leftovers from the other layer, but details of the "vase" are better seen.

References

[1] R. Jarvis, "A perspective on range-finding techniques for computer vision," *IEEE Trans. Patt. Anal. Machine. Intell*, vol. PAMI-3, pp. 122-139, 1983.

[2] T. Darrell and K. Wohn, "Pyramid based depth from focus," *Proc. CVPR*, pp. 504-509, 1988.

[3] S. K. Nayar, "Shape from focus system" *Proc. CVPR*, pp. 302-308, 1992.

[4] Y. Xiong and S. A. Shafer, "Depth from focusing and defocusing," *Proc. CVPR*, pp. 68-73, 1993.

[5] S. K. Nayar, M. Watanabe and M. Nogouchi, "Real time focus range sensor," *Proc. ICCV*, pp. 995-1001, 1995.

[6] T. T. E. Yeo, S. H. Ong, Jayasooriah and R. Sinniah, "Autofocusing for tissue microscopy," *Image and Vision Comp.* vol. 11, pp. 629-639, 1993.

[7] T. Darrell and E. Simoncelli, "Separation of transparent motion into layers using velocity-tuned mechanisms," TR-244, Media-Lab, MIT, 1993.

[8] M. Oren and S. K. Nayar, "A theory of specular surface geometry," *Proc. ICCV*, pp. 740-747, 1995.

[9] N. Ohnishi, K. Kumaki, T. Yamamura and T. Tanaka, "Separating real and virtual objects from their overlapping images," *Proc. ECCV*, vol. 2, pp. 636-646, 1996.

[10] J. R. Bergen, P. J. Burt, R. Hingorani and S. Peleg, "Computing two motions from three frames," *Proc. ICCV*, pp. 27-32, 1990.

[11] M. Irani, B. Rousso and S. Peleg, "Computing occluding and transparent motions," *Int. J. Comp. Vis.*, vol. 12, pp. 5-16, 1994.

[12] M. Shizawa and K. Mase, "Simultaneous multiple optical flow estimation," *Proc. ICPR*, pp. 274-278, 1990.

[13] M. Shizawa, "On visual ambiguities due to transparency in motion and stereo," *Proc. ECCV*, pp. 411-419, 1992.

[14] K. Itoh, A. Hayashi and Y. Ichioka, "Digitized optical microscopy with extended depth of field," *App. Opt.* vol. 28, pp. 3487-3493, 1989.

[15] D. A. Agard and J. W. Sedat, "Three-dimensional structure of a polytene nucleus," *Nature*, vol. 302, pp. 676-681, 1983.

[16] J. A. Conchello and E. W. Hansen, "Enhanced 3-D reconstruction from confocal scanning microscope images. I: Deterministic and maximum likelihood reconstructions," *App. Opt.*, vol. 29, pp. 3795-3804, 1990.

[17] F. Marcias-Garza, A. C. Bovik, K. R. Diller, S. J. Aggarwal and J. K. Aggarwal, "The missing cone problem and low-pass distortion in optical serial sectioning microscopy," *Proc. ICASSP*, vol. 2, pp. 890-893, 1988.

[18] D. N. Sitter and W. T. Rhodes, "Three dimensional imaging: a space invariant model for space variant systems," *App. Opt.*, vol. 29, pp. 3789-3794, 1990.

[19] Y. Y. Schechner, N. Kiryati and R. Basri, "Separation of transparent layers using focus," EE-PUB-1086, Technion - Israel Institute of Technology, 1997.