A Hybrid Stereo Matching Algorithm Guided by the 2D Affine Transform for the Short Baseline Stereo

Aiqiu Zuo, Kevin Stanley and Q.M. Jonathan Wu

Vision and Sensing Group
National Research Council Canada
3250 East Mall, Vancouver
BC V6T 1W5, Canada
aiqiu.zuo@nrc.ca

William A. Gruver

Intelligent Robotics & Manufacturing Systems Laboratory, School of Engineering Science
Simon Fraser University, Burnaby,
BC V5A 1S6, Canada

***Abstract*--In this paper we present a hybrid stereo matching algorithm for short baseline stereo. With this coarse-to-fine algorithm, the stereo matching is first guided by the 2D affine transform and then completed using correlation based on the intensity, and nonparametric transforms including the rank transform and the census transform. With only three known or detected reference points in the stereo images, the 2D affine transform can be recovered and used to model the stereo system providing the approximate disparity information for any other point. To deal with the perspective, a vote, which is based on the correlation of the intensity and the nonparametric transform, is carried out for precise stereo matching. Since it is guided by the 2D affine transform and limited to small neighbor windows, the stereo matching can be completed precisely. This stereo matching algorithm is especially efficient in situations where reference points are available and either the image scene is nearly planar or the scene is sufficiently far away from the cameras compared with the baseline. Experimental results proving the performance of our algorithm are presented.**

***Keywords*: stereo matching, affine transform, nonparametric transform**

I. Introduction

Stereo matching is a fundamental task for many applications of computer vision. Although it has been studied thoroughly, it is still an open problem.

This paper focuses on the stereo matching problem given a short baseline, either calibrated or uncalibrated. First we assume the conventional convergent stereo system configuration, and the scene is nearly planar or the scene is far away sufficiently from cameras compared with the baseline. Under these assumptions, it is reasonable to use the affine camera projection model to get rough disparity information, which could be used to guide finer stereo matching finished by the correlation of intensity and a nonparametric transformation. Three matched feature points are assumed detected or known, which could be used to recover the affine model. All these points do not necessarily have to appear on each pair of stereo images.

The rest of this paper is organized as follows. Section 2 reviews the related work in stereo matching. In section 3, we describe the relevant terminology such as the 2D affine transform and the nonparametric transform. Our hybrid stereo matching algorithm is detailed in section 4. Experiments are carried out and results are also presented in section 5. In section 6, we draw the conclusion and discuss the application of this stereo matching algorithm.

II. Related work

The stereo matching problem can be broken into several categories: calibrated and uncalibrated, short baseline and wide baseline, image stereo matching and feature stereo matching.

Calibrated systems usually simplify stereo matching because corresponding points can be found on the known epipolar line or a band. A lot of research has been devoted on this subject [1][2]. With feature stereo matching, many invariants can be used. In [3], Gouet used first order differential rotation invariants to do the stereo feature matching with a relaxation technique. A scale invariant feature transformation method is described in [4]. Baumberg describes affine texture invariant with detected features. The feature matching process is optimized, ignoring unreliable matches at the expense of reducing the number of the feature matches in [5]. R. Deriche tried to recover the epipolar geometry with detected features for the stereo matching [6]. In [7], Ebroul Izquierdo M. used a global to local hierarchical matching procedure, which joined motion and disparity estimation. Dynamic programming methods are extensively reported in [8][9]. Most of these approaches have a high computational complexity.

III. Terminology

A. 2D Affine Ttransform

The affine camera model introduced by Mundy and Zisserman[10] is described by the 3D-2D transformation

$$p = MP + t \tag{1}$$

where
P: point in three dimension space;
p: the image of P;
M: an arbitrary 2×3 matrix;
t: a 2-vector.
For stereo images, we have

$$p = MP + t \tag{2}$$

$$p' = M'P + t' \tag{3}$$

from which we get

$$p' - t' = M'M^{+}(p - t) \tag{4}$$

Let

$$M'M^{+} = U \tag{5}$$

Combing Eq.(4) and Eq.(5) yields

$$p' - t' = U(p - t) \tag{6}$$

which results in

$$p' = Up + t_0 \tag{7}$$

where

$$t_0 = -Ut + t' \tag{8}$$

Eq.(7) can be rewritten as

$$\begin{bmatrix} p' \\ 1 \end{bmatrix} = \begin{bmatrix} U & t_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p \\ 1 \end{bmatrix} \tag{9}$$

This is the 2D 6-parameter affine transform. In [11], it was used to model both motion and disparity information describing the warping transformation between a frame and sprite image. This model is efficient in situations where either the scene is sufficient far from the camera or the structure of the imaged scene is approximately planar. However, perspective still exists especially when the system configuration does not satisfy the above assumption. In our algorithm, we only use this 2D affine transform for coarse stereo matching to guide the fine stereo matching, which will be detailed later.

B. Nonparametric Ttransformation

The nonparametric transformation has been proposed for the stereo matching problem as the basis for correlation [12]. It does not depend on the intensity values, but on the comparison of the intensity of the center pixel to that of the neighborhood. Defining

$$N(L,L') = \begin{cases} 0, I(L) <= I(L') \\ 1, I(L) > I(L') \end{cases} \tag{10}$$

where

$N(L,L')$: comparison of intensities of pixels L and L';

L: one pixel;

$I(L)$: intensity of pixel L;

L': pixel in the neighbor area $A(L)$ of L with the diameter of d around.

then we get the nonparametric transformation of pixel L, which is the ordered sequence

$$\mathrm{E}(L,d) = \bigcup_{L' \in A(L)} N(L,L') \tag{11}$$

The nonparametric transformation is invariant to certain types of image distortion and noise, resulting in improved performance near object boundaries.

1. Rank Transformation

The rank transformation is a nonparametric measure of the local intensity. Its reliability was well studied in [13]. The rank transformation of any pixel is the number of pixels whose intensity is less than that of the center pixel in the local area.

$$R(L) = \sum_{L' \in A(L)} N(L,L') \tag{12}$$

where

$R(L)$: the rank transformation of the pixel L;

Obviously $R(L)$ is not an intensity at all, but an integer.

2. Census Transformation

The census transformation is a nonparametric structure of the intensities of pixels around the center pixel. It maps them into a bit string.

$$C(L) = \bigotimes_{L' \in A(L)} N(L,L') \tag{13}$$

where $\otimes$ means concatenation.

The census transformation shows the distribution of the intensity in the neighborhood of the central pixel.

IV. Algorithm

The block diagram of this algorithm for short baseline stereo matching is illustrated in Fig. 1. With only three known or detected reference points in the stereo images, the 2D affine transform can be recovered. Then it is used to model the stereo system and provide the approximate disparity information for any other point. To deal with weak perspective, the correlation and the nonparametric transformation are used to vote for the precise stereo matching. Since they are guided by the 2D affine transform and limited to small neighbor windows, the stereo matching

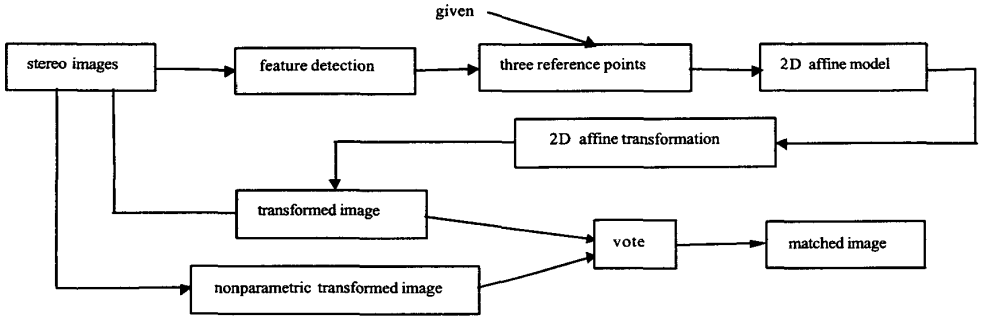


Fig. 1 Block diagram of the proposed stereo matching algorithm

can be completed precisely.

A. Coarse Stereo Matching with the 2D Affine Transform

From the 2D affine model, we get Eq.(9) in section 3.1, which can be rewritten as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad (14)$$

where

$p = [x \quad y]^t$, image of point P in the first stereo image;

$p' = [x' \quad y']^t$, image of point P in the second stereo image;

$U = \begin{bmatrix} a & b \\ d & e \end{bmatrix}$;

$t_o = [c \quad f]^t$;

With three reference points across the stereo images, P_1, P_2 and P_3, Eq. (14) can be rewritten as

(x_i', y_i') : image of P_i in the second stereo image;

The six unknowns, *a, b, c, d, e,* and *f*, can be determined. Therefore, the 2D affine transform can be recovered. The three corresponding reference points can be obtained by feature detection described in [6][7] [12][13]. They can also be given as the knowledge in some systems where the cameras are static. These three reference points are not necessary to appear on each pair of stereo images.

Since these six parameters only depend on the stereo system configuration and remain the same for all the pixels across the stereo images, for any pixel in the first stereo image, we can get the position of the corresponding pixel in the second stereo image with Eq.(14). However, because no perspective is considered in the affine camera model, stereo matching obtained only by this 2D affine model is not precise. Even with the assumption described previously, with this algorithm we can only get the coarse stereo matching.

B. Fine Stereo Matching Finished by the Correlation and the Nonparametric Transformation

The block diagram of the fine stereo matching is shown in Fig.2. After performing the coarse stereo matching using the 2D affine transform described in section 4.1, we perform fine stereo matching with nonparametric transformation followed by correlation around every rough matched pixel within a small window, which makes the stereo matching more precise.

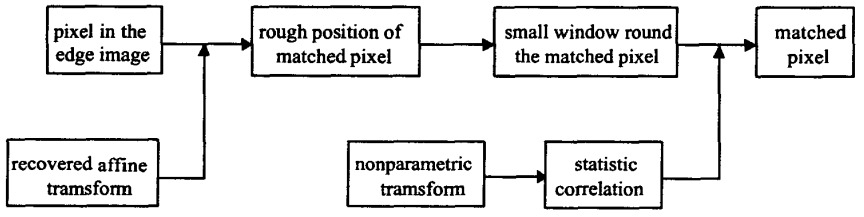


Fig. 2 Block diagram of the fine stereo matching by the nonparametric transform

$$F\vec{u} = \vec{k} \qquad (15)$$

where

$$F = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_2 & y_2 & 1 \\ x_3 & y_3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_3 & y_3 & 1 \end{bmatrix}$$

$\vec{u} = [a \quad b \quad c \quad d \quad e \quad f]^t$

$\vec{k} = [x_1' \quad y_1' \quad x_2' \quad y_2' \quad x_3' \quad y_3']^t$

(x_i, y_i) : image of P_i in the first stereo image;

The fine stereo matching is illustrated in Fig. 3 where p is any pixel in the first image of the stereo images. After the 2D affine transformation we can get its rough matched pixel p_r' in the second stereo image. With the assumption of the stereo system mentioned before, the exactly matched pixel p' should be within a small neighbor window around p_r', which we call the search window. Then the correlation based on the result of the nonparametric transformation and intensity is performed within the search window, and the precisely matched pixel p' can be obtained. To compare the similarity of the census transformed images, the correlation should be based on the Hamming distance, which means the number of bits that differ between the two bit strings.

Generally the size of the search window depends on the degree of similarity between the physical stereo system and the affine camera model. If the stereo system is calibrated, the fine stereo matching can also be performed along a band inside the search window in the same direction as the epipolar line shown in Fig. 3, leading to a smaller search scope than the search window.

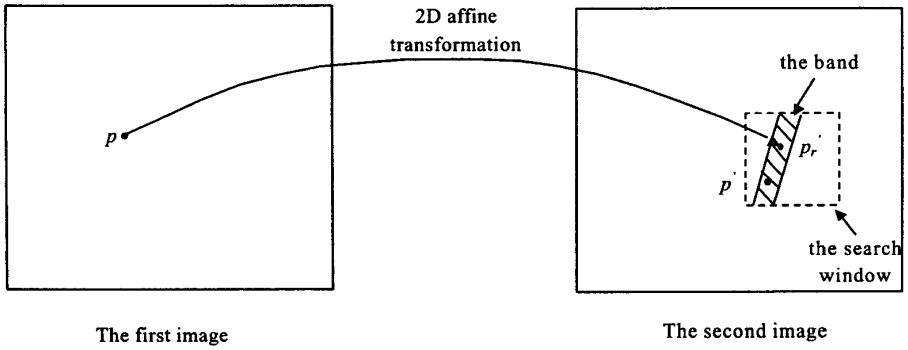


Fig. 3 The fine stereo matching

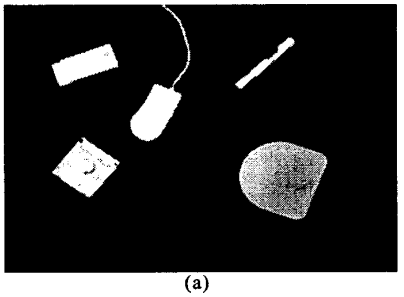
One shortcoming of the nonparametric transformation is that it depends heavily on the center pixel but the information associated with it is not very large [12]. However, in our algorithm because the fine stereo matching is guided by the 2D affine transform and limited to a small window or a band, the stereo matching can be more precise.

V. Experiments

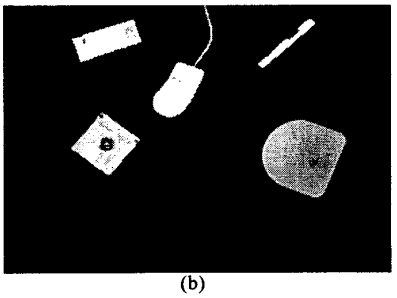
The performance of the algorithm, which is for stereo matching with short baselines described in section 3 and 4, is shown by stereo matching with real images named as STATIONERY. They are a set of real images, where the distance between the scene and cameras is about 2 meters and the baseline between the cameras is about 0.25 meters, which is a short baseline compared to the distance between the scene and cameras. The scene is also somewhat planar and the resolution of images is 640×480 pixels.

In our example, we perform the stereo matching from left to right using the binarized edge images. We use the binarized edge images because they enhance the highly textured area and retain most of the information.

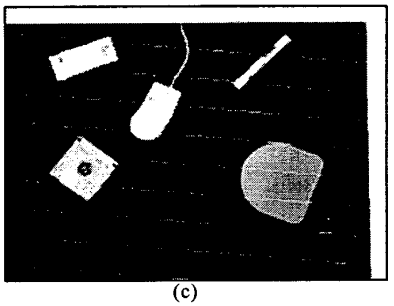
The original gray stereo images are showed in Fig.4 (a) and (b). Fig .4 (c) is the transformed left gray image based on the recovered 2D affine transform. In Fig.4 (c) there are some blank areas because some points are only visible in the right image and no matched pixels exist in the left image. Some discontinuity exists because of the calculation. The transformed binarized edge image of the left stereo image with the 2D affine transform is compared to the binarized edge image of the right image, and the result is shown as Fig.4 (d). Mainly because of the perspective there is little but obvious difference between them. Therefore only rough stereo matching can be obtained using the 2D affine transformation and fine stereo matching is required. Fig.4 (e) is the matched image, which is first guided by the 2D affine model and finished by the correlation based on the intensity and the nonparametric transform. The comparison between the matched image and the binarized edge image of the right image is shown in Fig.4 (f). We can see that they overlap almost perfectly, demonstrating good stereo matching.



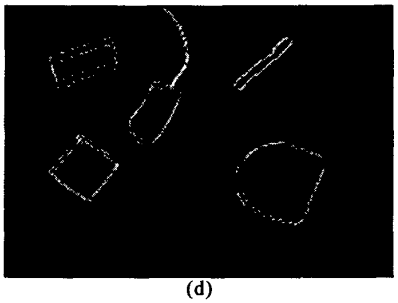
(a)



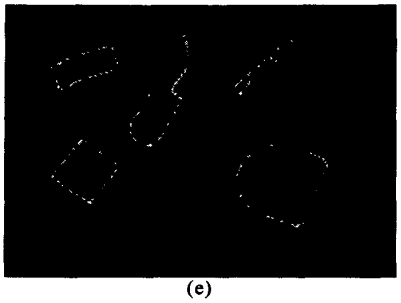
(b)



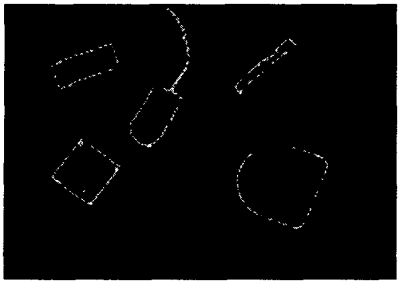
(c)



(d)



(e)



(f)

Fig. 4 STATIONERY

VI. Conclusion

In this paper, a new algorithm for stereo matching witha short baseline is presented. The stereo matching is first guided by the 2D affine transform, and then finished using correlation based on the intensity and the nonparametric transform.

With only three given or detected reference points, the 2D affine transform can be recovered. Coarse stereo matching can be performed and a rough matched image can be obtained. The nonparametric transformation and correlation can refine the stereo matching within a search window around the rough matched position. This stereo matching algorithm is not a general approach. However, it works well especially in the situation where either the structure of the scene is planar or the scene is far enough from cameras. Experiments show that good quality stereo matching can be obtained.

[1] L. Falkenhagen, "Depth estimation from stereoscopic image pairs assuming piecewise continuos surface," *Image Processing for Broadcast and Video Production*, pp. 115–127, Springer Great Britain, 1994.

[2] O. Faugeras, *Three-dimensional computer vision*, MIT Press, 1993.

[3] V. Gouet, P. Montesinos and D. Pel, "A fast matching method for color uncalibrated images using differential invariants," *British Machine Vision Conference*, 1998, Vol. 1, pages 367–376.

[4] D. Lowe, "Object recognition from local scale-invariant features," In *ICCV99*, pp. 1150–1157.

[5] Adam Baumberg, "Reliable feature matching across widely separated views," in *Proc. IEEE Conf. on Comput. Vis. and Pat. Recog.*, 2000, Vol: 1, pp: 774-781,

[6] R. Deriche, Z. Zhang, Q. Luong and O. Faugeras, "Robust recovery of the epipolar geometry for an uncalibrated stereo rig," In *ECCV94*, pp: 567–576.

[7] Ebroul Izquierdo M, "Stereo matching for enhanced telepresence in three-dimensional videocommunications," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No. 4, pp. 629-643, August 1997.

[8] H. H. Baker and T. O. Binford, "Depth from edge and intensity based stereo," in *Proc. 7th Int. Joint Conf. Artificial Intelligence*, Aug. 1981, pp. 631–636.

[9] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 139–154, Mar.1985.

[10] J. L. Mundy and A. Zisserman (eds.), *Geometric invariance in computer vision*, MIT Press, 1992.

[11] Nikos Grammalidis, "Sprite generation and coding in multiview image sequences," IEEE Transaction on Circuits and Systems for Video Technology, Vol. 10, No. 2, pp. 302-310, March 2000.

[12] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Euro. Conf. Comput. Vis.*, May 1994, pp. 151–158.

[13] Jasmine Banks, Mohammed Bennamoun, "Reliability analysis of the rank transform for stereo matching," *IEEE Transactions on Systems, Man, And Cybernetics–Part B: Cybernetics*, Vol. 31, No. 6, pp. 870-880, December 2001.