Improved depth map estimation in Stereo Vision

Hajer Fradi and and Jean-Luc Dugelay
EURECOM, Sophia Antipolis, France

ABSTRACT

In this paper, we present a new approach for dense stereo matching which is mainly oriented towards the recovery of depth map of an observed scene. The extraction of depth information from the disparity map is well understood, while the correspondence problem is still subject to errors. In our approach, we propose optimizing correlation based technique by detecting and rejecting mismatched points that occur in the commonly challenging image regions such as textureless areas, occluded portions and discontinuities. The missing values are completed by incorporating edges detection to avoid that a window contains more than one object. It is an efficient method for selecting a variable window size with adaptive shape in order to get accurate results at depth discontinuities and in homogeneous areas while keeping a low complexity of the whole system. Experimental results using the Middlebury datasets demonstrate the validity of our presented approach. The main domain of applications for this study is the design of new functionalities within the context of mobile devices.

Keywords: Stereo vision, disparity-map, matching, depth, occlusion, correspondence, edges

1. INTRODUCTION

Researchers have been giving special attention to stereo vision systems capable of perceiving accurate depth information of an observed scene. Most stereo vision implementations are based on two forward-facing cameras, where each camera delivers a 2D projection of a scene. The main difficulty encountered in this context is stereo matching which determines the spatial displacement between each two corresponding pixels in a stereo pair. This process is termed as correspondence problem and it aims at estimating a disparity map which is the set of disparity values of all the pixels in a reference image.

Intensive research[1, 2] has been conducted in the recent decades to solve the problem of finding the correspondences between the right and the left images. For more details, a good taxonomy of dense two-frame stereo matching algorithms can be found in.[3] In general, stereo matching algorithms can be categorized into two major classes: local and global methods. Global methods formulate the problem in terms of an energy function, which is subject to optimization. Then, all disparities are determined simultaneously by applying energy minimization techniques. Global methods usually achieve high matching accuracy such as graph cuts,[4] dynamic programming[5] and belief propagation.[6] However, most of these methods are computationally expensive.

Compared with them, local methods[7] have higher efficiency and they are more suitable for real-time application. To retain more smoothness in the disparity map, local methods based on correlation are utilizing the color or intensity values within a finite window. These methods have been widely employed, where a cost function is evaluated over a window around the pixel of interest. Correlation based methods fail in most points because they are strictly based on the resemblance constraint by assuming that the intensities of corresponding points are similar which is not robust to changes in illumination and contrast variation. Also, these methods are not able to deal with the problems[8] of occluded areas and discontinuities where some parts of one image are hidden in the second image. Moreover, pixels inside textureless regions and repetitive patterns are hard to be properly matched since the horizontal variation in such areas is very low.

Added to that, another central problem in correlation technique lies on selecting an appropriate window size: large window increases the reliability by averaging over a big area and increasing the effects of noise. However, using large window may affect the result by blurring the objects borders. On the other hand, if the window is

Further author information: (Send correspondence to Hajer Fradi)
Hajer Fradi: E-mail: hajer.fradi@eurecom.fr, Telephone: +33 (0)4 93 00 82 54
Jean-Luc Dugelay: E-mail: Jean-Luc.Dugelay@eurecom.fr, Telephone: +33 (0)4 93 00 81 41

very small, it does not cover intensity variation and it gives a poor disparity estimate. Thus, as window size is increased from small to large, the resulting disparity map changes from accurate boundaries but noisy in low texture areas to more reliable in low texture areas but blurred disparity boundaries. Generally, the choice of the correlation window size is a tradeoff between increasing reliability in low textured areas and decreasing the blurring effects in boundaries.[9] In addition, local support window is used to reduce the image ambiguity, it implicitly assumes that all pixels belonging to the support are from similar depth in a scene and therefore, they share similar disparity. The problem is that this assumption is not always validated, since the support window located on depth discontinuities represent pixels from different depths.

By ignoring these problems, most methods use a rectangular window of a fixed size, which may give bad performance and erroneous results. Therefore, earlier researchers attempt to overcome these limitations by allowing the window to vary across the image. Most of these methods[9] vary the size depending on local variation of intensity, but they still restrict the window to rectangular shape. However, there have been few works on varying the shape of the window.[10]

These ambiguities left by correlation-based methods make the matching process more challenging and the recovery of accurate disparity still difficult to be properly addressed. Therefore, in this paper we present a new dense stereo matching algorithm based on correlation and showing progress in handling problems of mismatched points. It is a three steps framework; first, an appropriate cost matching is used to avoid the drawbacks of possible ambiguous matches caused by the violation of the resemblance constraint. Then, a bidirectional matching is applied to detect and to reject mismatches. A matching is valid only if after a return (e.g. right-left-right) the final position is the same as the initial one. Third, the created holes at this stage will be filled in by incorporating edges detection to obtain full disparity map that can be converted to depth map using simple triangulation.

The rest of the paper is organized as follows: Section 2 details the development of the proposed approach for stereo matching. The extraction of depth information is described in Section 3. Section 4 shows the experimental results to demonstrate the effectiveness and robustness of our approach. Finally, we give a brief conclusion in Section 5.

2. PROPOSED APPROACH FOR STEREO MATCHING

The proposed method for stereo matching employs correlation-based technique to estimate the disparity map on a stereo pair and it is performed in three consecutive steps:

2.1 First step: Finding matching candidates

Using a pair of rectified stereo images I_1 and I_2, for a given pixel in the reference image, the matching candidates in the second image belong not only on the same scanline but also to a well defined disparity range $[0...d_{max}]$.[11] To find the best corresponding point, a matching cost should be adopted. The most common measure is SAD[12] which assumes similar intensity values for two matching pixels. Other matching measures like gradient based are more robust to changes in illumination and camera gain.

In our proposed approach, we use a matching metric that combines SAD and a gradient measures[13] which are defined as:

$$C_{SAD}(x,y,d) = \sum_{(i,j)\in W(x,y)} |I_1(i,j) - I_2(i,j-d)| \tag{1}$$

$$C_{GRAD}(x,y,d) = \sum_{(i,j)\in W(x,y)} |\nabla_x I_1(i,j) - \nabla_x I_2(i,j-d)| + \sum_{(i,j)\in W(x,y)} |\nabla_y I_1(i,j) - \nabla_y I_2(i,j-d)| \tag{2}$$

where $W(x,y)$ is a window surrounding the position (x,y), d is a disparity value, ∇_x is the forward gradient to the right and ∇_y is the forward gradient to the bottom. These measures (C_{SAD} and C_{GRAD}) are performed with color model by summing the differences for all channels. The resulting matching cost C is defined by:

$$C(x,y,d) = \lambda C_{SAD}(x,y,d) + (1-\lambda) * C_{GRAD}(x,y,d) \tag{3}$$

where λ is a weight parameter that is chosen to maximize the number of reliable correspondences. So, for every pixel of interest in the left image, a surrounding block of pixels slide across a row from the right image and the function C is computed at each position, the selected disparity value corresponds to the smallest matching cost and it is given by:

$$disp(x,y) = \arg \min_{d \in [0 \ldots d_{max}]} C(x,y,d) \tag{4}$$

2.2 Second step: Rejecting mismatches

The existence of the problems described above suggests that the matching problem has not been properly approached. Therefore, after estimating the disparity at each pixel in the image, this step attempts to keep only disparities of reliable points by rejecting points that are mismatched. It is at the same time a way to detect occluded and low textured areas since almost of false matched points appear in such parts.

2.2.1 Bidirectional matching

A first idea to deal with occluded and low textured areas is bidirectional matching;[14] unlike most of stereo matching algorithms that compute disparity map with left image as reference, bidirectional matching creates two disparity maps relative to each image. It consists of applying the procedure described above twice: once with left image as reference and once with right one. Then, we compare left-to-right and right-to-left disparity maps to reject false matches. As it is shown in Figure 1, after producing two disparity maps $disp_1$ and $disp_2$, the process is as follow: we pick up pixel at position (i,j) from the second map, the disparity value is $disp_2(i,j)$. We establish first match between pixel (i,j) from the second map and pixel $(i, j + disp_2(i,j))$ from the first map. Then, at position $(i, j + disp_2(i,j))$, we pick up pixel from the first map, the disparity value is $disp_1(i, j + disp_2(i,j))$. We establish second match between pixel $(i, j + disp_2(i,j))$ from the first map and pixel $(i, j + disp_2(i,j) - disp_1(i, j + disp_2(i,j)))$ from the second map. We define E as:

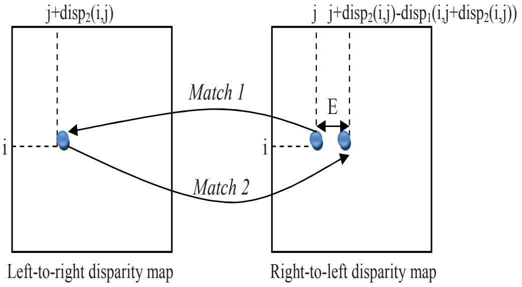
j+disp2(i,j)
j
j+disp2(i,j)-disp1(i,j+disp2(i,j))
Match 1
E
i
i
Match 2
Left-to-right disparity map
Right-to-left disparity map

Figure 1. Bidirectional matching.

$$E(i, j + disp_2(i,j)) = disp_2(i,j) - disp_1(i, j + disp_2(i,j)) \tag{5}$$

When $E(i,j)$ is equal to zero, we start from a point on the right hand-side image to reach another point on the left hand side image and we go back to the right handside image to reach the initial point; it means the same match is established whatever we used left or right image as reference.

To conclude, bidirectional matching is a typical method to detect occluded and low textured areas. For the same purpose, we have also proposed a set of constraints and rules.

2.2.2 Constraints

We propose to impose traditional assumptions for matching in order to be able to augment the confidence of a given match and to decrease the number of mismatches.[2] These constraints are:

- *Uniqueness constraint :* Every point can only be coupled to at most one point[15] from the other image. This constraint is violated in the case of transparent objects.[16] Moreover, it has been extensively applied to explicit detect occlusions.

- *Ordering constraint :* The second constraint is about ordering, it also has been able to successfully detect occlusions. This additional assumption states that if an object O_1 is on the left of an object O_2 in the left image, then the object O_1 will also appear on the left of the object O_2 in the right image.[17]

- *Continuity constraint:* This assumption states that disparity values are continuous almost everywhere.[17] It means neighboring points must have consistent match values. In most scenes, the continuity assumption is valid since physical surfaces are locally considered smooth. But, at edges, this assumption should be canceled.

2.3 Third step: Filling in the disparity maps

After the validation step, we have as input resulting disparity map with known and missing values. The task is to complete the missing data problem. To achieve this goal, the easiest solution is to use low pass filter. But, we have chosen to address the problem differently. The idea is to keep the same values of known disparities and to give smooth values for missing ones. Since we start by using rectangular window with fixed size, this choice will give incorrect results in two cases:

- Depth discontinuities:[10] if the correlation window overlaps a depth discontinuity, it manifests when a part of the window contains pixels from different depth with the pixel under consideration, which affects the results arbitrarily.

- Homogeneous areas: if the support window does not cover intensity variation, the matching process yields wrong results.

Considering the existence of these problems, this step aims at correcting the false matches to get accurate results at depth discontinuities as well as in homogeneous regions. To properly deal with the image ambiguity problem, we describe an efficient method for selecting a variable window size with adaptive shape such that the support window varies at each pixel (of incorrect disparity value). To go a little bit on details, we start by detecting the edges in the reference image. Then, we use large window to cover sufficient intensity variation. After that, for each missing value in the disparity map, we define the window of the pixels that encountered the pixel of interest. By using the edge detection result, we retain only pixels that belong to the same connected component to retrieve the proper object shape and to avoid crossing the boundaries. It is efficient way to restrict the amount of pixels used for correlation and to avoid that a window contains more than one object. The choice of this adaptively reshaped window reduces the influence of errors called outliers. We should notice that we did not use variable window in the whole image, because it is time consuming process, also, it results noisy disparity map. So, the best way is to vary the shape and the size of the window only at the points where the matching is not correct.

3. DEPTH MAP ESTIMATION

We proposed a novel approach to compute the disparity map. This information can be converted to depth map[1] which indicates the distances from surfaces to the imaging system. At position (i, j), the depth Z is defined by:

$$Z(i,j) = f.\frac{B}{d(i,j} \tag{6}$$

where f denotes the focal length, B is the baseline and d is the disparity value.
Uncertainties in finding disparity values lead to errors in depth map. By denoting $\hat{Z}$ and Z the estimated and the actual depths respectively, we get:

$$Z = \frac{B * f}{d} \tag{7}$$

$$\hat{Z} = \frac{B * f}{\hat{d}} \tag{8}$$

Where $\hat{d}$, d are the estimated and the actual disparity values respectively. By considering Δd the uncertainty in finding a disparity value and Δz the error of computing the depth map at that range, the equation which relates both of these variables is:

$$\Delta Z = |\hat{Z} - Z| = \frac{\hat{Z}}{\frac{\hat{d}}{\Delta d} - 1} \tag{9}$$

As a result, depth error vs. disparity error can be plotted in a graphical form as it is shown in Figure 2:

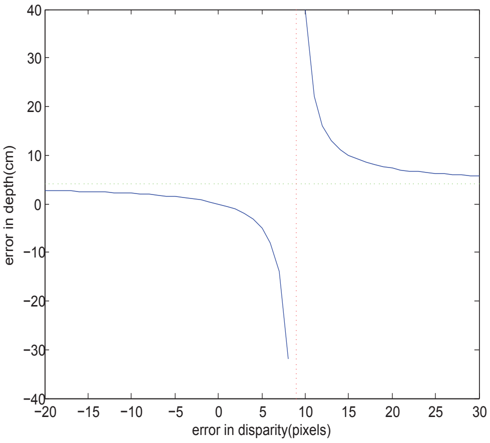

Figure 2. Depth error vs. disparity error.

4. EXPERIMENTAL RESULTS

The experiments have been performed on a laptop Intel Core 2 Duo P7350, 2 GHZ, 4 Go RAM, the calculation of the disparity map for a stereo pair takes between 25 and 36 seconds. In this section, we present experiment results on various datasets with comparisons to the state-of-the-art stereo algorithms. In Figure 3, the proposed approach for stereo matching is evaluated using the Middlebury datasets (http://vision.middlebury.edu/stereo/), we see the resulting disparity maps obtained on four stereo pairs. By viewing gray level results, close up objects have bigger disparity values while further-away objects are lineup very close. For each stereo pair, in the same figure, we show the difference between the computed disparity map and the ground truth disparity. Quantitative results and comparisons with NCC (Normalized Cross Correlation) method, Daisy descriptor[18] and Local evidence[19] methods are reported in Table 1. In this table, to better evaluate the results, we computed the quality metric EQM (Error Quadratic Means) using the following formula:

$$EQM = \frac{1}{N * M} \sqrt{\sum_{i,j} (I_1(i,j) - I_2(i,j))^2} \tag{10}$$

where $N * M$ is the size of the image, I_1 is the computed disparity map and I_2 is the ground truth disparity map.
From this table, it is shown that our approach provides more accurate disparity maps (smaller EQM values) compared to the other cited stereo matching methods.

5. CONCLUSION

A new stereo matching approach has been introduced in order to estimate dense and accurate disparity map. It is capable of addressing the problems of different intensity values between two images of a stereo pair. In addition, it is able of avoiding the drawbacks of ambiguous matches by validating the resulting disparity map. Using the Middlebury datasets, our approach has been experimentally validated showing encouraging performance. Also, by means of comparisons with other approaches in the literature, we demonstrated the overall performance of our results.

Stereo pair	Proposed method	NCC method	LCC method	Daisy descriptor
Dolls	5.0225	39.0779	5.8945	9.7493
Moebius	4.5028	31.6552	7.59322	8.8978
Aloe	2.82462	22.359	3.2399	10.186
Cones	11.3624	37.1176	12.3804	12.2502
Reindeer	5.628	38.5676	11.1398	12.219

Table 1. Comparisons between stereo matching algorithms in terms of Error Quadratic Means on different Middlebury stereo pairs

REFERENCES

[1] Calin, G. and Roda, V., "Real-time disparity map extraction a dual head stereo vision system," *Latin American Applied Research* **37**, 21–24 (2007).

[2] Gutierrez, S. and Marroquin, J., "Robust approach for disparity estimation in stereo vision," *Image and Vision Computing* **22** (2004).

[3] Scharstein, D. and Szeliski, R., "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal on Computer Vision* **47**, 7–42 (April 2002).

[4] Boykov, Y., Veksler, O., and Zabih, R., "Fast approximate energy minimization via graph cuts," *PAMI* **23**(11), 1222–1239 (2001).

[5] Lei, C., Selzer, J., and Yang, Y., "Region-tree based stereo using dynamic programming optimization," *CVPR 2* , 2378–2385 (2006).

[6] Felzenszwalb, P. F. and Huttenlocher, D. P., "Efficient belief propagation for early vision," *CVPR* , 261–268 (2004).

[7] Veksler, O., "Fast variable window for stereo correspondence using integral images," *CVPR 1* , 556–561 (2003).

[8] Zitnick, C. L. and Kang, S. B., "Stereo for image-based rendering using image over-segmentation," *International Journal of Computer Vision* **75**(1), 49–65 (2007).

[9] Kanade, T. and Okutomi, M., "A stereo matching algorithm with an adaptive window: Theory and experiments," *IEEE Trans. Pattern Analysis and Machine Intelligence* **16**, 920–932 (September 1994).

[10] Boykov, Y., Veksler, O., and Zabih, R., "A variable window approach to early vision," *IEEE Trans. Pattern Analysis and Machine Intelligence* **20**, 1283–1294 (December 1998).

[11] Tomabari, T., Mattoccia, S., and Stefano, L. D., "Segmentation-based adaptive support for accurate stereo correspondence," *PSIVT* (2007).

[12] Sunyoto, H., der Mark, W., and Gavrila, D., "A comparative study of fast dense stereo vision algorithms," *IEEE Intelligent Vehicles Symposium* (2004).

[13] Klaus, A., Sormann, M., and Karner, K. F., "Segment based stereo matching using belief propagation and a self-adapting dissimilarity measure," *International Conference on Pattern Recognitions* **3**, 15–18 (2006).

[14] Fua, P., "A parallel stereo algorithm that produces dense depth maps and preserves image features," *Machine Vision and Applications* **6**, 35–49 (1993).

[15] Ogale, A. S. and Aloimonos, Y., "Shape and the stereo correspondence problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **65**, 147–162 (December 2005).

[16] Ouali, M. H., Lange, H., and Laurgeau, C., "An energy minimization approach to dense stereovision," *Proceedings. International Conference on Image Processing* **2**, 841–845 (September 1996).

[17] Zitnick, C. and Kanade, T., "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 675–684 (July 2000).

[18] Tola, E., Lepetit, V., and Fua, P., "Daisy: An efficient dense descriptor applied to wide baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **99** (April 2009).

[19] Ogale, A. S. and Aloimonos, Y., "A roadmap to the integration of early visual modules," *International Journal of Computer Vision* **72**, 9–25 (April 2007).

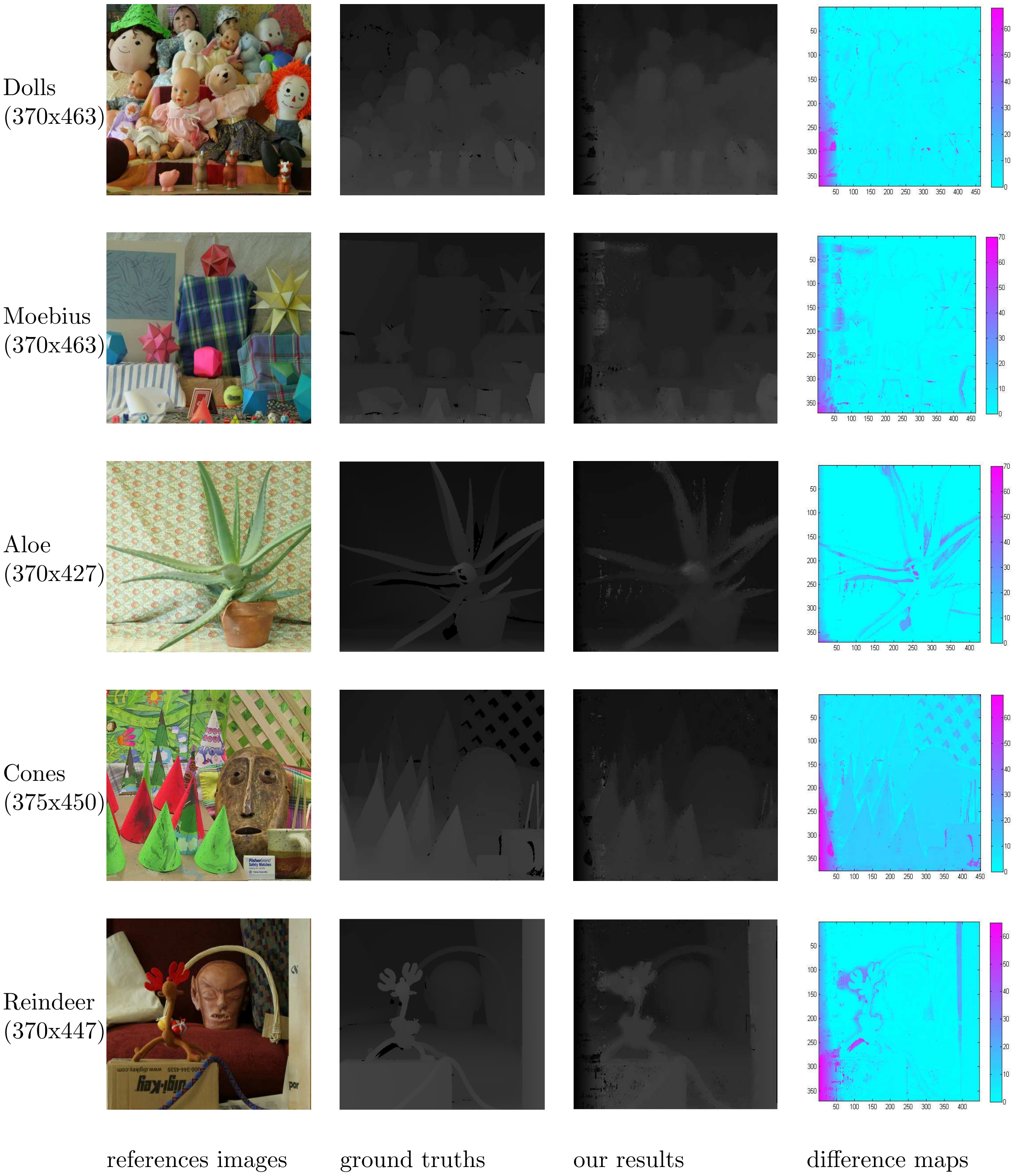

Figure 3. Our results on Middlebury datasets. From top to down order: Dolls, Moebius, Aloe, Cones and Reindeer stereo pairs. From left to right order: reference images, ground truth disparity maps, extracted disparity maps and difference maps between ground truth and extracted disparity maps (with cool colormap which varies smoothly from cyan to magenta)