Image Matching Using the Windowed Fourier Phase

JOHN (JUYANG) WENG
Department of Computer Science, Michigan State University, East Lansing, MI 48824-1027

Received January 25, 1991; Revised May 26, 1992 and April 8, 1993.

Abstract
A theoretical framework is presented in which windowed Fourier phase (WFP) is introduced as the primary matching primitive. Zero-crossings and peaks correspond to special values of the phase. The WFP is quasi-linear and dense; and its spatial period and slope are controlled by the scale. This framework has the following important characteristics: 1) matching primitives are available almost everywhere to convey dense disparity information in every channel, either coarse or fine; 2) the false-target problem is significantly mitigated; 3) the matching is easier, uniform, and can be performed by a network suitable for parallel computer architecture; 4) the matching is fast since very few iterations are needed. In fact, the WFP is so informative that the original signal can be uniquely determined up to a multiplicative constant by the WFP in any channel. The use of phase as matching primitive is also supported by some existing psychophysical and neurophysiological studies. An implementation of the proposed theory has shown good results from synthesized and natural images.

1 Introduction
General image matching is closely related to stereo matching. The former involves images that are different due to interframe displacement of the sensors or the motion of the objects in the scene. The latter concerns two images that are taken by two sensors, concurrently, with a known configuration. The objective of image matching, general or stereo, is to establish correspondence between the selected primitives so that the matched primitives arise from the same element in the scene. The basic difference between these two types of matching is that for stereo matching one can use the epipolar constraint, while for general matching such a constraint is not applicable. Although this difference makes general matching more difficult, both types of matching have to deal with many common problems.

One key question related to image matching is the selection of matching primitives. Differences in the selection may lead to quite different matching strategies and performances.

The random dot stereograms introduced by Julesz [22] have demonstrated that binocular combination need not happen after high-level monocular recognition. On the other hand, Julesz's experiments [22, 23] have shown that human stereopsis survives despite the remarkably different values of intensity, or the different micropatterns of intensity, between left and right eye images. Therefore, as far as human stereopsis is concerned, disparity cannot be mediated solely by matching intensity values themselves or their complex micropatterns.

Marr and Poggio [28] proposed a theory of stereoscopic matching specifically motivated by human psychophysics. The theory uses zero-crossings (short for zero-crossings of the Laplacian of Gaussian, unless stated otherwise) at different spatial resolutions as matching primitives and alleviates the false-target problem by trading off resolution against disparity range in a coarse-to-fine strategy. Marr and Poggio's stereoscopic matching theory is the first model of human stereopsis that incorporates coarse-to-fine strategy, although coarse-to-fine matching was independently used earlier by Moravec [30]. The use of zero-crossings as matching primitives constitutes a basic component of their theory, and it has been tested on random-dot stereograms as well as on natural images by Grimson [13]. Hildreth [15] extended this zero-crossing based scheme to general image matching.
The psychophysical experiments of Mayhew and Frisby [30] suggested that zero-crossings alone cannot explain the perception of a stereogram that is composed of saw tooth luminance gratings of the same period but with slightly different phases. The essence was put forward by Mayhew and Frisby [30]: "the zero-crossings will fail for those parts of an image where disparity variations are tied to luminance variations situated in between the parts of the image which produced zero-crossings." They proposed that peaks of the signal after the channel convolutions are also used by human stereopsis. Hoff and Ahuja [16] pointed out the difficulties of accurate surface reconstruction directly from raw zero-crossings and they developed an algorithm that integrates surface interpolation and contour detection with zero-crossing-based matching.

There are many matching algorithms that use other types of primitives. Some of those primitives are sparse, such as edges (e.g., Baker & Binford [21], straight line segments, e.g., Ayache & Faverjon [11], isolated points (e.g., Moravec [30], Dreschler & Nagel [8]), and some aspects of scene structure (e.g., Lim & Binford [27]). The others are dense, such as intensity (e.g., Horn & Schunck [19], Waxman [43]), correlation (e.g., Glazer et al. [12]), partial derivatives (e.g., Kass [25]), energy of the Gabor filters (e.g., Heeger [14]), phase information (e.g., Sanger [39], Burt et al. [6], Jepson & Jenkin [21]), and continuous measures of features (e.g., Weng et al. [44], Weng [46]).

Since zero-crossings of a bandpass signal represent the fastest transition positions in the signal passing through the channel, they are closely related to edges, at least in a fine channel. Therefore, the matching of zero-crossings is very desirable, especially in fine channels. Peaks may also be additional primitives to match, since they may represent salient intensity variations like those in Mayhew and Frisby's experiments [30]. However, are they the only primitives used in human stereopsis?

In this article, we will see that zero-crossings and peaks are poor measurements for some stereograms in which the shape of the signal, other than zero-crossings and peaks, is critical to the stereo fusion. Although Logan's theorem states that a certain class of signals can be determined, up to a multiplicative constant, from the zero-crossings of the signal, the Laplacian-of-Gaussian filtered signals generally do not belong to this class. Instead of using the zero-crossings in one channel, Hummel and Moniot [18] studied the problem of reconstructing signals, up to a multiplicative constant, from the Laplacian-of-Gaussian scale space of zero-crossings (using zero-crossings through an entire continuous scale range). It was found that reconstruction from zero-crossings scale is possible, but it is unstable (Hummel & Moniot [18]). After all, even if theoretically the signal can be reconstructed from the scale space of zero-crossings, it does not mean that stereo matching based solely on zero-crossings is stable, because only a relatively small number of scale channels (each channel corresponds to a cross-section of the scale space) are used in practice. The absence of information about the zero-crossings from between channels makes inferring signals impossible, and causes difficulties in finding correct matches, especially in the presence of depth discontinuities. Here, instead of finding other ad hoc discrete features to remedy the existing framework of zero-crossings plus peaks, we propose a different framework in which the windowed Fourier phase (WFP) profile is introduced as the major matching primitive. Each channel has a phase profile with a different spatial period. The period decreases from a coarse channel to a finer channel. We will see that the zero-crossings and the peaks correspond to special values of the WFP. A series of important properties of this primitive is demonstrated, and this theoretical framework is tested with synthetic and natural images.

In the next section, we will discuss the need for information other than the zero-crossings and peaks. Section 3 introduces the WFP. The properties of the WFP are discussed in section 4. Section 5 discusses a two-dimensional extension to deal with general 2-D image matching. An actual matching algorithm is described in section 6. The relationships between the proposed theory and the psychophysical and neurophysiological findings are discussed in section 7. The experimental results are presented in section 8; and section 9 presents a summary.

2 Beyond Zero-Crossings and Peaks

In fact, the techniques used in the interesting design of Mayhew and Frisby's psychophysical experiments [30] can also be used to demonstrate that, at least in some situations, zero-crossings and peaks are poor primitives for the purposes of extracting disparity information. For example, consider the stereogram shown in figure 1, which is composed of tent-shaped intensity gratings. The intensity profile, as shown in figure 1(b), has a dominant component of a sine wave while
other components are very small. In fact, the profile can be smoothed further to reduce the higher-order harmonics. The stereogram exhibits disparities on the slope of the waves. Since the local distortion is symmetric with respect to each zero-crossing, the disparity at zero-crossings is zero. Our experiments have shown that this phenomenon persists through all channels, until the channel is so fine that the zero-crossings cover a continuous interval (due to the linear slope). Because zero-crossing-based stereo matching detects disparity only at zero-crossings and the disparity at zero-crossings here is zero, the zero-crossing cannot detect the disparity in this case. The result after a typical Laplacian-of-Gaussian convolution is shown in figure 1(c). Although the peaks show some disparity at coarse channels, the disparity diminishes in the fine channels. The zero-crossings and peaks will converge, from coarse to fine channels, to the correct ones as shown in figure 1(d), but they will not show the disparity between them, which is indicated by the true disparity profile in figure 1(e). This example indicates a dilemma facing image matching schemes that are based on only zero-crossings and peaks.\(^1\) From figure 1(a), most viewers can see that the second and the fourth vertical bright columns are closer to the eyes than the other two, which is consistent with figure 1(e).\(^2\) This example seems to indicate that human stereopsis uses primitives other than zero-crossings and peaks. In the stereogram shown in figure 1(a), the disparity information resides in the slope shape of the waves. As can be seen in figure 1(f), this information is kept in the profiles of the windowed Fourier phase which will be discussed later.

---

\(^1\) Homola, J., \\
Kirlinger, S., \\
and \\
Dobrowiecki, W. \\
(2000) \\
"Stereoscopic \\
impression \\
of \\
random \\
yellow-green \\
fringes."

\(^2\) Homola, J., \\
Kirlinger, S., \\
and \\
Dobrowiecki, W. \\
(2000) \\
"Stereoscopic \\
impression \\
of \\
random \\
yellow-green \\
fringes."

---

"Fig. 1. (a) A stereogram composed of tent-shaped intensity gratings. (Fusion is better from a gray-level screen than from half-tone here.) (b) Intensity profiles. (c) Profiles after a typical Laplacian-of-Gaussian convolution. (d) Matching zero-crossings and peaks through scale multichannels leads to correct matches at zero-crossings and peaks, and all render a zero-disparity. (e) The actual disparity profile used to generate the stereogram. (f) The profiles of the windowed Fourier phase to be discussed."
Originally, one major motive for using zero-crossings is that they are sparse in a coarse channel, so that matching candidates are few to avoid the false-target problem. However, sparse matching candidates need not require sparse primitives. One can still have dense primitives without making matching candidates dense. For example, the primitives can be associated with different attributes and only those primitives that have the same attribute value can be matched. As long as nearby primitives do not have the same attribute value, the matching candidates can be sparse while the primitives are densely available to provide dense disparity information.

3 The Windowed Fourier Phase

The example of tent-shaped gratings suggests that one must also consider the shape of the signal between the zero-crossings and peaks to extract disparity information. In fact, the shape of the signals is well represented by the phase, and the phase has many properties desirable to stereopsis.

3.1 The Importance of Phase

Fourier phase contains very rich information about a signal. Oppenheim and Lim [35] demonstrated the following example about recovering an original image using inverse Fourier transform from partial data of its Fourier transform. If the phase of the Fourier transform is retained but the magnitude of the transform is replaced by a unity, the inverse Fourier transform can recover the original image to a fairly large extent, particularly, the objects in the recovered image being clearly recognizable. On the other hand, if the magnitude of the Fourier transform is retained but the phase of
the transform is replaced by zero, the inverse transform gives a completely indiscernible image. In fact, the sign of the phase of the continuous Fourier transform alone can determine a broad class of signals up to a multiplicative constant (Curtis & Oppenheim [7]), and the discrete version is also true (Huang & Sanz [41]). From the definition and properties of the Fourier transform, one can see that the phase characterizes the position of the structure (sine- and cosine-wave components) in the signals while the magnitude characterizes the contrast of the structure. Therefore, the phase difference indicates image disparity while the magnitude reflects image contrast which is of little use for stereo matching because two stereo images with quite different contrasts (same sign) can be easily fused by the human stereopsis (Kulesz [23]).

3.2 The Fourier Phase and a Global Shift

Consider an infinite signal represented by \( f(x) \) whose Fourier transform is \( F(u) \), and its shifted version given by \( f_s(x) = f(x - s) \). It is well known that a shift in the spatial domain corresponds to a shift of phase in the frequency domain:

\[
F_s(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_s(x)e^{-2\pi iux} \, dx = e^{-2\pi ius} \cdot F(u)
\]

Since the images to be matched are not generally related by a simple shift, in the next subsection we need to relate the phase in the frequency domain to the spatial domain. To do this, we first define here a spatially modulated Fourier transform, which is obtained by convolving the signal with the Fourier kernel \( e^{2\pi iux} \), with \( u \) as a frequency parameter and \( x \) a spatial variable:

\[
b(x) = e^{2\pi iux} * f(x)
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)e^{2\pi i(x-t)u} \, dt
\]

\[
e^{2\pi iux} \cdot F(u)
\]

\[
\equiv \arg[b(x)] + \arg[F(u)]
\]

where,

\[
\arg[b(x)] = 2\pi x + \arg[F(u)]
\]

In this article, the relational sign \( \equiv \) always indicates equality under modulation of \( 2\pi \), unless stated otherwise. From the last expression in (1), we can see that \( b(x) \) is a Fourier transform modulated by a spatial variable via \( e^{2\pi iux} \). For convenience, we define the range of the modulo operation to be symmetric with respect to the origin. For example, the range of \( a \) mod \( 2\pi \) is \((-\pi, \pi]\).

Equation (2) indicates that, with \( u \) fixed, the phase of the convolved signal \( b(x) \) is linear in \( x \), and has a slope \( 2\pi u \). The spatially modulated Fourier transform of the shifted signal \( f_s(x) \) is

\[
b_s(x) \equiv e^{2\pi iux} * f_s(x)
\]

\[
= e^{2\pi ius(x-s)} \cdot F(u)
\]

\[
\equiv |b_s(x)| \cdot e^{2\pi i\arg[b_s(x)]}
\]

where,

\[
\arg[b_s(x)] = 2\pi u(x-s) + \arg[F(u)]
\]

(3)

From (2) and (3), we can see that the phases of the shifted and the original signals are related by

\[
\arg[b(x)] - \arg[b_s(x)] = 2\pi us
\]

(4)

Letting \( u > 0 \), from (4), the amount of shift \( s \) can be determined from the difference of these two phases to within the period \( u^{-1} \):

\[
s = \frac{\arg[b(x)] - \arg[b_s(x)]}{2\pi u} \quad (\text{mod } u^{-1})
\]

One can see that as long as the frequency \( u \) is small enough so that \( 2|su| < 1 \), the exact shift \( s \) can be completely determined from the phase difference. Since the shift of the function is constant everywhere along the \( x \)-axis and the phase is exactly linear, it makes no difference at which point \( x \) the phase difference \( \arg[b(x)] - \arg[b_s(x)] \) is evaluated.

3.3 The Windowed Fourier Phase and a Local Shift

Because generally the disparity field of a stereo image pair is not constant everywhere, the above analysis is not directly applicable to stereo matching. However, the disparity of the images, seen from a window whose size corresponds to the scale of the channel, can be approximated by a shift. The precision of such an approximation increases from a coarse channel to a finer channel. So, we define a rectangular window with a parameter \( M \) which corresponds to the scale of the channel:

\[
w_M(x) = \begin{cases} 
1 & \text{if } |x| \leq M/2 \\
0 & \text{otherwise}.
\end{cases}
\]

(5)
The Fourier kernel is then windowed:
\[ h(x) = w_M(x)e^{j2\pi xu} \]  
(6)

The convolution in (1) becomes
\[ g(x) = h(x) * f(x) \]
\[ = \int_{-\infty}^{\infty} w_M(t)f(x-t)e^{j2\pi xu} \, dt \]  
(7)
\[ = \int_{-M/2}^{M/2} f(x-t)e^{j2\pi xu} \, dt \]
\[ = R(x) + jI(x) \]
\[ \overset{M}{=} \int_{-M/2}^{M/2} f(x+s)e^{-j2\pi xu} \, ds \]  
(8)

where \( R(x) \) is the real component of \( g \),
\[ R(x) = \int_{-M/2}^{M/2} f(x-t) \cos(2\pi xu) \, dt \]  
(9)
and \( I(x) \) corresponds to the imaginary component,
\[ I(x) = \int_{-M/2}^{M/2} f(x-t) \sin(2\pi xu) \, dt \]  
(10)

The effect of windowing the Fourier kernel is clearly indicated in (8). \( g(x) \) is the Fourier transform, with a frequency \( u \), of a piece of the signal with a length \( M \) centered at \( x \), as illustrated in figure 2. The magnitude of \( g(x) \) is then
\[ |g(x)| = \sqrt{R(x)^2 + I(x)^2} \]  
and the windowed Fourier phase (WFP) is defined as
\[ \phi(x) = \arg\{g(x)\} = \arctan2(I(x), R(x)) \]  
(11)

where \( \arctan2(y,x) \) is the same as \( \arctan(y/x) \) except that it maps to the range \((-\pi, \pi]\).

It is important for the WFP to be measured from a slightly smoothed signal instead of directly to the original signal. This pre-convolution removes high-frequency noise in the signal, and results in a smooth phase profile. If the pre-convolution is done by, for example, a Gaussian filter,
\[ n(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/(2\sigma^2)} \]  
(12)

the resulting signal becomes
\[ g_n(x) = h(x) * [n(x) * f(x)] = [h(x) * n(x)] * f(x) \]  
(13)

and the combined filter
\[ h_n(x) = h(x) * n(x) \]  
(14)

is called windowed Fourier of Gaussian (WFG). The pre-convolution version of the WFP is the phase of \( g_n(x) \). Unless stated otherwise, we use this pre-convolution WFP. Equivalently, we can consider the pre-convolved WFP as the phase of \( g(x) \) obtained by convolving \( h(x) \) with a pre-convolved signal \( n(x) * f(x) \). For example, the input signal in figure 2 can be considered as a pre-convolved signal \( n(x) * f(x) \). In the remainder of the article, wherever \( g(x) \) is used, we assume that the input signal has been pre-convolved.

The coupling of \( M \) and \( \sigma \) is important. The scale of the Gaussian should be significantly smaller than the size of the window \( M \). It was found in the experiments that a good choice of \( \sigma \) should be such that \( 6\sigma \), the major support interval span, be equal to a quarter of the window size \( M/4 \), which leads to \( 24\sigma = M \). An overly large \( \sigma \) makes the filter insensitive to the details.

---

**Fig. 2.** Windowed Fourier phase is the phase of the Fourier transform of a piece of the signal centered at \( x \).
of the disparity variation (especially bad at disparity discontinuities). An overly small $\sigma$ allows high frequency components to pass through the filter, resulting in undesirable sharp spikes in the WFP profile.

The real part of the WFG filter in (14) is the convolution of the real part of $h(x)$ in (6) with the Gaussian $n(x)$ in (12) which gives the even component of the filter. The imaginary part (odd component) is the convolution of the imaginary part of $h(x)$ in (6) with the Gaussian $n(x)$. Both components are shown in figure 3. Interestingly, the even component looks like the Mexican-hat-shaped filter of Marr and Hildreth [29], except that the central portion is lower here. This is certainly not an unexpected coincidence. As we will see in the next section, a vanished even-component convolution corresponds to a phase of $\pi/2$ or $-\pi/2$, which indicate a rising and falling step edge, respectively. The odd component is missing from Marr and Poggio’s theory of stereopsis. It is in fact closely related to the peaks advocated by Mayhew and Frisby [30] as information used in human stereopsis. A peak is indicated by a zero of the first derivative, while the odd component of the filter in figure 3 is roughly a differential operator in a scale channel. We will see below that step edges and peaks are included as special cases in the phase concept.

4 Properties of the WFP

The WFP has several desirable properties, as we explain in the following.

4.1 WFP: Zero-Crossings, Peaks, and Slopes

The WFP depends on the size of the window $M$, as does the zero-crossing of the Laplacian-of-Gaussian on the scale of Gaussian. Consider several waveforms in figure 4, assuming that the size of the window is roughly equal to the portion of the wave shown in the figure. The WFP values of these waves can be directly determined based on the following polar equation about the Fourier transform in (8). Consider the integral in (8) as the sum of vectors, each of which is represented by the polar direction $e^{-j2\pi u}$ and has a signed magnitude $f(x-t)$; as $t$ runs from $-M/2$ to $M/2$, the polar vector $e^{-j2\pi u}$ circles around the origin. The WFP is the polar angle of this vector sum. From this viewpoint, it is easy to qualitatively give the WFP values for the basic waveforms listed in figure 4. Rising and falling zero-crossings correspond to phases $-\pi/2$ and $\pi/2$, respectively. These values of the WFP distinguish the two types of zero-crossings. The positive and negative zero-crossings are represented, respectively, by $0$ and $\pi$. As to slopes that contain the disparity information in the example shown in figure 1, four types are illustrated in figure 4: concave rising, convex rising, convex falling, and concave falling. The concave rising slope on the left of the rising zero-crossing corresponds to a WFP value of about $-3\pi/4$. It is a concave slope because the second derivative is positive on the left of the zero-crossing. On the right side of the zero-crossing, the second derivative becomes negative and this gives a convex rising slope with a WFP value around $-\pi/4$. The corresponding WFP values for falling slopes, first convex and then concave, are similar.
It is worth noting here that the purpose of figure 4 is not to pin down exact phase values for certain types of features, but rather to indicate rich geometrical meanings of the WFP. For example, we know that generally zero-crossings of the Laplacian of Gaussian are not found near a step edge except in a very fine channel. For the same reason, a step edge will not always correspond to an exact \( \pi/2 \) WFP in every channel. Due to its particular profile, an edge in a real image can be represented, in a certain channel, by a WFP slightly off from \( \pi/2 \). But matching based on the WFP still can cope with this edge, regardless of the actual WFP value it has in a particular channel.

With the phase polar representation shown in figure 4, the phase value of each feature is represented by a point on the unit circle. The arc length between any two points on the circle indicates the degree of difference between the two features. Thus, we can define the distance between two features as the difference of their WFP values modulo \( 2\pi \). For example, a rising zero-crossing can hardly be matched to a falling zero-crossing since they are far apart on the unit circle as shown in figure 4. This distance measurement is important for stereo matching in the presence of noise, depth discontinuities, and a moderate amount of occlusion, since it makes "soft" matching possible and is easy to implement.

Of course, the scale of the features identified in figure 4 is related to the size of the window on which computation of the WFP is based. If the input signal has strong high-frequency components so that the waveforms are very rough compared to those in figure 4, the Gaussian presmoothing by \( h(x) \) will remove most of the high-frequency components. Thus, the waveforms in figure 4 can be considered as those of a presmoothed signal.
4.2 A Bandpass Filter

In the frequency domain, the window \( w_M \) in (5) corresponds to a sinc function,

\[
W_M(u) = \int_{-\infty}^{\infty} w_M(x) e^{-j2\pi xu} \, dx = M \text{sinc}(Mu)
\]

where \( \text{sinc}(x) = \sin(\pi x)/\left(\pi x\right) \), and the Fourier kernel \( e^{j2\pi xu} \) corresponds to a delta function (Bracewell [3]):

\[
E(u) = \int_{-\infty}^{\infty} e^{j2\pi u x} e^{-j2\pi xu} \, dx = \delta(u - u_0)
\]

The Fourier transform of the product \( h(x) = w_M(x)e^{j2\pi xu} \) is obtained by convolving the two Fourier transforms, \( W_M(u) \) and \( E(u) \),

\[
H(u) = W_M(u) \ast E(u) = M \text{sinc}(Mu) \ast \delta(u - u_0) = M \text{sinc}[M(u - u_0)]
\]  \hspace{1cm} (15)

which is a sinc shifted to \( u = u_0 \). So clearly \( h(x) \) is roughly a bandpass filter.

The long tails of ripples in the profile of sinc function account for the need for presmoothing in practice. In equation (13), convolution in the spatial domain corresponds to a multiplication in the frequency domain. So, multiplying \( H(u) \) with the Fourier transform of Gaussian \( n(x) \) (which is another Gaussian)

\[ N(u) = e^{-2\pi^2 u^2} \]

gives the Fourier transform of \( h_r(x) \):

\[ H_r(u) = H(u)N(u) = M \text{sinc}[M(u - u_0)]e^{-2\pi^2 u^2} \]

As can be seen from the above equation, in the frequency domain, presmoothing using the Gaussian convolution reduces the magnitude of the ripples of sinc via the Gaussian profile. Figure 5 shows a profile of the magnitude of \( H_r(u) \), which has very short tails. The major lobe of the profile is centered at \( u = u_0 \), the central spatial frequency of this bandpass filter. This bandpass property of \( h(x) \) and \( h_r(x) \) is consistent with the concept of spatial frequency channels (Julis & Miller [24], Wilson & Bergen [49]).

Note that \( H(u) \) is the spectrum of the filter from input \( f(x) \) to \( g(x) = h(x) \ast f(x) \), not to the phase \( \phi(x) \) which we will use. The overall filter from input \( f(x) \) to \( \phi(x) \) is no longer a linear filter. Therefore, the classical spectrum analysis for linear shift-invariant filters falls short of explaining the behavior of our nonlinear filter. To investigate the property of phase space, we need a different method. In the next subsection, we will see that the WFP has a very desirable character: it is almost linear.

4.3 WFP: Quasi-Linear

As we have seen from equation (8), the WFP is the phase of the Fourier transform of a piece of signal cut from the window centered at the current point. When the window moves, say, to the right, the signal appears to move to the left with the same speed. Just imagine sitting in a moving train and looking outside through a window. As the scene moves to the left at a constant speed, what is the speed of the phase change in the Fourier transform of the windowed scene? Since only the left margin of the scene is moving out of view and the right margin is showing new areas, and the central

![Frequency Response of the Filter](image)

*Fig. 5. A plot of \(|H_r(u)|\), the magnitude of the frequency response of the filter \( h_r(x) \).*
part of the scene is moving to the left with the uniform speed, we can immediately speculate that the speed is roughly constant, or more specifically, around $2\pi u$, where $u$ is the frequency of the Fourier transform. If the window size $M$ is an integral multiple of the Fourier period $1/u$, and the content of the scene leaving from the left is the same as that coming in from the right (the scene is wrapped around or periodic), the speed of the phase change is then exactly $2\pi u$. In fact, this is the speed we saw in the case of an infinitely large window in (2).

To examine the quasi-linearity of the phase $\phi(x)$, we would like to investigate the phase slope, since a linear phase profile has a constant slope. To obtain the phase slope, we take the derivative of $\phi(x)$ in (11):

$$\phi'(x) = \frac{I'(x)R(x) - R'(x)I(x)}{R^2(x) + I^2(x)}$$

(16)

where $R(x)$ and $I(x)$ are defined in (9) and (10). Using the rules for the derivative of a parametric integration and integration by parts, we get the following expressions for $R'(x)$ and $I'(x)$ from the definitions in (9) and (10):

$$R'(x) = f^-(x) \cos(\pi u M) - 2\pi u I(x)$$

$$I'(x) = -f^-(x) \sin(\pi u M) + 2\pi u R(x)$$

where

$$f^-(x) = f(x + M/2) - f(x - M/2)$$

$$f^+(x) = f(x + M/2) + f(x - M/2)$$

Substituting them into (15) gives

$$\phi'(x) = 2\pi u - \eta(x)$$

(17)

where

$$\eta(x) = \frac{I'(x)R(x) \sin(\pi u M) + f^-(x)I(x) \cos(\pi u M)}{R^2(x) + I^2(x)}$$

(18)

Notice that the denominator is squared magnitude of the Fourier transform, $|g(x)|^2$. Applying the Schwarz inequality to (17) yields

$$\eta^2(x) \leq \frac{|I'^2(x)|^2 \sin^2(\pi u M) + |f^-(x)|^2 \cos^2(\pi u M)}{R^2(x) + I^2(x)}$$

Then using $|\sin(x)| \leq 1$ and $|\cos(x)| \leq 1$ and some algebraic simplifications, we have

$$\eta^2(x) \leq \frac{2|f^2(x + M/2) + f^2(x - M/2)|}{R^2(x) + I^2(x)}$$

So, the magnitude of $\eta(x)$ is small as long as the magnitude of the Fourier transform is relatively large. From here on we let the window size $M$ be an integral multiple of the Fourier period $1/u$, that is, $uM$ is an integer $k$. The above $\eta(x)$ in (17) is simplified to give

$$\eta(x) = \frac{(-1)^k[f(x + M/2) - f(x - M/2)]}{R^2(x) + I^2(x)}$$

Now it is clear that if the signal is wrapped around or is periodic with a period $M$, $\eta(x) = 0$ and $\phi(x)$ in (16) is exactly linear. Generally the signal is not periodic and by integrating (17) with respect to $x$ we can get an expression of the WFP around a point $x_0$.

$$\phi(x) = \phi(x_0) + 2\pi u(x - x_0) + \int_{x_0}^{x} \eta(t) \, dt$$

Since $\eta(t)$ is generally small and changes signs from one point to the other, the last integral is generally small, and $\phi(x)$ is nearly linear in $x$. So we say that the WFP, $\phi(x)$, is a quasi-linear function.

The quasi-linearity of the WFP can also be explained qualitatively by the fact that $h(x)$ is a bandpass filter. In the extreme case, suppose that the passband is so narrow that the filter passes only a single frequency $u = M^{-1}$. Then, according to the Fourier series expansion, the signal is a sinusoidal signal with a period $M$. Thus, the WFP $\phi(x)$ of this signal is exactly linear.

Why do we need the WFP to be linear? To answer this question, we need to consider the problem of matching images based on the WFP. Let $\phi_l(x)$ and $\phi_r(x)$ be the WFPs of the left and right images, respectively. Given any point $x_l$ in the left image, we want to find a point $x_r$ in the right image such that they have the same WFPs, $\phi_l(x_l) = \phi_r(x_r)$. Suppose $\phi(x)$ is linear, as shown in figure 6(a). If $x_r$ is not far from $x_l$, that is, their distance is less than a half period of the phase profile in figure 6(a), we have

$$\phi'_l(x_l) = \frac{\phi_r(x_r) - \phi_l(x_l)}{x_r - x_l}$$

Substituting $\phi_r(x_l) = \phi_r(x_r)$ into the above equation, it follows that

$$x_r = x_l + \frac{\phi_l(x_l) - \phi_r(x_r)}{\phi'_l(x_l)}$$

Therefore, $x_r$ can be directly computed from local derivative of $\phi_l(x)$ and so, no iterations are needed.

Suppose $\phi(x)$ is not linear, and thus, this may create another false-matching target within the period shown
Image Matching Using the Windowed Fourier Phase

![Diagram of a linear \( \phi(x) \) enabling matching without iteration but a nonlinear \( \phi(x) \) creates false matches and requires iterative search.](image)

(a) A linear \( \phi(x) \). (b) A nonlinear \( \phi(x) \).

in figure 6(b). An iterative search could be used to find matching but the search may get trapped in a wrong local extremum. Both the possibility of leading to wrong matches and the number of iterations needed increase with the nonlinearity of \( \phi(x) \).

We define the WFP scale space to be the WFP in the \((x, M)\) domain. The WFP profile of a row of a natural image is shown in figure 7 in which we let \( M = \nu \).

We can see from the figure that the WFP is roughly linear in spatial coordinate \( x \), and the period (or slope) of the WFP is steadily controlled by the scale \( M \). The matching candidates \( x \) with the same \( \phi(x) \) are sparse in a coarse channel (large \( M \)), and dense in a fine channel. Since \( \phi(x) \) is densely defined, primitives are dense in every channel, whether coarse or fine.

If the window is not constant in the domain of support, the phase linearity could be damaged, as shown in figure 8, where the window in (5) is a Gaussian function. The Gaussian windowed sine or cosine filter is commonly called the Gabor filter, since it was originally introduced by Gabor [11]. The odd component of the Gabor filter is

\[
G_{e}(x) = e^{-(x^2/2\sigma^2)} \sin(2\pi\sigma x)
\]

and since it is odd, its DC residual is equal to zero

\[
\int_{-\infty}^{\infty} G_{e}(x) \, dx = 0
\]

The even component is

\[
G_{o}(x) = e^{-(x^2/2\sigma^2)} \cos(2\pi\sigma x)
\]

Represented in the polar plane, the Gabor phase is the phase of the polar vector whose horizontal component is the result of convolution using \( G_{o}(t) \) and whose vertical component is that using \( G_{e}(t) \). During 1989, while this research (Weng [45]) of the WFP was independently motivated and developed from the well-known phase-shift property of the Fourier phase, we became aware of the work by Jepson and Jenkin (Jepson & Jenkin [21]) in using the Gabor phase for computing stereo image disparities. Since 1988, the Gabor phase has been used by several researchers, including Sanger [39], Langley et al. [26], Fleet & Jepson [9], and Jenkin et al. [20] for computing one-component disparity information (either stereo disparity or one component along the gradient direction).

The two components of the Gabor filter are plotted in figure 9. Comparing figures 3 and 9, we can see the similarities and differences between the WFG and the Gabor filters. The odd component of the former has one pair of ripples while that of the latter has many small ripples in addition to the major pair. We can also see that, with the same frequency \( \nu \), the WFG filter is shorter in the spatial domain than the Gabor filter, which indicates that, with the same spatial frequency \( \nu \), the former is more suited for fast disparity transitions. Another major difference between the two is the DC in the even component as explained below.

The central lobe in the even component of the Gabor filter is larger than the two smaller lobes beside it. In fact, the even component of the Gabor filter has a positive DC residual:

\[
\int_{-\infty}^{\infty} G_{e}(x) \, dx = a e^{-2\pi^2 \sigma^2}
\]

We can also see from figure 8 that the phase is virtually constant in a coarse channel and hardly ever completes a 2\( \pi \) circle. This can be mostly attributed to the DC residual in the Gabor filters. If the image pixel intensities are represented by positive integer gray values, the result from the even-component convolution is mostly positive, which implies that the Gabor phase value will mostly fall in \([-\pi/2, \pi/2]\) as shown in figure 8.

A few methods can be used to remedy this DC problem. The first method is to use a narrower bandwidth with a small \( \sigma \). Figure 8 shows that the nonlinearity is not very severe for small \( \sigma \), and the DC residual in (21) is small when \( \sigma \) is very small. But this method does not eliminate the DC imbalance. The second method is to subtract the DC residual from the input signal so that the new input signal has a zero DC. But this cannot completely solve the problem since a strong DC residual can still exist locally (e.g., the values of the signal may be above the average on a large interval).
Fig. 7. The WFP space of a row of a natural image. (a) Intensity profile of the row \( f(x) \). (b) the WFP scale space \( \phi(x; \tau) \). Period \( \tau = 1/u = M \). The intensity value represents the WFP from \(-\pi\) (the darkest) to \(\pi\) (the brightest). So “black” \((-\pi\) \) is a neighbor of “white” \((\pi)\). (c) The WFP profile \( \phi(x) \) with \( M = 1/u = 65 \). (d) The magnitude of output \( g(x) \) after WFG filtering, with \( M = 1/u = 65 \).
Fig. 8. The Gabor phase space from the same input as in Figure 7. (a) The Gabor phase scale space. Period $T = 1/\pi$. The intensity value represents the phase from $-\pi$ (the darkest) to $\pi$ (the brightest). So "black" ($-\pi$) is a neighbor of "white" ($\pi$). $\sigma = 5/(4\pi\alpha)$ which implies that at the frequency that corresponds to the first zero in the frequency response of the WPG, the frequency response of the Gabor filter falls to 4.4% of its peak value. Thus, the bandwidth of the WPG in Figure 7 is comparable with that of the Gabor filter here. (b) The Gabor phase profile with $1/\alpha = 65$. The vertical axis represents the Gabor phase in degrees. As can be seen, the Gabor phase is very nonlinear, and is almost constant in coarse channels.

Fig. 9. The two components of the Gabor filter. The left side is the odd component $G_1(x)$, and the right side is the even component $G_2(x)$. $\alpha = 65^{-1}$. 
The third method is to subtract the DC residual from the filter by using \( \hat{G}_c(t) = G_c(t) - a \) where \( a \) is such that \( \int_{-\infty}^{\infty} \hat{G}_c(t) \, dt = 0 \). But since the domain of the filter \( G_c(t) \) is infinite, this method is not possible unless the filter \( G_c(t) \) is approximated by setting it to zero when \( \epsilon \) is sufficiently large. Other methods alter the basic definition of the Gabor filter and lose some of its basic properties to various degrees. For example, the method used by Fleet and Jepson [9] is to replace \( \cos(2\pi \omega x) \) in (20) by a lowered cosine: \( \cos(2\pi \omega x) - \beta \) where \( \beta \) is a positive constant. Consequently, the even part has a smaller central positive lobe and two larger major negative lobes than that in figure 9, and thus, it becomes more like its WFG counterpart shown in figure 3.

Note that the WFP does not suffer from this DC problem since the DC residuals of the two components shown in figure 3 are all equal to zero.

However, since both the WFP and the Gabor phase belong to the local phase information of a signal, the WFP or the Gabor phase all share the basic principle of using phase information. Furthermore, the properties of the WFP that are intrinsic for phase information are also applicable to the Gabor phase.

It should be noted that the global quasi-linear behavior of \( \phi(x) \) does not exclude the possibility that \( \phi(x) \) may vary slightly away from the linear trajectory. This variation is in fact exactly what we need to estimate the disparity profile in each channel. Disparity between two images is indicated by the difference in their WFPs. Taking the stereo case as an example: compared to the WFP around a point in the left image, a faster change of the WFP around the corresponding point in the right image indicates a “squeeze” in the right image and a slower change indicates a “stretch.”

4.4 Uniqueness of Signals from the WFP in One Channel

A question may be raised here: it may be fine to match phase because of what we have raised so far; but why do we throw away the information in the magnitude? An immediate answer to this question is that magnitude represents the contrast of images and, according to Julesz’s experiments [23], images of different contrast (same sign) can be easily fused. But this answer still seems not satisfactory because discarding magnitude might lose some important information, other than the vague terms such as “contrast,” that might be important to stereopsis. However, I found that the WFP in one channel is enough to uniquely determine the original signal up to a multiplicative constant.

**Proposition.** Suppose a signal \( f(x) \), defined on an interval \( D \), is equal to zero (or known) on a known interval of length \( M \) in \( D \), and the WFP \( \phi(x) \) is defined in \( D \) except for some isolated points. Then the signal \( f(x) \) is determined up to a multiplicative constant by its WFP \( \phi(x) \) with a window of size \( M \).

The proof is presented by Weng [48]. Although the complete treatment of this problem (Weng [46], [48]) is beyond the scope of this article, an outline of the approach is helpful here. We define a new function from \( f(x) \)

\[
\hat{f}(x) = f(x) - f(x - M)
\]

(22)

Suppose that \( \hat{f}(x) \) can be determined from the WFP \( \phi(x) \). From (22) we can see that the values of \( f(x) \) on an interval of length \( M \) determine the values of \( f(x) \) on the neighboring intervals of the same length via the “incremental” function \( \hat{f}(x) \). For example, let \( f(x) \) be defined in \([0, \infty)\) and then, extend it to \([-M, \infty)\) by setting \( f(x) \) to zero in \([-M, 0)\). Now we can use

\[
f(x) = f(x - M) + \hat{f}(x)
\]

(23)

to determine the values of \( f(x) \) in \([0, M)\). Similarly, we can recursively determine the values of \( f(x) \) on interval \([iM, (i + 1)M)\) using \( \hat{f}(x) \) and the value of \( f(x) \) on \([(i - 1)M, iM)\).

Therefore, the key problem is to investigate how \( \hat{f}(x) \) is determined by the WFP \( \phi(x) \). In fact, it has been proved (Weng [46], [48]) that \( f(x) \) is determined by the WFP \( \phi(x) \) as

\[
\hat{f} \left( x + \frac{M}{2} \right) = C e^{-\int_{\phi(x)}^{\phi(x)} p(\xi) \, d\xi}
\]

(24)

where \( C \) is a constant (scale factor), and

\[
p(x) = \frac{\cos(\phi(x)) \left( \frac{\phi'(x)}{\sin(\phi(x))} + 1 \right)}{\phi'(x) - 2\pi u}
\]

The requirement that \( f(x) \) be zero on an interval is very easily satisfied for finite images, because one can always assume that images are zero outside the image frame. Similar requirements are imposed by Curtis and Oppenheim [7] and Huang and Sanz [41] where the region of support of an image must be restricted to a nonsymmetric half-plane in order for the sign of Fourier phase (in all frequencies) to determine the signal, up to a
5 A Two-Dimensional Extension

A two-dimensional (2-D) image-matching problem is to match two images taken at different positions, with respect to the scene. For example, a pair of stereo images may have some moderate vertical displacement. It has been reported by Nielson and Poggio [31] that human stereo fusion survives about 4' - 7' of vertical disparity. But most needs for general 2-D image matching arise from motion analysis. Aside from engineering applications of motion analysis (Huang [17]), human ability to process visual motion information is also an interesting and important issue (Braddick [4], Ullman [42], Braddick [5]).

The Fourier kernel in 2-D space is $e^{j2\pi(\alpha x + \beta y)}$. The 2-D window is defined as

$$w_M(x, y) = \begin{cases} 1 & \text{if } |x| \leq M/2 \text{ and } |y| \leq M/2 \\ 0 & \text{otherwise} \end{cases}$$

The 2-D windowed Fourier kernel is

$$h(x, y) = w_M(x, y) e^{j2\pi(\alpha x + \beta y)}$$

Given the input $f(x, y)$, the convolution in 2-D is

$$g(x, y) = h(x, y) * f(x, y)$$

$$= \int_{-M/2}^{M/2} \int_{-M/2}^{M/2} f(x - t, y - s) \times e^{j2\pi(\alpha t + \beta s)} \, dt \, ds$$

(25)

We note that the filter $h(x, y)$ is separable. The 2-D WFP $\phi(x, y)$ is the phase of $g(x, y)$:

$$\phi(x, y) = \arg[g(x, y)]$$

with the parameters $\alpha$ and $\beta$, or, to write explicitly,

$$\phi(x, y; \alpha, \beta) = \arg[g(x, y; \alpha, \beta)]$$. If the image disparity field can be locally approximated by a shift (image plane translation), we can consider phase in two directions separately. So we define the phase in $x$-direction by

$$\phi_x(x, y) = g(x, y; \alpha, 0)$$

(26)

with a parameter $\alpha$, and the phase in $y$-direction by

$$\phi_y(x, y) = g(x, y; 0, \beta)$$

(27)

with a parameter $\beta$. As indicated in (25), when one of the frequency parameters $\alpha$ or $\beta$ is set to zero, the convolution in the corresponding direction becomes a simple integration.

As an example, in figure 10 a pair of synthesized random-dot stereo images is shown together with the...
WFP images of horizontal direction. For this pair of 512 × 512 stereo images, the window size of the computed WFP images shown here is $M = 33$. As can be seen from the figure, these two WFP images have the following characteristics. 1) The horizontal period, which is controlled by the window size, is spatially stable. 2) The spatial gradient of the WFP is mostly in horizontal direction. 3) The WFP is quasi-linear in horizontal direction. 4) The disparity of these two WFP images indicates the disparity of the original stereogram. Due to the good properties of the WFP, disparity information can be easily extracted from the WFP images. Thus, one can avoid dealing with original complex intensity patterns directly. The results of image matching for this pair of images will be presented later in figure 13.

6 Image Matching Based on the WFP

Supported by the properties established in section 4, the matching algorithm to be presented in this section is based on the WFP, that is, the matching points should have roughly the same phase.

The computational structure and data flow used in this algorithm are shown in figure 11. For details of
this scheme, the reader is referred to Weng et al. [44], [47] and also the related control strategies (e.g., Marr & Poggio [28], Mayhew & Fishy [30], Ohta & Kanade [32], Hoff & Ahuja [16]). The algorithm uses a multiscale multigrid computational structure. Each level corresponds to a scale, with higher levels corresponding to larger scales (with coarse resolution). Zero disparity is used to initialize the disparities at the highest level. At a coarse resolution, the displacement field only needs to be computed along a relatively coarse grid, since the displacement computed at a coarse resolution is not accurate and a low sample rate suffices. The coarse displacement field is projected to the next-finer level with a linear interpolation, where it is refined further. This process of projection and refinement continues down to finer levels successively until we get the final result at the original resolution. Each resolution level corresponds to a channel we discussed previously.

At a level with spatial period $1/u$, if a point is matched to the nearest point that has the same phase, the recoverable disparity is roughly a half of the period, $1/(2u)$. Therefore, as long as the error in the estimated disparity from the previous coarse level does not exceed $1/(2u)$, it is possible for the current level to further improve the disparity estimate. If the two successive levels are separated by an octave, each level should guarantee that the error in its disparity estimates is bounded by a quarter of the channel's spatial period. This is the precision allowed for the channel. If the frequency gap between two successive levels is reduced, the error limit imposed on each channel can be relaxed accordingly.

The window size $M$ can be naturally made to equal an integral multiple of the spatial frequency $1/u$, that is, $M = k/u$ where $k$ is an integer. I have found that $k > 1$ makes the matching near depth discontinuities more difficult because of the excessively large extent in which the phase is computed. So we select $M = 1/u$. In the family of filters, the window size $M = 1/u$ assumes the values $2^{i+1} + 1$ at level $l$, $l = 0, 1, 2, \ldots, L$.

We consider the case in which rotation in the image plane is small so that in the window the image can be approximated, to within the precision allowed at this
resolution level, by a local shift. In the case of large rotations, local rotational information can be transmitted down from coarse channels. The direction of local phase is then accordingly rotated. Therefore, the phase might also be used for the cases where image-plane rotation is not small.

The magnitude of \( g(x) \) may sometimes vanish so that the phase is not defined. With natural images, a vanishing \( g(x) \) rarely occurs, since it requires both \( I(x) \) and \( R(x) \) to vanish at the same time (see, e.g., figure 7(d)). But it does occur in uniform regions. The phase computed from a very small \( g(x) \) is also unreliable due to noise. For the case where phase estimates are undefined or unreliable, an interpolation should be automatically triggered. A framework to incorporate both phase matching and interpolation follows:

Let the first image \( i_1(x) \) have WFP \( \phi_1(x) \), and the second image \( i_2(x) \) have WFP \( \phi_2(x) \). Suppose a point \( x = (x_1, x_2) \) in the first image is matched to a point \( x + d \) in the second image, where \( d(x) = (d_1(x), d_2(x)) \) is the disparity at \( x \). Then their phase difference

\[
\phi_2(x + d(x)) - \phi_1(x)
\]

should be small. We want to minimize a term of controlled smoothness in order to take into account interpolation for regions with low magnitude in \( g(x) \):

\[
d(x) - \tilde{d}(x)
\]

where \( \tilde{d}(x) \) is the average neighbor disparity around \( x \), which takes discontinuity into account (Weng et al. [44], [47]). Specifically, \( \tilde{d}(x) \) is the average disparity of the neighboring points that belong to the same region as point \( x \) and thus, points on the other side of the intensity discontinuities are not included.

First, suppose that the disparity is purely horizontal. Then \( x \) and \( d \) can be considered as scalars which correspond to the horizontal component in a pixel row. From an initial \( d \) computed at a coarse level, we want to determine an increment \( \delta \) of \( d \) so that

\[
e(\delta) = [\phi_2(x + d + \delta) - \phi_1(x)]^2 + \lambda^2 |d + \delta - \tilde{d}|^2
\]

is minimized, where \( \lambda \) is a regularization parameter (Poggio et al. [38], Grimson [13], Hoff & Ahuja [16], Weng et al. [44], [47]). The last term corresponds to a controlled smoothness constraint which imposes a smooth surface at part of an image where variation of intensity and disparity is small, but suppresses smoothing across discontinuous surface where a large image intensity transition occurs or a significant disparity variation is detected based on the disparities estimated from the coarse channels. The value of \( \lambda \) should be related to the reliability of the computed phase value. For example, it can be made inversely proportional to the magnitude of \( g(x) \) in (7) when the magnitude is sufficiently small, so that the regularization term is dominant in a uniform region to automatically perform disparity extrapolation.

Expanding \( e(\delta) \) in (30) at \( x + d \) gives

\[
e(\delta) = [\phi_2(x + d) + \phi_2'(x + d)\delta - \phi_1(x + d)]^2 + \lambda^2 (d + \delta - \tilde{d})^2
\]

Notice that the right-hand side is quadratic in \( \delta \). Minimizing the right-hand side with respect to \( \delta \) gives

\[
\delta = \frac{[\phi_2(x + d) - \phi_1(x + d)]\phi_2'(x + d) + \lambda^2 (d - \tilde{d})}{[\phi_2'(x + d)]^2 + \lambda^2}
\]

The more linear \( \phi_2(x) \) is, the more accurate (31) is. If \( \phi_2(x) \) is exactly linear, (31) is exact and (32) gives the disparity directly and no iterations are needed. To take into account slight nonlinearity in \( \phi_2(x) \), replacing \( d \) by \( d + \delta \) gets an improved disparity. Then we get next \( \delta \) according to (32). Such iterations based on (32) are performed a few times at every grid point in the refinement for each level. Since \( \theta_2(x) \) is almost linear, very few iterations suffice (e.g., 2 or 3).

For general image matching, we need to determine two components of the displacement field, horizontal and vertical. The expressions (26) and (27) represent the WFP in horizontal and vertical directions respectively. Equation (30) becomes

\[
e(\delta) = [\phi_{iw}(x + d + \delta) - \phi_{iw}(x)]^2 + [\phi_{jw}(x + d + \delta) - \phi_{jw}(x)]^2 + \lambda^2 |d + \delta - \tilde{d}|^2
\]

where \( \phi_{iw}(x) \) is the WFP of the \( i \)th image defined in (26) and \( \phi_{jw}(x) \) is the WFP of the \( j \)th image defined in (27). Analogous to the stereo case, using Taylor expansion for two variables \( \delta = (\delta_1, \delta_2) \), we get an expression similar to (31). Taking the derivatives with respect to \( \delta_1 \) and \( \delta_2 \), respectively, and setting them to zero yields two linear equations in the two unknowns \( \delta_1 \) and \( \delta_2 \). Solving these two equations for \( \delta_1 \) and \( \delta_2 \) gives the expressions for updating \( \delta_1 \) and \( \delta_2 \).
Now let us see what types of computation are needed. $R(x)$ and $I(x)$ are computed by simple convolutions with the image. Then the WFP can be computed from $R(x)$ and $I(x)$ using a two-value lookup table. Disparity updating at each pixel need only evaluate (32) for stereo, or a similar equation in the case of general motion, which can be computed by simple byte-wide ALUs and lookup tables in commercially available frame-rate image-processing hardware.

Expressing this approach as a neural network, each neuron would correspond to a pixel, and be linked with 8 of its immediate neighbors. Phase signals are transmitted to corresponding neurons after the odd and even convolutions and one step of phase mapping (implemented by a lookup table). A layer of neurons corresponds to a channel. The disparity fired by each neuron is transmitted directly to the corresponding neuron in the finer layer. The layer of nodes as well as images are sampled at coarse levels, which significantly reduces the size of the odd and even convolution kernels. The whole network can be completely implemented using $1 \times 17$ convolvers and computational nodes each linked with its $3 \times 3$ neighbors. We have simulated this network using a C program on a SUN SPARC workstation. Real-time implementation on frame-rate hardware is possible. Figure 12 shows an overview of the network.

Fig. 12. An image matcher implemented as a network. Note that this figure does not show the fact that lower channels may have fewer nodes than the higher channels.

7 Psychophysical, Neurophysiological, and Computational Considerations

Although the primary objective of this study concerns the computational aspects of image matching instead of modeling human stereopsis, a brief discussion on the psychophysical, neurophysiological, and computational considerations may link this study to the existing results in those related areas.

Some aspects of Marr and Poggio’s pioneering theory on human stereopsis (Marr & Poggio [28]) have been supported by the accumulated psychophysical and neurophysiological evidence as well as computer vision research results. Among these, the multiscale representation and the coarse-to-fine strategy have been used, to various degrees, in the computer vision community. Zero-crossings have also been used at matching primitives by several stereo-matching algorithms. Mayhew and Frisby’s interesting experiments led to important extensions of Marr-Poggio’s theory, including the need for peaks, cross-channel correspondence, figural continuity, and edge connectivity. There are many other proposed improvements, for example, Baker and Binford’s ordering constraint (Baker & Binford [21]) and Hoff and Ahuja’s contour detection (Hoff & Ahuja [16]). As pointed out by Poggio and Poggio [37], the scheme of Marr and Poggio will have to be modified in several respects. While more and more constants and methods are proposed, the computational steps become more sophisticated. Many of these steps require case-by-case checking, complicated bookkeeping, and extensive search. However, a generally accepted notion is that stereopsis has a knowledge-free, low-level character (Poggio & Poggio [37]). It is increasingly evident that we need a general framework that is natural, effective, and less expensive to be implemented, either by natural neural networks or by artificial computer architectures. The study presented here is a step in these latter directions.

In this article the WFP concept is introduced for image matching. It seems that many proposed discrete features fall into this phase space and all those can be treated in a simple and integrated manner. The computational implementation of the matching scheme is low level, uniform, and local. The matching algorithm does not need case-by-case checking, sophisticated bookkeeping, or extensive search, which seem to be difficult, or at least expensive to accomplish with either the visual cortex or current computer hardware. This framework is consistent with several existing psychophysical observations. Among others, 1) stereograms with different contrast (same sign) can be fused (Julesz [23]); 2) fusion is impossible if the contrast sign is reversed (the maximum phase change of $\pi$) (Julesz [22]); 3) the need for peaks as demonstrated in Mayhew and Frisby’s experiments; 4) the concept
of spatial frequency channels (Julesz & Miller [24], Wilson & Bergen [49]).

The correspondence problem seems to reach at least an initial solution at a set of complex cells in the visual cortex (Poggio & Poggio [37]). It has been reported that the complex cells in the cat's visual cortex show a considerable spatial-phase dependence of their response, and phase-specific binocular interaction (Spitzer & Hochstein [40], Ohzawa & Freeman [34]). And Ohzawa and Freeman [33] report that most simple cells also show phase-specific binocular interaction. A linear summation model has been proposed, based on the data of neurophysiological experiments, to explain the extraction of interocular Fourier phase difference in visual cortex (Ohzawa & Freeman [34]). However, neurophysiological data are still lacking for concluding that these interactions in the cells are conveying information about depth. Based on cell response, an interesting scheme of sine and cosine filtering has also been proposed by Pollen and Ronner [36] and Foster et al. [10], which is helpful for further understanding how monocular and binocular phase information is coded in visual cortex.

8 Experimental Results

A series of synthesized images as well as natural images has been tested to evaluate the theory and the algorithm. Due to the quasi-linear nature of the WFP, just two to three iterations suffice for each node at each level. Further iterations yield little improvement. All images shown here are composed of about 512 x 512 pixels. Although the epipolar constraint can be used to constrain the matching in the case of stereo, we used the general matching algorithm for all the cases, both stereo and motion images. This arrangement is to test the theory and the implementation under more difficult 2-D matching situations. If a general matching algorithm performs well for those tougher 2-D matching problems, it should work at least as well for a simpler 1-D matching problem (stereo matching with the epipolar constraint).

8.1 Synthesized Images

The first example shown here in figure 13 is a random-dot stereogram whose underlying depth shape is called "wedding cake" by Grimson [13]. There are four layers of depth planes. The disparity from the first layer to the last is 0, 8, 16, 24 pixels, respectively. Cross-eye stereo fusion can immediately bring this structure under examination. The general image matcher was applied to this stereogram, and the computed disparity map is shown as intensity image in figure 13(b), and plotted as a surface on a 128 x 128 grid in figure 13(c). The quality of the surface looks very good. Sharp transitions between disparity layers have been preserved relatively well using the simple method we discussed in section 6. It is worth mentioning that no surface-fitting step is needed for this already very dense depth map (512 x 512 pixels) and some of the well-known artifacts of surface fitting are avoided.

The second example is a random-dot stereogram shown in figure 14 whose underlying structure is a half-ball. The underlying depth is symmetric from the view point of the left image. So when the stereogram is stereoscopically fused, one can perceive that the ball skews to the right a little bit. The maximum disparity is 32 pixels. This example is designed to test the capability of constructing curved surfaces from stereograms. The disparity obtained from the image matcher is shown as an intensity image and a surface plot, respectively, in figure 14. The clear ball shape of the disparity can be seen. Notice that the automatic scaling of the surface plot makes the ball look taller than it really is.

The third example uses random-dot motion images shown in figure 15. A half-ball structure is rotated about an axis (1, 1, 2) by an angle of -5 degrees. A sampled displacement field, which is shown in the figure, was obtained from the computed dense 512 x 512 displacement field. In this motion case, disparity shape is not directly related to the depth, and therefore, the error of the displacement field is presented instead. To prevent the display of errors from being distracted by many dots, only error vectors that are longer than two pixels are shown in figure 15. As expected, errors are mostly restricted to the regions of motion discontinuities. The overall displacement field is still good.

8.2 Natural Images

With natural images, the surface of the scene can be complex. It is usually very difficult to accurately obtain these surfaces through surface fitting from sparse feature disparities.

The next example is a natural stereogram called "ruts," shown in figure 16. It is interesting to let the
general 2-D matcher to determine the irregular surface of the ruts from the stereograms. The computed disparity is shown as an intensity image as well as plotted as a surface in the figure. It can be seen that details of the surface are well recovered.

The final example is a pair of natural motion images shown in figure 17. The scene contains a row of bushes planted on the shoulder of a wall. The upper left and the upper right corners are two windows, which correspond to the largest depth. The upper wall is closer than the lower wall. The bush constitutes complex surface in the foreground. The samples of the computed displacement field is shown in figure 17. Since the motion to the lower right, the horizontal disparity is directly related to the depth. The horizontal disparity shown in the figure does display well the different layers of depths and, interestingly, the irregular surface of the bush.

Due to the large number of stereo algorithms in the literature, it seems impractical to compare performances among many existing algorithms. However, many of those algorithms have been tested on "wedding cake" stereograms, and the algorithm presented here appears to have given better results.
9 Summary

Despite the alluring application of automatic systems for stereoscopic vision, the stereo-matching problem has been proven to be surprisingly difficult (Poggio & Poggio [37]). Although a complete solution is unlikely in the near future, the theory and the algorithm introduced in this article seem to display an interesting dimension in stereo matching as well as motion matching.

The main contributions of the theory are introduction of the windowed Fourier phase to image matching and a series of exposition of its properties related to the matching. The windowed phase is supported by 1) the example in figure 1 which indicates the need for more information in addition to zero-crossings and peaks; 2) the inclusion of zero-crossings and peaks, as well as other shape information (convex rising, convex falling, concave rising, concave falling) as special cases; 3) the establishment of the desirable quasi-linear property of
the WFP; 4) the completeness of the WFP in representing signals; 5) the easier, faster, uniform, parallel, network-like matching that is made possible by the use of the WFP; and 6) experiments on synthetic and natural images, which showed good results.

Although the question is still open as to whether human stereoscopic fusion performs sparse discrete feature matching or dense phase matching, some of the neurophysiological studies have suggested close links between stereopsis and phase information. A comparison between the existing discrete feature-based methods and the network-based method here seems to indicate that many sophisticated case-by-case procedures can be naturally replaced by parallel, simple, uniform, and node-based operations in the WFP-based matching.

Acknowledgment

This work was supported by the Natural Science and Engineering Research Council of Canada under grant number OGP0046283, while the author was with Centre de Recherche Informatique de Montréal and École Polytechnique de Montréal, Montreal, Canada.
Fig. 16. Ru's stereogram and the result. (a) The stereo gram. (b) Computed disparity map as intensity image. (c) Plot of the computed disparity surface on a 128 × 128 grid, where the border is set to zero.

Notes

1. This does not rule out the possibility of combining, in a complicated way, the zero-crossings and peaks in many scales to detect the disparity. For example, it is possible to reconstruct the original images using zero-crossings and peaks in many scales and then use any stereo-matching algorithm to perform matching from the reconstructed images. But such a combination contradicts the basic motives of using matching primitives—detecting disparity directly at the position of the primitive so that the matching is easier, more efficient, and more accurate. In this context, the zero-crossing and peak are not regarded as codes that implicitly encode the disparity information but, rather, they are considered as matching primitives for a coarse-to-fine scheme as originally proposed.

2. In the experiment, 16 subjects who have good vision (either uncorrected or corrected) viewed this stereo gram displayed on film slides through a stereoscope. The subjects were instructed to pay attention to the depth difference of the four bright columns. Among them, 12 subjects reported that they saw depth difference and they all correctly indicated which columns are closer. The remaining 4 subjects reported that they did not see noticeable depth difference.
Fig. 17. Bush-wall-window motion images and the result. (a) The pair of images. (b) Samples of the computed displacement field. (c) Computed horizontal disparity as intensity image.

References


