Chapter 2

Stereo Correspondence of Face Images Based on Wavelets
Chapter 2

Stereo Correspondence of Face Images Based on Wavelets

In this chapter, we present an approach for estimating dense disparity map for stereo face images using coarse to fine technique based on wavelets. To find the corresponding point in a stereo pair of face images often consider intensity or color differences as a local matching metric, which is sensitive to contrast changes. In order to find disparity, we employed the multiwavelet based coarse to fine algorithm for stereo matching, which depends on multiple spatial frequency channels for matching of a stereo pair of face image. The coarser-scale basis function has larger support while the finer-scale basis function has smaller support. The stereo test images of face database and middlebury database are used to generate experimental results. The performance of the proposed approach is compared with other popular area-based approaches such as SAD, SSD and NCC, experimental result shows that our approach yields better results.

2.1 Introduction

3D reconstruction is the process of capturing the shape and appearance of real world objects. 3D reconstruction is the process of estimating the depth
from the pair of images. 3D reconstruction has two steps: the first step is to find the disparity map between the corresponding points in the image pairs (stereo matching) and second step is to estimate the depth map from a computed disparity map. The displacement of position in the two corresponding image points in a stereo image pair is called disparity. The stereo matching algorithms are mainly categorized into two types: area (Local) based and energy (Global) based algorithms. In area-based methods, a disparity vector for each pixel within a window search area is calculated using a matching algorithm. In energy based methods, the disparity vector is determined using a global cost function minimization technique. Global algorithms for stereo matching are having some of the disadvantages. The smoothness parameters used by global algorithms, may lead to poor performance at image discontinuities. Because of computational complexity, global algorithms are unsuitable for real-time applications (Stefano, 2004). Despite the demand for rapid and reliable stereo matching techniques, only few rapid algorithms are available (Vander, 2001; Stefano, 2004; Stefano, 2005), and most of them are area-based algorithms.

Scharstein and Szeliski (2002) presented the taxonomy of stereo correspondence methods. They provided an exhaustive comparison of the recently best performing dense stereo correspondence algorithms. Most of the stereo correspondence methods working in the spatial domain, assume that the projection of an object will have the same area in both images. However this condition is violated by perspective projection and causes the ambiguous mismatches. This will greatly reduce the accuracy of a stereo vision system and can be avoided by considering the stereo matching in the frequency domain. The wavelet analysis has been used for stereo matching (Sarkar, 2007; Begheri, 2010; Bhatti, 2010). I. Sarkar and M. Bansal (2007) using a mutual information to solve the correspondence problem. P. Begheri and C.V Sedan (2010) employed a global energy minimization technology to generate a disparity map for each baseband of the stereo pairs are combined used a fuzzy algorithm. A. Bhatti, S.Nahavandi (2010)
introduced a multiresolution based technique, they adopted wavelet transform modulus maxima to establish correspondences between the stereo pair of images. P. Bagheri, C.V. Serdean (2010) proposed a multiwavelet based multiresolution technique using fuzzy algorithm to combine a four disparity map at a coarsest level. Then adopted a coarse to fine strategy to refine the disparity map up to the finest level. A. Ogale et al. (2005) presented a contrast invariant stereo matching algorithm, which uses gabor filters for decomposition of images.

However, finding correct corresponding points is subject to a number of potential problems like occlusion, ambiguity, radial distortions and illuminative variations (Scharstein et al., 2002). A number of algorithms have been proposed to address some of the aforementioned issues in stereo vision however it is still relatively an open problem. Current research in stereo vision has attracted a lot of focus on multiresolution techniques, based on wavelets/multiwavelets scale-space representation and analysis, for correspondence estimation (Mallat et al., 1991). However, very little work has been reported in this regard. The main advantage of these algorithms is their hierarchical nature that exhibit behavior similar to iterative optimization algorithms. These algorithms generally operate on an image pyramid where results from coarser levels are used to constrain regional search at finer levels. The coarse-to-fine techniques adopted in these algorithms are considered to be a middle approach unlike other existing algorithms where correspondences are established either purely on local search (Stefano et al., 2004) or as a global cost-function optimization problem (Scharstein et al., 2002).

In this chapter, we introduce a correspondence algorithm to estimate accurate and dense displacements between two face images. The wavelet-based method represents displacement by a linear combination of hierarchical basis functions generated by dilating and translating two prototype functions-a scaling function and a wavelet. A coarse-scale basis function has a large support while a fine-scale basis function has a small support.
Corresponding to these supports of various sizes, large-to-small regions in full resolution are concurrently used. This allows both global and local information to be utilized simultaneously for image correspondence. The disparity is estimated by a linear combination of coarse-to-fine wavelet basis functions and estimates the wavelet coefficients from the coarsest-scale coefficients by minimizing the sum of absolute intensity differences. Note that this coarse-to-fine wavelet transform is different from the conventional wavelet transform that is carried out by two stages: the first stage is to decompose the function from fine to coarse resolution, and the second stage is to reconstruct the function from coarse to fine resolution. The underlying coarse-to-fine representation enables us to use large-to-small windows simultaneously for accurate estimation.

The organization of the chapter is as follows: The multiwavelet transform is briefly discussed in section 2.2. Section 2.3, presents the rectification of stereo images using uncalibrated rectification method. The coarse to fine approach is presented in section 2.4. We present the brief discussion of matching strategy in section 2.5. The experimental results are illustrated in section 2.6. Finally, section 2.7 draws a chapter summary.

### 2.2 Wavelet Transform

The multiresolution analysis is generally performed by Wavelets. Shi et al. (1997) evolved a multi-wavelet theory from wavelet theory and enhanced. Wavelet analysis is relatively one way of scale space representation of the images and considered to be as fundamental as Fourier and a better alternative. One of the reasons that make wavelet analysis more attractive to researchers is the availability and simultaneous involvement of a number of compactly supported bases for scale-space representation of images, rather than infinitely long sine and cosine bases as in Fourier analysis. Approximation order of the scaling and wavelet filters provide better approximation capabilities and can be adjusted according to input image by selecting the
appropriate bases. Other features of wavelet bases that play an important role in image processing application are their shape parameters, such as symmetric, asymmetric, orthogonality and orthonormality.

Success of multi-wavelets bases over scalar ones, stems from the fact that they can simultaneously possess the good properties of symmetry, orthogonality, short support and high approximation order, which is not possible in the scalar case. A discrete wavelet transforms (DWT) is any wavelet transform for which the wavelets are discretely sampled. The wavelet transform, captures both frequency and location information (location in time). The 2D discrete wavelet transform (DWT) is used to decompose the given image into four subbands, namely: 1) LL; 2) LH; 3) HL; and 4) HH. After each level of decomposition, the size and resolutions of the image are reduced by a factor of 2 due to down sampling. The wavelet theory is based on the refinement equation is given below:

\[
\phi(t) = \sum_{k=-\infty}^{\infty} h_k \phi(mt - k) \quad (2.1)
\]

\[
\varphi(t) = \sum_{k=-\infty}^{\infty} c_k \phi(mt - k) \quad (2.2)
\]

Where \( \phi(t) \) is a scaling function, \( \varphi(t) \) is a wavelet function, and \( h_k \) and \( c_k \) are scalar filters and \( m \) represents the number of subband. The scaling function and wavelet have finite support, if and only if the most of coefficients and are finite. Multi-resolution can be generated not just in the scalar context, i.e. with just one scaling function and one wavelet, but also in the vector case where there is more than one scaling functions and wavelets are involved. A multi-wavelet basis is characterized by \( n \) scaling and \( n \) wavelet functions. Here \( n \) denotes the multiplicity of the scaling functions and wavelets in the vector setting with \( n > 1 \). Multi-wavelet consists of several wavelet and scaling functions and are defined as
\[
\phi(t) = [\phi_1(t), \phi_2(t) \ldots \phi_n(t)]^T,
\]
(2.3)

\[
\varphi(t) = [\varphi_1(t), \varphi_2(t) \ldots \varphi_n(t)]^T,
\]
(2.4)

where \(\varphi(t)\) and \(\phi(t)\) are the multi-scaling and multi-wavelet functions, with \(k\) scaling and \(k\) wavelet functions.

\[
\phi(t) = \sqrt{2} \sum_k H_k \phi(mt - k),
\]
(2.5)

\[
\varphi(t) = \sqrt{2} \sum_k C_k \phi(mt - k),
\]
(2.6)

where \(H_k\) and \(C_k\) are \(k \times k\) matrix filters. Because of multiple filters, multiwavelets can possess symmetry, orthogonality and approximation orders higher than one simultaneously.

The multiwavelets are used to increase the accuracy and reduce the number of erroneous matches in the disparity maps. 2D multiwavelet transforms are separable and can be calculated by two 1D transforms. The 2D multiwavelet transform generates sixteen subbands in two levels, where four of them are approximation subbands. Approximation subband consist of different spectral content of the input image, while the remaining subbands mainly contain a mixture of horizontal, vertical and diagonal details of input image. In addition to this, the information in the base bands is less sensitive to the shift variability of the wavelets.

\textbf{2.2.1 Bi-Orthogonal Wavelet Transform}

The scaling equations on the scaling functions and wavelets show that the decomposition and reconstruction of a signal from a resolution to the next one is implemented by perfect reconstruction filter banks and is shown in Figure 2.1. The orthogonality property of wavelet gives a strong limitation
on the construction of wavelets. The biorthogonal wavelets has been con­
sidered to gain more exibility, which has both a dual scaling function and
a dual wavelet function that generate dual multiresolution analysis with
subspaces.

\[ \text{Figure 2.1: Analysis/synthesis stage of one level multiwavelet transform.} \]

2.3 Uncalibrated Stereo Image Rectification

Given a pair of stereo images, rectification determines a transformation of
each image plane such that pairs of epipolar lines become parallel to one
of the image axes. There are two ways of rectifying stereo images (i) cali­
brated stereo image rectification and (ii) uncalibrated stereo image rectifi­
cation (Chapter 1). In case of calibrated stereo image rectification, estimate
cameras intrinsic (focal length, aspect ratio, skew and principal point) and
extrinsic (rotation and translation) parameters using camera calibration
toolbox with the help of calibration object (checkerboard pattern). In un­
calibrated stereo image rectification, only image features are extracted to
estimate a suitable transformation. The Quasi-Euclidean epipolar uncal­
ibrated rectification method proposed by Andrea Fusiello et al., (2008) is
used for rectification of stereo images. Rectification is achieved by a suitable
rotation of both image planes. The correspondent image transformation
is the collineation induced by the plane at infinity. As a result, the plane
at infinity is the locus of zero-disparity in the rectified stereo pair. This is signified by saying that Euclidean rectification is done with respect to the plane at infinity. In the uncalibrated case the reference plane is generic, as any plane can play the role of the infinity plane in the projective space. We assume that intrinsic parameters are unknown and that a number of corresponding points \( m_l \leftrightarrow m_r \) are available. The fundamental matrix of a rectified pair has a very specific form, namely it is the skew-symmetric matrix associated with the cross-product by the vector \( u_i = (1, 0, 0) \):

\[
[u_1] \times = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}
\] (2.7)

Let \( H_r \) and \( H_l \) be the unknown rectifying collineations. When they are applied to the corresponding tie-points \( m_l, m_r \) respectively, the transformed points must satisfy the epipolar geometry of a rectified pair. Collineations induced by the plane at infinity, namely

\[
H_r = K_{nr} R_r K_{or}^{-1}, \quad H_l = K_{nl} R_l K_{ol}^{-1}.
\] (2.8)

The old intrinsic parameters \( (K_{or}, K_{ol}) \) and the rotation matrices \( (R_l, R_r) \) are unknown, whereas the new intrinsic parameters \( (K_{nl}, K_{nr}) \) can be set arbitrarily, provided that vertical focal length and vertical coordinate of the principal point are same. Each collineation depends in principle on five (intrinsic) plus three (rotation) unknown parameters. The rotation of one camera along its \( X \)-axis, however, can be eliminated. Consider the matrix

\[
F = K_{or}^T R_r^T [u_1] \times R_l K_{ol}^{-1}.
\] (2.9)

Let \( R_r' \) and \( R_l' \) be the same matrices as \( R_r \) and \( R_l \) after pre-multiplying with an arbitrary (but the same for both) rotation matrix about the \( X \)-axis. It is easy to verify that \( R_r^T [u_1] \times R_l = R_r'^T [u_1] \times R_l' \).
We further reduce the number of parameters by making as educated guess on the old intrinsic parameters: no skew, principal point in the center of the image, aspect ratio equal to one. The only remaining unknowns are the focal lengths of both cameras. Assuming that they are identical and equal to $\alpha$, we get:

$$K_{or} = K_{ol} = \begin{bmatrix} \alpha & 0 & \omega/2 \\ 0 & \alpha & \eta/2 \\ 0 & 0 & 1 \end{bmatrix}$$

(2.10)

where $w$ and $h$ are width and height (in pixel) of the image. In summary, the two collineations are parameterized by six unknowns: five angles and the focal length $\alpha$. Focal length is expected to vary in the interval $[1/3(\omega + h), 3(\omega + h)]$, so we consider instead the variable $\alpha' = \log_3(\alpha/(\omega + h))$ which varies in $[-1, 1]$. When $\alpha'$ converges outside the boundaries of the interval $[-1, 1]$ a random restart is attempted. If the problem persists the minimization is carried out with fixed $\alpha' = 0$. Finally, the new intrinsic parameters ($K_{nr}$ and $K_{nl}$) are set equal to the old ones: $K_{nr} = K_{nl} = K_{ol}$, that might be necessary to center the rectified images in the customary image coordinate frame.

### 2.4 Coarse to Fine Approach

For area-based matching, it is difficult to select a suitable size of window. In general, the window should be large enough to cover sufficient intensity variation and small enough to avoid over smoothing problem. However, choosing a suitable size of the window is application dependent. Even in the same image, the optimal size of the window is different from area to area. Kanade and Okutomi (1973) introduced a method to select a window size adaptively according to local intensity variation and dense disparity. However, it is not cheap to adaptively select the window size in terms of computational cost. Instead of working on choosing a suitable window size,
we propose to use a large window size in the top layer of the pyramid to obtain a smooth initial estimation of the dense disparity map, and then refine the disparity map in the finer layer of the pyramid to recover the depth discontinuity.

The problem of ambiguous matches in stereo images can also be avoided by using a coarse to fine search strategy is shown in Figure 2.2. The least numbers of disparities are checked at a given resolution, so that the lower the likelihood of an ambiguous match. Another benefit is that a wide range of disparities may be checked for relatively small cost; the coarsest levels have the least data. In coarse-to-fine stereo correspondence method initially images are brought into a pyramid representation where the base level captures the original image, while successive levels capture coarser resolutions with smaller format images via spatial subsampling, applied after low-pass or band-pass filtering (to avoid aliasing). The most commonly used pyramids are Quadtree (Jahne, 1993), Gaussian and Laplacian (Burt et al., 1983). In coarse-to-fine stereo corresponding operates by initially estimating disparity for lower resolution images (images of smaller size), then taking these disparities as an offset for refinement using higher-resolution images. The search range is shorter for low resolution images, because the size of images are shorter. This procedure is performed recursively by doing progressive matching starting from the coarsest pyramid level to the finest level. The refinement process is cheaper than calculation from scratch, because the local search range is smaller. This procedure is performed iteratively by matching starting from the coarsest level and use the matches obtained at the coarser level to guide the matching process gradually up to the finest level.

The coarse-to-fine method also helps to remove local minima in correspondence search by their reduction at the coarse level and allows for variable support aggregation as support region of the same size (in terms of pixels) yields greater smoothness at coarser levels. The coarse-to-fine area-based methods significantly improve on single-scale local matchers in
textureless regions, as they are able to aggregate greater support at coarse levels, but they cannot solve this problem completely (Anandan, 1989).

2.5 Matching strategy

In empirical evaluation we will concentrate on the closely related SAD, which yields to efficient implementation and offers increased robustness to outliers. In SAD algorithm, homologous point of reference image is selected by searching along the corresponding scan line in the other image, and within a certain disparity range, for the point that minimize (maximize) an error (similarity) function. Unlike algorithms based on bidirectional matching, sum of absolute difference algorithm (Stefano et al., 2004) uses
only a direct matching phase. Let's assume that the left image \( L(r) \) is the reference image, that disparity, \( d \); it belongs to the interval \([0 \ldots d_{\text{max}}]\). To estimate the depth image for the left image, the pixel at position \( r \) in the left image corresponds to the pixel at the position \( r + d \) in the right image \( R(r) \). The sum of absolute difference (SAD) used to find the pixel correspondence \( D \):

\[
D = \sum_r |L(r) - R(r + d)|.
\]  

(2.11)

2.6 Experimental Results

In order to estimate the efficiency of the proposed stereo correspondence algorithm, we conducted experiments using test images of stereo face database. We compared the proposed method with existing methods based only on visual comparison, because the ground truth disparity map of faces of stereo face database is not available, in order to employ RMSE and BPM. Hence, we conducted experiments on some of the images of Middlebury stereo database and evaluated the results using Root Mean Square Error (RMSE) and Bad Pixel Map (BPM).

2.6.1 Experiments on Stereo Face Images

In order to evaluate the performance of the proposed method on stereo face images, the coarse to fine search strategy using the wavelet based algorithm was applied to the stereo face images, which were taken from the stereo face database. The stereo images are first rectified to suppress the vertical displacement is shown in Figure 2.3. The biorthogonal discrete multiwavelet transform is then applied to rectified stereo images to decorrelate them into their subbands. Figure 2.4 shows the resulting subbands after applying a
two-level multiwavelet transform. In coarse to fine method, disparity is estimated at coarsest level is used to generate the disparity map up to finest level. Median filter is applied to have smooth disparity.

To validate the performance of the proposed method, the results of proposed method are compared with results of classical stereo matching technique, which perform matching in a spatial domain. We are able to evaluate by comparing visually the results of our method with the results of popular conventional correlation methods such as SAD, SSD and NCC. Figure 2.5 shows the disparity maps obtained using our approach. From Figure 2.5, it can be observed that the obtained disparity map using our approach looks visually good and suitable for face images compared to conventional correlation based approaches. The experiments are conducted by varying disparity range from 10 to 20. The experimental results show that, the good disparity map is obtained for disparity range 20. In face recognition system only face region is considered for recognition, so that the facial data is extracted from the face image of stereo database, which may contain outliers such as clothing and neck. The facial data is extracted by detecting nose tip to crop the original image in the 2D domain. To show the efficacy of our approach for varying window size, we conducted experiments with the variation of window size 3 x 3, 5 x 5, 7 x 7 and 9 x 9. From Figure 2.6, it is observed that even after varying the window size the disparity map looks visually similar in nature. This is due to the fact that, disparity map is computed by using coarse to fine refinement process.

Pose Variation

Normally, the face images suffer from pose variation due to rotation of head. The images taken at two different viewpoints of the same subject may appear more different than two images taken from the same view point for two different subjects. We present the experimental result carried out on some of the subjects with eight different poses of stereo database to evaluate effects of our approach. There is a pose variation in left as well as
Chapter 2. *Stereo Correspondence of Face Images Based on Wavelets*

**Figure 2.3:** Rectified stereo face images.

right images. Figure 2.7 shows the stereo face images with pose variation and the depth map computed using our approach.
Illumination Variation

In order to acquire the faces in controlled environment is very difficult. The illumination problem arises, when the same face appears differently due to the change in lighting variation. We have conducted experiments on stereo face database with varying illumination. To show the effect of our approach for illumination variation, the illumination of left image is varied using Adobe Photoshop software and right image is kept unaltered. The Figure 2.8 shows the pair of face images of stereo database used in the experiments with illumination variation and the corresponding estimated disparity map using our approach.

2.6.2 Experiments on Middlebury database

Experiment is also performed on some of the stereo test images from the Middlebury stereo database and compared with the corresponding ground truths. We have used stereo images viz. Teddy, Cones and Aloe (First Column of Figure 2.9) of Middlebury database. The corresponding ground truth disparity map is shown in the second column of Figure 2.9. The
Figure 2.5: Disparity Map: first column - Reference Image; second column - SAD; third column - SSD; fourth column - NCC; fifth column - Our approach using $7 \times 7$.

Third column shows disparity map obtained using proposed method. To evaluate the proposed method, bad pixel map is estimated between a result of proposed method and corresponding images ground truth. The fourth column of Figure 2.9 shows the bad pixel map. The figure shows that mapping of bad pixel is very less. Root mean square error method is a quantitative way to estimate the quality of the computed correspondences. The RMS error is calculates absolute difference between the computed depth map and the ground truth map that is

$$R = \left( \frac{1}{N} \sum_{(x,y)} |d_r(x,y) - d_T(x,y)|^2 \right)^{1/2}. \quad (2.12)$$
FIGURE 2.6: Disparity Map computed for cropped face image using our approach with different window size: First column - $3 \times 3$; second column - $5 \times 5$; third column - $7 \times 7$; fourth column - $9 \times 9$.

The Table 1.1 shows the root mean square error estimated for results obtained for proposed method SAD, SSD and NCC algorithm.

2.7 Processing Time

The processing time is calculated for our approach and compared with processing time of four state-of-the-art correlation methods using a PC with processor Intel core i3 with speed 2.27 GHz and 4 GB RAM. The proposed
Figure 2.7: Disparity Map computed using our approach for images with pose variation: first row of (a),(b),(c)-left images with pose variation, second row of (a),(b) and (c)-right images with pose variation, third row of (a),(b) and (c)-obtained disparity map.
Figure 2.8: Disparity Map computed using our approach for images with illumination variation; first column—left image with illumination variation, second column—right image, third column—obtained disparity map.
Table 2.1: Root Mean Square Error

<table>
<thead>
<tr>
<th>Methods</th>
<th>Teddy</th>
<th>Cones</th>
<th>Aloe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method</td>
<td>0.0652</td>
<td>0.0642</td>
<td>0.0348</td>
</tr>
<tr>
<td>SAD</td>
<td>0.0745</td>
<td>0.0823</td>
<td>0.045</td>
</tr>
<tr>
<td>SSD</td>
<td>0.0692</td>
<td>0.0789</td>
<td>0.042</td>
</tr>
<tr>
<td>NCC</td>
<td>0.0691</td>
<td>0.0792</td>
<td>0.039</td>
</tr>
</tbody>
</table>

Figure 2.9: Results on Middlebury datasets: From top to bottom: teddy, Cones and Aloe. From left to right: reference images, ground truth disparities, the results of the proposed algorithm and the error images where the black regions represent the erroneous pixels.
algorithm is implemented using MATLAB R2010a. The computational costs of proposed stereo correspondence method is approximately 16.15 seconds while SAD, SSD, and NCC spends 20.12, 23.35, and 32.72 seconds respectively.

2.8 Observations

From all the experimental results, the following observations were made:

- The proposed stereo correspondence approach effectively produces smooth disparity maps while preserving depth discontinuities well compared to conventional correlation measures for stereo face images.
- An initial disparity estimation is based on the largest window size, leading to a good recovery of large displacements.
- At each finer level, all coefficients, from the coarsest level to the current level, are updated so that not only the finer displacement components are recovered but also the errors occurring in the previous coarser level estimate may have a chance to be corrected.
- Dense displacements are obtained via interpolation at each resolution level, and the process can be terminated whenever the matching requirements are satisfied.
- The proposed method, however, is computationally a less expensive than other correlation methods because the disparity of high resolution image is obtained via interpolation.

2.9 Chapter Summary

In this chapter, we introduced an wavelet based coarse to fine approach to estimate disparity map for face stereo images. The approximation sub-bands of stereo images, produced by four levels discrete wavelet decompositions. Then coarse to fine approach is used estimate a disparity map.
The disparity map at coarse level is obtained by using SAD algorithm. The proposed approach is tested for stereo face images and middlebury stereo images. The experimental results were evaluated using evaluation techniques such as bad pixel map and Root Mean Square Error. The experiments were conducted to show the effectiveness of our approach for variation in illumination and blur. The results show that the good results are achieved in terms of quality under variation in illumination and pose variations in less computation time.