Stereoscopic Ranging by Matching Image Modulations

Tieh-Yuh (Terry) Chen, Member, IEEE, Alan Conrad Bovik, Fellow, IEEE, and Lawrence K. Cormack

Abstract—We apply an AM–FM surface albedo model to analyze the projection of surface patterns viewed through a binocular camera system. This is used to support the use of modulation-based stereo matching where local image phase is used to compute stereo disparities. Local image phase is an advantageous feature for image matching, since the problem of computing disparities reduces to identifying local phase shifts between the stereoscopic image data. Local phase shifts, however, are problematic at high frequencies due to phase wrapping when disparities exceed $\pm \pi$. We meld powerful multichannel Gabor image demodulation techniques for multiscale (coarse-to-fine) computation of local image phase with a disparity channel model for depth computation. The resulting framework unifies phase-based matching approaches with AM–FM surface/image models. We demonstrate the concepts in a stereo algorithm that generates a dense, accurate disparity map without the problems associated with phase wrapping.

I. INTRODUCTION

In recent years there has been a growing interest in the utilization of range images to map scenes arising in industrial, medical, military and other applications. Computational stereopsis, where disparity is computed from two or more cameras with displaced image planes and then converted to range via triangulation, is a powerful means for the passive acquisition of range information. Passive sensing has the advantages of requiring less sensor sophistication and, in relevant applications, reduced sensor detectability. Stereoscopic ranging also is simple in concept and implementation. Further, very successful biological stereo ranging systems are available for study and emulation. Virtually identical stereo algorithms have evolved independently across classes (mammals and birds) and perhaps species (carnivores and primates), indicating that there may be a single, best stereo algorithm that the process of evolution inevitably uncovers [1].

Computational stereopsis algorithms can generally be broken into three processing steps [2]. First is image preprocessing, where the stereo pair is processed such that the second stage (matching) is made easier, either by enhancing features to be matched or by decomposing the images into subimages to be separately matched. In the second stage, matching features from each image are detected (either at sparse spatial locations [3]–[10], or at all the image points [11]–[13]), and it is decided which points in one image match to which points from the other. This solves the “correspondence problem,” the most difficult aspect of the stereo ranging problem. The means for solving it is what differentiates the many algorithms that have been proposed. The relative shift between the coordinates of the matched points is the disparity, the computation of which depends on the geometry of the cameras. Finally, the third step involves calculating range from disparity using the known camera geometries.

An important class of approaches that deliver dense depth results involves the use of measured image phase as matching features [14]–[16], [42]–[45]. Sanger [14] uses a bank of one-dimensional (1-D) Gabor filters (aligned along the epipoles) to decompose the stereo image data into different scale-subimages. Local instantaneous frequencies are approximated by the center frequencies of the Gabor filters (introducing an error). Pointwise phase differences are then computed between the stereoscopic images (without the need for matching). Disparities from each channel are computed from the phase differences and averaged over spatial scales to give a final result. The approach has advantages: search-free correspondence computation and, hence, depth computation without solving an intense optimization problem. Sanger’s algorithm has some limitations, owing to the coarse approximations used, and it produces oversmoothed disparity maps.

Jepson and Jenkin [15] also proposed a simple scheme for rapidly computing disparities from the output phase differences of 1-D Gabor filters. They formulated a multiscale disparity computation, rather than simply averaging over channels. Their first approach also used coarse phase approximations, failing to recover disparities densely, e.g., in simple random-dot stereograms. However, in a later effort [44], they utilized auxiliary methods for inferring local surface structures, which produced denser and more detailed results.

Langley et al. [16] developed an algorithm using the responses of both 1-D and two-dimensional (2-D) Gabor filters. Disparities were estimated by solving a first-order differential equation via a Newton–Raphson root-finding iteration. The approach suffered from problems related to unstable phase wrapping in the computation of local image phase; this was ameliorated to some degree by approximating the instantaneous phase to assist in the disparity computations.
In a deeper study of the problem of computing phase-based disparities, Fleet et al. [42], [43] examined the stability of stereoscopic phase computations, and the role of phase discontinuities, which are found to create errors (see also [26]). They proposed methods to ameliorate these effects by detecting phase discontinuities and using surrounding phase information to produce better disparity estimates. There has also been increasing recognition in the psychophysics community of the probable significance of phase information in biological visual processing. For example, Fleet et al. [45] propose a model of phase-based disparity processing in human visual cortex. They view phase shift information as being pooled across the responses of neurons tuned to multiple orientations and radial spatial frequencies, producing a stable, unambiguous disparity representation.

The view taken in this paper is that the use of local image phase to compute disparities in stereoscopic images (or motion images, for that matter) as advanced by these researchers is a powerful concept that has the potential, once fully developed, to free current matching-based image analysis problems from a dependence on solving difficult search-based correspondence calculations. However, phase-based approaches have lacked a clear connection to the scene that is projected into the image being processed.

It is a goal of this paper to better develop this connection. We employ a recently introduced and very natural image model, the AM–FM image model, wherein image information is captured by the amplitude and frequency variations of modulated 2-D sinusoidal functions. The model is quite successful for analyzing nonstationary, locally coherent image structures [17]. A multichannel Gabor filterbank is used to make localized measurements of the modulating functions, and to separate distinct modulating functions. Accurate multiscale measurements are made of local image phase and local image contrast. A more complete physical meaning for these measurements is developed by using the AM–FM model within the context of surface-to-image projection: image phase is directly related to the surface patterns. Given accurate measurements of local image phase, stereo disparities are computed from phase differences. We propose a new model for disparity computation utilizing the concept of disparity channels, which admits a systematic approach for coarse-to-fine disparity computation that melds seamlessly with phase disparity measurements made under the multiscale AM–FM image model. Through these developments, we create a new stereo algorithm that recovers disparity at every pixel and that preserves information near depth discontinuities. Intensity discontinuities (which typically accompany depth discontinuities) are well handled by the AM–FM image model, where they are represented as sudden phase shifts.

The paper is organized as follows. Section II describes a powerful AM–FM image model that is derived from the surface albedo and a stereoscopic imaging geometry. Section III considers potential modulation-based matching features. Among these, the local image phase is deemed most useful. Section IV develops a coarse-to-fine matching strategy using a model of disparity channels designed in concert with the Gabor filterbank. Simulation results appear in Section V. The paper concludes in Section VI.

II. FROM SURFACE MODULATION TO IMAGE MODULATION

We first describe the image modulation model that will be used in the development and analysis of a phase-based approach to computed stereopsis. The model uses a simple projective geometry applied to a model of modulated surface reflectivity. A simple, static nonconvergent stereo imaging geometry is used to demonstrate principle; however, the same approach may be applied (with modifications) to vergent or active stereo imaging geometries.

Most of the mathematical developments are cast in a continuous-domain setting for simplicity. Discretization will be considered in the context of algorithm. A sensed 2-D optical image will be modeled by general multicomponent amplitude- and frequency-modulated (AM–FM) complex functions \( (\mathbb{C}^2 = -1) \) of the form \( t : \mathbb{R}^2 \rightarrow \mathbb{C} \) indexed by 2-D coordinates \( \mathbf{x} = (x, y) \), as follows:

\[
\hat{t}(\mathbf{x}) = \sum_{n=1}^{N} A_n(\mathbf{x}) \exp[i\Omega_n(\mathbf{x})]
\]

where the amplitude modulation (contrast) functions \( A_n : \mathbb{R}^2 \rightarrow [0, 1] \) are the local peak-to-peak contrast of each component and the frequency modulation functions (phases) \( \Omega_n : \mathbb{R}^2 \rightarrow \mathbb{R}^+ \) capture locally significant frequency variations over diverse orientations and granulations. Optical images have real, positive intensity distributions, whereas (1) is complex. As explained in Section II-E, (1) can be regarded as the result of preprocessing a real, positive image function.

AM–FM models are becoming increasingly popular as analysis tools for a variety of signal and image processing tasks. As early as 1977, Moorer [20] used 1-D models similar to (1) to analyze and synthesize computer music. More recently, 1-D single component models \( (N = 1) \) have been studied [21], [22] and applied to speech signal analysis [23]. Two-dimensional models of image modulation were introduced and systematically developed by Bovik et al. in the context of image segmentation [24], [25], nonstationary image analysis [17], [26], [27], and shape-from-texture [28]. Multicomponent AM–FM image models were first proposed in [26] and have recently been deeply analyzed [29], [30]. These studies have shown that the model (1) can effectively capture physically meaningful signal structures that arise from natural and artificially occurring processes. Moreover, the concept of local frequency and amplitude variation, which presumes a specific class of modulated image structure, is quite direct and intuitive as opposed to, for example, very general statistical models, such as Markov random fields, that are described in terms of unspecified local interactions between samples. Recent work in multicomponent AM–FM modeling [29] has shown that very general images can be represented with good visual accuracy by a small sum of AM–FM functions, and that the essential structures of realistic image patterns can usually be captured with very few such components [29], [31].

The main restriction is that the AM–FM functions be regular in the sense that the modulating functions be smooth; more specifically, that they have small Sobolev norms [17], [22], [26], [29]–[31].
In this paper, the goal is primarily to model image “emergent” structures, viz., that dominate the local perception of an image, but that may also be nonhomogeneous in that the local spectrum varies. Various tools seek to capture spectral nonstationarities, e.g., short-time Fourier transforms (STFT’s), time-frequency distributions, and wavelet expansions [32]; however, the model (1) is quite well suited for modeling them. Nonstationary image structures may arise in a number of ways, including optical (geometric) distortion, but we are mostly interested in the following three types of physical processes, 1) Irregular, yet locally coherent patterns or subpatterns in surface reflectance, such as the markings on an animal, in wood, or in woven fabrics. 2) Variations in surface topography lead to apparent changes in pattern density and pose, even if the surface pattern is intrinsically regular. Smoothly-varying shapes project stationary surface patterns onto nonstationary image patterns. The division between the processes 1) and 2) is ambiguous, and depends to a great extent on the scale of observation and on the surface smoothness (or lack thereof). In 3), the process of imaging, or perspective projection, distorts patterns through foreshortening. All of these factors provide information that can be used to disambiguate image intensities between stereoscopic image frames.

A. Surface Reflectivity Model

Underlying the image modulation model is a similar surface reflectance model, whereby the surface reflectivity or albedo is assumed to take the form

$$t_a(x_s) = 1 + \sum_{n=1}^{N} \alpha_n(x_s) \cos[\omega_n(x_s)]$$

(2)

where $x_s = (x_s, y_s)$ are local surface coordinates at a macroscopic scale of observation. Models of the form (1) and (2) are related to recently emerging models of natural pattern formation. For example, partial differential equations that are used to describe naturally-forming surface patterns often have solutions that have coherent AM-FM forms. These include the Cross–Newell phase diffusion equations [47]–[49] and certain reaction-diffusion equations [50].

Note that

$$1 - \sum_{n=1}^{N} \alpha_n(x_s) \leq t_a(x_s) \leq 1 + \sum_{n=1}^{N} \alpha_n(x_s).$$

(3)

Since albedo is positive, we have the following additional restriction:

$$\sum_{n=1}^{N} \alpha_n(x_s) \leq 1.$$ (4)

This constraint can be obtained in another way. If $t_{\text{max}}$ and $t_{\text{min}}$ are the extremes of $t_a(x_s)$ over an arbitrary surface patch $\{x_s \in \Gamma\}$, the local contrast over $\Gamma$ is $C_{\Gamma} = (t_{\text{max}} - t_{\text{min}})/(t_{\text{max}} + t_{\text{min}})$, which must satisfy $0 < C_{\Gamma} < 1$. From (3),

$$C_{\Gamma} \leq \sum_{n=1}^{N} \alpha_n(x_s),$$

which again yields the restriction (4).

The notion of a surface coordinate system on which an albedo function is mapped is a subtle concept that can only be loosely defined. In (2), it is assumed that the surface(s) being imaged have a smooth geometry at the scale of observation, allowing for a smooth interpretation of the modulating functions. For example, if a field of grass on a flat plot of earth is imaged from a distance, the surface coordinate system may be regarded as flat, with the fine scale surface texture (the blades of grass and their shadows) being contained in the modulating functions. Of course, the local coordinate system $x_s$ will not necessarily be flat. However, the surface-to-image projection model that we will use assumes a local planarity on the surface, in the sense that a local coordinate system is defined on the surface tangent plane at every surface point. Since both the albedo model and the processing are highly local, this is not very restrictive.

B. Relating Surface Coordinates to World Coordinates

Any surface lies in the three-dimensional (3-D) space of world coordinates $x_w = (x_w, y_w, z_w)$. We will need to develop a relationship between world coordinates and local surface coordinates. The $z_w$-axis is usually defined to be orthogonal to the camera image planes, unless the cameras verge, in which case a nominal assumption is made. A visible surface patch can be expressed

$$F(x_w) = 0$$

with surface normal

$$\hat{n}(x_w) = \frac{\nabla F(x_w)}{|\nabla F(x_w)|}.$$ (6)

Assuming that $F$ is differentiable at a point $x_w$, the surface orientation at $x_w$ is described by the local slant $\sigma$ (the angle between the normal and the $z_w$-axis) and the tilt $\tau$ (the angle between the $x_w$-axis and the projection of the surface normal onto the $(x_w, y_w)$ plane).

The local surface coordinate system is defined similar to that used in the shape-from-texture work of Super and Bovik [28], [33]. At each surface point, the local coordinate system $x_s$ is defined so that the unit normal vectors in the $x_s$ and $y_s$ directions lie in the surface tangent plane, and are related to slant and tilt according to [33]:

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} \cos \sigma \cos \tau, & \cos \sigma \sin \tau, & -\sin \sigma \\ -\sin \sigma, & \cos \tau, & 0 \end{bmatrix}.$$ (7)

C. Relating Surface Coordinates to Binocular Image Coordinates

We take the origins of the image coordinate systems $x_L = (x_L, y_L)$ and $x_R = (x_R, y_R)$ (Left and Right, respectively) to be at world coordinates $(-D, 0, f)$, $(D, 0, -f)$, where $f$ is the camera focal length(s) and 2D is the baseline separation between the lens centers of the cameras.

As a demonstration of principle, we assume that the camera optical axes are parallel, in which case $(x_L, y_L) = (x_w + D, y_w)$ and $(x_R, y_R) = (x_w - D, y_w)$. The general ideas here can be applied to nonconvergent systems, but with complications in the geometric development [34].

We adopt the standard upright pinhole or perspective projection model for image formation [35], whence it can be shown
that for $A = L$ (Left) and $A = R$ (Right) [33], [36]

$$x_A = K(T x_s + kd).$$

Fig. 1. Binocular camera geometry.

In (8), $K = f / z_{aw}$, $d = [D \ 0]^T$, and $k = +1$ when $A = L$ and $k = -1$ when $A = R$. The matrix (9)

$$T = \begin{bmatrix}
\cos \sigma \cos \tau & -\sin \tau \\
\cos \sigma \sin \tau & \cos \tau
\end{bmatrix}$$

is a composition of the transformation from the surface-to-world coordinates and projection onto the $(x_w, y_w)$ plane. Fig. 1 depicts all of the relevant quantities.

The surface coordinates can be expressed in terms of image coordinates by inverting (8):

$$x_s = T^{-1}[(1/K)x_A - kd]$$

where

$$T^{-1} = \begin{bmatrix}
\cos \tau & -\sin \tau \\
\cos \sigma \sin \tau & \cos \sigma \cos \tau
\end{bmatrix}.$$ (11)

Given a surface coordinate $x_s$, and assuming the cameras are accurately positioned, its projections $x_L$ and $x_R$ are related by

$$x_R = x_L - 2fD / z_{aw}$$

and $y_R = y_L$. The epipolar line for such a nonconvergent geometry is simply the horizontal, along which a search may be made for matches. Once a match is determined, standard triangulation solves $z_{aw}$ from the known quantities $f$, $D$, and the measured disparity $\Delta x = (x_L - x_R)$.

D. Mapping Surface Reflectivity to Image Intensity

The patterns that lie on a surface are deformed by the local surface topography (variation in slant and tilt) and by perspective projection. From (2), the irradiance of sensed (left or right) image $s(x)$ may be expressed (x is meant to represent either $x_L$ or $x_R$)

$$s(x) = k_l(x) \cdot k_r(x) \cdot \left\{1 + \sum_{n=1}^{N} B_n(x) \cos[\Phi_n(x)]\right\}$$ (13)

where the AM–FM functions in (13) are related to those at the surface by the transformations $B_n(x) = \alpha_n [x_s(x)]$, $\Phi_n(x) = \omega_n [x_s(x)]$. In (13), $k_l(x)$ is an illumination function, and $k_r(x)$ is the reflectance function that depends on the local surface orientation [36]. Assume that $k_l(x)$, $k_r(x)$ vary smoothly with $x$, which will simplify later developments. This is reasonable in most instances, although not at gross intensity discontinuities.

E. Image Preprocessing

Since local image phase will be used to compute disparities, a mechanism for separating the modulated albedo functions from illumination and reflectance effects is used. A simple device for doing this is a logarithmic point operation [17]

$$\hat{x}(x) = \log[s(x)]$$

$$= \log[k_l(x) \cdot k_r(x)] + \log\left\{1 + \sum_{n=1}^{N} B_n(x) \cos[\Phi_n(x)]\right\}$$

$$\approx \log[k_l(x) \cdot k_r(x)] + \left\{\sum_{n=1}^{N} B_n(x) \cos[\Phi_n(x)]\right\}$$ (14)

using the approximation (or nonlinear transformation) $\log(1 + x) \approx x$ for $|x| < 1$, which is reasonable except for images with very high local contrasts, which are infrequent in nature.

The separation in (14) is done since $k_l(x) \cdot k_r(x)$ contains low-frequency information that is not well captured by the AM–FM model. Separate stereo processing of $k_l(x) \cdot k_r(x)$ could yield additional three-dimensional (3-D) data; however,
the assumption of smoothness suggests that they are poor matching primitives.

In the next section, a method for isolating the phase responses is described, which involves passing the preprocessed image \(t(x)\) through a bank of narrowband frequency- and orientation-sensitive spatial filters \(h_{\sigma}(x;\omega_c)\) indexed by scale \(\sigma\) and center frequency vector \(\omega_c\). The filters are selected so that, among other things, the response near DC is small, hence assume

\[
h_{\sigma}(x;\omega_c) \times \log [i(x)h_{\sigma}(x)] \approx 0. \tag{15}
\]

Since the filters \(h_{\sigma}(x;\omega_c)\) are effectively one-sided in their frequency responses, it is true that

\[
h_{\sigma}(x;\omega_c) \times \sum_{n=1}^{N} B_n(x) \cos [\Phi_n(x)] \approx (1/2)h_{\sigma}(x;\omega_c) \times \sum_{n=1}^{N} A_n(x) \exp [j\Phi_n(x)] \tag{16}
\]

where \(A_n(x)\) is the analytic image associated with each term \(B_n(x) \cos [\Phi_n(x)]\), where \(H\{\cdot\}\) is a 2-D Hilbert transform defined in the positive horizontal direction [30], [51], [52]. The approximation (16) is the same as assuming that the Hilbert transform of the image was taken first, and then the filters applied; equality would hold exactly if the filters were strictly half-plane. The factor \((1/2)\) appears since the right-half frequency plane coefficients are not doubled prior to filtering; for convenience, this factor will be ignored (absorbed) in the sequel. The Hilbert transform interpretation removes any ambiguity in the definition of image phase, and in the definition of instantaneous frequency. In-depth discussions of the 2-D Hilbert transform and instantaneous frequencies are available in [30], [51], and [52]. With this, we may use the image model (1), which will be used throughout the rest of the paper as a convenient representation of the essential modulated structure of the input stereoscopic images.

### III. DISPARITY FROM FILTERED RESPONSES

The model (1) contains an unspecified number \(N\) of AM–FM components. Indeed, for any \(N \in \mathbb{N}\), there are uncountably infinite possible selections of the AM–FM functions \(\alpha_n(x)\) and \(\omega_n(x)\) so that (2) represents the surface pattern exactly. Small values of \(N\) can usually be found such that (1) captures the image efficiently. The most interesting interpretations arise from components with large average amplitudes and smooth phase gradients. These are more likely to have arisen from meaningful physical processes, since they are “emergent” in the image, and they are easier to extract, particularly in the presence of noise [17], [31].

The degree of variation (deviation from linearity) of each component phase \(\Omega_n(x)\) may change substantially over the image, so the instantaneous frequency vectors \(\nabla \Omega_n(x)\) may traverse much of the frequency plane. Variations in the local surface frequencies \(\nabla \alpha_n(x)\), the surface topography, and the projection (14) all contribute to changes in \(\nabla \Omega_n(x)\). Therefore, it is necessary to introduce a device to extract local image modulations. Since the distribution of modulation energy will vary with the local frequencies, an effective way to capture and identify locally significant components is by a bank of spatially localized image filters that cover the frequency plane. Since the image may contain multicomponents \((N \geq 1)\), the filters need to be spectrally localized to avoid cross-component interference. These familiar joint constraints are often used to support the choice of low-uncertainty filters such as the optimal Gabor functions [17], [24], [31], [37], [41], compactly supported wavelets [38], prolate functions [39], etc.

#### A. Definition of Filterbank

We will use multiple 2-D Gabor functions \((RS)\) filters doubly-indexed by \(1 \leq r \leq R\) and \(1 \leq s \leq S\) as the optimizing minimum-uncertainty choice. They have the general form

\[
h_{rs}(x) = g_{rs}(x) \exp (2\pi i \omega_{rs} x) \tag{17}
\]

where \(\omega_{rs} = (u_{rs}, v_{rs})\) is the modulation center frequency and the \(g_{rs}(x)\) are 2-D Gaussians

\[
g_{rs}(x) = \frac{1}{2\pi \sigma_r^2} \exp \left[ -\frac{(x^2 + y^2)}{2\sigma_r^2} \right] \tag{18}
\]

with scale parameter \(\sigma_r\) (fixed by \(r\)) (octaves) and (uniform) aspect ratio \(\lambda\). The frequency response is

\[
H_{rs}(\omega) = G_{rs}(\omega - \omega_{rs}) = \exp [-2(\pi \sigma_r)^2((u - u_{rs})^2 + (v - v_{rs})^2)] \tag{19}
\]

which is bandpass with aspect ratio \(1/\lambda\). The Gabor filters are parametrized by radial center frequencies \(F_r = (u_{rs}^2 + v_{rs}^2)^{1/2}\), by filter orientations \(\theta_r = \tan^{-1}(v_{rs}/u_{rs})\), and by half-peak radial frequency bandwidth \(B_r = \log_2[(\pi F_r \lambda \sigma_r + \alpha)/(\pi F_r \lambda \sigma_r - \alpha)]\) (octaves), where \(\alpha = [(\ln 2)^2/4]^{1/2}\).

Fig. 2 depicts a sampling of the spatial frequency domain by 12 Gabor filters (four scales in three orientations). The parameters used are \(B_r = B = 1\) octave over all scales, orientations \(\theta_r = -30^\circ, 0^\circ,\) and \(+30^\circ\), and aspect ratio \(\lambda = 2\). In the experiments given later, this filter sampling (with \(\lambda = 2\)) is used to capture horizontal local frequencies with a higher resolution than vertically oriented frequencies, since disparities will only be computed along horizontal scanlines.

The Gabor filters form a set of spatial frequency channels, or \(s\)-channels, through which images are decomposed. These are to be distinguished from disparity channels, or \(d\)-channels, which will be tuned to specific stereo disparities rather than specific spatial frequencies.

#### B. Computing Local Image Phase from a Single Component

Obtaining a closed-form expression for the response of a linear \(s\)-channel to an arbitrary AM–FM function is generally difficult or impossible. However, under certain assumptions, the response of the \(s\)-channel filter indexed \((r,s)\) to component \(n\) of (1)

\[
o_{rs}(x) = h_{rs}(x) \times A_n(x) \exp [j\Omega_n(x)] \tag{20}
\]
can be closely approximated by

\[ o_{n,r,s}(x) \approx A_n(x) \exp[j \Omega_n(x)] H_{r,s}[\nabla \Omega_n(x)]. \]  

(21)

The approximation (21), termed quasieigenfunction approximation (QEA), has been utilized in numerous studies involving filtering of AM–FM signals and images [17], [22], [29], [31]. The absolute error in (21) relative to (20) is tightly bounded by functionals expressing both the smoothness of \( A_n(x) \) and \( \nabla \Omega_n(x) \) (Sobolev norms) and the spatial durations (energy variances) of the \( s \)-channel filters \( H_{r,s}(x) \). The theory behind these approximations has been explored in detail elsewhere [17], [22], [29], [31]. In the current context, the theory applies well since the \( s \)-channel filters are low-uncertainty and the modulating functions are assumed smooth.

Using (21), the output amplitude responses \( M_{n,r,s} : \mathbb{R}^2 \rightarrow \mathbb{R}^+ \) can be approximated by

\[ M_{n,r,s}(x) = |o_{n,r,s}(x)| \approx A_n(x) H_{r,s}[\nabla \Omega_n(x)]. \]  

(22)

while the phase responses \( \psi_{n,r,s} : \mathbb{R}^2 \rightarrow [-\pi, \pi] \) have the simple approximation

\[ \psi_{n,r,s}(x) = \omega_{n,r,s}(x) \approx \Omega_n(x). \]  

(23)

Direct computation of \( \psi_{n,r,s} \) using (23) yields a wrapped phase with distances between phase discontinuities depending on the spatial distribution of instantaneous frequencies. When the instantaneous frequencies are smooth, the phase wrapping period can be locally approximated by

\[ \Delta \psi_{n,r,s}(x) \approx \frac{2\pi}{|\nabla \Omega_n(x)|}. \]  

(24)

In (23), the local phase of component \( n \) is extracted independent of the \( s \)-channel being used. In practice, QEA approximations hold best for \( s \)-channels having (relatively) strong responses, since the AM–FM signal energy is strong in those channels. An \( s \)-channel with a weak response at some coordinate \( x \) likely is not responding strongly to an AM–FM component, hence its response may be composed primarily of multiple (out-of-band) components or by noise. This strategy is used later when selecting \( s \)-channels to be used in the disparity computations.

C. Disparity from Phase Difference

Now we determine disparities from \( s \)-channel responses. The aggregate response of \( s \)-channel \( m \) to the left image is

\[ O_{L,m,s}(x_L) = h_{r,s}(x_L) * \sum_{n=1}^{N} A_{L,n}(x_L) \cos[\Omega_{L,n}(x_L)] \]  

(25)

where \( A_{L,n}(x) \), \( \Omega_{L,n}(x) \) are defined analogous to (16). By the QEA approximation (21), we write

\[ O_{L,m,s}(x_L) \approx \sum_{n=1}^{N} A_{L,n}(x_L) \exp[j \Omega_{L,n}(x_L)] H_{r,s}[\nabla \Omega_{L,n}(x_L)]. \]  

(26)

and likewise the approximation for the right image:

\[ O_{R,m,s}(x_R) \approx \sum_{n=1}^{N} A_{R,n}(x_R) \exp[j \Omega_{R,n}(x_R)] H_{r,s}[\nabla \Omega_{R,n}(x_R)]. \]  

(27)

We assume that the response to each \( s \)-channel \( h_{r,s}(x) \) is dominated by at most one AM–FM component, indexed \( n(r,s) \). Then, (26) and (27) become

\[ O_{L,n(r,s),m}(x_L) \approx A_{L,n(r,s)}(x_L) \exp[j \Omega_{L,n(r,s)}(x_L)] \times H_{r,s}[\nabla \Omega_{L,n(r,s)}(x_L)]. \]  

(28)

\[ O_{R,n(r,s),m}(x_R) \approx A_{R,n(r,s)}(x_R) \exp[j \Omega_{R,n(r,s)}(x_R)] \times H_{r,s}[\nabla \Omega_{R,n(r,s)}(x_R)]. \]  

(29)

This assumption is not very restrictive, since we are not interested in separating and identifying the multiple components. If multiple components fall within the passband at some coordinate, they may be merged into a single component. Applying (22) and (23) to the merged component yields a larger AM estimate but there will be little effect on the FM estimate.

Two merged components \( A_1 \exp[j \Omega_1] \) and \( A_2 \exp[j(\Omega_1 + \Delta \Omega)] \) (dropping dependence on \( x_L, x_R \)), have approximate magnitude \( |A_1 + A_2| \) with error less than \( 2\sqrt{A_1 A_2} \sin \frac{\Delta \Omega}{2} \), and approximate phase \( \Omega_1 \) with error less than \( |\tan^{-1}(\frac{\sin(\Delta \Omega)}{A_1/A_2 + \cos(\Delta \Omega)})| \), which we may regard as negligible when \( \Delta \Omega \) is small, viz., both components lie within the same passband. Hence, we may regard multiple components within a passband as approximated by a single component, which may be used to compute an aggregate phase disparity between channels. Merged AM–FM components typically represent instances where two (or more) otherwise globally distinct AM–FM components become indistinct at some spatial location.

There are three simple modulation-based functions which could be used to compute disparity: the AM estimate, the FM estimate, and the instantaneous frequencies. We will examine each separately. From (22) and (23), the magnitude and phase
responses of s-channel indexed \((r,s)\) (applied to the left image), can be computed with approximate values

\[
M_{L_{r,s}}(x_L) \approx A_{L_{r,s}}(x_L)H_{r,s}[\nabla \psi_{L_{r,s}}(x_L)] \tag{30}
\]

\[
\psi_{L_{r,s}}(x) \approx \Omega_{L_{r,s}}(x_L). \tag{31}
\]

1) Stereo from AM Matching: The AM function is estimated by dividing \(M_{L_{r,s}}\) by \(H_{r,s}\) evaluated at \(\nabla \psi_{L_{r,s}}\), as follows:

\[
A_{L_{r,s}}(x_L) \approx \frac{M_{L_{r,s}}(x_L)}{H_{r,s}(x_L)} \tag{32}
\]

The goal of stereo disambiguation is then to find the disparity function \(\Delta x(x_L) = (\Delta x(x_L), 0)\) such that the left and right AM functions match:

\[
A_{L_{r,s}}(x_L) - A_{R_{r,s}}(x_L) = 0 \tag{33}
\]

Of course, it is not generally possible to find \(\Delta x(x_L)\) such that (33) is everywhere satisfied; however, the difference should be small. Unfortunately, the difference might be small for many definitions of \(\Delta x(x_L)\) at most locations \(x_L\), since the amplitude \(A_{L_{r,s}}(x_L)\) often will vary only slightly over substantial image regions. For example, in the case of a fronto-parallel plane with a periodic texture overlaid on it, the extracted AM-FM components will be nearly sinusoidal, with constant amplitudes. At best, solving (33) should be regarded as an ancillary constraint. In fact, this is done later here. However, it should be observed that it is possible that the amplitudes may correlate imperfectly with the phases due to various surface distortions, an observation that was also made by Fleet and Jepson [43].

2) Stereo from Instantaneous Frequency Matching: For the same reason, the instantaneous frequencies are also ambiguous. Solving

\[
\nabla \Omega_{L_{r,s}}(x_L) - \nabla \Omega_{R_{r,s}}(x_L) = 0 \tag{34}
\]

will also fail over regions that are locally homogeneous in frequency. As in (33), severe, unresolvable ambiguities will arise if \(\nabla \psi_{L}\) is used as the primary disambiguation primitive.

3) Stereo from Phase Matching: The s-channels phases, which are also the FM functions, provide ample information for disambiguation. Attempting to solve

\[
\Omega_{L_{r,s}}(x_L) - \Omega_{R_{r,s}}(x_L) = 0 \tag{35}
\]

over the disparity function is a much better-posed problem, since unresolvable ambiguities will not often arise. Exceptional instances occur where there isn’t any intensity variation over a region, so the dominant AM-FM components are near-constant with zero phases. We are not aware of any approach to stereo disambiguation that claims to resolve this situation, other than through some kind of interpolation.

Thus the local s-channel phase responses, or FM functions, have the best potential for computing disparities, with local measurements of AM or instantaneous frequencies being useful as secondary pieces of information. Therefore, we may regard solving (35) as the fundamental problem to be addressed in the remainder of the paper.

One approach that is simple enough is to observe that by Taylor’s theorem

\[
\Omega_{R_{r,s}}(x_L) = \Omega_{L_{r,s}}(x_L) - \Delta x(x_L)^T \nabla \Omega_{R_{r,s}}(x_L) + R_1[\Delta x(x_L), \Delta x(x_L)] \tag{36}
\]

where \(R_1\) is the first-order Taylor remainder [40]. If the phase function is assumed to vary smoothly, then a simple equation for the disparity can be obtained from a single s-channel [16]:

\[
\Omega_{L_{r,s}}(x_L) - \Omega_{R_{r,s}}(x_L) \approx \Delta x(x_L)^T \nabla \Omega_{R_{r,s}}(x_L). \tag{37}
\]

Of course, tracking \(\nabla \Omega_{R_{r,s}}(x_L)\) to solve (37) would require switching s-channels over space.

However, (36) is ineffective since (37) fails wherever the instantaneous frequencies change tangibly. Reasonable results might be had only from observations at a very small scale, viz., using filters having a very small effective spatial support. However, image modulations occur over a wide range of scales; indeed, the model is intrinsically and naturally multiscale. For this reason we next propose and motivate a multiscale, coarse-to-fine mechanism for disambiguating stereo disparities from measurements of the s-channel phases.

IV. A DISPARITY CHANNEL MODEL

Here, we explore a natural, multiscale method for solving the stereo problem (35) using the s-channel filter phase responses. The development uses the concept of disparity channel, or d-channel, which enables a search-free correspondence mechanism for extracting disparities and which melds seamlessly with the s-channel phase-based disparity model. The approach extends prior methods of multiscale disparity calculation, e.g., [3], [6], [14]–[16], [42], [44].

A. Disparity Channels

For each s-channel \(h_{r,s}(x)\) with radial frequency \(F_r\), and orientation \(\theta_s\), define a finite set of d-channels \(Ch_{k_{r,s}}\), indexed \(0 \leq k \leq K_{r,s} - 1\), by computing threshold responses from the s-channels (with notations defined below):

\[
C_k(x_L, \Delta d_{k_r}, \theta_s) = \begin{cases} 
P_{r,s}(x_L), & \text{if } \Delta \psi(x_L, k\Delta d_{r,s}) < 0.5\Delta d_{r,s}, \tag{38} \\
0, & \text{otherwise}.
\end{cases}
\]

The maximum of these threshold responses is used to select the particular disparity channel that is used at each coordinate, as described below. In (38), \(\Delta d_{r,s}\) is a predefined disparity.

Fig. 3. Model of the disparity channel mechanism.
channel width (pixels), fixed for each s-channel, but varying with \( r \) and \( s \) (Section IV-D). In (38), the s-channel phase difference relative to disparity \( k \Delta d_{rs} \) is

\[
\Delta \psi(x_L, k \Delta d_{rs}) = |\psi_{L, r, s}(x_L) - \psi_{R, r, s}(x_L - k \Delta d_{rs})|
\]

\[
\approx |\Omega_{L, n(r, s)}(x_L) - \Omega_{R, n(r, s)}(x_L - k \Delta d_{rs})|
\]  

(39)

and where \( P_{rs}(x_L), 0 \leq P_{rs}(x_L) \leq 1 \), is the d-channel intensity, given by

\[
P_{rs}(x_L) = \frac{\max[M_{L, r, s}(x_L), M_{R, r, s}(x_L - k \Delta d_{rs})]}{\min[M_{L, r, s}(x_L), M_{R, r, s}(x_L - k \Delta d_{rs})]}
\]

\[
= 1 - \frac{[M_{L, r, s}(x_L) - M_{R, r, s}(x_L - k \Delta d_{rs})]}{[M_{L, r, s}(x_L) + M_{R, r, s}(x_L - k \Delta d_{rs})]}
\]

(40)

\( P_{rs}(x_L) \) is identical to a confidence measure used in [14] that is used to define a weighted disparity measurement across scales. Here, it is used to define the disparity channels. If the phase difference \( \Delta \psi(x_L, k \Delta d_{rs}) \) is small, then

\[
P_{rs}(x_L) \approx \frac{\min[A_{L, n(r, s)}(x_L), A_{R, n(r, s)}(x_L - k \Delta d_{rs})]}{\max[A_{L, n(r, s)}(x_L), A_{R, n(r, s)}(x_L - k \Delta d_{rs})]}
\]

(41)

which is maximized when the shifted amplitude modulation functions are equal. Care is needed when \( \Omega_{L, n(r, s)}, \Omega_{R, n(r, s)} \) fall in the range \( \pm \pi \), e.g., by performing arithmetic modulo \( \pm \pi \).

Thus, (38) functions as a phasor matching mechanism across disparity space. Matching between s-channel responses is triggered by a small phase difference at disparity shift \( k \Delta d_{rs} \), and is subsequently graded by the similarity in the shifted s-channel magnitude responses. As shown in Fig. 3, the outputs of the d-channel (38) are the disparity \( k \Delta d_{rs} \) and the threshold response \( C_k \).

B. Computing Candidate Disparity from a Single S-Channel

Each s-channel \( h_{rs}(x) \) has an associated population of d-channels \( C_{k, r, s} \) indexed \( k = 0, \ldots, K_{r, s} - 1 \), each supplying a candidate disparity \( k \Delta d_{rs} \). From these, a single candidate disparity

\[
d_{rs}(x_L) = \hat{k} \Delta d_{rs}
\]

(42)

is selected according to

\[
\hat{k} = \arg \max[C_k(x_L, \Delta d_{rs}; F_r, \theta_s); k = 1, \ldots, K - 1]
\]

(43)

This strategy yields the disparity giving the largest threshold response (38). Alternately, one could form a disparity estimate weighted across channels (e.g., by \( C_k \)), as in [14]. However, low threshold channels have greater likelihood of contributing errors from between-image distortions, hence a weighted strategy is more likely to contribute erroneous data. The process used here is depicted in Fig. 4, where the maximum channel intensity is used as the control signal to a multiplexer (Mux) which transfers the disparity output from the maximizing disparity channel.

C. Computing Candidate Disparity Across a Set of Oriented S-Channels

Although different s-channel filter tessellations of the frequency plane are possible, we choose to parse the plane by radial frequency and orientation, as depicted in Fig. 2. Filters are aligned along contours of equal radial frequency and of equal orientation. Separations between s-channels along frequency/orientation axes are closely tied to the width of the d-channels that should be used, as discussed later. This tessellation allows for a coarse-to-fine processing schedule that efficiently directs the computation of disparity to successively higher resolutions.
At a fixed scale, the best possible disparity is selected from those computed at different orientations. Thus, the disparity responses from multiple $s$-channels sharing a fixed radial frequency $F_r$, but across all orientations $\theta_s$; $s = 1, \ldots, S$ are compared. The simplest way of doing this is by taking the disparity computed from the most active $s$-channel, i.e., having the largest $s$-channel magnitude response to the left or right images, or possibly the product of responses. For simplicity, the left magnitude response is used here.\footnote{An effective way of verifying correspondences would then be to match right image to left, using the right image magnitude responses. Inconsistent correspondences, possibly arising from occlusion, can then be discarded.} Hence, from (42), for $r$ fixed:

$$d_r(x_L) = d_{rs_{\max}}(x_L)$$  

(44)

where

$$s_{\max} = \arg\max[M_{L,\theta_s}(x_L); s = 1, \ldots, S].$$  

(45)

In this way, disparities computed from local frequencies/phases that are out-of-band are rejected. The overall process is depicted in Fig. 5.

**D. Relating $D$-Channels to $S$-Channels**

The AM–FM model (1) is naturally multiscale, but it is not quantum. In fact, it is smooth across the continuum of scales—local image frequencies/phases may vary from “low” to “high” and smoothly over orientations. Passing an image through $s$-channels quantizes local frequencies into orientation and scale bins. This is convenient, since disparity computations can be segregated by local orientation and scale. Likewise, defining discrete $d$-channels makes it possible to localize the computations according to disparity scale. Since disparity has a direct spatial interpretation, it is natural to control the action of the $d$-channels by the $s$-channel responses.

1) **Disparity Channel Sensitivity:**

A natural and essential strategy of a coarse-to-fine disparity computation that measures local phase across a range of frequency scales is that finer disparity measurements be made from higher frequency measurements. One reason for this is that $s$-channel phase responses fall within a dynamic range $\pm \pi$, and so there is an intrinsic periodic ambiguity introduced by the wrapped phase. As indicated by (24), the ambiguity occurs at finer scales with increased local frequency.

Therefore, the $d$-channel widths $\Delta d_{rs}$ should decrease with (be inversely proportional to) the center frequency of the associated $s$-channel filter. The question arises whether $\Delta d_{rs}$ should increase with $F_r$, or with $\theta_s$. While both strategies
are interesting, stereoscopic disparity (phase) computation in a Left-Right camera system is primarily sensitive to horizontally oriented phase variations. Then define the \textit{disparity channel sensitivity} by $\delta_{rs} = \Delta d_{rs} u_{rs}$. Forcing the $d$-channel sensitivity to be constant: $\delta_{rs} = \delta$ across all scales and orientations, thus requires the $d$-channel widths to fall off inversely with the horizontal $s$-channel center frequencies:

$$\Delta d_{rs} = \delta / u_{rs}.$$  

\textit{2) Channel Shutdown Mechanism:}

Another important consideration is that the $s$-channel phase responses be used only when the instantaneous frequencies fall in-band, otherwise the signal-to-noise ratio (SNR) will suffer. This may be assured by setting the output of the $d$-channel to zero whenever it is not true that

$$||\nabla \psi_{L,r,s}(x_L) - \omega_{rs}|| \leq ||\nabla \psi_{R,r,s}(x_L - k\Delta d_{rs}) - \omega_{rs}||$$  

where $||\cdot||$ is Euclidean distance and $\alpha / \pi \sigma_r \lambda$ is the distance to $s$-channel half-peak. Thus, if at coordinate $x_L$ there is no AM–FM component that falls within the passband of a particular $s$-channel, then that channel is excluded from disparity computation at $x_L$. 

Fig. 7. Stereo pair of turf and disparity map.

Fig. 8. Stereo pair of tree trunk and disparity map.

Fig. 9. Stereo pair of baseball on newspaper and disparity map.
3) **S-Channel Sampling Tessellation**: In a coarse-to-fine processing framework that uses local $s$-channel phase responses, it is natural that the channels be arranged in a dyadic, wavelet-like tesselation with doubling of linear bandwidth and channel centers at higher frequencies. However, we propose a twist: instead, at a fixed orientation the $s$-channel horizontal center frequencies are doubled from coarse-to-fine scales. Thus, for $s$ fixed, the horizontal center frequencies relate as follows across scales:

$$\frac{u_{r+1,s}}{\Delta u_{r,s}} = 2^B$$ \hspace{1cm} (48)

where the octave bandwidth $B = 1$ is assumed here. Therefore, the $d$-channel widths satisfy

$$\frac{\Delta d_{r+1,s}}{\Delta d_{r,s}} = 2^{-B} = 1/2.$$ \hspace{1cm} (49)

Since the number of disparity channels is finite, a maximum disparity $d_{\text{max}}$ may be defined, making it possible to determine the number of $d$-channels $k \Delta d_{r,s}$; $k = 1, \ldots, K_{r,s} - 1$ associated with each $s$-channel. This can be set to the image width, or determined in a simple way by assuming a closest possible object distance to the camera baseline. In any case,
it follows that
\[(K_{r,s} - 1) \Delta d_{r,s} = d_{\text{max}} \] (50)
since the \( k = 0 \) channel has zero width (unit width in discrete implementation).

Fig. 6 depicts a layered representation of the coarse-to-fine disparity structure that will be used in a coarse-to-fine disparity computation procedure that is described next.

E. Coarse-to-Fine Processing

The coarse-to-fine phase correspondence process proceeds from the largest scale \((r = 1)\) to the finest \((r = R)\), in a manner not unlike that proposed by Langley et al. [16]. For \( 1 \leq r \leq R \), compute the disparity map \( \Delta x_r(x_L) \) at scale \( r \) according to
\[
\Delta x_r(x_L) = \Delta x_{r-1}(x_L) + d_r(x_L) 
\]
where \( d_r(x_L) \) is computed across orientations at iteration (scale) \( r \) as in (44), and where the initial disparity map \( \Delta x_0(x_L) = 0 \) is assumed. Because of the channel shutdown mechanism, it is possible that \( d_r(x_L) \) will not have an assigned value. In this case, one can simply set \( d_r(x_L) = 0 \).

V. Simulation

We now demonstrate a phase-based stereo algorithm defined by the developments in Section IV and earlier. The Gabor filterbank used is depicted in Fig. 2; 12 s-channels distributed over four scales and along three orientations. The s-channel bandwidth is one octave at every scale. The s-channels were designed with aspect ratio \( \lambda = 2 \); thus, the channels have twice the spectral resolution along the horizontal (disparity) direction. In each example, no effort was made to correct or obscure meaningless results obtained at the extreme left and right edges of each disparity map, where there is no possible correspondence. In each display, disparity is coded as intensity, with nearer objects appearing brighter. The images used are 256 \( \times \) 256 pixels.

The first example (Fig. 7) depicts a slightly mounded area of turf receding into the distance. The computed depth map successfully captures both the progression in depth as well as the rounding of the surface. Fig. 8 depicts a gnarled tree trunk next to a creek. The algorithm was able to capture most of the depth variation, including detail in the smaller root on the right. The disparity map obtained from a stereo pair of a baseball (Fig. 9) appears at first to suffer discrepancies near the ball's contour. However, close inspection reveals that the eye is fooled at these same locations, since there are patterns that are difficult to disambiguate. The next example (Fig. 10) is a stereo pair of a doorway. The algorithm does an excellent job at capturing abrupt changes in disparity, and even succeeds in computing depth from the reflected background. Fig. 11 shows a large rock having considerable surface variation, most of which is successfully captured and represented in the computed disparity map. Finally, Fig. 12 shows the result of applying the algorithm to the well-known stereo pair of the Pentagon in Washington, DC. The result obtained is quite comparable to the results obtained by computationally expensive algorithms that rely on simulated annealing-type optimization to disambiguate.

VI. Conclusion

We applied a powerful AM–FM surface albedo model to analyze the projection of surface patterns viewed through a binocular camera system. This led to the development of a modulation-based stereo matching framework where local image phase was used to compute stereo disparities. A coarse-to-fine matching strategy was utilized based on a new concept of disparity channels which melds naturally with the natural multiscale AM–FM modeling paradigm and Gabor filterbank decomposition. It is envisioned that the combined concepts of phase-based matching and AM–FM image modeling will lead to further improvements in the future. We expect that one valuable direction of research would be the development of more specific quantitative models of modulated albedo profiles on different surfaces.

REFERENCES

CHEN et al.: STEREOSCOPIC RANGING 797


Tieh-Yuh (Terry) Chen (M’95) was born in Pen-Hu, Taiwan. R.O.C. on Feb. 17, 1958. He received the B.S. and M.S. degrees in electrical engineering from National Cheng-Kung University, Tainan, Taiwan, in June 1980, and June 1982, respectively. In 1995, he received the Ph.D. degree from the University of Texas at Austin, Austin, TX.

From 1992 to 1990, he was an Assistant Scientist in the Chung-Shan Institute of Science and Technology (CSIST), Taiwan. He is currently a Senior Research Engineer at CSIST.

Dr. Bovik has been involved in numerous professional society activities, including, currently, Board of Governors, IEEE Signal Processing Society, Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, and Editorial Board, PROCEEDINGS OF THE IEEE. He was the Founding General Chairman of the First IEEE International Conference on Image Processing, Austin, TX, November, 1994. He is a recipient of the IEEE Signal Processing Society Meritorious Service Award (1998), a winner of the University of Texas Engineering Foundation Halliburton Award, and a two-time Honorable Mention winner of the International Pattern Recognition Society Award for Outstanding Contribution (1988 and 1993).