A New Approach for Stereo Matching in Autonomous Mobile Robot Applications

Pasquale Foggia
Dipartimento di Informatica e Sistemistica
Università di Napoli, Via Claudio 21, I80125,
Napoli, ITALY
foggiapa@unina.it

Jean-Michel Jolion
Lyon Research Center for Images and Information
Systems, UMR CNRS 5205 Bat. J. Verne INSA
Lyon 69621, Villeurbanne Cedex, France
Jean-Michel.Jolion@insa-lyon.fr

Alessandro Limongiello **Mario Vento**
Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica
Università di Salerno, Via Ponte don Melillo, I84084 Fisciano (SA), ITALY
[mvento, alimongiello]@unisa.it

Abstract

We propose a new approach for stereo matching in Autonomous Mobile Robot applications. In this framework an accurate but slow reconstruction of the 3D scene is not needed; rather, it is more important to have a fast localization of the obstacles to avoid them. All the methods in the literature are based on a punctual correspondence, but they are inefficient in realistic contexts for the presence of uniform patterns, or some perturbations between the two images of the stereo pair. Our idea is to face the stereo matching problem as a matching between homologous regions, instead of a point matching. The stereo images are represented as graphs and a graph matching is computed to find homologous regions. We present some results on a standard stereo database and also on a more realistic stereo sequence acquired from a robot moving in an indoor environment, and a performance comparison with other approaches in the literature is reported and discussed. Our method is strongly robust in case of some fluctuations of the stereo pair, homogeneous and repetitive regions, and is fast. The result is a semi-dense disparity map, leaving only a few regions in the scene unmatched.

1 Introduction

A pair of images acquired from a stereo camera contains depth information about the scene: this is the main assumption of stereo vision, based on the binocular parallax property of the human visual system. The main difficulty is to establish a correspondence between points of the two images representing the same point of the scene; this process is called *disparity matching*. The set of displacements between matched pixels is usually indicated as *disparity map*.

All the approaches, in the literature, are based on this punctual definition of the disparity. In this paper we propose an extension of that concept, namely we define a disparity value for a whole region of the scene starting from the two homologous views of it in the stereo pair. The main reason of this extension is that a punctual approach is redundant for Autonomous Mobile Robot (AMR) applications. In fact, in AMR applications, it is not very important to have a good reconstruction of the surfaces, but it is more important to identify adequately the space occupied by each object in the scene (as soon as possible to avoid collisions), even by just assigning to it a single disparity information. Moreover the punctual approaches are lacking in robustness in some realistic frameworks, especially for video acquired from a mobile platform. The algorithms based on correlation, which are available in off-the-shelf systems, are unable to deal with large uniform regions or with vibration of the cameras. On the contrary, some efforts have been done in the literature to improve the robustness of the algorithms, but in despite of the running time. Our method estimates the average depth of the whole region by an integral measure, and so has less problems with uniform regions than other methods have. The estimate of the position of the regions is sufficiently accurate for navigation, also in the mentioned cases, and it is fast enough for real time processing.

Now, we report a brief description of several methods to show better the limits of punctual approaches. For more details, there is a good taxonomy of dense two-frame stereo correspondence algorithms, recently proposed by Scharstein and Szeliski [2002], and a survey on stereo vision for mobile robot by Zhang [2002]. There are two major types of techniques, in the literature, for disparity matching: the area-based (also known as correlation-based) and feature-based techniques. Moreover, the area-based algorithms can be classified in local and global approaches.

The local area-based algorithms [Kanade and Okutomi, 1994; Fusiello and Roberto, 1997; Veksler, 2001] provide a correspondence for each pixel of the stereo pair. They

assume that a pixel is surrounded by a window of pixels with the same disparity, and these windows of pixels are matched. They produce a dense disparity map, excessive for AMR aims. Furthermore, they can be quite unreliable not only in homogeneous regions, but also in textured regions for an inappropriately chosen window size.

On the other side, the global area-based approaches try to propagate disparity information from a pixel to its neighbors [Marr and Poggio, 1976; Zitnick and Kanade, 2000], or they define and minimize some energy function over the whole disparity map [Geiger *et al*., 1995; Roy, 1999; Boykov et al., 2001]. They have a better performance in homogeneous regions, but they frequently have parameters which are difficult to set, and are highly time-consuming.

The feature-based approaches [Marr and Poggio, 1979; Grimson, 1981; Candocia and Adjouadi, 1997] detect and match only "feature" pixels (as corner, edges, etc.). These methods produce accurate and efficient results, but compute sparse disparity maps (only in correspondence to the feature points). Therefore, AMR applications require more details, in fact some information about the size; also a rough shape of the objects is needed for guiding a robot in the environment or for basic recognition tasks (e.g. in industrial applications, or for platooning of robots).

All the proposed methods, as already said, look for a punctual matching in the stereo pair. Therefore, some constraints both on the scene and on the input images have been introduced, since the first works on the stereopsis by Marr and Poggio [1976;1979], in order to guarantee good results and to reduce the complexity. To guarantee these constraints, the stereo pair is supposed to be acquired from a sophisticated system, so that the energy distributions of the two images are as similar as possible. Moreover, a pre-processing phase is needed, before the correspondence finding step, to compensate the hardware setup (*calibration* phase), or to assume an horizontal epipolar line (epipolar *rectification*). Unfortunately, in realistic applications of mobile robot these constraints are not easy to guarantee. The two images of the stereo pair could have a different lighting, the motion of the mobile platform on a rough ground should produce mechanical vibrations of the cameras, and consequently local or global perturbations between the two images, that could undermine the initial phases of calibration and rectification. We want to relax some constraints on the input images in order to consider a more realistic acquiring system, and consequently we add some constraints on our goal.

In this paper we propose, as said before, an extension of the disparity concept. The main idea is to determinate a unique disparity value for a whole region of the scene and not for a pixel. In fact, even if we can suppose an unique correspondence between each pixel in the left and right images from an optical point of view (as said by the *uniqueness constraint*), in some cases we can not have enough information to find this correspondence looking just at a single pixel. Let us consider, for example, pixels inside homogeneous areas, or belonging to repetitive patterns, or pixels suffering from perspective or photometric distortions. We prefer to have a 2D and ½ representation of the scene, assigning a constant disparity value to a whole region and only to some significant regions. Therefore, we propose a new stereo matching approach based on a region segmentation of the two images and a graph representation of these regions, to face the matching problem as a graph matching problem. The computational process is simpler and faster, because we consider only some significant regions, i.e. big areas, or some areas selected by a specific target. The result is more robust in a realistic context, because an integral measurement of the disparity for the whole region can mitigate some local and global fluctuations between the two images.

This paper is organized as follows: Section 2 presents an overview of our approach; Section 3 is devoted to the segmentation and the graph representation of the stereo pair; Section 4 shows the graph matching between the left and right image and Section 5 shows the disparity computation. Finally, in Section 6 there is a discussion of experimental results on standard stereo database and also on our stereo video sequence. Conclusions are drawn in Section 7.

2 Overview of our approach

The main idea of our approach is to obtain a disparity map looking at the distance between homologous regions (instead of pixels) in the stereo images. Let these regions be called *blobs*. In this way the computation of the disparity map is carried out on a set of pixels having the same spatial and color proprieties, producing a more robust performance with respect to local and global perturbations in the two images.

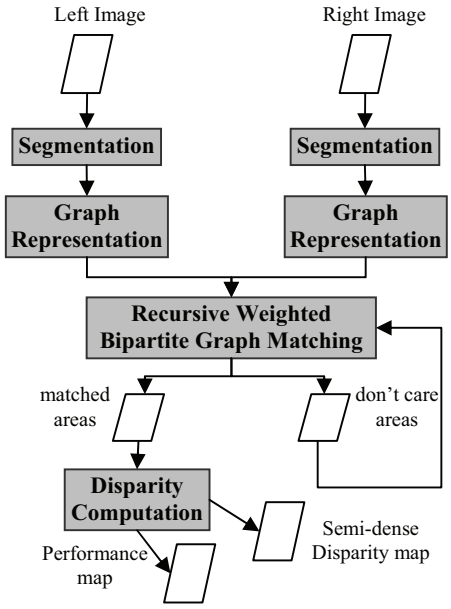

Figure 1: A scheme of our approach.

It should be noted that a blob is not an object; objects are decomposed into several blobs, so the overall shape of the object is however reconstructed, except for uncommon

pathological cases. An example of pathological case can be a uniform object almost along the line of sight, but it has been satisfactorily dealt with only by global criteria optimization, which is extremely time consuming.

In our approach (see Figure 1), the left and right images are segmented and each area identifies a node of a graph (see section 3). The segmentation process is simple and very fast. In fact, we are not interested in a fine segmentation, because we do not pursue a reconstruction aim. Anyway, we need similar segments between the left and right image in order to correctly find homologous regions. This objective is possible, in fact the stereo images are likely similar because they represent two different view points of the same scene. Moreover, the segmentation process does not influence the rest of algorithm, because a recursive definition of the matching (see section 4) and a performance function (see section 5) guarantee a recovery of some segmentation problems.

A bipartite graph matching between the two graphs is computed in order to match each area of the left image with only one area of the right image. This process yields a list of reliably matched areas and a list of so-called *don't care* areas. By calculating a vertical displacement between the corresponding areas, a depth is found for those areas of the reference image (i.e. left image). The list of the don't care areas, instead, could be processed in order to refine our result (see section 4).

As it is clear, this approach is robust even in case of uniform texture and it does not need a strong calibration process because it looks for area correspondence and not pixel correspondence. On the other hand, an effort is required in graph matching to assure real-time requirements. The application time is reduced using some constraints for a quicker computation of the bipartite graph matching (see section 4). Our method can be classified as a systemic approach [Jolion, 1994], in fact we consider constraints coming from the scene, from the objective and from the observer. In particular, with regard to scene constraints, we assume a strong continuity constraint for each selected region, and the compatibility and the uniqueness constraints are applied on the whole region and not longer on each pixel. The horizontal epipolar line constraint is generalized in a *horizontal epipolar band* (see section 4), to take the nature of the mobile observer into account. Moreover, the observer is supposed to move in an indoor environment and not too fast. Finally, the objective is considered to be real-time and highly related to the AMR applications. Therefore, all these constraints are taken into account to achieve our goal.

3 Segmentation and Graph representation

The first phase of the algorithm is the segmentation of the stereo images and their graph representation. We need a very fast segmentation process that produces similarly segmented areas between the left and right images. We have used a simple multi-threshold segmentation. It is essentially based on the quantization of the histogram in some color ranges (of the same size). The left and right segmentations are very similar, considering an adaptive quantization for each image according to its lighting condition. A connected component detection procedure is applied on each segmented image to obtain 4-connected areas of the same color. Each connected area (blob) is then represented as a node of an attributed graph. Each node has the following attributes:

- ***colMean***: the RGB mean value of the blob (*m_r, m_g, m_b*);
- ***size***: the number of pixels in a connected area;
- ***coord***: the coordinates of the box containing the blob (*top, left, bottom, right*);
- ***blobMask***: a binary mask for the pixels belonging to the blob.

It is easy to understand that a segmentation yielding many segments can be more accurate but creates lots of nodes, consequently requiring a more expensive graph matching process. On the other hand, a rougher segmentation process generates matching nodes that are very dissimilar in size and shape. As a compromise, we consider a segmentation process tuned to over-segment the image, and subsequently we filter the image in order to discard small noisy areas.

4 Graph Matching

Formally our matching algorithm can be described in the following way. A number of nodes is identified in each frame (left and right image) and a progressive label is associated to each node (blob). Let $G^L = \{N_0^L,\ldots,N_n^L\}$ and $G^R = \{N_0^R,\ldots,N_m^R\}$ be the two graphs representing the left and right image respectively. The solution of the spatial matching problem, between two stereo frames, is an injective mapping between a subset of G^L and a subset of G^R. The problem at hand can be represented by using a matrix whose rows and columns are respectively used to represent the nodes of the set G^L, and the nodes of the set G^R (correspondence matrix). The element (i,j) of the matrix is 1 if we have a matching between the element N_i^L with the element N_j^R, it is 0 otherwise. Each row contains no more than one value set to 1. If the j-th row or the i-th column contains only zeros, it means that it is a don't care node. The bijective mapping τ: $G^L \rightarrow G^R$ solves a suitable Weighted Bipartite Graph Matching (WBGM) problem. A Bipartite Graph (BG) [Baier and Lucchesi, 1993] is a graph where nodes can be divided into two sets such that no edge connects nodes in the same set. In our problem, the first set is G^L, while the second set is G^R. Before the correspondence is determined, each node of the set G^L is connected with each node of the set G^R, thus obtaining a Complete BG. In general, an assignment between two sets G^L and G^R is any subset of $G^L \times G^R$, i.e., any set of ordered pairs whose first elements belongs to G^L and whose second elements belongs to G^R, with the constraint that each node may appear at most once in the set. A maximal assignment, i.e. an assignment containing a maximal number of ordered pairs is known as a matching (BGM) [Kuhn, 1955].

A cost function is then introduced, so that each edge (N_i^L, N_j^R) of the complete bipartite graph is assigned a cost. This cost takes into account how similar are the two nodes N_i^L and N_j^R. The lower is the cost, the more suitable is that edge. If the cost of an edge is higher than a threshold (thrMatch), the edge is considered unprofitable and is removed from the graph (its cost is considered to be ∞).

Let us now introduce the cost function:

$$Cost = \frac{colCost + d\,i\,mCost + posCost}{3}$$

Where:

$$colCost = \frac{\sum_{i\in\{m\_r,m\_g,m\_b\}} \left| colMean_i^L - colMean_i^R \right|}{3*256}$$

$$d\,i\,mCost = \frac{\sum_{\substack{i\in\{bottom,right\}\\ j\in\{top,left\}}} \left| (i^L - j^L) - (i^R - j^R) \right|}{width + height}$$

$$posCost = \frac{\sum_{i\in\{bottom,right,top,left\}} \left| i^L - i^R \right|}{2*(width + height)}$$

where *width* and *height* are the dimensions of the frame. The matching with the lowest cost among the ones with maximal cardinality is selected as the best solution. The problem of computing a matching having minimum cost is called Weighted BGM (WBGM). This operation is generally time-consuming; for this reason the search area (that is the subset of possible couples of nodes) is bounded by the ***epipolar*** and ***disparity bands*** (see Figure 2).

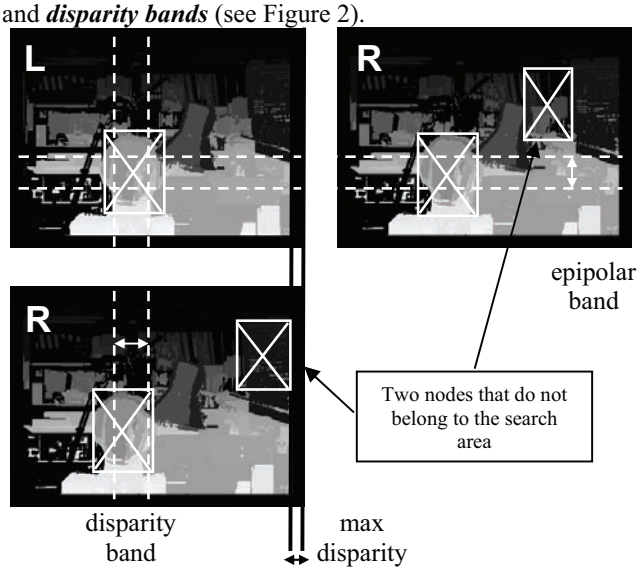

Figure 2: Epipolar and disparity bands: some constraints to optimize the WBGM.

These constraints come from stereo vision geometry, but in our case they represent a generalization. The epipolar band is a generalization for epipolar line, that is the maximum horizontal displacement of two corresponding nodes (generally its value can be a few pixel). Disparity band, instead, is a vertical displacement, so a node of the right image can move on the left almost of α*maxdisparity pixels (with α a small integer). These two displacements are computed with respect to the centers of the bounding box of the two blobs.

The graph matching process yields a list of reliably matched areas and a list of so-called *don't care* areas. The matched areas are considered in the following section for the disparity computation. The list of the don't care areas, instead, is processed in order to group adjacent blobs in the left and right image and consequently reduce split and merge artifacts of the segmentation process. Finally, a new matching of these nodes is found. The recursive definition of this phase assures a reduction of the don't care areas in few steps, but sometimes this process is not needed because don't care areas are very small.

5 Disparity Computation

The disparity computation is faced superimposing the corresponding nodes until the maximum covering occurs. The overlapping is obtained moving the bounding box of the smallest region into the bounding box of the largest one; precisely, the bounding box with the minimum width is moved horizontally into the other box, and the bounding box with the minimum height is moved vertically into the other box. The horizontal displacement, corresponding to the best fitting of the matched nodes, is the disparity value for the node in the reference image (left image).

Figure 3: Some examples of matched regions. In grey color the region from the right image and in white color the region from the left image.

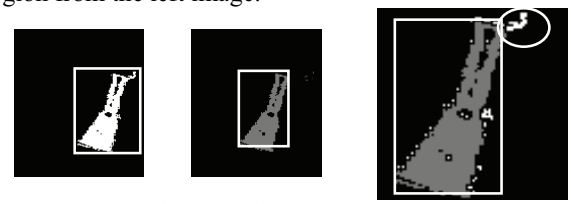

Figure 4: The overlapping process minimizes the mismatching between the two matched regions.

A lot of objects have some appendices (see Figure 4) depending on the different segmentation between left and right image. However, this process finds the correct value for the disparity, minimizing the mismatching between the two matched regions. Moreover, we propose a performance measurement for the disparity computation in order to

consider also some cases with larger errors coming from both segmentation and matching process.

$$performance = \frac{\max Fitting}{\max(sizeL, sizeR)}$$

It is the percentage value of the best fitting area size (*maxFitting*) with respect to the maximum size of the two matched regions (*sizeL* and *sizeR*).

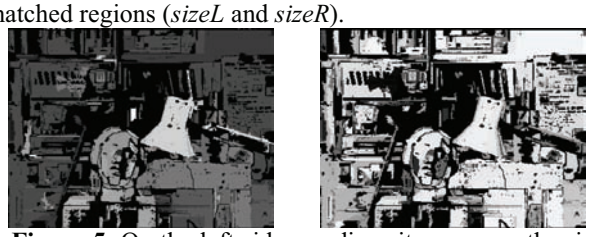

Figure 5: On the left side our disparity map; on the right side a graphical representation of the performance function (a brighter region has the upper value of performance).

The result of our algorithm can be represented in a graph, the so-called *disparity graph*, and, as it is clear from the Figure 5, the nodes of this graph can have a *don't care* attribute or, alternatively, the couple of disparity and performance attributes. Therefore, we could select a minimum performance value, and label the regions below this value as don't care. All these don't care areas could be processed again in the WBGM, as say in the section 3, if we should need to refine the result. Anyway, in our experimental results, we use a simple post-filtering in order to reduce don't care regions. Each 4-connected don't care area is labeled choosing the most frequent among the disparities of the adjacent regions. This assumption comes from the continuity constraint, but it is clear that it is applicable only inside a region and not between two different regions, so it is checked that most of the adjacent regions have the same disparity value. An example of the post-filtering use is shown in the Figure 6.

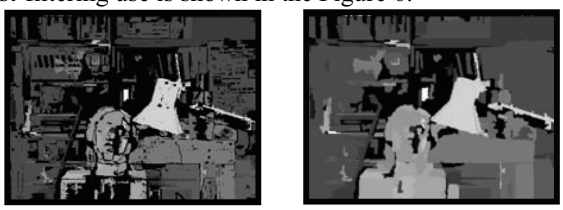

Figure 6: On the left side the original disparity map; on the right side the result after applying the post-filtering.

6 Experimental results

In the literature, tests are usually performed with standard databases composed of static images, well-calibrated and acquired in uniform lighting. The web site of Scharstein and Szeliski [Middlebury web site] is a good reference for some stereo images and to compare some stereovision algorithms. In this section we want to show our qualitative results and discuss some errors of the best algorithms in the literature, when applied to real cases. Nowadays, in AMR applications it is not defined a quantitative measurement for performance evaluation. On Middlebury web site it is proposed a quantitative performance evaluation for disparity map but in case of reconstruction aims. The following figure presents a stereo pair from the Tsukuba data set.

Figure 7: The reference image (on the left) and the ground truth (on the right), from Tsukuba data set.

The second image in Figure 7 represents the ground truth of the disparity map. An object has a higher grey level (corresponding to a high disparity between the two images) the closer it is to the camera, i.e. the lamp is in front of the statue that it in front of the table, etc.

The following Figure 8 shows our result on the Tsukuba DB and a comparison with other approaches. We have selected the best methods in the literature: squared differences (SSD), dynamic programming (DP) and graph cuts (GC) [Scharstein and Szeliski, 2002]. The first is a local area-based algorithm, the other two ones are global area-based algorithm. The experiments have been performed on a notebook Intel P4 1.5 GHz, 512 Mb RAM, and we have considered a resolution of 384x288 pixel.

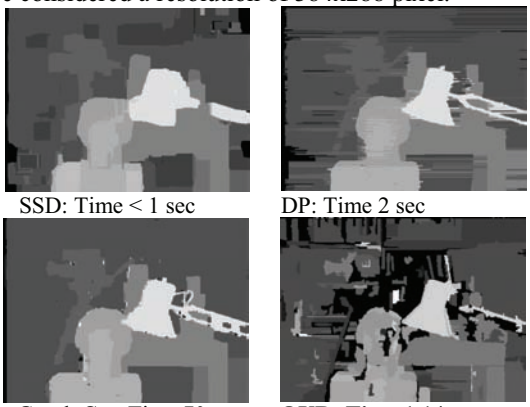

Figure 8: A comparison with other approaches.

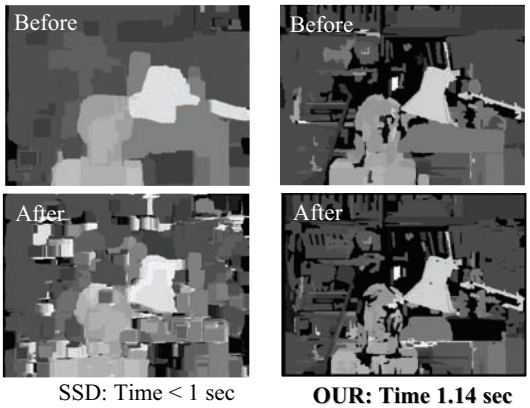

Figure 9: SSD and Our approach after a vertical translation of 2 pixels.

In Figure 9 it is clear the robustness of our approach in relation to the loss of the horizontal epipolar constraint.

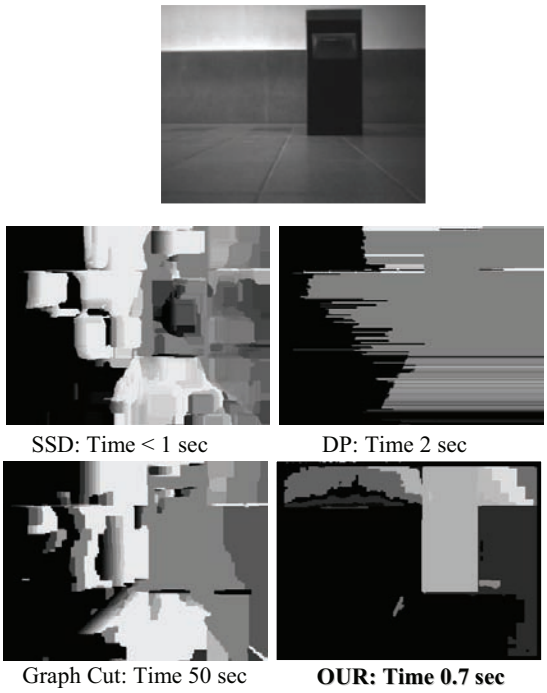

Figure 10: Results on our stereo pair: it is characterized by only one homogeneous object.

The presence of texture-less regions (very frequent in real contexts) causes serious problems to the best algorithms of the literature.

7 Conclusions

We have presented a stereo matching algorithm that is especially oriented towards AMR applications, providing a fast and robust detection of object positions instead of a detailed but slow reconstruction of the 3D scene. The algorithm has been experimentally validated showing an encouraging performance when compared to the most commonly used matching algorithms, especially on real-world images. Furthermore, we are working on a further improvement of the performance by defining an hybrid approach, i.e. we can apply our algorithm only to the large homogenous regions, and use traditional dense stereo matching for the rest.

References

[Scharstein and Szeliski, 2002] D. Scharstein, R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1):7-42, May 2002.

[Zhang, 2002] C. Zhang. A Survey on Stereo Vision for Mobile Robots. *Technical report, Dept. of Electrical and Computer Engineering, Carnegie Mellon University*, Pittsburgh, PA, 15213, USA, 2002.

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

[Fusiello and Roberto, 1997] A. Fusiello and V. Roberto. Efficient stereo with multiple windowing. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 858–863, Puerto Rico, June 1997.

[Veksler, 2001] O. Veksler. Stereo matching by compact windows via minimum ratio cycle. In *Proceedings of the International Conference on Computer Vision*, I: 540–547, Vancouver, Canada, July 2001.

[Marr and Poggio, 1976] D. Marr and T.A. Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, October 1976.

[Zitnick and Kanade, 2000] C.L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(7):675–684, July 2000.

[Geiger *et al*., 1995] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[Roy, 1999] S. Roy. Stereo without epipolar lines: A maximum-flow formulation. *International Journal of Computer Vision*, 34(2/3):1–15, August 1999.

[Boykov et al., 2001] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[Marr and Poggio, 1979] D. Marr and T.A. Poggio. A computational theory of human stereo vision. *RoyalP*, B-204:301–328, 1979.

[Grimson, 1981] W.E.L. Grimson. A computer implementation of a theory of human stereo vision. *Royal*, B-292:217–253, 1981.

[Candocia and Adjouadi, 1997] F. Candocia, and M. Adjouadi. A similarity measure for stereo feature matching. *IEEE Transaction on Image Processing*, 6:1460-1464, 1997.

[Baier and Lucchesi, 1993] H. Baier and C. L. Lucchesi. Matching Algorithms for Bipartite Graphs. *Technical Report DCC-03/93, DCC-IMECC-UNICAMP*, Brazil, March 1993.

[Kuhn, 1955] H.W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2: 83-97, 1955.

[Middlebury web site] http://cat.middlebury.edu/stereo/

[Jolion, 1994] J.M. Jolion. Computer Vision Methodologies. CVGIP: Image Understanding, 59(1):53–71, 1994.