Scene Segmentation Assisted by Stereo Vision

Carlo Dal Mutto and Pietro Zanuttigh and Guido M. Cortelazzo
Department of Information Engineering (DEI)
University of Padova
Padova, Italy
Email: dalmutto,zanuttigh,corte@dei.unipd.it

Stefano Mattoccia
Department of Electronics, Computer Science
and Systems(DEIS), University of Bologna
Bologna, Italy
Email: stefano.mattoccia@unibo.it

***Abstract*—Stereo vision systems for 3D reconstruction have been deeply studied and are nowadays capable to provide a reasonably accurate estimate of the 3D geometry of a framed scene. They are commonly used to merely extract the 3D structure of the scene. However, a great variety of applications is not interested in the geometry itself, but rather in scene analysis operations, among which scene segmentation is a very important one. Classically, scene segmentation has been tackled by means of color information only, but it turns out to be a badly conditioned image processing operation which remains very challenging. This paper proposes a new framework for scene segmentation where color information is assisted by 3D geometry data, obtained by stereo vision techniques. This approach resembles in some way what happens inside our brain, where the two different views coming from the eyes are used to recognize the various object in the scene and by exploiting a pair of images instead of just one allows to greatly improve the segmentation quality and robustness. Clearly the performance of the approach is dependent on the specific stereo vision algorithm used in order to extract the geometry information. This paper investigates which stereo vision algorithms are best suited to this kind of analysis. Experimental results confirm the effectiveness of the proposed framework and allow to properly rank stereo vision systems on the basis of their performances when applied to the scene segmentation problem.**

***Keywords*-scene segmentation; stereo; 3D;**

I. Introduction

Stereo vision systems provide estimates of the 3D geometry of a framed scene from two or more views of it. In this field there has been sizable amount of research and current methods are able to give dense and reliable depth information. Exhaustive surveys of stereo vision algorithms can be found in [3] and [12]. There are many different algorithms with different trade-offs between accuracy and computational requirements. Simple and fast stereo algorithms are usually not very well performing in terms of 3D geometry estimation, especially in presence of non-textured regions, while more complex algorithms (e.g. the global algorithms) have better performances. The main application of a stereo algorithm is of course the estimation of the framed scene 3D geometry, but many other applications can benefit from the information extracted by stereo vision systems. Even when the estimated 3D geometry is not very accurate, it can still be very useful in many scene analysis applications. Following this rationale, in this paper we propose to exploit stereo vision algorithms in scene segmentation. Scene segmentation is the well-known problem of identifying the image regions corresponding to the different scene elements. In light of the rapid evolution and progress of image capture technology, the vast attention received by scene segmentation directly on the basis of the scene image itself does not come as a surprise. Unfortunately scene segmentation from a single image is an ill-posed problem, still lacking robust solutions, despite a huge amount of research. Many segmentation techniques based on different insights have been developed, such as methods based on graph theory [6], methods based on clustering algorithms, e.g. [4], [13], and methods based on many other different techniques (e.g. region merging, level sets, watershed transforms and many others). The intrinsic limit of the classical approach is that the information contained in an image does not always suffice to completely understand the scene composition. A straightforward way to overcome the limits of the color information of an image is to consider it together with some 3D geometry information of the framed scene. The 3D scene geometry nowadays can be obtained by a variety of hardware and software techniques, the simplest and the most inexpensive of which is certainly stereo vision. Indeed it suffices to frame the scene just by two cameras, forming a stereo system, in order to have both 3D geometry and color information for each point seen by both the cameras (i.e., points not occluded and in the field of view of both cameras). This resembles what happens inside the human vision system, where the brain exploits the differences betweent the views of the two eyes as a clue to estimate the three dimensional structure of the scene and to recognize the different objects. In [7] the recognition of the foreground from the background is solved by exploiting two images and analyzing also the stereo match likelihood togehter with color information. This paper follows a similar rationale but frames scene segmentation as a three step procedure, characterized by: *a)* a stereo vision algorithm used to provide 3D geometry; *b)* a way of jointly representing 3D geometry and color information; *c)* a suitable clustering technique. While a single 3D geometry and color representation is proposed, various stereo vision algorithms and clustering methods are evaluated on the basis of the scene segmentation results. Section II presents an overall synthesis of the proposed scene segmentation framework. Section II-A reports all the considered stereo algorithms. In Section II-B, the proposed joint representation scheme for 3D geometry and color information is presented. Section II-C presents the considered segmentation algorithms based on clustering. Section III provides a comprehensive set of experimental results, and finally Section IV draws the conclusions.

II. Proposed framework

The proposed scene segmentation method requires as input two views of the same scene acquired by a standard stereo setup. The proposed scene segmentation method can be subdivided into four steps:

- Estimation of the 3D scene geometry by a stereo vision algorithm
- Construction of a new scene representation that jointly considers both geometry and color information
- Application of a clustering algorithm on the combined color and geometry data
- Final refinement stage in order to remove artefacts due to noise or errors in the geometry extraction.

The scheme in Fig. 1 shows a detailed overview of the architecture of the proposed scene segmentation method.

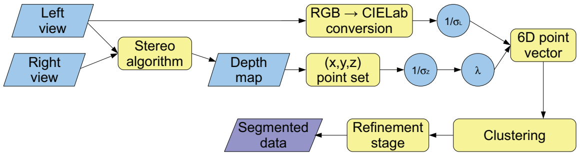

Figure 1. Architecture of the proposed scene segmentation method (segmentation of the left view, the same procedure can be applied to the rigth one).

A. Considered stereo vision algorithms

As already stated, the goal of the proposed method is to perform scene segmentation by exploiting both 3D geometry and color information acquired by a stereo vision system. Of course, different stereo vision algorithms produce different depth maps of the scene and such differences have an impact on the segmentation which needs to be examined. Indeed in order to verify the capability of the proposed method to perform scene segmentation, it is necessary to test how the various stereo vision algorithms affect segmentation. The depth map produced by each algorithm is then provided as an input for the clustering methods used for segmentation, described in Section II-C. Different artefacts in the depth maps cause different behaviours of the various clustering algorithms. The rest of subsection II-A presents the stereo algorithms used in our analysis.

1) Fixed Window: The Fixed Window (FW) algorithm is the basic local approach. It aggregates matching costs over a fixed square window and uses, as most local algorithms, a simple winner-takes-all (WTA) strategy for disparity optimization. Similarly to most local approaches, the aggregation of costs within a frontal-parallel support window implicitly assumes that all the points within the support have the same disparity. Therefore, FW does not perform well across depth discontinuities. Moreover, as most local algorithms, FW performs poorly in textureless regions. Nevertheless, thanks to incremental calculation schemes [5], [11], FW is very fast. For this reason, despite its notable limitations, this algorithm is widely used in practical applications. In our implementation the cost function is the Sum of Absolute Differences (SAD).

2) Adaptive Weights: The AdaptiveWeights (AW) algorithm [15] is a very accurate local algorithm that uses a fixed squared window but weights each cost within the support window according to the image content. The weights are first computed, on the left and right image, similarly to a bilateral filter (i.e. deploying a spatial and a color constraint) and then multiplied to obtain a symmetric weight assigned to each cost within the support window. This method uses the WTA strategy for disparity optimization and the sum of Truncated Absolute Differences metric for the costs. This method provides very accurate disparity maps and preserves depth discontinuities. However, as for other local approaches, this method performs poorly in textureless regions. Moreover, the support windows shrinks to few points (or equivalently, AW set very small weights for several points) in presence of highly textured regions making this method error prone. The AW algorithm is computationally expensive; to process a typical stereo pair it requires minutes (authors report 1 minute for small size images).

3) Segment Support: The Segment Support (SS) algorithm [14] is a local algorithms that aims at improving the AW approach by explicitly deploying segmentation. Similarly to AW, it aggregates weighted costs within a square support window of fixed size. Starting from the stereo pairs and the corresponding segmented stereo pairs, SS computes the weights on each image according to the following strategy. The weights of the points belonging to the same segment in which the central points lies is set to 1. The weight of the points outside such a segment are set according to color proximity constraint only and discarding the spatial proximity constraint. The overall weight assigned to each point is computed similarly to AW. In [14] it was shown that this strategy allows SS to improve the effectiveness of AW, near depth discontinuities and in presence of repetitive patterns and highly textured regions. However, similarly to other local approaches, this methods performs poorly in textureless regions. Although the segmentation of the stereo pairs can be quickly performed, SS has an execution time higher than AW. It is very interesting to apply SS in the proposed segmentation framework, because there is a segmentation step both before computing disparity and after the stereo matching calculation.

4) Fast Bilateral: The Fast Bilateral Stereo (FBS) approach [10] combines the effectiveness of the AW approach with the efficiency of the traditional FW approach enabling results comparable to AW much more quickly. In this algorithm the weights are computed on each image and on a block basis with respect to the central point according to a strategy similar to AW. The weight assigned to each block is related to the difference between the color intensity of the central point and the average color intensity of the block. The costs within each block are computed, very efficiently, on a point basis by means of incremental calculation schemes. Therefore, at each point within a block, this method assigns the same weight and its point-wise matching cost. Disparity optimization is based on the WTA strategy. With block of size 3×3, FBS obtain results comparable to AW, well preserving depth discontinuities, in a fraction of the time required by AW. Increasing the block size decreases the accuracy of the disparity maps but reduces the execution time further. Moreover, in [10] it was shown that computing weights on block basis makes this method more robust to noise compared to AW. Similarly to other local algorithms describes so far, FBS performs poorly in textureless regions.

5) Semi-Global: The Semi Global Matching (SGM) algorithm [9] explicitly models the 3D structure of the scene by means of a point-wise matching cost and a smoothness term. However, this method is not a traditional global approach since the minimization of the energy function is computed, similarly to Dynamic Programming or Scanline Optimization approaches, in a 1D domain [12]. That is, several 1D energy functions computed along different paths are independently and efficiently minimized and their cost summed up. For each point, the disparity corresponding to the minimum aggregated cost is selected. In [9] the author propose to use 8 or 16 different independent paths. The SGM approach works well near depth discontinuities, however, due to its (multiple) 1D disparity optimization strategy, produces less accurate results than more complex 2D disparity optimization approaches. Despite its memory footprint, this method is very fast (the fastest algorithm between those considered) and potentially capable to deal with poorly textured regions.

6) Graph Cut: The Graph Cut stereo vision algorithm (GC) introduced in [2] is a global stereo vision method, that explicitly accounts for depth discontinuities by minimizing an energy function that combines a point-wise matching cost and a smoothness term. The GC algorithm models the 3D scene geometry with a Markov random field in a Bayesian framework and determines the stereo correspondence solving a labeling problem. The energy function is represented as a graph and its minimization is done by means of graph cut, an efficient algorithm that relies on the Min-Cut/Max-Flow theorem. As most global methods, GC is computational expensive and has a large memory footprint. However, global algorithms can deal with depth discontinuities and textureless regions.

With the exception of GC, in our implementations of all the considered algorithms there is a standard sub-pixel refinement step based on the fitting of a parabola in proximity of the best disparity, Moreover, occlusions are explicitly computed by cross-checking the disparity maps computed according reference and target images. We implemented all the stereo vision algorithms, except for GC and SGM, for which we used the OpenCV implementation.

B. Joint representation of 3D geometry and color information

The application of one of any considered stereo vision algorithms returns a geometrical description of the scene. This estimated 3D geometry can be used together with color information in order perform scene segmentation. The combined use of geometry and color information in scene segmentation allows to obtain better results than using geometry or color information only. The usage of geometry only may allow good segmentation performances, but there are situations that cannot be dealt on the basis of such information alone. A typical example is the case of objects of different colors but placed close one to the other (e.g., two people wearing different clothes but very close each other). At the same time color information can not distinguish objects with similar colors even if they are distant from each other. To exploit both types of information at the same time it is first of all necessary to build an unified representation that includes both color and 3D geometry data. Given a scene $\mathcal{S}$, after applying one of the stereo algorithms, both 3D geometry and color information are available for all the scene points $p_i \in \mathcal{S}, i = 1, ..., n$ visible in both images (non occluded points in the stereo vision system field of view). All such points can be represented by 6-dimensional vectors $\mathbf{p}_i = [L(p_i), a(p_i), b(p_i), x(p_i), y(p_i), z(p_i)]^T$, where the first three components of $\mathbf{p}_i$ represent color information and the other three components represent geometry. The color information vector is built as follows: first of all the available color data are converted from the RGB to the CIELab color space. A uniform color space ensures that the Euclidean distance between points is close to the perceptual difference between the various colors and allows to compare the distances in the three color channels. Color information is then normalized by the standard deviation σ_L of the L component in order to simplify the comparison with geometric data that are defined in a completely different space. The 3D geometry information of each scene point p_i is represented by the 3D vector $[x(p_i), y(p_i), z(p_i)]^T$ that contains the point position in the three dimensional space. The coordinates $x(p_i), y(p_i)$ and $z(p_i)$ of the points $p_i, i = 1, ...N$ can be computed from the depth map provided by the stereo vision system and the parameters of the acquisition systems (i.e., intrinsic parameters of the two cameras that form the stereo pair, and the baseline of the stereo pair). After representing each point p_i by its 3D coordinates $x(p_i), y(p_i)$ and $z(p_i)$, the resulting vectors are normalized by the standard deviation σ_z of the z coordinate[1]. The purpose of normalization is to help the comparison of data belonging to completely different physical cues (geometry and color). The trade-off between the relevance of color and depth information is controlled by a factor λ. The value of λ selects the relative importance of the two kinds of information. Hence, the final representation of each non-occluded point $p_i, i = 1, ..., N$ in the field of view of the acquisition system is the 6-dimensional vector $\mathbf{p}_i, i = 1, ..., N$, defined as in Eq. 1.

$$\mathbf{p}_i \triangleq \begin{bmatrix} \bar{L}(p_i) \\ \bar{a}(p_i) \\ \bar{b}(p_i) \\ \lambda\bar{x}(p_i) \\ \lambda\bar{y}(p_i) \\ \lambda\bar{z}(p_i) \end{bmatrix} = \begin{bmatrix} L(p_i)/\sigma_L \\ a(p_i)/\sigma_L \\ b(p_i)/\sigma_L \\ \lambda x(p_i)/\sigma_z \\ \lambda y(p_i)/\sigma_z \\ \lambda z(p_i)/\sigma_z \end{bmatrix}, i = 1, ..., N \quad (1)$$

High values of λ give more importance to geometry, low values of λ give more importance to color.

C. Scene Segmentation

The set $\mathcal{V}$ of the 6D vectors $\mathbf{p}_i, i = 1, ..., N$ is a unified description of the framed scene $\mathcal{S}$ that accounts for both geometry and color information. In the assumption that scene $\mathcal{S}$ is formed by different meaningful parts (i.e., different objects) $\mathcal{S}_k, k = 1, ..., K$, the segmentation is the task of finding the different groups of points representing the different objects. This can be formulated as the problem of clustering the vectors $\mathbf{p}_i, i = 1, ..., N \in \mathcal{V}$ into the clusters $\mathcal{V}_i, i = 1, ..., K$ that represents the various objects. Nowadays, there is a great variety of clustering algorithms that can be used in this task, each one with pros and cons. In order to test the correctness of our approach, and in order to test which stereo vision algorithm produces more effective data, three different clustering techniques were considered, namely k-means clustering, mean-shift and spectral clustering with Nyström method.

1) Segmentation by k-means clustering: K-means is a classical central grouping clustering algorithm. It is very simple to implement and it is pretty fast. It is not very precise when applied to scene segmentation, because it assumes that the distribution of the considered feature vectors $\mathbf{p}_i$ that represent the points $p_i, i = 1, ..., N$ is a mixture of Gaussians. This assumption is not generally verified in the scene segmentation contest and for this reason, this clustering method applied to the set $\mathcal{V}$ may give poor results.

2) Segmentation by mean-shift: The mean-shift algorithm [4], is a standard non-parametric feature-space analysis technique, that can be used as a clustering algorithm. It aims at locating the maxima of a density function, given some samples drawn from the density function itself. It is useful for detecting the modes of a density, and therefore for clustering the feature vectors in a very efficient way. Mean-shift clustering is very fast, but prone to return outliers. However it is worth considering this clustering technique, since it is very fast, quite reliable and widely used in computer vision and image analysis.

3) Segmentation by spectral clustering with Nyström method: This method, proposed in [13], is a state-of-the-art clustering algorithm. It is based on pairwise affinity measures computed between all possible couples of points in $\mathcal{S}$. It does not impose any model or distribution on the points $p_i, i = 1, ..., N$, and therefore its results in practical situations are more accurate and robust than those of k-means and mean-shift. Spectral clustering alone is very expensive for both CPU and memory resources. This characteristic is intrinsic to the nature of the algorithm, because the computation of a pairwise affinity measure between all the points $p_i \in \mathcal{S}$ requires to build a graph that has a node for each point and an edge between each couple of points. Such graph is usually very large. However, one may obtain an approximated version of such a graph by imposing that not all the points are connected. The Nyström method, proposed in [8], is a way to approximate the graph, based on the integral eigenvalue problem. Spectral clustering with Nyström method provides a nice framework to incorporate the fact that $\mathcal{S}$ has to be partitioned into subsets where color and 3D geometry are homogeneous. The resulting speed of spectral clustering with Nyström method is comparable with the ones of k-means and mean-shift.

III. Experimental results

The feasibility of a robust and effective scene segmentation exploiting both 3D geometry acquired by a stereo vision setup and color clues has been experimentally verified. All the combinations of above mentioned algorithms have been tested on various scenes from the standard Middlebury dataset [1] and on some scenes acquired in our laboratory. On the Middlebury dataset, some camera parameters are not available (e.g., the camera principal point), and were estimated in order to obtain a realistic 3D reconstruction from the ground truth. In the acquired scenes all the stereo vision system parameters were known. The limits of the usage of color only or geometry alone in segmentation is exemplified by the results of Fig. 2. A photo of the scene can be seen in Fig. 3a. As it is shown in Fig. 2a, classical scene segmentation algorithms, such as mean-shift, based on color information only, fail in the detection of many regions of the considered scenes because of complex texture patterns or similar colors. The same algorithm feeded with

[1] We assume the z axis to be parallel to the optical axis, i.e., $z(p_i)$ correspond to the depth of the point p_i

geometry data only gives better results, as shown in Fig. 2 b, but it cannot disambiguate close objects (such as the baby's feet on the open book). Before concentrating on the results of the combined usage of geometry and color, let's note that the occlusions are estimated via cross-checking by the stereo vision algorithm, and the occluded points are discarded in the clustering step. In Fig. 4, 6 and 8, the estimated occlusions are reported in black in the segmented image. In the scenes presented in Fig. 3 and 5 it is interesting to notice how the use of color allows the identification of the separation of the baby's feet from the book in Fig. 4 and of the plant from the vase in Fig. 6. These objects can not be separated by means of segmentation based on geometry only, as shown in Fig. 2b. Fig. 4 shows the results on the *Baby 2* image while Fig. 6 refers to the *Aloe* image. Finally Fig. 8 shows an example on the data acquired with our setup. Six different stereo vision algorithms and three clustering algorithm were tested. The superiority of the results based on both color and geometry versus the ones obtainable by just color or geometry is clear. What is more difficult to determine is which stereo and clustering algorithms combination turns out to be more effective. Such an evaluation was performed on the basis of a supervised metric which computes the percentage of missclassified pixels with respect to a ground truth segmentation, obtained from the ground truth depth map, which is available for each scene of the Middlebury data-set. This computation excludes the occluded pixels.

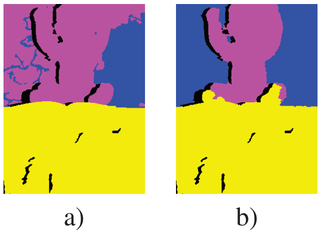

Figure 2. A classical scene segmentation method applied to color information only (a) and to geometry only (b).

The percentages of missclassified pixels for all the 18 combinations of stereo vision algorithms and clustering methods on the *Baby 2* image are reported in Table I, together with the execution time of the stereo algorithms. Almost all the scene segmentations, with the exception of the ones obtained by applying k-means on the *person* scene are robust and effective, far way better that what is delivered by classical scene segmentation algorithms based on color information only. According to the considered metric, the most effective combination is given by SS stereo vision and spectral clustering with Nyström method. It is interesting to notice that the usage of global or semi-global stereo algorithms versus the usage of local stereo algorithms does not lead to significant performance improvements. For example GC-based segmentation, specially combined with mean-shift clustering, on the person image (Fig. 8) leads to more artefacts than local stereo vision algorithm. FBS appears to be a very good trade off between computational efficiency and segmentation precision and robustness, even if though at times it may introduce false occlusions, as shown in Fig. 4. K-means clustering does not work properly in the *person* scene (Fig. 8). The more reliable clustering algorithm is spectral clustering with Nyström method, because it works robustly in all the scenes and with all the stereo vision algorithms.

	FW	AW	SS	FBS	SGM	GC
k-means	2.21	0.87	0.92	0.92	1.60	0.97
Mean-s.	2.31	1.33	1.02	0.98	1.62	1.02
Spectr.	2.03	0.84	0.81	0.93	1.45	0.97
St. Time	2.05s	207s	508s	29.7s	0.91s	28.4s

Table I
COMPARISON WITH THE SEGMENTATION PERFORMED ON THE MIDDLEBURY *baby 2* GROUND TRUTH: PERCENTAGE OF INCORRECTLY ASSIGNED PIXELS. THE LAST ROW SHOWS THE EXECUTION TIME (RELATIVE TO THE STEREO ALGORITHMS ONLY) ON A SINGLE CORE 2.53 GHZ MACHINE. GC AND SGM ARE HIGHLY OPTIMIZED ALGORITHMS, GC DOES NOT HAVE SUBPIXEL OPTIMIZATION.

Scene	Baby 2	Aloe	Person
λ	3	1.8	7
N. of clusters (K-m. and S. C.)	3	3	3
Kernel (Spectr. Cl.)	Gaussian	Gaussian	Gaussian
Spectr. Cl. Bandwidth	1	1	1
Meanshift Bandwidth	1.2	1.5	1

Table II
VALUES OF THE CLUSTERING ALGORITHMS PARAMETERS ADOPTED IN THE CONSIDERED SCENES, I.E., MIDDLEBURY BABY 2, MIDDLEBURY ALOE AND "PERSON".

Finally Fig. 9 shows how the segmentation results depend on the selected value of the λ parameter. A large value of λ gives more trust to depth information and allows to easily separate the foreground object from the background. A small value of λ forces the use or more color information thus allowing to better separate the baby from the box but it is also more noisy. By properly balancing the two clues it is possible to obtain optimal results (e.g. in the baby scene by setting $\lambda = 4$ it is possible to properly recognizing all the objects in the scene).

In terms of speed, mean-shift clustering is slightly faster than the other two algorithms (that are comparable). Our MATLAB implementations of the clustering algorithms takes less than 7 seconds, therefore the proposed methods are well suited to real time applications, once the proper hardware and software provisions are taken.

IV. CONCLUSIONS

This paper proposes a novel scene segmentation framework based on both 3D geometry and color information. Stereo vision techniques allow to extract 3D geometry from a pair of views of the scene and to use it together with color

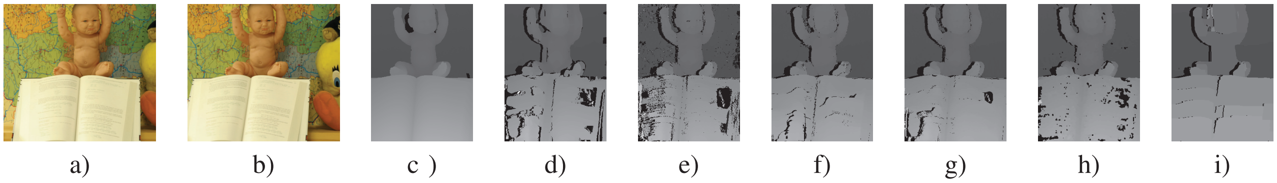

Figure 3. Middlebury Baby 2 dataset and stereo vision algorithms results: a) *View1* image, b) *View5* image, c) *disp1* ground truth disparity map, d) *disp1* disparity map obtained with FW, e) *disp1* disparity map obtained with AW, f) *disp1* disparity map obtained with SS, g) *disp1* disparity map obtained with FBS, h) *disp1* disparity map obtained with SG, i) *disp1* disparity map obtained with GC

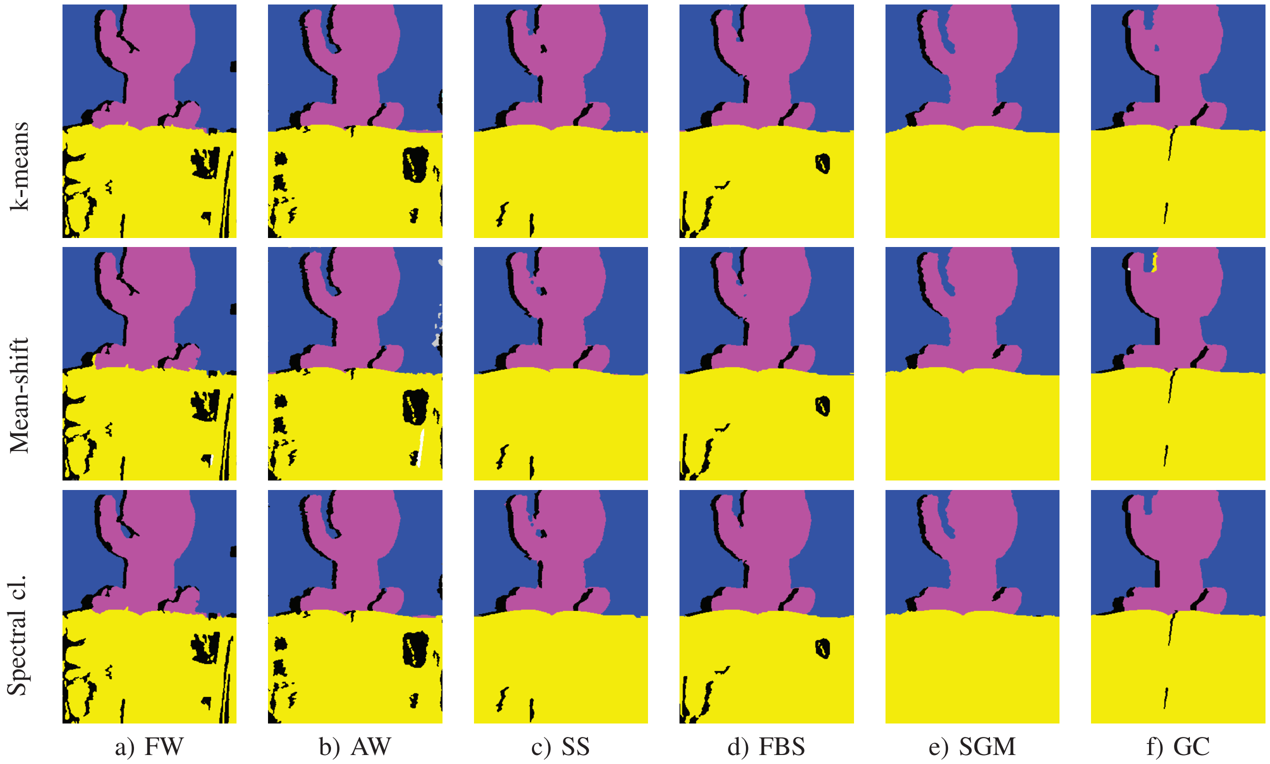

Figure 4. Middlebury Baby 2 dataset segmentation results: (first row) from k-means clustering; (second row) from mean-shift clustering; (third row) from spectral clustering with Nyström method; The columns refers to the different stereo vision methods: a) FW stereo algorithm, b) AW stereo algorithm, c) SS stereo algorithm, d) FBS stereo algorithm, e) SGM stereo algorithm, f) GC stereo algorithm.

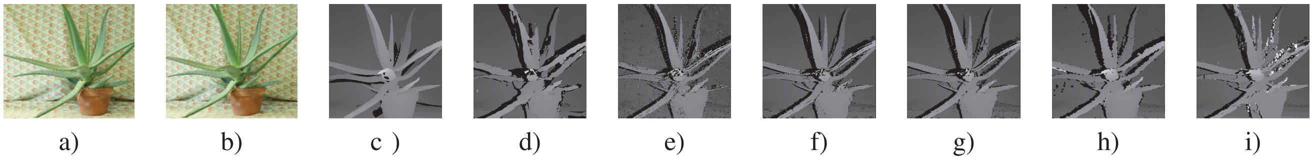

Figure 5. Middlebury Aloe dataset and stereo vision algorithms results: a) *View1* image, b) *View5* image, c) *disp1* ground truth disparity map, d) *disp1* disparity map obtained with FW, e) *disp1* disparity map obtained with AW, f) *disp1* disparity map obtained with SS, g) *disp1* disparity map obtained with FB, h) *disp1* disparity map obtained with SG, i) *disp1* disparity map obtained with GC

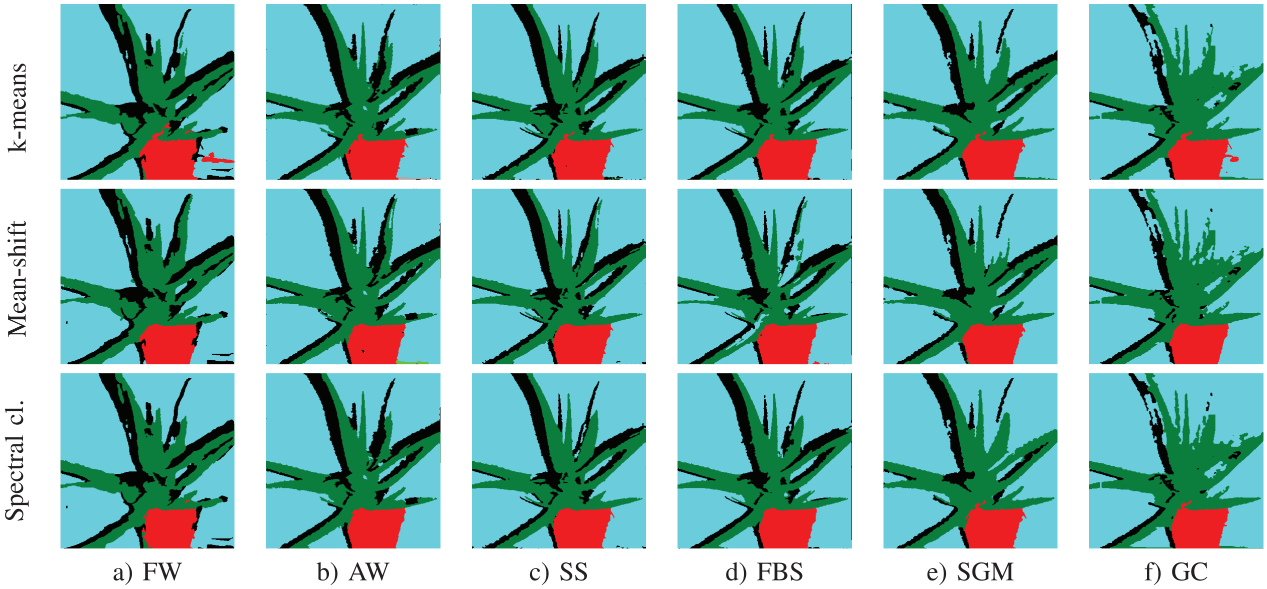

Figure 6. Middlebury Aloe dataset segmentation results in the same order of Fig. 4

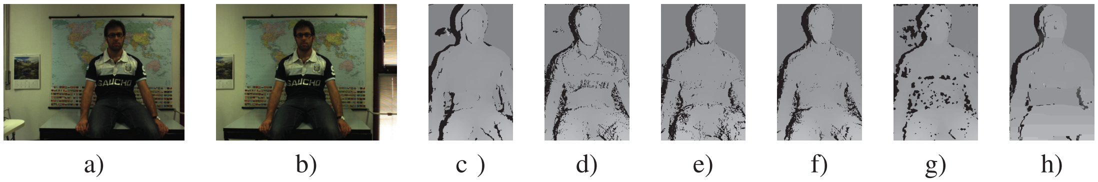

Figure 7. “Person” dataset and stereo vision algorithms results: a) *Left view* image, b) *Right view* image, c) Disparity map obtained with FW, d) Disparity map obtained with AW, e) Disparity map obtained with SS, f) Disparity map obtained with FB, g) Disparity map obtained with SG, h) Disparity map obtained with GC

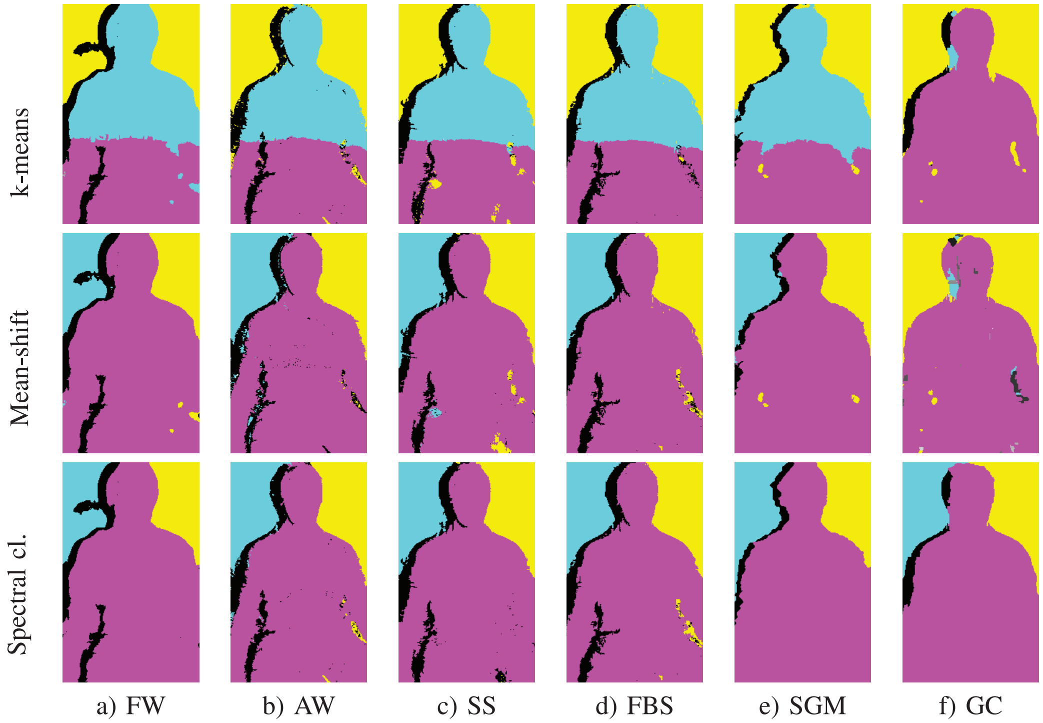

Figure 8. “Person” dataset and stereo vision algorithms results in the same order of Fig. 4

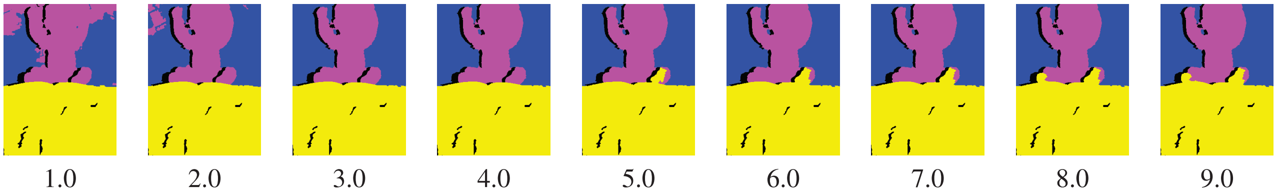

Figure 9. Segmentation results on the Middlebury Baby 2 scene corresponding to different values of the parameter λ (SS stereo vision algorithm).

information to segment it. The experimental results show that the proposed approach can provide a better segmentation than the methods based on just color or just geometry. Since the main ingredients of the proposed approach are specific stereo vision and clustering algorithms, this paper examines the results of the combinations between six different stereo vision and three different clustering algorithms. Among the various solutions the SS stereo vision algorithm combined with spectral clustering with Nyström method provides the best performances. This result is a little unexpected since it indicates that global stereo vision algorithms may not be needed in order to properly segment a scene. The acquisition system needed for the proposed scene segmentation approach is a regular stereo vision system, essentially requiring two cameras instead of a single camera, as the standard color based segmentation methods. It is certainly true that two cameras form a more complex set-up than a single camera, but new applications, advocated by 3DTV and gesture game interfaces, are making increasingly common 3D acquisition systems, among which stereo vision are the most inexpensive and popular. The overall quality of the obtained results is good enough to justify such a modest complication of the acquisition system. Future research will be devoted to the exploitation of the proposed scheme into stereo vision methods based on segmentation in order to improve both the segmentation and the quality of the extracted depth data, thus introducing an interesting coupling between the two problems. We will also explore the optimization of stereo vision algorithms for the segmentation task and the use of different acquisition methods for 3D geometry, like matricial Time-Of-Flight cameras and structured light cameras (e.g., Microsoft Kinect).

REFERENCES

[1] Middleburey, http://vision.middlebury.edu/stereo

[2] Y. Boykov and V. Kolmogorov, “An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, volume 26

[3] M. Z. Brown and D. Burschka and G. D. Hager, “Advances in Computational Stereo”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, volume = 25

[4] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, volume = 24 603-619

[5] F. C. Crow, “Summed-area tables for texture mapping”, SIGGRAPH, 1984

[6] P. F. Felzenszwalb. and D. P. Huttenlocher,“Efficient Graph-Based Image Segmentation”, International Journal Computer Vision, 2004, volume 59

[7] V. Kolmogorov and A. Criminisi and A. Blake and G. Cross and C. Pother, “Probabilistic fusion of stereo with color and contrast for bilayer segmentation,”IEEE Transactions on Pattern Analysis and Machine Intelligence , vol.28, no.9, pp.1480-1492, Sept. 2006

[8] C. Fowlkes and S. Belongie and F. Chung and J. Malik,“Spectral grouping using the Nystrm method”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, volume = 26

[9] H. Hirschmuller, “Stereo Vision in Structured Environments by Consistent Semi-Global Matching”, CVPR, 2006

[10] S. Mattoccia and S. Giardino and A. Gambini, “Accurate and Efficient Cost Aggregation Strategy for Stereo Correspondence Based on Approximated Joint Bilateral Filtering”, ACCV, 2009

[11] M. J. McDonald, “Box-Filtering Techniques”, Comp. graphics and image processing, 1981, volume 17

[12] D. Scharstein and R. Szeliski, “A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms”, International Journal of Computer Vision, 2001, volume 47

[13] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation ”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000

[14] F. Tombari and S. Mattoccia and L. Di Stefano, “Segmentation-based adaptive support for accurate stereo correspondence”, Pacific-Rim Symp. on Image and Video Tech., 2007

[15] K. J. Yoon and I. S. Kweon, “Adaptive Support-Weight Approach for Correspondence Search”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, volume 28