IMAGE-GUIDED NON-LOCAL DENSE MATCHING WITH THREE-STEPS OPTIMIZATION

Xu Huang[a], Yongjun Zhang[a, *], Zhaoxi Yue[a]
[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China

Commission Ⅲ, WG Ⅲ/1

KEY WORDS: HOG; Image Guided Matching; Non-local Dense Matching; SGM; Image Guided Interpolation

ABSTRACT:

This paper introduces a new image-guided non-local dense matching algorithm that focuses on how to solve the following problems: 1) mitigating the influence of vertical parallax to the cost computation in stereo pairs; 2) guaranteeing the performance of dense matching in homogeneous intensity regions with significant disparity changes; 3) limiting the inaccurate cost propagated from depth discontinuity regions; 4) guaranteeing that the path between two pixels in the same region is connected; and 5) defining the cost propagation function between the reliable pixel and the unreliable pixel during disparity interpolation. This paper combines the Census histogram and an improved histogram of oriented gradient (HOG) feature together as the cost metrics, which are then aggregated based on a new iterative non-local matching method and the semi-global matching method. Finally, new rules of cost propagation between the valid pixels and the invalid pixels are defined to improve the disparity interpolation results. The results of our experiments using the benchmarks and the Toronto aerial images from the International Society for Photogrammetry and Remote Sensing (ISPRS) show that the proposed new method can outperform most of the current state-of-the-art stereo dense matching methods.

1. INTRODUCTION

Stereo dense matching, which aims to find pixel-wise correspondences between a stereo pair, has attracted increasing attention in the photogrammetry and computer vision communities for decades. Although stereo matching is maturing, research gaps remain in the aspects related to improving the performance of cost computation, matching accuracy in homogeneous intensity regions, and disparity interpolation. Most stereo matching algorithms currently consist of four steps: 1) cost computation, 2) cost aggregation, 3) disparity computation, and 4) disparity refinement (Scharstein and Szeliski, 2002).

1.1 Review of previous work

Cost is defined as the measure to describe the similarity between correspondences. The performance of cost computation methods are affected by the radiometric conditions of the image. For example, when the image radiometric condition is good, the absolute difference (AD), the insensitive measure of Birchfield and Tomasi (BT), or the gradient measure can achieve accurate matching results (Meiet et al, 2011). When the image radiation varies, zero-based normalized cross correlation (ZNCC) and normalized gradient (Zhou and Boulanger, 2012) are often used to compensate for linear radiation distortions between correspondences, while Census (Zabih and Woodfill, 2005; Jiao et al., 2014; Kordelas et al, 2015), mutual information (Paul et al, 1997), and image radiation correction (Jung et al, 2013) are insensitive to nonlinear radiation distortion. Hirschmuller evaluated the popular cost computation methods and concluded that Census and mutual information measures can achieve the best matching results under varying radiometric conditions (Hirschmueller and Scharstein, 2009). In addition to image radiation, cost computation also is influenced by the vertical parallax of the epipolar stereos, whereby the intensity distribution of the central pixel may be different from that of its correspondence in rich texture areas, thereby making the corresponding cost computation based on local metrics potentially unreliable. In order to make the cost computation more robust, most stereo matching methods choose to enlarge the window of cost computation. However, there is no consensus on the appropriate size. In addition, the run time of cost computation will increase with the growing size of the window. The disparities of the pixels in a large window may not be consistent, which may blur the edges in depth discontinuity regions. Thus, investigating a new cost computation method that can weaken the influence of vertical parallax is worthwhile.

In 2002, Scharstein and Szeliski (2002) divided stereo matching methods into two types, local methods and global methods, according to the cost aggregation pattern. In recent years, many researchers have developed local methods into non-local methods successfully. As local methods and the later non-local methods are essentially image-guided, and global methods or semi-global methods (SGM) are based on the minimization of energy function (Taniai et al, 2014; Yang et al, 2009; Hirschmuller, 2008). Stereo matching methods also can be divided into image-guided methods and energy function-guided methods. Image-guided stereo matching methods suppose that all the pixels with similar intensities in the support window or homogeneous intensity regions have the same disparities. The cost aggregation is guided by the image intensity. Local methods, such as the bilateral filter (Yoon and Kweon, 2006) and the image-guided filter (He et al, 2013), have achieved matching results as good as those of global methods since 2006. However, the local methods need to define the window size first. Large windows take more time for cost aggregation while small windows perform badly in textureless areas. In order to avoid the window definition, the non-local methods, which are based on recursion, were proposed (Yang, 2015; Pham and Jeon, 2013; Cigla and Alantan, 2013; Sun et al, 2014; Cheng et al, 2015), which differed from the local methods in that the cost aggregation of every pixel is supported by the remaining pixels in the whole image for non-local methods. The supports from the remaining pixels depend on the intensity similarity and the cost aggregation path. The non-local methods are fast and perform well in depth discontinuity regions or textureless regions with consistent disparities. However, none of the current methods consider that the intensities of pixels may be similar but the corresponding disparities may change smoothly or sharply, in which case the non-local methods may perform badly. Thus, the non-local methods need to be improved in order to avoid the problem of inconsistent disparities. Image-

guided matching works well in depth discontinuity areas, and energy function-guided matching can achieve more robust matching results. If both methods are combined, more accurate matching results can be achieved. The literature indicates that the local methods and the Global/SGM methods have been combined successfully (Mei et al, 2011; Žbontar and LeCun, 2015; Mozerov and van de Weijer, 2015). However, combining non-local methods and Global/SGM methods has yet to be achieved.

Initial disparity images, after cost aggregation and disparity computation, still have many mismatches that must be eliminated by outlier detection, such as a left-right consistency check and removal of peaks. The outlier detection may invalidate some disparities and may lead to holes in the disparity image, which then needs to be interpolated for a dense result. Interpolation methods can be divided into disparity image-based interpolation and intensity image-guided interpolation. Disparity image-based interpolation methods classify invalid disparities into mismatches and occlusions. The interpolation of invalid disparities is based on the neighbourhood valid pixels (Mei et al., 2011; Hirschmuller, 2008). This method is fast and can achieve good interpolation results for good initial disparity images. However, if the valid pixels are insufficient, the interpolation results will be unsatisfactory. Intensity image-guided interpolation supposes that pixels with similar intensities have consistent disparities (Yang, 2015). The valid disparities are propagated from valid pixels to invalid pixels with similar intensities. Although there are less valid pixels, this method can still acquire satisfactory interpolation results. However, intensity image-guided methods may be invalidated in the homogenous intensity region with inconsistent disparities. Instead, it is more reasonable to allow disparity changes for pixels with similar intensities. Thus, a new intensity image guided interpolation method which can satisfy the new hypothesis is needed.

1.2 Contribution of this paper

This paper proposes a new image-guided non-local dense matching method with a three-step optimization based on the combination of image-guided methods and energy function-guided methods. The image-guided non-local method with a three-step optimization (INTS) consists of the following steps: (1) a new non-local method is adopted to strengthen the cost propagation in the homogenous intensity regions; (2) the semi-global matching method (SGM) (Hirschmuller, 2008) is used to guarantee cost propagation in the area with rich textures; and (3) a new intensity image-guided interpolation method is utilized to interpolate invalid disparities, from which the final matching results are acquired. The main contributions of this paper are as follows:

- This paper improves the histogram of the oriented gradient (HOG) feature, which is capable of being linear radiometric invariant. The improved HOG feature is used to compute costs, and this is the first time the HOG feature is used as the cost metric, which is able to reduce the influence of vertical parallax to cost computation.
- This paper proposes a new image-guided non-local matching method where penalty terms are introduced during matching in homogenous intensity regions. A new kernel function is proposed, which is more robust than the Gaussian kernel function. The costs propagated from depth discontinuity regions are limited. New cost propagation paths are available, which can guarantee the connectivity of the path between the homogenous pixels. The new method can avoid the mismatches in homogenous intensity regions with inconsistent disparities.
- This paper proposes a new intensity image-guided interpolation method. The new method defines new rules of propagation from valid pixels to invalid pixels, which can improve the interpolation accuracy.

2. PROPOSED METHOD

2.1 Cost Computation

HOG is a well-known feature descriptor for object detection, which has been successfully used in image recognition (Triggs and Rhone-Alps, 2005). As HOG can describe object features accurately, it also has been used as a cost metric. However, computing the complete HOG features for every pixel is very time-consuming. HOG features also are radiometric variant, which may be unreliable due to varying radiometric conditions. Thus, in this paper, an improved HOG feature is proposed which is not only fast but also is linear radiometric invariant.

This paper supposes that the radiation distortion is linear in a very small window, such as a 3×3 neighbourhood. The linear radiation distortion can be expressed as the following equation:

$$g_r(\boldsymbol{p}_r) = c \cdot g_l(\boldsymbol{p}_l) + \boldsymbol{t} \tag{1}$$

where, $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$ represent the correspondences in left and right images, respectively; when stereo pairs are epipolar images, the relationship between their horizontal ordinates is $p_{rx} = p_{lx} - d$, where d represents the disparity; g_l and g_r represent the intensities of the correspondences respectively; c and $\boldsymbol{t}$ are the coefficients of the linear radiation distortion model.

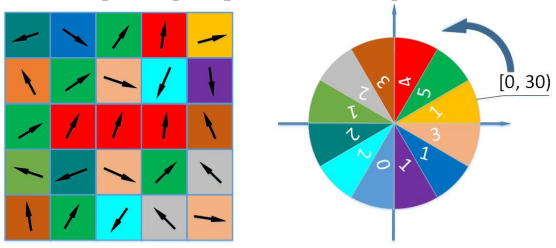
The translation factor $\boldsymbol{t}$ can be eliminated by gradient computation in a local small window. The Sobel operator is used in this paper. In order to eliminate the scale factor c further, gradient directions are calculated as follows:

$$\begin{aligned} \theta_r(\boldsymbol{p}_r) &= \theta_l(\boldsymbol{p}_l) \\ \theta_r(\boldsymbol{p}_r) &= \arctan\left(Gy_r(\boldsymbol{p}_r) / Gx_r(\boldsymbol{p}_r)\right) \\ \theta_l(\boldsymbol{p}_l) &= \arctan\left(Gy_l(\boldsymbol{p}_l) / Gx_l(\boldsymbol{p}_l)\right) \end{aligned} \tag{2}$$

where, $Gx_l(\boldsymbol{p}_l)$ and $Gx_r(\boldsymbol{p}_r)$ represent the horizontal gradients of $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$ respectively; $Gy_l(\boldsymbol{p}_l)$ and $Gy_r(\boldsymbol{p}_r)$ represent the vertical gradients of $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$ respectively; $\theta_l(\boldsymbol{p}_l)$ and $\theta_r(\boldsymbol{p}_r)$ represent the gradient directions of $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$ respectively. The gradient direction is a linear radiometric invariant metric; and the range of gradient directions is [0, 360°).

We define a $W\times W$ window centred at pixel $\boldsymbol{p}$ as a basic description cell. All the gradient directions in the cell are counted, and then a gradient direction histogram is built, as shown in Figure 1. The range of gradient directions is divided into 12 bins. The initial count of every bin is set as zero. When the gradient direction in the cell belongs to a certain bin, add 1 to the corresponding count. Figure 1(a) represents the description cell, where the rectangle represents a pixel in the cell; the direction of the arrow represents the gradient direction; the background colours of the pixels correspond one to one with the bins of the gradient direction histogram. Figure 1(b) represents the gradient direction histogram, where the number in each bin represents the corresponding count. Finally, the

count in each bin is divided by the sum of the pixels in the cell for normalization. According to the normalized gradient direction histogram, a 12×1 vector can be constructed as the feature descriptor of pixel $\boldsymbol{p}$, as shown in Equation (3).



(a) (b)

Figure 1. Gradient Direction Histogram

$$V_{HOG}(\boldsymbol{p})=(b_0,b_1,b_2,b_3,b_4,b_5,b_6,b_7,b_8,b_9,b_{10},b_{11})^T \quad (3)$$

where, b_i (i = 0~11) represents the value in each bin of the normalized gradient direction histogram; and $V_{HOG}(\boldsymbol{p})$ represents the feature descriptor of pixel $\boldsymbol{p}$.

When compared with the traditional HOG feature, the improved HOG neglects the gradient norm during the construction of the histogram, which makes it possible for the improved HOG to be linear radiometric invariant. Also, the traditional HOG needs to combine several basic description cells into a block to construct a larger feature descriptor. As the cost aggregation step of stereo matching can aggregate the cost in a neighbourhood together, the improved HOG only describes pixel features at the level of cells instead of in blocks, which can avoid repeated calculations.

This paper regards the distance between the HOG feature descriptors of the correspondences as cost metrics, as shown in Equation (4). In order to describe the pixel features more precisely, this paper combines HOG and Census as the final cost metric, as shown in Equation (5).

$$C_{HOG}(\boldsymbol{p}_l,d)=\left\|V^l_{HOG}(\boldsymbol{p}_l)-V^r_{HOG}(epl(\boldsymbol{p}_l,d))\right\| \quad (4)$$

where, $epl(\boldsymbol{p}_l,d)$ represents the suspected corresponding pixel of $\boldsymbol{p}_l$ with d as a disparity, namely, $\boldsymbol{p}_r=epl(\boldsymbol{p}_l,d)$; $C_{HOG}(\boldsymbol{p}_l,d)$ represents the HOG cost metric between pixel $\boldsymbol{p}_l$ and $\boldsymbol{p}_r$; $V^l_{HOG}(\boldsymbol{p}_l)$ represents the HOG feature descriptor of $\boldsymbol{p}_l$ in the left image; and $V^r_{HOG}(epl(\boldsymbol{p}_l,d))$ represents the HOG feature descriptor of $\boldsymbol{p}_r$ in the right image.

$$\begin{aligned}&C(\boldsymbol{p}_l,d)\\&=q\cdot\min\left(C_{Census}(\boldsymbol{p}_l,epl(\boldsymbol{p}_l,d)),t_{Census}\right)/t_{Census}\cdot t_{HOG}\\&\quad+(1-q)\cdot\min\left(C_{HOG}(\boldsymbol{p}_l,epl(\boldsymbol{p}_l,d)),t_{HOG}\right)\end{aligned} \quad (5)$$

where, $C(\boldsymbol{p}_l,d)$ represents the cost of $\boldsymbol{p}_l$ with d as a disparity; C_{Census} represents the cost calculated by Census metrics; C_{HOG} represents the cost calculated by HOG metrics; t_{Census}, t_{HOG} represent the truncation thresholds; and q represents the weight coefficient whose range is [0, 1]. In order to make C_{Census} and C_{HOG} in the same range, C_{Census} is scaled by t_{HOG}/t_{Census}.

The difference between the y coordinates of correspondences in stereo epipolar pairs is called vertical parallax. The vertical parallax has a corrupt influence on the cost computation, especially in the areas with rich textures. The HOG features can reduce the influence of vertical parallax to a certain extent. Each gradient direction computation, except for the central gradient, is independent of the central pixel in the basic cell of HOG; and the vertical parallax of the central pixel therefore cannot affect the computation of other gradient directions. Each bin in the gradient direction histogram also has a 30-degree range. When the gradient direction changes because of vertical parallax, the computation of the HOG features is not affected as long as the change is no more than the range. In Section 3.1, experiments on stereo epipolar images with vertical parallax are discussed.

2.2 Image-guided Non-local Matching

In recent years, several image-guided non-local methods were proposed, of which the basic mathematic models are consistent essentially (Yang, 2015; Pham and Jeon, 2013; Cigla and Alantan, 2013; Sun et al, 2014; Cheng et al, 2015):

$$L_r(\boldsymbol{p},d)=C(\boldsymbol{p},d)+T\cdot L_r(\boldsymbol{p}-1,d) \quad (6)$$

where, $L(\boldsymbol{p},d)$ represents the aggregated cost of pixel $\boldsymbol{p}$ at disparity d in the current path; r represents the direction of the path; $C(\boldsymbol{p},d)$ represents the cost of pixel $\boldsymbol{p}$ at disparity d; $\boldsymbol{p}-1$ represents the previous pixel in the current path; and T represents the intensity similarity of pixel $\boldsymbol{p}$ and $\boldsymbol{p}-1$, which is used to constrain the cost propagation between $\boldsymbol{p}$ and $\boldsymbol{p}-1$. Generally, T is computed by the Gaussian kernel function:

$$T=T_G(\boldsymbol{p},\boldsymbol{p}-1)=\exp\left(-|g(\boldsymbol{p})-g(\boldsymbol{p}-1)|/\sigma\right) \quad (7)$$

where, T_G represents the constraint term T which is computed by the Gaussian kernel function; g represents the image intensity; and σ represents the smooth term.

The value of T is close to 1 in homogenous intensity regions, which means that the non-local methods force disparities in the homogenous regions to be consistent. It is obviously improbable because the surface of textureless areas in the real world may be uneven so the corresponding disparities should be inconsistent.

In order to solve the matching problem caused by textureless areas with inconsistent disparities, this paper introduces penalty term P_1 and P_2 into Equation (6), as shown in Equation (8). Compared to Equation (6), Equation (8) considers the change in disparities in homogenous intensity regions. When the neighbour disparity changes smoothly, a lower penalty P_1 is used for slanted or curved surfaces. When the disparity changes sharply, a larger penalty P_2 is used for depth discontinuities.

$$L_r(\boldsymbol{p},d)=C(\boldsymbol{p},d)+T\cdot\min\begin{Bmatrix}L_r(\boldsymbol{p}-1,d),\\L_r(\boldsymbol{p}-1,d-1)+P_1,\\L_r(\boldsymbol{p}-1,d+1)+P_1,\\\min\limits_k L_r(\boldsymbol{p}-1,k)+P_2\end{Bmatrix} \quad (8)$$

When pixel $\boldsymbol{p}$ and $\boldsymbol{p}+1$ is located in the same homogenous region, traditional non-local methods fully trust the cost propagated from $\boldsymbol{p}$ and add a large constraint term $T(\boldsymbol{p}+1,\boldsymbol{p})$. However, the cost propagated from the pixel in the same region

may be unreliable due to the depth discontinuities. The cost computation is unreliable if the disparities in the cost metric window are discontinuous (e. g., object edges). The improper cost of pixels around depth discontinuity region is the cause of the fattening problem. This paper introduces a new approach to compute constraint term T, which can reduce the fattening problem greatly. When the cost is propagated from $\boldsymbol{p}$ to $\boldsymbol{p}+1$, the absolute differences of image intensities between pixel $\boldsymbol{p}+1$ and the previous $s+1$ pixels are calculated, respectively, as shown in Equation (9). If every difference is small, the cost propagation $\boldsymbol{p}$ is reliable and a larger $\boldsymbol{T}$ is added. Otherwise, if one of the differences is large, the cost is unreliable because the cost may be propagated from depth discontinuities. A smaller $\boldsymbol{T}$ is added to reduce the propagation. This paper defines s as half of the cost metric window in all the experiments.

$$T=\begin{cases} T_q(\boldsymbol{p}+1,\boldsymbol{p}) & \forall_{i=1}^{s+1}\left|g(\boldsymbol{p}+1)-g(\boldsymbol{p}+1-i)\right|\le Q \\ T_q(\boldsymbol{p}+1,\boldsymbol{p}+1-t) & \exists_{t}\left|g(\boldsymbol{p}+1)-g(\boldsymbol{p}+1-t)\right|>Q \end{cases} \quad (9)$$

where, the symbol $\forall$ s means any, namely, that any one of $s+1$ previous pixels along the cost aggregation path; the symbol $\exists$ means exist, namely, that there exists at least one of the previous $s+1$ pixels which satisfy the threshold condition. t is decided by the first one of the previous pixels that satisfy the threshold condition. Q is the threshold of the intensity difference. T_q represents the constraint term T, which is computed by a quadratic-based kernel function. The quadratic-based kernel function is described below.

Traditional image-guided methods adopt the Gaussian kernel function to compute constraint term T, as shown in Equation (7). The Gaussian kernel function is a decreasing function, whose absolute slope is also decreasing. The Gaussian function value T_G with a small smooth term σ will decrease greatly, when the intensity differences only change slightly around zero. In fact, it is impossible that the intensities of pixels in the same region are exactly the same. In order to strengthen the cost propagation in homogenous intensity regions, traditional methods choose a larger σ (σ=15~25) to acquire a larger T (Yang, 2015; Pham and Jeon, 2013; Cigla and Alantan, 2013; Sun et al, 2014; Cheng et al, 2015). However, when the cost is propagated between two different regions, T may remain large with a large σ, which may bring mismatches in the depth discontinuities. This paper introduces a new kernel function based on a quadratic term, whose slope is increasing in the range [0, 2σ], as follows:

$$T_q(\boldsymbol{p}+1,\boldsymbol{p})=\begin{cases} a\cdot\Delta g^2+1 & \Delta g\le 2\sigma \\ T_G(\boldsymbol{p},\boldsymbol{p}+1) & \Delta g>2\sigma \end{cases} \quad (10)$$
$$\Delta g=\left|g(\boldsymbol{p}+1)-g(\boldsymbol{p})\right| \quad a=(e^{-2}-1)/4\sigma^2$$

where, T_q represents constraint term T based on the quadratic-based kernel function; a represents the coefficient of the quadratic term; and σ represents the smooth term.

The absolute slope of the quadratic-based kernel function is increasing in the range [0, $+\infty$]. When intensity differences change a little bit, the value of T is still large. When the intensity differences change greatly, the value of T decreases sharply. The range of T is [0, 1]. In order to make the new kernel function meet the range, the Gaussian kernel function is used when the intensity difference is larger than 2σ. When σ is equal to 5 or 20, respectively, the performances of the Gaussian kernel function and the quadratic-based kernel function are shown in Figure 2.

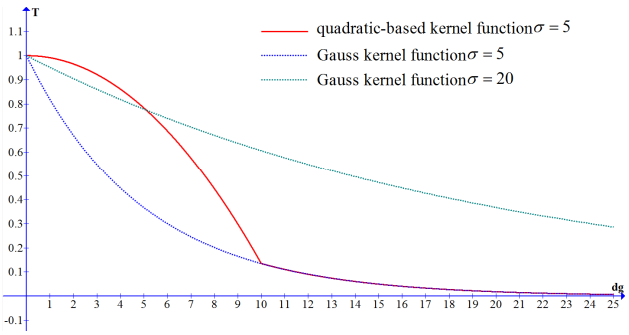


Figure 2. Comparison of Gaussian Kernel Function and Quadratic-based Kernel Function

In Figure 2, the horizontal axis represents the absolute value of intensity difference Δg; and the vertical axis represents the value of constraint term T. $T_q(\sigma)$ is defined as the value of T computed by the quadratic-based kernel function. $T_G(\sigma)$ is defined as the value of T computed by the Gaussian kernel function. It can be seen from Figure 5 that when $\sigma=5$, the quadratic-based kernel function can acquire a larger $T_q(5)$ with a small intensity difference ($\Delta g\le 5$). However, the corresponding Gaussian kernel function value $T_G(5)$ decreases sharply. If a larger smooth term σ (e.g., $\sigma=20$) is chosen, the Gaussian kernel function can acquire a larger $T_G(20)$ when the intensity difference Δg is small. However, it does not mean that the corresponding $T_G(20)$ will become small when the intensity difference is large ($\Delta g>10$). It is also worth noting that when the intensity difference is small ($\Delta g\le 5$), the quadratic-based function value $T_q(5)$ with $\sigma=5$ is still larger than the Gaussian kernel function value $T_G(20)$ with $\sigma=20$, which shows the robustness of the quadratic-based kernel function in the case of small smooth term σ.

The matching results of the non-local methods are related to the cost aggregation path. This paper classifies the cost propagation paths into connected paths and unconnected paths. If all the pixels from the beginning to the end belong to a same region, the path is called a connected path; otherwise, the path is called an unconnected path. A good cost aggregation path should be connected. Most non-local methods define paths through horizontal/vertical scanlines (Pham and Jeon, 2013; Cigla and Alantan, 2013; Sun et al, 2014; Cheng et al, 2015). However, the path, which is described by horizontal/vertical scanlines, may be unconnected even though the beginning and the end belong to the same region, as shown in Figure 3.

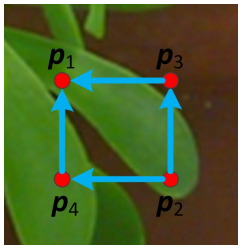


Figure 3. Paths Described by Horizontal/Vertical Scanlines

In Figure 3, the red circles represent pixels; the blue lines represent paths; and the arrows represent the directions of the cost aggregation. Pixels $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ belong to the same region. Pixels $\boldsymbol{p}_3$ and $\boldsymbol{p}_4$ belong to other regions. If the cost of $\boldsymbol{p}_2$ is propagated to $\boldsymbol{p}_1$, there are two paths: $\boldsymbol{p}_2$—$\boldsymbol{p}_3$—$\boldsymbol{p}_1$ or $\boldsymbol{p}_2$—$\boldsymbol{p}_4$—$\boldsymbol{p}_1$. Regardless of the path chosen, passing through the pixel beyond the region of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ cannot be avoided. Thus, the support from $\boldsymbol{p}_2$ to $\boldsymbol{p}_1$ is very small, namely, the path is unconnected.

In order to solve the above problem, this paper proposes a new cost aggregation method based on eight directions, which contain not only the horizontal/vertical directions, but four diagonal directions as well. The new method needs two iterations. In the first iteration, the cost aggregation results from the eight directions are summed, as follows:

$$S(\boldsymbol{p},d)=C(\boldsymbol{p},d)+\sum_{r=1}^{8}\left(L_r(\boldsymbol{p},d)-C(\boldsymbol{p},d)\right) \tag{11}$$

where, L represents the aggregated cost calculated by Equation (8); r represents the path directions, including 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°; and S represents the sum of the aggregated cost from the eight directions.

After the first iteration, the path only exists along scanlines of the 8 directions for every pixel. There is no path to link the pixel beyond the scanlines. Thus, regard the aggregation result S in the first iteration as the new cost in the second iteration, and then aggregate the new cost again along the scanline. Finally, sum the cost aggregation results from the 8 directions. S^2 is defined as the new aggregation result in the second iteration. After two iterations, there exists paths between any two pixels in the image. The path is connected or unconnected.

The final cost aggregation results may vary greatly in magnitude for each pixel, which will make it necessary to normalize aggregation result S^2. After two iterations, the path between arbitrary two pixels is shown in Figure 4.

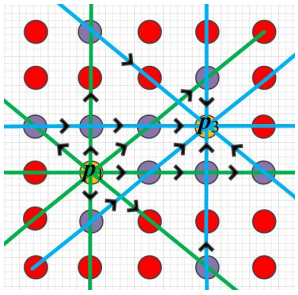


Figure 4. Paths after Two Iterations

Figure 4 shows the paths where the cost is propagated from pixel $\boldsymbol{p}_1$ to $\boldsymbol{p}_3$. The circles represent pixels; the green lines represent the eight scanlines of $\boldsymbol{p}_1$, and the blue lines represent the eight scanlines of $\boldsymbol{p}_3$. The paths of cost propagation are defined by the intersections between the two sets of scanlines, which are represented by purple circles; and the arrows represent the directions of the cost propagation. It can be seen from Figure 4 that the cost aggregation method based on eight directions could provide several paths for cost propagation from $\boldsymbol{p}_1$ to $\boldsymbol{p}_3$, including horizontal paths, vertical paths, and diagonal paths. Ordinarily, at least one of these paths must be connected; but it is possible that none of the paths are connected in the ring or U-shaped regions.

2.3 Semi-global Matching (SGM) based on Non-local Aggregated Cost

The above image-guided non-local method performs well in homogenous intensity regions, but the matching result is not robust in the texture regions because the non-local methods do not consider the cost propagation between different regions. As the energy function-based matching methods perform well in texture regions, the proposed method combines the two. First, the non-local matching method in Section 2.2 was adopted. Then, the cost aggregation result S^2 were regarded as the new cost and the semi-global method (Hirschmuller, 2008) based on the new cost was used to guarantee performance in texture regions. Finally, the Winner Takes All (WTA) strategy was adopted to achieve the initial disparity image, and the left-right consistency check method was used to eliminate outliers.

2.4 Image-guided Disparity Interpolation

After the left-right consistency check, some invalid pixels were present in the initial disparity image. In order to achieve better matching results, these invalid pixels needed to be interpolated. This paper therefore proposes a new image-guided disparity interpolation method as well. In this method, the valid pixels are regarded as the reliable pixels, and the invalid pixels as unreliable pixels. The disparities of the unreliable pixels are interpolated by the reliable pixels in the same region. The interpolation method is similar to the non-local method described in Section 2.2. First, the cost of every pixel is computed according to the initial disparity image, as shown in Equation (12).

$$C(\boldsymbol{p},d)=\begin{cases}\min\left(|\,d-M^0(\boldsymbol{p})\,|,t\right) & if\ M^0(\boldsymbol{p})=valid\\ 0 & if\ \ M^0(\boldsymbol{p})=invalid\end{cases} \tag{12}$$

where, $C(\boldsymbol{p},d)$ represents the cost of pixel $\boldsymbol{p}$ at disparity $\boldsymbol{d}$; M^0 represents the initial disparity image; and t represents the truncation threshold.

Then, the cost is aggregated in a manner similar to the non-local method described in Section 2.2. Different from the non-local method, the interpolation method makes full use of the reliable pixels and unreliable pixels to constrain the cost propagation path. Utilizing the reliable pixels and unreliable pixels, this new method defines a new constraint term T, as follows.

$$T(\boldsymbol{p},\boldsymbol{p}+1)=\begin{cases}T(\boldsymbol{p},\boldsymbol{p}+1) & \begin{array}{c}if\ \boldsymbol{p}=reliable\ and\\ \boldsymbol{p}+1=reliable\end{array}\\ T(\boldsymbol{p},\boldsymbol{p}+1) & \begin{array}{c}if\ \boldsymbol{p}=unreliable\ and\\ \boldsymbol{p}+1=unreliable\end{array}\\ 0 & \begin{array}{c}if\ \boldsymbol{p}=unreliable\ and\\ \boldsymbol{p}+1=reliable\end{array}\\ u^{T(\boldsymbol{p},\boldsymbol{p}+1)}-1 & \begin{array}{c}if\ \boldsymbol{p}=reliable\ and\\ \boldsymbol{p}+1=unreliable\end{array}\end{cases} \tag{13}$$

Equation (13) represents the value of T when the cost is propagated from $\boldsymbol{p}$ to $\boldsymbol{p}+1$. When both $\boldsymbol{p}$ and $\boldsymbol{p}+1$ are reliable pixels or unreliable pixels, T is still computed by Equation (10). When $\boldsymbol{p}$ is an unreliable pixel but $\boldsymbol{p}+1$ is a reliable pixel, T is set as equal to 0, which cuts off the path between $\boldsymbol{p}$ and $\boldsymbol{p}+1$. When $\boldsymbol{p}$ is a reliable pixel but $\boldsymbol{p}+1$ is an unreliable one, T is strengthened by an exponential function, which encourages the

propagation from $\boldsymbol{p}$ to $\boldsymbol{p}$ + 1. Ordinarily, the value of u should be larger than 2. The detailed paths during cost propagation are shown in Figure 5.

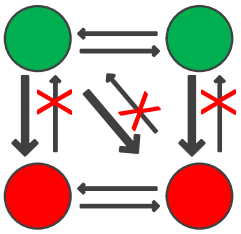


Figure 5. Cost Propagation Path Defined by Reliable Pixels and Unreliable Pixels

In Figure 5, the green circles represent reliable pixels; the red circles represent unreliable pixels; the arrows represent the directions of cost propagation; the lines represent the cost propagation paths; and the thickness of the lines is related to the value of T. The line is thicker for a larger value of T. A red cross on the line means the cost is forbidden to pass through the path. After cost aggregation, the WTA strategy is adopted again to achieve the final disparity image.

3. EXPERIMENTS

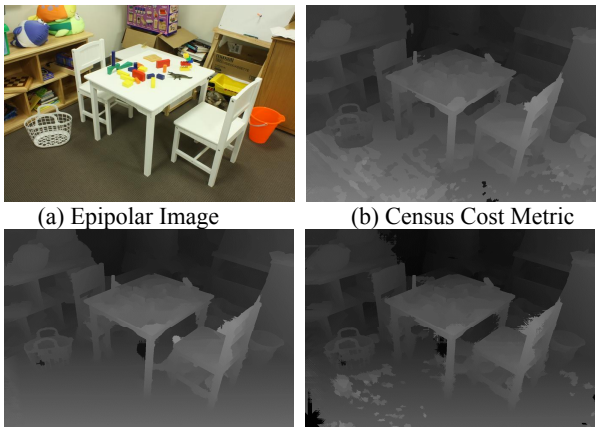
In order to test the performance of the proposed method, a series of experiments were carried out. The experiments can be divided into five parts: (1) experiment on the stereo pair with vertical parallax; (2) experiment on the stereo pair with inconsistent disparities in homogenous intensity regions; (3) experiment on the famous Middlebury Stereo Vision data sets; (4) experiment on the KITTI benchmark; (5) experiment on the actual aerial imagery of Toronto. The first experiments aimed at testing the robustness of the proposed HOG cost metric in stereo pairs with vertical parallax. The second experiments compared the traditional non-local method and the proposed non-local method. The third experiments compared the proposed method with the state-of-the-art matching methods. The fourth experiment tested the proposed method in street view images. The fifth experiments tested the reliability of the proposed method in outdoor images. In these five experiments, all the matching parameters were fixed, as shown in Table 1.

Step	Parameter	Value
Cost Computation	Window Size	5 (pixel)
	Weighting Coefficient q	0.3
Image-guided Non-local Matching	Smooth Term σ	6
	Penalty Term P_1	0.3
	Penalty Term P_2	6
Disparity Interpolation	Truncation Threshold t	5
	Smooth Term σ	3
	Function Base u	5

Table 1. Matching Parameters

3.1 Epipolar Stereo Pair with Vertical Parallax Existing

This paper chose PlayTable data provided by Middlebury Stereo Vision for experiments, as shown in Figure 6(a). The average vertical parallax was 0.481 pixels. The maximum vertical parallax was 1.333 pixels. The Census cost metric, the HOG cost metric, and a combination of Census and HOG were used in the proposed matching method. The corresponding matching results are shown in Figure 6(b), Figure 6(c) and Figure 6(d), respectively.



(a) Epipolar Image (b) Census Cost Metric

(c) HOG Cost Metric (d) Census+HOG Cost Metric

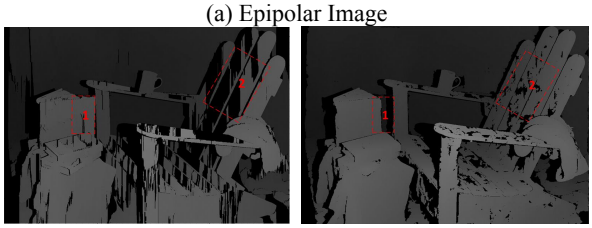
Figure 6. Comparison of Cost Metric

Comparing Figure 6(b) and Figure 6(c) indicates serious mismatches, especially in the floor region when the Census cost metric was used, which was due to the floor region being rich in fine-textures. Therefore, the Census cost computation may be wrong even though the vertical parallax is small. On the other hand, the HOG cost metric was able to reduce the influence of vertical parallax, which can achieve a better matching result. Figure 6(d) shows the results of the Census and HOG combination. The weight of Census is 0.3, as shown in Table 1. The matching result of the combination metric was better than that of the Census metric but was worse than that of the HOG metric. Although the Census metric was sensitive to the vertical parallax, it performed well in stereo pairs with nonlinear radiometric distortions. Thus, the combination metric of Census and HOG was used in the proposed new method.

3.2 Epipolar Stereo Pair with Inconsistent Disparities in Homogenous Region

Adirondack data provided by Middlebury Stereo Vision were used, as shown in Figure 7(a). Both of the regions in the red rectangles are homogenous intensity regions. The disparity changed sharply in Rectangle 1. The disparities in Rectangle 2 changed smoothly. Figure 7(b) is the result of the traditional non-local method (Cigla and Alantan, 2013). Figure 7(c) is the result of the proposed method. In order to compare the performances of both methods solely, only the initial disparity images were given.

(a) Epipolar Image



(b) traditional method (c) proposed method

Figure 7. Comparison of Matching in Homogenous Regions

Figure 7(b) clearly shows serious mismatches in both Rectangle 1 and Rectangle 2 if their disparities were forced to be

consistent. The proposed method considers the changes in the disparities in the homogenous intensity areas, which achieved a better matching result for Figure 7(a).

3.3 Test on Middlebury Stereo Vision Data Sets

The proposed method was used to match stereo pairs in Middlebury benchmark. The average absolute error in the pixels was the accuracy metric. The performance of the proposed method is shown in Table 2, which lists the final matching results of the top 8 algorithms in the Middlebury Stereo Vision Benchmark. In Table 2, the first row lists the name of every algorithm, and the second row lists the general average absolute error. The centred number represents the matching accuracy, and the superscript number represents the rank. The third row lists the running time. The fourth row list the running environment. The corresponding cell is made up of two parts: the left part represents the processor (CPU or GPU), and the right part represent the number of processor cores used for running the algorithms (serial or parallel).

It can be seen from Table 2 that the general matching accuracy of the INTS method ranks top five in the Middlebury Stereo Vision by the end of 2016-04-09. The INTS method uses a single i7 CPU core for matching, and the average running time is 104s. Although INTS does not have an obviously superior run time, it was the fastest compared to four other top methods when all the run time was converted to the same running environment. In addition, the cost aggregation of eight directions in Sections 2.2 and 2.3 was independent, which would be easy to implement to parallel processing for acceleration.

Name	NTDE	MCCNN_Layout	MCCNN+RBS	SOU4P-net	**INTS**	LCU	MC-CNN	Mesh Stereo
Avg (pixel)	2.20 [1]	2.35 [2]	2.61 [3]	2.62 [4]	**2.86** [5]	3.63 [6]	3.82 [7]	4.38 [8]
Time (s)	128	300	157	688	**104**	4764	106	62.0
CPU /GPU Count	Geforce GTX TITAN X	i7 1	NVIDIA GTX TITAN X	GTX 980 GPU	**i7 1**	E5-2690 1	NVIDIA GTX TITAN Black (GPU)	i7 8

Table 2 Rank in Middlebury Stereo Vision Benchmark (2016.04.09)

3.4 Test on KITTI Benchmark

INTS method was tested on stereo 2015 data sets provided by KITTI. A pixel is considered to be correctly estimated if the disparity is less than 3 pixels. The performance of INTS is shown in Table 3. In table 3, D1 represents percentage of stereo disparity outliers in first frame; bg represents percentage of outliers averaged only over background regions; fg represents percentage of outliers averaged only over foreground regions; all represents percentage of outliers averaged over all ground truth pixels.

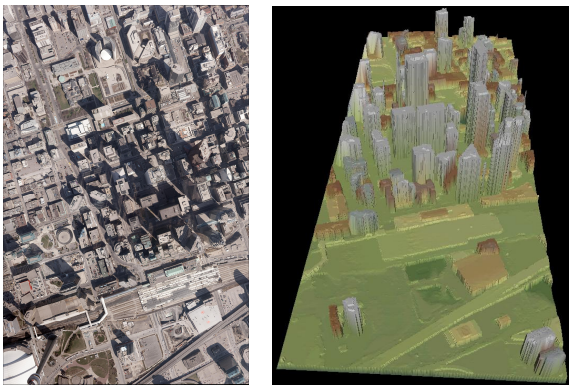
Error	D1-bg (%)	D1-fg (%)	D1-all (%)
All/All	5.88	14.11	7.25
All/Est	5.84	14.11	7.22
Noc/All	5.29	13.04	6.57
Noc/Est	5.27	13.03	6.56
Rank	22		

Table 3. Accuracy Evaluation

It can be seen from Table 3 that only about 7% pixels are outliers for street view image matching with INTS method. INTS method performed well in foreground regions, but it performed badly in background regions. It may be caused by deep depth of field in background regions. The rank in KITTI benchmark is only 22th. The rank is not high for two reasons: 1. optical flow information was not used, which has already been used in top algorithms in KITTI benchmark; 2. INTS method performed badly in background regions. Future work will focus on how to improve the matching results in background regions.

3.5 Test on Aerial Imagery of Toronto

This paper applied the INTS method to aerial images of Toronto to test the performance of matching outdoor images, as shown in Figure 8(a). The stereo pair was captured by the Microsoft Vexcel's UltraCam-D (UCD) camera. The image size is 7500×11500. In order to test the matching accuracy, the corresponding LiDAR point set was used as control information, which was captured by the Optech airborne laser scanner ALTMORION M. The point density is approximately 6.0 points/m^2. The the matching result is shown in Figure 8(b).



(a) Epipolar Image (b) Reconstruction Results

Figure 8. Reconstruction of Toronto

It can be seen from Figure 8(b) that the INTS method can achieve good reconstruction results in urban areas. Occlusion regions and shadow regions do exist, but the tall buildings were reconstructed well. In addition to the aerial images of Toronto, ISPRS also provided the corresponding LiDAR point sets which had been registered rigorously. In order to test the actual matching accuracy of INTS in Toronto, the LiDAR point sets were regarded as truth values and the matching point sets were compared to the LiDAR point sets. However, some outliers still remained in the LiDAR point sets and some LiDAR points lay in occlusion regions that were not visible in the images. Thus, the LiDAR point sets were filtered to eliminate the outliers and occlusion points before comparison. Then, the filtered LiDAR points were projected onto the epipolar stereo pairs, which acquired a series of correspondences. Finally, the disparities from dense matching were compared with the disparities from the LiDAR points to evaluate the performance of the INTS method. In order to comprehensively evaluate the accuracy results, the average matching accuracy, the percent of pixels with matching accuracy below 0.5 pixels, 1 pixel, 2 pixels and 4 pixels were chosen as the accuracy metrics, as shown in Table 4.

Average (pixels)	% ≤ 0.5 pixels	% ≤ 1 pixels	% ≤ 2 pixels	% ≤ 4 pixels
0.888	0.571	0.818	0.912	0.958
Number of LiDAR Points	1225753			
Running Time	13079.67 s			

Table 4. Accuracy Evaluation

It can be seen from Table 4 that the matching points of INTS are very similar to the LiDAR points. The average matching accuracy is only 0.888 pixels. The matching accuracies of most points (57.1%) are below 0.5 pixel, which shows the robustness of the INTS method. The points with matching accuracies above 4 pixels were regarded as outliers, which were caused by not only the INTS method itself, but also by the LiDAR points because the filtering method cannot guarantee total elimination the outliers and occlusions in LiDAR point sets. The running time was low when only a single i7 CPU core was used. However, the running time could be expected to be improved greatly with parallel processing.

4. CONCLUSION

This paper proposed a new image-guided non-local dense matching method. An improved HOG feature was introduced into the cost computation for the first time, which can reduce the influence of vertical parallax. A new non-local cost aggregation strategy was presented, which guarantees that the paths between any two pixels were connected. Varied disparities in homogenous intensity regions were considered, a more robust kernel function was introduced, and a new disparity interpolation method was proposed, which defined new propagation rules to guarantee the interpolation accuracy. Experiments on benchmarks and the actual aerial imagery showed that INTS method performed well in matching accuracy and reliability. INTS is one of the present state-of-the-art matching methods. Its running time is not exceptional at this time with only one single CPU, but INTS can be accelerated greatly with parallel processing in future work.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 41571434, 41322010), and the Academic Award for Excellent Ph.D. Candidates funded by the Ministry of Education of China (Grant No. 5052012213002).

REFERENCES

Cheng, F. Y., Zhang, H., Sun, MG., et al, 2015. Cross-trees, edge and superpixel priors-based cost aggregation for stereo matching. *Pattern Recognition*, 48(7), pp. 2269-2278.

Cigla, C., Alantan, A. A., 2013. Information permeability for stereo matching. *Signal Processing-Image Communication*, 28(9), pp. 1072-1088.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, California, pp. 886-893.

He, K., Sun, J., Tang, X., 2013. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6), pp. 1397-1409.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 30(2), pp. 328-341.

Hirschmueller, H., Scharstein, D., 2009. Evaluation of stereo matching costs on Images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9), pp. 1582-1599.

Jiao, J. B., Wang, R. G., Wang, W. M., et al., 2014. Local stereo matching with improved matching cost and disparity refinement. *IEEE Multimedia*, 21(4), pp.16-27.

Jung, I. L., Chung, T. Y., Sim, J. Y., et al, 2013. Consistent stereo matching under varying radiometric conditions. *IEEE Transactions on Multimedia*, 15(1), pp. 56-69.

Kordelas, G. A., Alexiadis, D. S., Daras, P., 2015. Enhanced disparity estimation in stereo images. *Image And Vision Computing*, 35, pp. 31-49.

Mei, X., Sun, X., Zhou, M., Jiao, S., Wang, H., Zhang, X., 2011. On Building an Accurate Stereo Matching System on Graphics Hardware. In: *2011 IEEE International Conference on Computer Vision Workshops*, Barcelona, pp. 467-474.

Mozerov, M. G., van de Weijer, J., 2015. Accurate stereo matching by two-Step energy minimization. *IEEE Transactions on Image Processing*, 24(3), pp. 1153-1163.

Paul, V., William, M., Wells, III, 1997. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2), pp. 137-154.

Pham, CC., Jeon, J. W., 2013. Domain transformation-based efficient cost aggregation for local stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7), pp. 1119-1130.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3), pp. 7-42.

Sun, X., Mei, X., Jiao, SH., et al, 2014. Real-time local stereo via edge-aware disparity propagation. *Pattern Recognition Letters*, 49, pp. 201-206.

Taniai, T., Matsushita, Y., Naemura, T., 2014. Graph cut based continuous stereo matching using locally shared labels. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, pp. 1613-1620.

Yang, Q. X., Wang, L., Yang, R. G., et al, 2009. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3), pp. 482-504.

Yang, Q. X., 2015. Stereo matching using tree filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4), 834-846.

Yoon, K. J., Kweon, IS., 2006. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), pp. 650-656.

Zabih, R., Woodfill, J., 2005. Non-parametric Local Transforms for Computing Visual Correspondence. *Lecture Notes in Computer Science*, 801, pp. 151-158.

Žbontar, J., LeCun, Y., 2015. Computing the stereo matching cost with a convolutional neural network. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1593-1599.

Zhou, X. Z., Boulanger, P., 2012. Radiometric invariant stereo matching based on relative gradients. In: *IEEE International Conference on Image Processing*, Lake Buena Vista, FL, pp.2989-2992.