Stereo Image Matching Using Wavelet Scale-Space Representation

Asim Bhatti
Intelligent Systems Research Lab
Deakin University
asimbh@ieee.org

Saeid Nahavandi
Intelligent Systems Research Lab
Deakin University
nahavand@deakin.edu.au

Abstract

A multi-resolution technique for matching a stereo pair of images based on translation invariant discrete multi-wavelet transform is presented. The technique uses the well known coarse to fine strategy, involving the calculation of matching points at the coarsest level with consequent refinement up to the finest level. Vector coefficients of the wavelet transform modulus are used as matching features, where modulus maxima defines the shift invariant high-level features (multiscale edges) with phase pointing to the normal of the feature surface. The technique addresses the estimation of optimal corresponding points and the corresponding 2D disparity maps. Illuminative variation that can exist between the perspective views of the same scene is controlled using scale normalization at each decomposition level by dividing the details space coefficients with approximation space and then using normalized correlation. The problem of ambiguity, explicitly, and occlusion, implicitly, is addressed by using a geometric topological refinement procedure and symbolic tagging.

1. Introduction

Finding correct corresponding points from more than one perspective views in stereo vision is subject to a number of potential problems like occlusion, ambiguity, illuminative variations. A number of algorithms have been proposed to address at least some of these problems in stereo vision, the majority of them can be categorized into either area based or feature based algorithms. Area based approaches such as [18, 12] are based on the correlation of two image functions over locally defined regions, whereas feature based algorithms [2, 7] attempt to establish correspondences between the selected features, which are extracted from the images usually by using some explicit feature extraction algorithms. Both area-based and feature-based algorithms are considered local algorithms (LA). There exists another category of stereo vision algorithms, which are considered as global algorithms (GAs) [17, 15]. These GAs deal with the correspondence estimation process as global cost function optimization problem. These algorithms usually do not perform local search but rather try to find a disparity assignment that minimizes a global cost function. There is a clear consensus in the computer vision community that algorithms belonging to GAs group has overall better performance over the algorithms of LA group. However, GA algorithms are not free of shortcomings. GAs are very much dependent on how well the cost function represents the relationship between the disparity and some of its attributes, like smoothness and regularity. Furthermore, how close the cost function representation reflects the true nature of the disparity maps. However, the smoothness parameters, used by GAs, make disparity map smooth everywhere which may lead to poor performance at image discontinuities. Another disadvantage of GA algorithms is their computational complexity, which makes them unsuitable for real-time applications.

A promising way to improve the performance of the image-matching algorithms is to combine the best features of both techniques, i.e. LAs and GAs. This led to the development multi-resolution concept, which involves the matching of the two images at different resolutions and scales. In multi-resolution analysis, as is obvious from the name, the input signal is divided into different resolutions, i.e. scales and spaces [10], before the estimation of the correspondences. Multi-resolution algorithms can be considered as the middle way between these two broad classes of local search based algorithms, i.e. area based and feature based are considered as one extreme and global algorithms i.e. graph cuts, simulated annealing [16]. These algorithms do not explicitly state a global function that is to be minimized, but exhibits behavior close to that of the iterative optimization algorithms [5, 15]. The algorithm described in this paper is a variant of this class of techniques.

Considering the potential of the multiresolution analysis, especially in the context of multi-wavelet theory, a novel robust algorithm is presented. The proposed algorithm performs the coarse to fine search strategy using the multi-

Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)
0-7695-2606-3/06 $20.00 © 2006 IEEE

wavelet based scale space representation of the input signals. The presented work is the continuation of the work by the authors [3, 11], where multi-wavelet coefficients were used as corresponding features. In the proposed work the concept of Multi-wavelet Modulus Maxima (**MWMM**) is introduced with significant improvement in the disparity estimation. The proposed algorithm involves the matching of coefficients, with magnitude and phase, pointing to the normal of the edge surfaces. The selection of the potential candidates is performed using a robust selection criterion which involves the contribution of normalized correlation, probability of occurrence, symbolic tagging and geometric refinement. This selection criterion ensures the involvement of only the most consistent corresponding candidates throughout the iterative procedure.

2. Wavelets and Multiwavelets Fundamentals

Classical wavelet theory is based on the refinement equations as given below

$$\phi(t) = \sum_k c_k \phi(Mt - k) \tag{1}$$

$$\psi(t) = \sum_k w_k \phi(Mt - k) \tag{2}$$

where c_k and w_k represents the scaling and wavelet coefficients. 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 function and wavelet are involved. The latter case leads to the notion of multi-wavelets. A multi-wavelet basis is characterized by r scaling and r wavelet functions. Here r denotes the multiplicity in the vector setting with $r > 1$. Multi-scaling functions satisfy the matrix dilation equation as

$$\Phi(t) = \sum_k C_k \Phi(Mt - k) \tag{3}$$

Similarly for the multi-wavelets the matrix dilation equation can be expressed as

$$\Psi(t) = \sum_k W_k \Phi(Mt - k) \tag{4}$$

where C_k and W_k are real $r \times r$ and $(M-1)r \times r$ matrices of multi-filter coefficients, respectively, whereas M represents the number of bands. In this work we are dealing with only two band multiwavelets, i.e., $M = 2$, defining equal number of multi-wavelets as multi-scaling functions. For more information, about the generation and applications of multi-wavelets with, desired approximation order and orthogonality, the interested readers are referred to [4]. Wavelet transform results in the information belonging to the approximation space A_k and detail space D_k possessing the property

$$A_{k-1} = A_k \oplus D_k \tag{5}$$

One of the most widely used discrete wavelet transform representation is introduced by Mallat [10]. Using Mallat's representation, the details space, i.e. D_k in (5), consists of three components, which are horizontal, vertical and diagonal. In general terms, the wavelet transform modulus maxima (**WTMM**) can be described at any point (s, k) such that the magnitude of wavelet transform modulus (**WTM**), i.e. $|WT_{s,k}|$ is a local maximum, i.e.

$$\frac{\partial WT_{s,k}}{\partial k} = 0 \tag{6}$$

Furthermore, the vector coefficients of the **WTM** can be expressed as

$$WT_{s,k} = |WT_{s,k}| \angle \Theta_W \tag{7}$$

where $|WT_{s,k}|$ represents the magnitude of the **WTM**, where as s represents the scale and k for the coefficient under consideration. Furthermore, the magnitude of **WTM** can be expressed in terms of horizontal and vertical details spaces as

$$WT_{s,k} = \sqrt{|W_{h,s}^2| + |W_{v,s}^2|} \tag{8}$$

where $W_{h,s}$ and $W_{v,s}$ are horizontal and vertical detail spaces, respectively. Similarly, the phase of the **WTM** can be expressed as

$$\Theta_W = \begin{cases} \alpha(k) & \text{if } W_{h,s} > 0 \\ \pi - \alpha(k) & \text{if } W_{h,s} < 0 \end{cases} \tag{9}$$

where

$$\alpha(k) = tan^{-1}\left(\frac{W_{v,s}}{W_{h,s}}\right) \tag{10}$$

The vector $\vec{n}(k)$ points to the direction normal to the edge surface as

$$\vec{n}(k) = [cos(\Theta_W), sin(\Theta_W)] \tag{11}$$

An edge point is a point p at some scale s such that **WTMM** is a local maximum at $k = p$ and $k = p + \epsilon\vec{n}(k)$ for ϵ small enough. These points are called wavelet transform modulus maxima (**WTMM**), and are shift invariant features [10]. **WTMM** are also defined as multi-scale edges, which are the discontinuities of the decomposed signals that appear in different scales and resolution. Furthermore, **WTMM** represent multi-scale features that are invariant to shifts between the input signals.

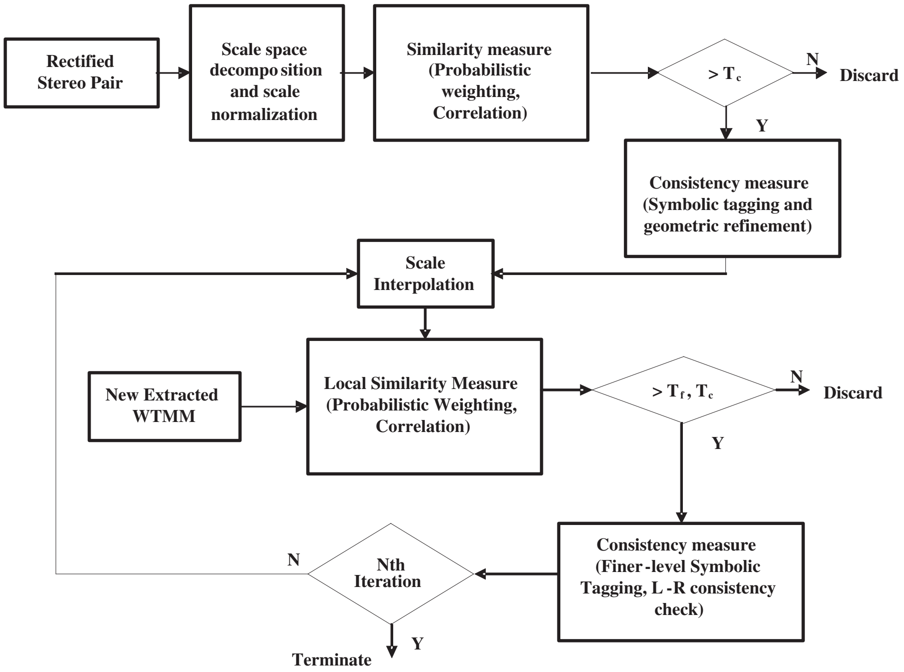
Rectified Stereo Pair
Scale space decompo sition and scale normalization
Similarity measure (Probabilistic weighting, Correlation)
> T$_c$
N
Discard
Y
Consistency measure (Symbolic tagging and geometric refinement)
Scale Interpolation
New Extracted WTMM
Local Similarity Measure (Probabilistic Weighting, Correlation)
> T$_f$, T$_c$
N
Discard
Y
Consistency measure (Finer -level Symbolic Tagging, L -R consistency check)
N
Nth Iteration
Y
Terminate

Figure 1. The block diagram of the algorithm

3 Stereo Matching

Before starting the matching process scale normalization is performed in order to minimize the effect of illuminative variation that can exist between the stereo image pairs. The scale normalized wavelet coefficients can be defined as

$$NW_{s,k} = \frac{W_{s,k,i}}{|A_{s,k}|} \quad \forall i \,\epsilon\, \{h, v, d\} \qquad (12)$$

where $\{h, v, d\}$ represents horizontal, vertical and diagonal details, respectively, s represents the scale of decomposition, k represents the k th coefficient and A represents the approximation space. A block diagram of the complete stereo matching algorithm is presented in Figure 1 for better understanding.

3.1 Similarity Measure

The scale normalization step results in r^2 search spaces for each image of the stereo pair. These search spaces are scaled down versions of the original images. For detailed explanation of the scale-space representation using wavelet transform, readers are kindly referred to [**?**, 10]. To find the similarity between the stereo image pair, normalized correlation is performed for each **WTMM** in the reference image.

3.2 Consistency Measure

To keep the matching process consistent, a symbolic tagging procedure is introduced based on probability of occurrence and three different thresholds. Probability of occurrence (*POC*) is the probability of selection of a candidate from any of the search space out of r^2 search spaces. Probability of occurrence can be defined as

$$P_c(C_i) = n_{C_i}/r^2 \qquad \text{where} \qquad 1 \leq n_{C_i} \leq r^2 \quad (13)$$

where n_{C_i} is the number of times a candidate C_i is selected and r is the multiplicity of multi-filter coefficients. As all matching candidates have equal probability of being selected so the probability of occurrence for any candidate through one search space is $1/r^2$. The correlation score for each candidate is then weighted with the occurrence probability, which can be expressed as

$$CS_{C_i} = P_c(C_i) \sum_{n_{C_i}} NC_{C_i}(x, d) \;\forall_{\, n_{C_i} \in \mathbb{Z} \,:\, n_{C_i} \leq r^2} \qquad (14)$$

After that step, all candidate features points have a correlation score attached and are then divided into two pools based on three different thresholds T_i possessing the criteria $T_1 < T_2 < T_3$. The values of these thresholds is

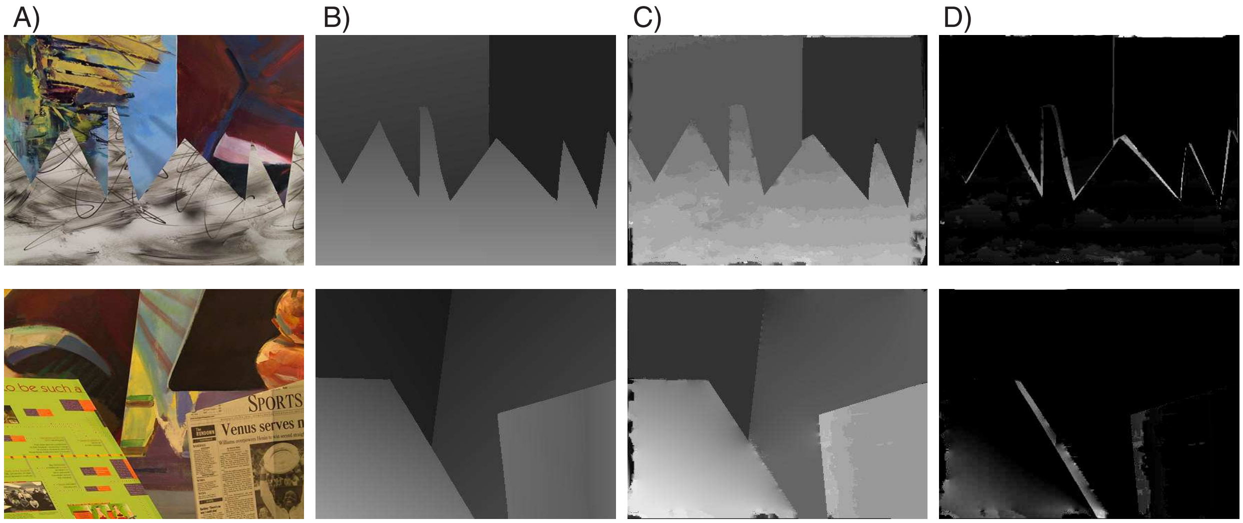

Figure 2. a: right images of the Sawtooth (up) and Venus (down) stereo pair, b: Respective ground truth disparity maps, c: Respective estimated disparity maps, d: Respective disparity error

usually within the ranges of [0.5 0.6], [0.75 0.85] and [0.9 0.95] respectively. First threshold T_1 is applied just after the correlation step to filter out the bad matches. Remaining candidate feature points are then assigned symbolic tags depending on there consistency as given below

$$\begin{array}{r}NC(x,d) \;\geq\; T_{c_1}, \quad \text{and} \quad P_c(C_i) \;=\; 1, \Longrightarrow \boldsymbol{Op} \\ NC(x,d) \;\geq\; T_{c_1}, \quad \text{and} \quad 0.5 \;\leq\; P_c(C_i) \;<\; 1, \Longrightarrow \boldsymbol{Cd} \\ NC(x,d) \;\geq\; T_{c_2}, \quad \text{and} \quad 2/r^2 \;\leq\; P_c(C_i) \;<\; 0.5, \Longrightarrow \boldsymbol{Cr}\end{array} \tag{15}$$

It can be seen from the first expression in (15), there is no ambiguity for the matches with tag ***Op*** as the **POC** is 1, whereas ambiguity does exist for the matches with tags ***Cd*** and ***Cr***. These tags helps in addressing the problem of ambiguity. The candidates with tag ***Op*** are considered to be the optimal candidates and are used as reference for the remaining candidates.

3.3 Geometric Refinement

To deal with this ambiguity problem, a simple geometric topological refinement is performed to pick the most suitable or optimal match out of the pool of candidate matches. For that purpose, the geometric orientation of the candidates with reference to the optimal candidates, i.e. with tag ***Op***, is checked and the candidate pair having the closest geometric topology with respect to these points is selected as an optimal candidate pair. Three geometric features: relative distance, absolute distance and the slope, are calculated with respect to the reference points to check the topological similarity. The topological measure is then weighted with the correlation score to consider the previous achievement of each candidate. The score for each candidate pair PS_i is then calculated as given below

$$Gc_i \;=\; CS_{C_i}\,\overline{(e^{-d_{A_{C_i}}} + e^{-d_{R_{C_i}}} + e^{-d_{S_{C_i}}})} \tag{16}$$

The term $\overline{(.)}_n$ defines the average over n repetitions where n is usually taken within the range of $[3\ 5]$. The reason of selecting these geometric features for addressing the problem of ambiguity is there invariance through many geometric transformations, as Projective, Affine, Matric and Euclidean.

3.4 Scale Interpolation

The matching process at the coarsest level ends up with a number of matching pairs, which needs to be interpolated to the finer level. The constellation relation between the coefficients at coarser and finer levels can be visualized by taking the decimation of factor 2 into consideration. The following relation updates the disparity from coarser disparity d_c to finer disparity d_F, as follows

$$d_F \;=\; d_L \;+\; 2d_c \tag{17}$$

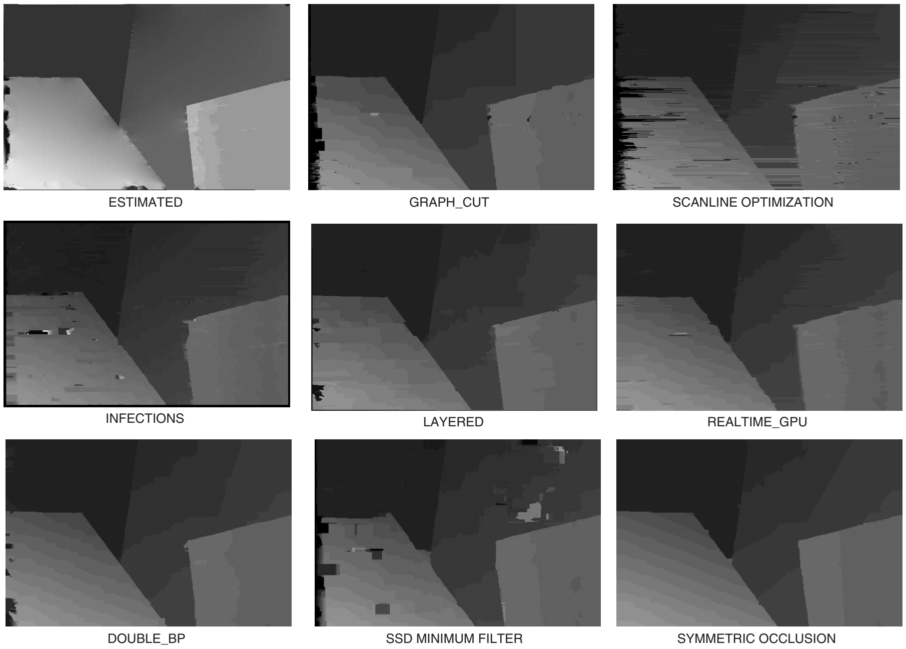

Figure 3. Comparison of estimated disparity map with a number of existing algorithms for image Venus

Table 1. A comparison of the estimated disparity with a number of existing well known algorithms

Algorithms	Est.	dB [13]	Gc [16]	If [6]	Ld [9]	RG [1]	SO [16]	SF [16]	SO [8]
R	1.9885_1	2.2114_3	3.3977_5	4.4952_8	3.1955_4	2.0780_2	4.2491_7	3.7333_6	15.7478_9
B	0.0262_1	0.2860_3	0.3065_5	0.3119_7	0.3186_8	0.2609_2	0.3090_6	0.2960_4	1.0000_9

where d_L is the local disparity obtained within the interpolated area. After the matches are interpolated to the finer level, correlation process is performed locally, only for the locations having their corresponding pairs at the coarser level. Hence refining the matches up to the finest level and leaving most consistent matches at the end of the process.

4 Results

The algorithm presented in this work, is evaluated in terms of the disparity estimation performance on the number of images taken from [14]. These images are designed exclusively for the purpose of performance evaluation of the stereo matching algorithms. All calculated disparities, given in this chapter, use right images as the reference images. Therefore all calculated disparities are right image disparities. As an example two images are shown in Figure 3, with related ground truth disparity maps, estimated disparity maps and the error between the ground-truth and the estimated ones. These images are known as *Sawtooth* and *Venus* images. As it can be seen in Figure 3 the edges of the discontinuities are extracted to high accuracy and estimated disparity is very much similar to the ground truth disparity.

To further validate the claims about the performance of the proposed algorithm a comparison is performed. This comparison is performed between the proposed algorithm and a number of selected algorithms from the literature, which are very well known algorithms in terms of their disparity estimation performance. Disparity maps related to the chosen algorithms are also shown as a comparative measure, taken from [14]. The estimated disparity map is related to the image **Venus**. The chosen algorithms for comparison purpose are *Double-bp* [13], *Graph Cuts* [16], *Infection* [6], *Layered* [9], *Realtime-Gpu* [1], *Scanline Optimization* [16], *SSD min. Filter* [16] and *Symmetric-Occlusion* [8]. To create a better understanding of the comparison, statistic R, root mean square error, and B, percentage of bad disparities, are calculated as

$$R = \sqrt{\frac{1}{N}\sum_{x,\,y} |d_E(x,y) - d_G(x,y)|^2} \qquad (18)$$

and

$$B = \frac{1}{N} \sum_{(x,y)} |d_E(x,y) \,-\, d_G(x,y)|^2 \,>\, \xi \qquad (19)$$

where d_E and d_G are the estimated and ground truth disparity maps, N is the total number of pixels in an image whereas ξ represents the disparity error tolerance (DET) and is taken as 1. The statistics R and B related to all of the above algorithms is presented in Table 1.

5 Conclusions

The proposed algorithm uses the stereo vision capabilities and multi-resolution analysis to estimate disparity maps and the concerned 3D depths. The proposed technique addresses the estimation of optimal corresponding points leading towards the construction of optimal disparity maps. The algorithm introduced a new comprehensive selection criterion called the strength of the candidate, which involves the contribution of probabilistic weighted normalized correlation, symbolic tagging and geometric refinement. The geometric features used in the geometric refinement procedure are carefully chosen to be invariant through many geometric transformations, such as affine, matric, Euclidean and projective. This comprehensive selection process helps in addressing the problem of ambiguity explicitly and occlusion implicitly and helps to extract the most optimal corresponding points from the two perspective views.

References

[1] Anonymous. High-quality real-time stereo using graphics hardware. In *CVPR*, 2006.

[2] H. Baker and T. Binford. Depth from edge and intensity based stereo. In *Int. Joint Conf. on Artificial Intelligence*, pages 631–636, Vancouver, Canada, 1981.

[3] A. Bhatti and S. Nahavandi. Accurate 3d modelling for automated inspection: A stereo vision approach. 2005.

[4] A. M. Bhatti. *A NEW METHOD FOR GENERATING MULTIWAVELETS WITH APPROXIMATION ORDER*. Ms thesis, Eastern Mediterranean University, 2000.

[5] T. K. C. L. Zitnick. A cooperative algorithm for stereo matching and occlusion detection. *IEEE PAMI*, 22(7):675–684, 2000.

[6] C. P. G. Olague, F. Fernndez and E. Lutton.

[7] W. E. L. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE Trans. Pattern Analysis and Machine Vision*, 7(1):17–34, 1985.

[8] S. K. J. Sun, Y. Li and H.-Y. Shum.

[9] M. U. S. W. L. Zitnick, S.B. Kang and R. Szeliski.

[10] S. Mallat. *A Wavelet Tour of Signal Processing*, volume 2nd edition. Academic Press, 1999.

[11] A. B. Nahavandi and Saeid. A multi-wavelet based technique for calculating dense 2d disparity maps from stereo. In WAC, editor, *World Automation Congress (WAC2004)*, 2004.

[12] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 7(2):139–154, 1985.

[13] R. Y. H. S. Q‘?ngxiong Yang, Liang Wang and D. Nister. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In *CVPR*, 2006.

[14] D. Scharstein and R. Szeliski. www.middlebury.edu/stereo.

[15] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *Int. J. of Computer Vision*, 28(2), 1998.

[16] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. of Computer Vision*, 47:7–42, 2002.

[17] O. V. Y. Boykov and R. Zabih.

[18] O. F. Z. Zhang, R. Deriche and Q. T. Luong. Technical report.