Stereo From Flickering Caustics

Yohay Swirski , Yoav Y. Schechner, Ben Herzberg
Dept. of Electrical Engineering
Technion - Israel Inst. Technology
Haifa 32000, Israel
yohays@tx.technion.ac.il, yoav@ee.technion.ac.il,
sbenh@t2.technion.ac.il

Shahriar Negahdaripour
Electrical and Computer Eng. Dept.
University of Miami
Coral Gables, FL 33124-0640
shahriar@miami.edu

Abstract

Underwater, natural illumination typically varies strongly temporally and spatially. The reason is that waves on the water surface refract light into the water in a spatiotemporally varying manner. The resulting underwater illumination field is known as underwater caustics *or* flicker. *In past studies, flicker has often been considered to be an undesired effect, which degrades the quality of images. In contrast, in this work, we show that flicker can actually be useful for vision in the underwater domain. Specifically, it solves very simply, accurately, and densely the stereo correspondence problem, irrespective of the object's texture. The temporal radiance variations due to flicker are unique to each object point, thus disambiguating the correspondence, with very simple calculations. This process is further enhanced by compounding the spatial variability in the flicker field. The method is demonstrated by underwater in-situ experiments.*

1. Introduction

Interest in underwater computer vision is growing significantly [2, 14, 19, 21]. It includes research of new principles as well as applications. The latter include underwater robotic vision [3], inspection of pipelines [9], communication cables, ports, ship hulls [24] and swimming pools [20]. Underwater imaging is also applied to archaeological documentation [18] and observation of wildlife [4, 5, 23]. Computer vision methods that are sought for this environment are sometimes versions of open-air methods, such as mosaicing and stereo. They are adapted to be more robust to the environment. Other methods try to tackle poor visibility conditions [27, 28], which often exist [32] in such a medium.

One effect which can be rather strong in this domain is *sunlight flicker*. Here, submerged objects and the water volume itself are illuminated by a natural random pattern [17, 32] which is spatially and temporally varying. An example is shown in Fig. 1. So far, this phenomenon has been considered mainly as a significant *disturbance* to vision.[1] Thus, attempts were made to reduce this effect by postprocessing [13, 26].

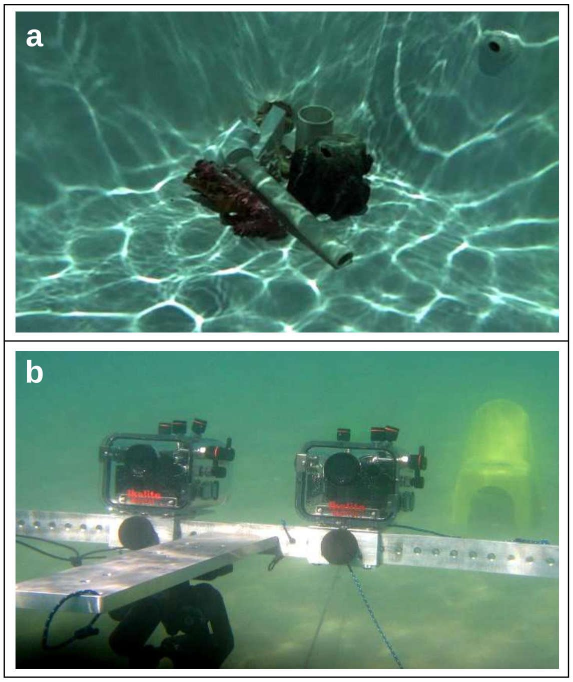

Figure 1. [a] Sunlight flicker irradiating a scene in a pool. [b] An underwater stereoscopic video setup in the Mediterranean.

In this paper, we show that the spatiotemporal variations of this flicker can actually be very beneficial to underwater vision. Specifically, flicker disambiguates stereo correspondence. This disambiguation is very simple, yet it yields very accurate results. Beyond computer vision, the problem of stereo under flicker may be relevant to biological vision, as we discuss in Sec. 7.

We begin by modeling underwater image formation, in the context of our recovery problem. Past studies focused either on non-flickering models [27] in scattering media, or

[1] Studies in computer graphics synthesized this effect [11, 16, 17].

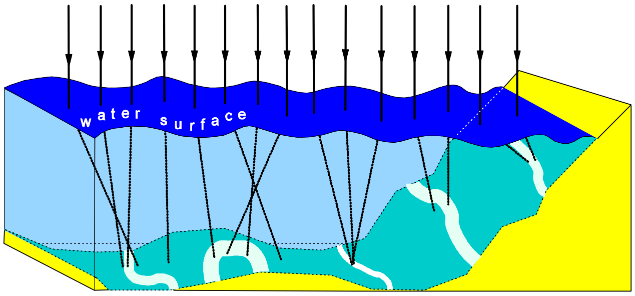

Figure 2. A wavy water surface refracts natural illumination in a spatially varying way. Underwater, the refracted rays create a spatial pattern of lighting. The pattern varies temporally, due to the motion of the surface waves [26].

on computer vision with variable lighting without accounting for scattering effects that occur underwater. Our model combines flicker, scattering effects along the line of sight (LOS) and stereo formulation. The model is significantly simplified using several approximations, which we detail. These approximations are partly based on statistical properties of natural flicker patterns [11]. Based on this model, it becomes clear that it is justified to use the temporal variations of flicker as a tool to establish unambiguous correspondence by local (even pointwise) calculations. Moreover, the approach is capable of functioning under degraded visibility effects encountered underwater. The approach is sufficiently simple and robust to field conditions. It has the ability to identify failure cases, i.e, pixels in which it cannot reliably establish correspondence.

2. Theoretical Background

2.1. Shallow Water Sunlight Flicker

Consider a short period of time. In every single moment, the water surface is generally not flat but rather wavy [10, 12]. Therefore, the refraction of sunlight through this wavy surface creates inhomogeneous lighting as illustrated in Fig. 2. Concave regions on the surface diverge the light rays refracting underwater, while convex regions converge them. This results in uncontrolled patterns of variable irradiance.[2] When reaching the sea floor or other underwater objects, the light rays create bright regions termed *caustic networks* [22]. Due to the natural motion and evolution of the surface waves, this light pattern changes in time, and is thus known as *sunlight flicker* [13]. Consequently, the irradiance in the water changes as a function of space and time. Thus, it is denoted by $I^{\text{lighting}}(\mathbf{x}, z, t)$, where t is the temporal frame index, $\mathbf{x} = (x, y)$ is an image coordinate, which corresponds to a specific LOS in the scene, and z is the distance between a point in space and the camera housing.

2.2. Effects Along a Line of Sight

Let the object radiance at a point be $I^{\text{obj}}(\mathbf{x})$. Due to attenuation in the water, the *signal* originating from this object [27] is

$$S(\mathbf{x}) = I^{\text{obj}}(\mathbf{x})e^{-\eta z(\mathbf{x})} \quad , \tag{1}$$

where η is the attenuation coefficient of the water. Here $z(\mathbf{x})$ is the distance to the object at $\mathbf{x}$.

In addition, the water scatters the ambient illumination into the LOS, creating *veiling light*, also termed *backscatter* [14]. It is given [27] by an integral over the LOS,

$$B(\mathbf{x}) = \int_0^{z(\mathbf{x})} I^{\text{lighting}}(\tilde{z})p(\theta)e^{-\eta\tilde{z}}d\tilde{z} \ . \tag{2}$$

Here, p is the phase function and θ is the lighting angle relative the LOS, at that point. According to Ref. [27], Eq. (2) can be approximated as

$$B(\mathbf{x}) = B_\infty[1 - e^{-\eta z(\mathbf{x})}] \quad , \tag{3}$$

where B_∞ is the veiling light in a LOS reaching infinity. Overall, the radiance measured by the camera is

$$I(\mathbf{x}) = S(\mathbf{x}) + B(\mathbf{x}) \quad . \tag{4}$$

3. Image Formation Model

The image formation model is simple. It combines the spatiotemporal irradiance field described in Sec. 2.1 with the LOS effects described in Sec. 2.2. We adapt the formulation to stereo. In addition, in this work we employ approximations that simplify the model and the resulting shape recovery.

3.1. Flicker Signal

We use stereoscopic vision. Denote the left camera by $\mathtt{L}$. We align the global coordinate system with this camera, i.e, the position of a point in the water volume or an object is uniquely defined by the left spatial coordinate vector $\mathbf{x}_{\mathtt{L}}$ and the distance z from the housing of the left camera. The signal corresponding to Eq. (1) is

$$S_{\mathtt{L}}(\mathbf{x}_{\mathtt{L}}, t) = I^{\text{lighting}}(\mathbf{x}_{\mathtt{L}}, z, t)r_{\mathtt{L}}(\mathbf{x}_{\mathtt{L}})e^{-\eta z(\mathbf{x}_{\mathtt{L}})} \quad , \tag{5}$$

where $r_{\mathtt{L}}$ denotes the reflectance coefficient of the object towards the left camera.

The right camera is denoted by $\mathtt{R}$. The object corresponding to $(\mathbf{x}_{\mathtt{L}}, z)$ in the left camera is projected to pixel $\mathbf{x}_{\mathtt{R}}$ in the right camera. The corresponding disparity vector is

$$\mathbf{d} = \mathbf{x}_{\mathtt{R}} - \mathbf{x}_{\mathtt{L}} \quad . \tag{6}$$

The viewpoints of the two cameras are different, separated by a baseline of length b. This leads to two consequences.

[2] The wavy water surface also refracts lines of sight passing through the water surface, as described in [7].

First, the object distance z to the left camera is different than the distance to the right one. This may affect the signal attenuation. Second, the direction from an object point to the left camera is different than the direction to the right one. This may affect the reflectance coefficient r.

For reasons detailed in the following, we assume that these differences do not have significant consequences to the eventual shape recovery. Overall we use

Approximation 1: At both cameras, the signal is attenuated in the same manner, $\approx e^{-\eta z(\mathbf{x}_{\rm L})}$. Difference of attenuation has no significance.

Approximation 2: The reflectance coefficients are the same for both cameras: $r_{\rm L} = r_{\rm R}$. Differences in these coefficients have no significance.

Suppose for the moment that the attenuation to the right camera is significantly different than the left, by a factor $f_1(\mathbf{x}_{\rm R})$. Furthermore, suppose the corresponding reflectance coefficients are different by a factor $f_2(\mathbf{x}_{\rm R})$. Then, the signal in the right-camera is

$$\begin{aligned} S_{\rm R}(\mathbf{x}_{\rm R}, t) &= I^{\rm lighting}(\mathbf{x}_{\rm L}, z, t) r_{\rm L}(\mathbf{x}_{\rm L}) e^{-\eta z(\mathbf{x}_{\rm L})} f_1(\mathbf{x}_{\rm R}) f_2(\mathbf{x}_{\rm R}) \\ &= S_{\rm L}(\mathbf{x}_{\rm R}, t) f_1(\mathbf{x}_{\rm R}) f_2(\mathbf{x}_{\rm R}) \quad . \end{aligned} \tag{7}$$

The factors f_1 and f_2 are a function of space, but are *temporally invariant*. Hence, they are *canceled out* if matching between $\mathbf{x}_{\rm L}$ and $\mathbf{x}_{\rm R}$ is established by *normalized temporal correlation* of the signals. Furthermore, these factors typically change very gradually. Hence, even matching by normalized spatial (non-temporal) correlation is rather insensitive to these factors.

These approximations are supported by other reasons. The distance difference to the two cameras is typically much lower than b, since the baseline is approximately perpendicular to the optical axis, and $b \ll z$. Anyway, the distance difference is bounded by b. Thus, $f_1(\mathbf{x}_{\rm R})$ is bounded by $e^{-\eta b}$. Typically, $\eta b \ll 1$. For example, in water of visibility-distance $(1/\eta)$ of 10m and baseline of 30cm, $e^{-\eta b} \approx 0.97$. Hence, $f_1(\mathbf{x}_{\rm R}) \approx 1$. The assumption that $f_2(\mathbf{x}_{\rm R}) \approx 1$ is common in the stereo matching literature, and is known as the *brightness constraint*. Nevertheless, as explained above, temporal matching (by normalized temporal correlation) is less sensitive to brightness differences due to reflectance and attenuation differences in the two viewpoints. Recapping, based on the above approximations. the signals in the two cameras are related by

$$S_{\rm L}(\mathbf{x}_{\rm L}, t) \approx S_{\rm R}(\mathbf{x}_{\rm R}, t) \quad \forall\ t \ . \tag{8}$$

3.2. Backscatter

Models such as Eqs. (2,3) have been used for recovery in non-flickering environments [27]. Here we discuss our model. Now, Eq. (2) depends both on time and the viewpoint. Specifically, in the left camera,

$$B_{\rm L}(\mathbf{x}_{\rm L}, t) = \int_0^{z(\mathbf{x})} I^{\rm lighting}(\tilde{z}, t) p(\theta_{\rm L}) e^{-\eta \tilde{z}} d\tilde{z} \ . \tag{9}$$

Is the spatiotemporal variation of the integral significant? Note that flicker does not change the overall energy irradiated into the water. At each frame, the caustic pattern only changes the spatial *distribution* of this fixed energy. In this random spatial pattern, some points on the LOS have higher irradiance than others. However, different points have different weight in Eq. (9): a closer water voxel contributes more to the overall backscatter than a distant voxel. If the caustic distributes much of its energy at nearby voxels, the backscatter can be significantly higher than if this energy is focused only at very distant voxels. So, an important question is, how wide are the features of the caustic pattern? In other words, what is the *correlation length* of the random pattern $I^{\rm lighting}(z)$?

In this work, we make the following assumption:

Approximation 3: The correlation length of $I^{\rm lighting}(z)$ is much smaller than the attenuation length $1/\eta$.

This means, that the caustic pattern changes from bright to dark features within short mutual distances along the LOS, while the attenuation weight does not change much in such mutual proximity. According to [11], a typical flicker correlation distance may vary between a few centimeters to a few tens of centimeters, depending on the underwater depth and the water surface waves. On the other hand, the attenuation distance typically varies [32] between a few meters to a few tens of meters. Therefore, there is at least one order of magnitude between the flicker correlation length and $1/\eta$.

Under *approximation 3*, the spatial variations of the caustic are filtered out by the integral in Eq. (9). Consequently, Eq. (3) holds for the left-camera. For the same reasons, Eq. (3) holds also for the right-camera. Hence,

$$B_{\rm R}(\mathbf{x}_{\rm R}, t) \approx B_{\rm L}(\mathbf{x}_{\rm L}, t) = B_\infty(t)[1 - e^{-\eta z(\mathbf{x}_{\rm L})}] \ \ \forall\ t \ . \tag{10}$$

The temporal variations in $B_\infty(t)$ are small, for the reason mentioned above: the flicker changes the distribution of energy, but not its spatial integral. Hence, in different frames, the pattern changes, bright voxels dim and vice versa, but the integral over the LOS is insensitive to these temporal changes. Compounding Eqs. (4,8,10), the overall scene radiance, as measured by the two stereo cameras can be formulated as

$$I_{\rm R}(\mathbf{x}_{\rm R}, t) \approx I_{\rm L}(\mathbf{x}_{\rm L}, t) \quad \forall\ t \ . \tag{11}$$

4. Correspondence From Flicker

Equation (11) claims intensity similarity at points $\mathbf{x}_{\rm R}$ and $\mathbf{x}_{\rm L}$ at time t. However, this similarity is generally not unique, at time t. A set of pixels $\Omega_{\rm R}(t) = \{\mathbf{x}_k^{\rm incorrect}\}$ in $I_{\rm R}$ have intensities that are very close to, or equal to $I_{\rm L}(\mathbf{x}_{\rm L})$. One reason why this can happen is that objects at such non-corresponding pixels may have the same reflectance, irradiance and backscatter. This situation leads to the classic correspondence problem in non-flickering environments. A more general reason is that the reflectance, irradiance and backscatter in each $\mathbf{x}_k^{\rm incorrect}$ are all different than the ones in $\mathbf{x}_{\rm R}$, but their combination in Eq. (4) yields the same overall intensity, at time t.

Fortunately, in flicker, such ambiguities are completely resolved with high probability, since the lighting is dynamic.[3] Due to the lighting dynamics, non-corresponding pixels in $\Omega_{\rm R}(t)$ are generally different than those at $\Omega_{\rm R}(t')$, at time $t' \neq t$. A coincidence of matching intensities at t has rare chances of re-occurring at t'. Considering a large number of frames N_F,

$$\bigcap_{t=1}^{N_F} \Omega_{\rm R}(t) \longrightarrow \emptyset \ , \tag{12}$$

where in practice, even a small N_F suffices to eliminate the non-corresponding pixels.

4.1. Temporal Correlation

In practice, correspondence is solved in our work using mainly simple *temporal* normalized correlation. Define the vector

$$\mathbf{I}_{\rm L}(\mathbf{x}_{\rm L}) \equiv \begin{bmatrix} I_{\rm L}(\mathbf{x}_{\rm L}, 1) \\ I_{\rm L}(\mathbf{x}_{\rm L}, 2) \\ \vdots \\ I_{\rm L}(\mathbf{x}_{\rm L}, N_F) \end{bmatrix} \ . \tag{13}$$

Now, in the right image, there is a set of pixels Ψ, each of which is a candidate for correspondence with $\mathbf{x}_{\rm L}$. Without calibration of the stereo setup, Ψ is the whole field of view (all the pixels in the image). If calibration of the system had been done, then Ψ is the epipolar line [15, 31] corresponding to $\mathbf{x}_{\rm L}$. For a candidate pixel $\mathbf{x}_{\rm R}^{\rm cand} \in \Psi$, define

$$\mathbf{I}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}) \equiv \begin{bmatrix} I_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}, 1) \\ I_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}, 2) \\ \vdots \\ I_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}, N_F) \end{bmatrix} \ . \tag{14}$$

Subtracting the mean of each vector, we obtain

$$\tilde{\mathbf{I}}_{\rm L}(\mathbf{x}_{\rm L}) = \mathbf{I}_{\rm L}(\mathbf{x}_{\rm L}) - \langle \mathbf{I}_{\rm L}(\mathbf{x}_{\rm L}) \rangle , \tag{15}$$

$$\tilde{\mathbf{I}}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}) = \mathbf{I}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}) - \langle \mathbf{I}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand}) \rangle . \tag{16}$$

The empirical normalized correlation [8] between $\mathbf{x}_{\rm L}$ and $\mathbf{x}_{\rm R}^{\rm cand}$ is

$$C(\mathbf{x}_{\rm R}^{\rm cand}) = \frac{\tilde{\mathbf{I}}_{\rm L}(\mathbf{x}_{\rm L})^T \tilde{\mathbf{I}}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand})}{\|\tilde{\mathbf{I}}_{\rm L}(\mathbf{x}_{\rm L})\|_2 \|\tilde{\mathbf{I}}_{\rm R}(\mathbf{x}_{\rm R}^{\rm cand})\|_2} , \tag{17}$$

where T denotes transposition. For pixel $\mathbf{x}_{\rm L}$ in the left image, the corresponding pixel in the right image is then estimated as

$$\hat{\mathbf{x}}_{\rm R} = \arg \max_{\mathbf{x}_{\rm R}^{\rm cand} \in \Psi} C(\mathbf{x}_{\rm R}^{\rm cand}) . \tag{18}$$

As described in Sec. 3.1, the criterion of normalized temporal correlation is rather insensitive to inter-camera differences in signal attenuation and object reflectance. It is also insensitive to differences in camera exposure parameters (gain and shutter speed) [30].

Experimental Example

To see the effectiveness of the approach, consider our first example, which is an experiment done in a swimming pool. The scene includes several objects at $z \in [1\text{m}, 2\text{m}]$, near the corner of the pool. The depth at the bottom of the pool was $\sim$ 1m. The stereo setup was a *Videre Design* head shooting at 7fps, with $b = 25$cm. A sample frame-pair appears in Fig. 3a,b. Temporal correlation was performed using $N_F = 35$. Here, as in all the following experiments, the setup was not calibrated, hence the search domain Ψ includes the entire field of view.[4] As common in studies dealing with stereo correspondence [25], the result is displayed as a *disparity map*, rather than a range map.[5] The disparity map is derived based on Eq. (6):

$$\hat{d}(\mathbf{x}_{\rm L}) = \|\hat{\mathbf{x}}_{\rm R} - \mathbf{x}_{\rm L}\| \ . \tag{19}$$

It is shown in Fig. 3d. One example of the temporal match in corresponding points is shown in Fig. 3c.

There are a few small regions with clearly outlying results. These regions were in constant shadow, hence without any flicker. This is discussed in Sec. 5.

4.2. Spatiotemporal Correlation

The method described in Sec. 4.1 is very simple and accurate. It does not blur range edges, since it involves no spatial operations. However, it requires that correlation be established over a length of time. In static scenes, this is

[3] This is related to a method described in Ref. [6]. There, man-made light patterns illuminate the scene using a projector in order to establish correspondence between stereo video cameras. In scattering media, artificial illumination is problematic, since it cannot irradiate distant objects [27]. Artificial structured illumination is often designed to be narrow, to reduce excessive backscatter [14]. In our case, lighting variations are natural and are anywhere along the LOS.

[4] Since epipolar geometry was not exploited to limit the match search, a few erroneous matches appeared, which would have been bypassed with epipolar search constraints. These singular errors were effectively eliminated from the disparity map using a 3 × 3 median filter.

[5] A range map can be derived from the correspondences, once the setup is calibrated. Underwater, such a calibration may not match the calibration model in air, since the water interface introduces a non-single viewpoint geometry [29] to each camera.

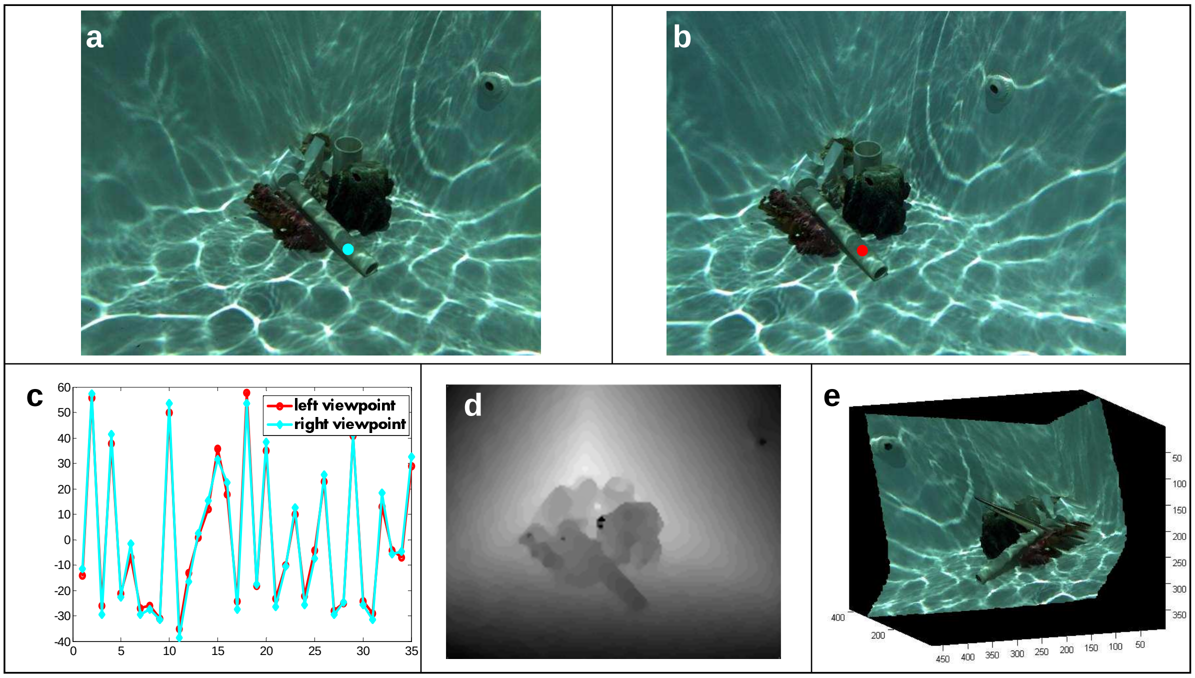

Figure 3. Left [a] and right [b] frames at one instance in the sequence. [c] Temporal plots of $\tilde{\mathbf{I}}_{\mathrm{L}}(\mathbf{x}_{\mathrm{L}})$ and $\tilde{\mathbf{I}}_{\mathrm{R}}(\hat{\mathbf{x}}_{\mathrm{R}})$ extracted from corresponding pixels. These pixels are marked by cyan and red in the respective frames [a,b]. The estimated disparity map $\|\hat{d}\|$ is shown in [d]. Its reciprocal, which is similar to the range map, is used for texture mapping a different viewpoint in [e].

not a problem. However, if speed is desired, it can be enhanced by using spatiotemporal, rather than just temporal correlation, as explained in this section. Here, the comparison is not pixel-wise, but using spatial blocks. This enables the use of smaller N_F, at the price of loss of spatial resolution and consequent range errors, particulary in range edges. Fig. 4 illustrates the possibilities of correlation support.

Let $\beta(\mathbf{x}_{\mathrm{L}})$ be a block of $l \times l$ pixels centered around $\mathbf{x}_{\mathrm{L}}$. The pixel values in this block change during the N_F frames. Thus, the video data cube corresponding to these pixels has dimensions of $l \times l \times N_F$. Concatenate this video data cube into a vector

$$\mathbf{I}_{\mathrm{L}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{L}}) \equiv [\mathbf{I}_{\mathrm{L}}(\mathbf{x}_1)^T,\ \mathbf{I}_{\mathrm{L}}(\mathbf{x}_2)^T, \ldots \mathbf{I}_{\mathrm{L}}(\mathbf{x}_{l^2})]^T \quad , \qquad (20)$$

where $\{\mathbf{x}_m\}_{m=1}^{l^2} \in \beta(\mathbf{x}_{\mathrm{L}})$. Analogously, an $l \times l$ block $\beta(\mathbf{x}_{\mathrm{R}})$ is centered around $\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}}$. Use the same concatenation as in Eq. (20) over the video in $\beta(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}})$ of the right camera. This yields

$$\mathbf{I}_{\mathrm{R}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}}) \equiv [\mathbf{I}_{\mathrm{R}}(\mathbf{y}_1)^T,\ \mathbf{I}_{\mathrm{R}}(\mathbf{y}_2)^T, \ldots \mathbf{I}_{\mathrm{R}}(\mathbf{y}_{l^2})]^T \quad , \qquad (21)$$

where $\{\mathbf{y}_m\}_{m=1}^{l^2} \in \beta(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}})$.

Now, Eqs. (15,16) are redefined as

$$\tilde{\mathbf{I}}_{\mathrm{L}}(\mathbf{x}_{\mathrm{L}}) = \mathbf{I}_{\mathrm{L}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{L}}) - \langle \mathbf{I}_{\mathrm{L}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{L}}) \rangle, \qquad (22)$$

$$\tilde{\mathbf{I}}_{\mathrm{R}}(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}}) = \mathbf{I}_{\mathrm{R}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}}) - \langle \mathbf{I}_{\mathrm{R}}^{\mathrm{cube}}(\mathbf{x}_{\mathrm{R}}^{\mathrm{cand}}) \rangle. \qquad (23)$$

Eqs. (22,23) are then used in Eqs. (17,18).

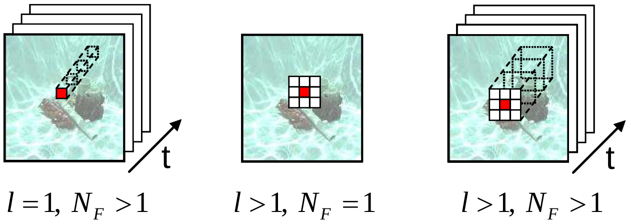

Figure 4. Support used for [left] temporal, [middle] spatial and [right] spatiotemporal correlation.

Spatial Correlation

A degenerate case is a solution based on a single stereo frame-pair, i.e, $N_F = 1$, while $l > 1$. Here, only spatial correlation is performed. This is the common method [1, 31] for establishing stereo correspondence. Even in this case, the flicker usually helps [24]. Matching that is based solely on spatial correlation requires significant spatial texture. Thus, the spatial variations in the caustic lighting field provides some texture over areas having textureless albedo. This had been used [24] in underwater experiments to enhance the correspondence, independently per each individual stereo frame-pair.

5. Knowing Where It Can't Work

It is possible to assess in which image locations, correspondence estimation using Eq. (17) is unreliable. As in any stereo setup, occluded regions cannot be directly matched. More particular to this approach, however, is that some pixels simply correspond to object points that reside in the shadow of downwelling lighting, due to objects above them. Points in the shadow are unaffected by flicker. Similarly, for objects which are very far away, the signal is attenuated (Eq. 5), thus it is difficult to sense the temporal variations due to flicker there in short periods.

The set of pixels in $\mathbf{I}_{\mathrm{L}}$ that are occluded in $\mathbf{I}_{\mathrm{R}}$ have a low value of C even in the "optimal" match $\hat{\mathbf{x}}_{\mathrm{R}}$. Hence, if $C(\hat{\mathbf{x}}_{\mathrm{R}})$ is below a threshold τ_C, it indicates an unreliable correspondence. What about the set of pixels in which flicker is absent or too weak to detect (due to shadow or long distance)? In each of these pixels, the standard deviation of the pixel value $\|\tilde{\mathbf{I}}_{\mathrm{L}}(\mathbf{x}_{\mathrm{L}})\|_2$ is very low. Hence, this set can be assessed by thresholding the field $\|\tilde{\mathbf{I}}_{\mathrm{L}}(\mathbf{x}_{\mathrm{L}})\|_2$ by a parameter τ_{STD}.

Thus, we may define a set ρ of reliable pixels.

$$\rho = \{\mathbf{x}_{\mathrm{L}} : \; [C_{\mathbf{x}_{\mathrm{L}}} > \tau_C] \text{ AND } [\|\tilde{\mathbf{I}}_{\mathrm{L}}(\mathbf{x}_{\mathrm{L}})\|_2 > \tau_{\mathrm{STD}}]\} \; . \quad (24)$$

Range information in pixels outside ρ should be filled-in using other mechanisms, such as regularization or inpainting.

6. Additional Experiments

We conducted a set of in-situ field experiments. Different sceneries and cameras were used, in the ocean and in a pool.

6.1. Swimming-Pool Experiment

A swimming pool experiment is described in Sec. 4.1, where results of temporal correlation are shown in Fig. 3. Here, water visibility was good. This allowed us to extract quantitative performance measures based on manual matching. In the field of view, 100 points were randomly selected in I_{L}. These points were manually matched in I_{R}. This match served as ground truth in the tests. First, Fig. 5 plots the required N_F as a function of the required reliability of matching, where epipolar constraints were not put to use.

Then, using the same video data, we re-ran the recovery using spatiotemporal correlation, as described in Sec. 4.2, using various values of l. Qualitatively, the resulting disparity maps resemble those of Fig. 3. The quantitative plots in Fig. 5 lead to interesting conclusions. Using spatial support for the correlation, a moderate success rate of $\approx 80-85\%$ can be achieved with much less frames than if using only temporal correlation. However, widening the spatial support stagnates the success rate below $\approx$ 90% even when the number of frames grows. Possibly, this is caused by true violations of spatial smoothness in the range map. This problem does not occur when only temporal correlation is used pointwise. With pointwise analysis, the success rate increases monotonically with time and eventually surpasses the results achieved using spatial matching windows.

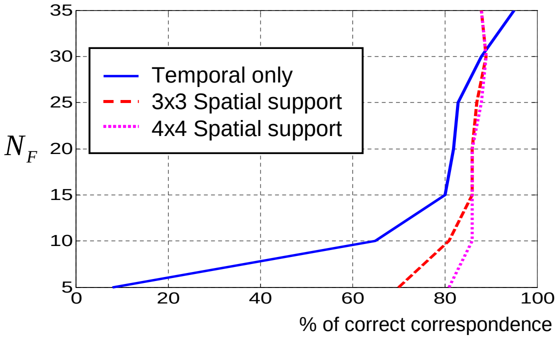

Figure 5. The number of frames required to achieve a certain rate of successful match in the experiment corresponding to Fig. 3.

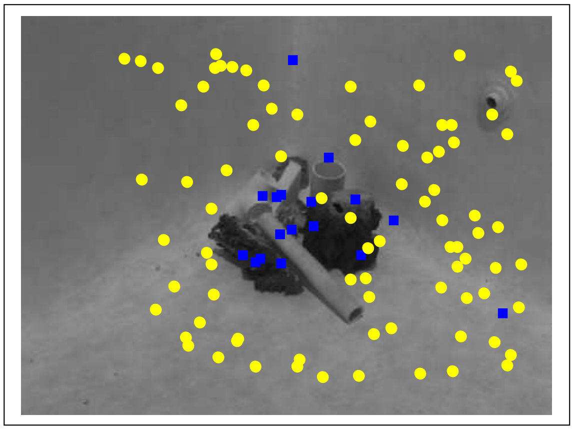

Figure 6. A rendered "flicker-free" image of the scene. Stereo correspondence without caustics yields a few successful results (blue squares) in textured regions, and many wrong results (yellow circles) in textureless regions.

As mentioned above, correspondence may also be sought using only spatial correlation, in a single stereo pair ($N_F = 1$) of a flicker scene. As an example, we applied this known method [24] using $l = 7$. Here, the success rate was $\approx 60\%$. It is much lower than using temporal or spatiotemporal matching. However, it is still valuable, compared to matching without flicker. To see this, we first created an image of the scene in which the flicker is reduced by temporal median filtering [13]. The resulting images appear as if the scene is illuminated nearly uniformly (Fig. 6). Now, large smooth areas do not support well stereo matching. Hence, spatial correlation matched only 17% of the points.

6.2. Oceanic Experiments

We conducted field experiments in the Mediterranean, aided by scuba diving. The experiments were conducted at depths of $3-5$m. Photographs of the stereo setup are shown in Figs. 1 and 7. Here, we used Canon HV-30 high-definition PAL video cameras within Ikelite underwater housings. To synchronize the video sequences, brief

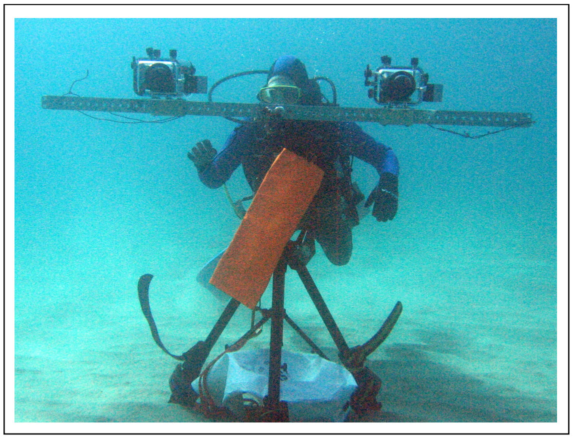

Figure 7. The oceanic experiment setup. Here, $b \in [30, 70]$ cm.

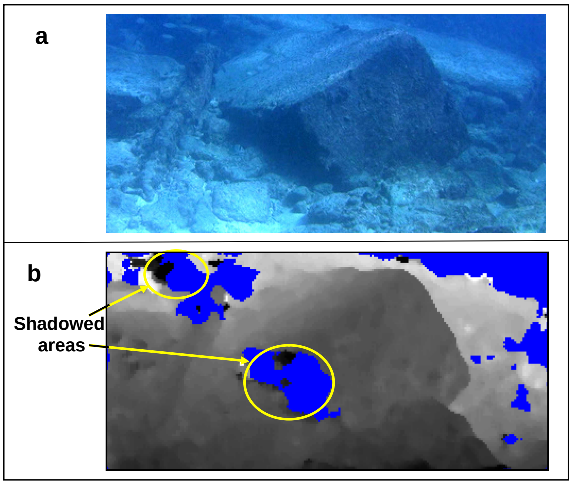

Figure 8. [a] A raw left frame from an experiment in a marine archaeological site (Caesarea). [b] The estimated disparity map. Blue areas represent low correspondence reliability.

light flashes were shined into the running cameras before and after each experiment. These flashes were later detected in postprocessing and used to temporally align the videos.

In the sea, the visibility was much poorer than in the pool. Hence, the flicker pattern had lower contrast. This required somewhat longer sequences to reliably establish the correlation, and thus correspondence. In any case, the sequences were just a few seconds long. In one experiment, a natural scene in an underwater archeological site was captured using a $b = 70$cm baseline and $N_F = 66$. The resulting disparity map is presented in Fig. 8. The distance of the large cube from the cameras was ~ 5m. As explained in Sec. 5, there is automatic determination of pixels having low reliability of the match. Such pixels are marked in blue in Fig. 8. They appear mainly in shadowed areas.

Another oceanic experiment done in a different day is depicted in Fig. 9. Here visibility was poorer, leading to shorter objects distances. The distance of the chair was 2m. Consequently the baseline was shorter ($b = 30$cm) and $N_F = 75$.

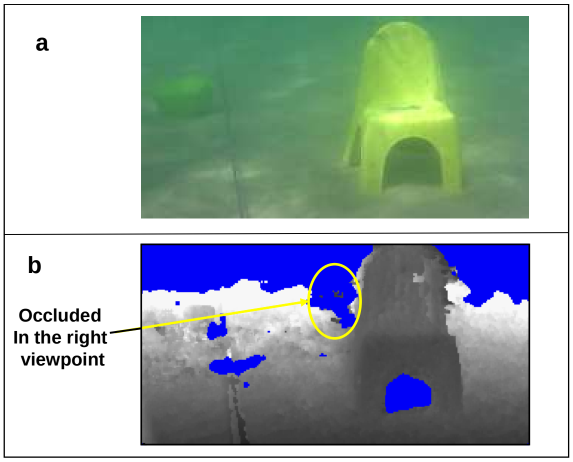

Figure 9. [a] A raw left frame from a second oceanic experiment. [b] The estimated disparity map. Blue areas represent low correspondence reliability.

7. Biological Relevance?

The problem of stereo vision under flicker may be relevant to biological vision. Consider marine animals that share three properties: (1) They live in shallow water, where there is abundance of natural light and thus flicker; (2) They have binocular vision, which can potentially enable them to assess distance by triangulating objects from different eye elements; (3) Their brain is very small. Can such small brains solve the complex problems associated with correspondence? Specifically, such animals include stomatopods and other crustaceans (e.g, lobsters) [4, 5, 23].

The mechanism proposed in this paper may potentially be a key to stereo vision in such animals: it suits their habitat, and requires very simple, pointwise calculations. Such an hypothesis may be verified in biological studies. For instance, the animals may undergo controlled tests in the lab, where their ability to succeed in a task which requires distance assessment (prey, navigation) under flicker (wavy water surface) is compared to the success in still water.

8. Conclusions

The method presented in this paper exploits the natural phenomenon of underwater flicker to create a dense correspondence map. The effectiveness of the method is demonstrated by in-situ experiments in the field. As mentioned in Sec. 4.2, there is a trade-off between the spatial and temporal supports needed to obtain a certain success rate. This trade-off may be optimized by an adaptive algorithm: it may determine the support of the spatial and temporal windows around each pixel, according to the spatiotemporal texture in its neighborhood.

The method establishes correspondence rather reliably even without epipolar constraints. So, in turn, the results of a sequence of such correspondence mappings can possibly establish the epipolar geometry of the system. This would be interesting to study.

Acknowledgements

We thank Tali Treibitz, Amin Sarafraz and Ali Taatian for their help in conducting the underwater experiments. We thank Nadav Shashar, Justin Marshall, Tom Cronin and Dan Nilsson for helpful discussions about potential stereo vision in marine invertebrates. Yoav Schechner is a Landau Fellow - supported by the Taub Foundation. The work was supported by the Israel Science Foundation (Grant 1031/08) and the US-Israel Binational Science Foundation (BSF) Grant 2006384. This work was conducted in the Ollendorff Minerva Center. Minerva is funded through the BMBF.

References

[1] R. Bolles, H. Baker, and M. Hannah. The JISCT stereo evaluation. In *Proc. DARPA Image Understanding Workshop*, pages 263–274, 1993.

[2] T. Boult. DOVE: Dolphin omni-directional video equipment. In *Proc. IASTED Int. Conf. Robotics and Autom.*, pages 214–220, 2000.

[3] M. Bryant, D. Wettergreen, S. Abdallah, and A. Zelinsky. Robust camera calibration for an autonomous underwater vehicle. In *Proc. Australian Conf. on Robotics and Autom.*, pages 111–116, 2000.

[4] T. W. Cronin and J. Marshall. Parallel processing and image analysis in the eyes of mantis shrimps. *Biol. Bull.*, 200:177–183, 2001.

[5] T. W. Cronin, J. N. Nair, R. D. Doyle, and R. L. Caldwell. Ocular tracking of rapidly moving visual targets by stomatopod crustaceans. *J. Exp. Biol.*, 138:155–179, 1988.

[6] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Trans. PAMI*, 27:296–302, 2005.

[7] A. A. Efros, V. Isler, J. Shi, and M. Visontai. Seeing through water. In *Proc. NIPS 17*, pages 393–400, 2004.

[8] R. Eustice, O. Pizarro, H. Singh, and J. Howland. UWIT: underwater image toolbox for optical image processing and mosaicking in Matlab. In *Proc. Int. Sympos. on Underwater Tech.*, pages 141– 145, 2002.

[9] G. L. Foresti. Visual inspection of sea bottom structures by an autonomous underwater vehicle. *IEEE Trans. Syst. Man and Cyber*, 31:691–705, 2001.

[10] A. Fournier and W. T. Reeves. A simple model of ocean waves. In *Proc. SIGGRAPH*, pages 75–84, 1986.

[11] A. Fraser, R. Walker, and F. Jurgens. Spatial and temporal correlation of underwater sunlight fluctuations in the sea. *IEEE J. Oceanic Eng.*, 5:195 – 198, 1980.

[12] M. Gamito and F. Musgrave. An accurate model of wave refraction over shallow water. *Computers and Graphics*, 26:291–307, 2002.

[13] N. Gracias, S. Negahdaripour, L. Neumann, R. Prados, and R. Garcia. A motion compensated filtering approach to remove sunlight flicker in shallow water images. In *Proc. MTS/IEEE Oceans*, 2008.

[14] M. Gupta, S. Narasimhan, and Y. Y. Schechner. On controlling light transport in poor visibility environments. In *Proc. IEEE CVPR*, 2008.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*, chapter 9-12. Cambridge University Press, 2003.

[16] K. Iwasaki, Y. Dobashi, and T. Nishita. An efficient method for rendering underwater optical effects using graphics hardware. *Computer Graphics Forum*, 21:701–711, 2002.

[17] N. G. Jerlov. *Marine Optics*, chapter 6. Elsevier, Amsterdam, 1976.

[18] Y. Kahanov and J. Royal. Analysis of hull remains of the Dor D vessel, Tantura lagoon, Israel. *Int. J. Nautical Archeology*, 30:257–265, 2001.

[19] D. M. Kocak, F. R. Dalgleish, F. M. Caimi, and Y. Y. Schechner. A focus on recent developments and trends in underwater imaging. *MTS J.*, 42(1):52–67, 2008.

[20] J. M. Lavest, F. Guichard, and C. Rousseau. Multiview reconstruction combining underwater and air sensors. In *Proc. IEEE ICIP.*, volume 3, pages 813–816, 2002.

[21] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas. Synthetic aperture confocal imaging. *ACM TOG*, 23:825–834, 2004.

[22] D. K. Lynch and W. Livingston. *Color and Light in Nature*, chapter 2.4,2.5,3.7,3.16. Cambridge U.Press, 2nd edition, 2001.

[23] J. Marshall, T. W. Cronin, and S. Kleinlogel. Stomatopod eye structure and function: A review. *Arthropod Structure and Development*, 36:420–448, 2007.

[24] S. Negahdaripour and P. Firoozfam. An ROV stereovision system for ship-hull inspection. *IEEE J. Oceanic Eng.*, 31:551–564, 2006.

[25] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.

[26] Y. Y. Schechner and N. Karpel. Attenuating natural flicker patterns. In *Proc. MTS/IEEE Oceans*, pages 1262–1268, 2004.

[27] Y. Y. Schechner and N. Karpel. Recovery of underwater visibility and structure by polarization analysis. *IEEE J. Oceanic Eng.*, 30:570–587, 2005.

[28] T. Treibitz and Y. Y. Schechner. Active polarization descattering. *IEEE Trans. PAMI*, 31(3):385–399, 2009.

[29] T. Treibitz, Y. Y. Schechner, and H. Singh. Flat refractive geometry. *In Proc. IEEE CVPR*, 2008.

[30] A. Troccoli, S. Kang, and S. Seitz. Multi-view multi-exposure stereo. In *Proc. 3DPVT06*, pages 861–868, 2006.

[31] E. Trucco and A. Verri. *Introductory Techniques For 3D Computer Vision*, chapter 6. Prentice Hall, New Jersey, 1998.

[32] R. E. Walker. *Marine Light Field Statistics*, chapter 10. John Wiley, New York, 1994.