Integration of Perceptual Grouping and Depth

Juan Andrade-Cetto and Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial
Universitat Politècnica de Catalunya
cetto,sanfeliu@iri.upc.es

Abstract

Different data acquisition methods are tailored at extracting particular characteristics from a scene and by combining their results a more robust scene description can be created. A method to fuse perceptual groupings extracted from color-based segmentation and depth information from stereo using supervised classification is presented. The merging of data from these two acquisition modules allows for a spatially coherent blend of smooth regions and detail in an image. Depth cues are used to limit the area of interest in the scene and to improve perceptual grouping solving subsegmentation and oversegmentation of the original images. The complexity of the algorithm does not exceed that of the individual acquisition modules. The resulting scene description can then be fed to an object recognition module for scene interpretation.

Keywords*: data fusion, color-based segmentation, depth from stereo.*

1. Introduction

The fusion of 3D information from different acquisition methods allows for more robust scene descriptions. The shortcomings of individual low level processing modules can be overcome in an integrated environment. However, the inherent variability of the acquisition methods and the data formats and noise levels they produce make sensor fusion a challenging task. In the field of computer vision several attempts have been made at coupling data from different sensors. Integration models proposed in the literature vary in the number and type of sensor inputs, in the level at which fusion takes place, and in the rules used for data fusion.

In [10] an integration framework that encompasses four vision modules is presented. The merging of data from stereo, shape from shading, perceptual organization, and line labelling is used to estimate accurate depth maps of a scene. Other contributions consider the fusion of 2D and 3D data in the form of intensity images and stereo or range data [1, 9], or only 3D data acquired from stereo and range data [7, 8]. Methods that merge data from stereo and shape from shading include [2, 3, 5].

Most of the methods that fuse depth information are designed to work at a point or pixel level, whereas those methods that include perceptual grouping in 2D images are mostly directed towards fusing higher level primitives [10], or more specifically, at labelling primitives such as contours or edges based on their low-level properties, i.e., depth estimates. In [1, 8] emphasis is made in that an integration architecture should be made at both the pixel and higher perceptual grouping levels.

The methods used for fusing different types of data vary extensively, from ad-hoc implementations that use empirical thresholds to select which sensor contributes to scene formation, to the more elegant techniques of Extended Kalman Filtering to update depth estimates [7, 8], or Bayesian networks to integrate top-down and bottom-up visual processes [11]. Some exploit the fact that one of the data acquisition modules can be viewed as a multiresolution system and have embedded their data fusion techniques in between each level of resolution [10].

We present a method to fuse perceptual groupings extracted from color-based segmentation and depth information from stereo using supervised classification providing a robust solution to the sub- and oversegmentation problems characteristic of most color-based segmentation algorithms. Our contribution is mostly related to the work presented in [9, 10], but has a critical difference. In [10], segmentation results of grey-scale images are used, namely the segmented boundaries, to limit the enforcement of smoothing constraints in the stereo module, and to prevent the propagation of depth values across uniform regions. We believe that the flow of information should be considered in the opposite direction. This is, depth cues should be considered as an aid to perceptual grouping rather than using perceptual boundaries to limit the adjustment of depth estimates.

The reasoning behind this assumption is based on the fact that segmentation of intensity images is prone to illumi-

15th IAPR International Conference on Pattern Recognition
Barcelona, September 2000, pp. 295-298

nation conditions and surface properties, and the perceptual groupings thus produced by most segmentation algorithm might have considerable error. On the other hand, depth estimation at scene discontinuities is less sensitive to the factors mentioned above and can be used more robustly to further group segments into higher level perceptual entities that could match an object model, or to discriminate from noisy or undesirable segmented regions.

A description of our data acquisition modules followed by their integration is presented next. Experiments on complicated scenes have been performed and the results are also shown.

2. Low level processing modules

2.1. Depth from stereo

Images are taken with a calibrated stereo rig. They are further rectified so that image rows correspond to epipolar lines in both the left and right image planes. These rectified images are the ones used in the segmentation module also. A stereo matching algorithm based on the sum of the absolute differences with left-to-right and right-to-left correspondence gives the disparity between corresponding pixels at image point (u, v)

$$d(u,v) = \arg\min_{d} \left(\sum_{(u,v)} |I_l(u,v) - I_r(u+d,v)| \right), \quad (1)$$

where $I(u,v) = w_r R(u,v) + w_g G(u,v) + w_b B(u,v)$ is a function of the color reflectance, and the weights w_r, w_g, and w_b can be modified to give relative importance to one or another color channel.

To eliminate noise and depth estimation errors due to occlusions and reflectance variations, the disparity values resulting from matching points in the left image to points in the right image are compared to the disparities obtained when matching in the opposite direction, i.e., from right to left. Only those disparity values that coincide in both directions are used for segment characterisation. The resulting disparity map is a set of points for which a disparity value $d_i = d(u_i^d, v_i^d)$ is associated to the image coordinates (u_i^d, v_i^d), and is inversely proportional to the distance z_i at which objects are located from the camera. Given both camera calibration matrices, the computation of $\vec{\mathbf{p}}_i = (x_i, y_i, z_i)^T$ is straightforward.

Disparity maps obtained from stereo methods based on the area of local regions (correlation, sum of squared differences, or sum of absolute differences) are dense where the scene is rich in detail, and sparse in smooth regions.

2.2. Perceptual grouping

Color images are segmented via a greedy algorithm that uses local variation information [6, 12]. A graph $G = (V, E)$ is generated for either the left or right original image, where the nodes $v_i \in V$ correspond to each pixel, and each edge $e_{jk} \in E$ is weighted by a color distance measure among neighbouring pixels.

The segmentation algorithm iterates over an increasingly sorted array of the elements in E to separate G into a forest of segmented surfaces based on a color smoothness criterion, where the external variation must be larger than the internal variation between regions. Then, a search for the maximum spanning trees in the remaining forest is performed to allow for a compact representation of the segmented regions. The label of the segment associated to the pixel with coordinates (u_i^s, v_i^s) in the segmented image is $l_j = s(u_i^s, v_i^s)$. The details of the algorithm are presented in [12].

3. Integration of depth and color information using supervised classification

The set of regions $\mathbf{R} = \{R_1, R_2 \ldots R_3\}$ with good color continuity provided by our color segmentation algorithm can be expressed as $R_j = \{(u_i^s, v_i^s) | l_j = s(u_i^s, v_i^s)\}$. Our segmentation algorithm differs from the one presented in [12] in that regions in $\mathbf{R}$ are divided in two subsets. If $|R_j| > t_A$ and $p_j^2/|R_j| < t_C \Rightarrow R_j \in \mathbf{\Omega}$, otherwise $R_j \in \mathbf{\Gamma}$. $\mathbf{R} = \mathbf{\Omega} \cup \mathbf{\Gamma}$. $\mathbf{\Omega}$ represents the regions with area greater than t_A and compactness smaller than t_C, $\mathbf{\Gamma}$ represents the detail in the image, and p_j is the perimeter of R_j.

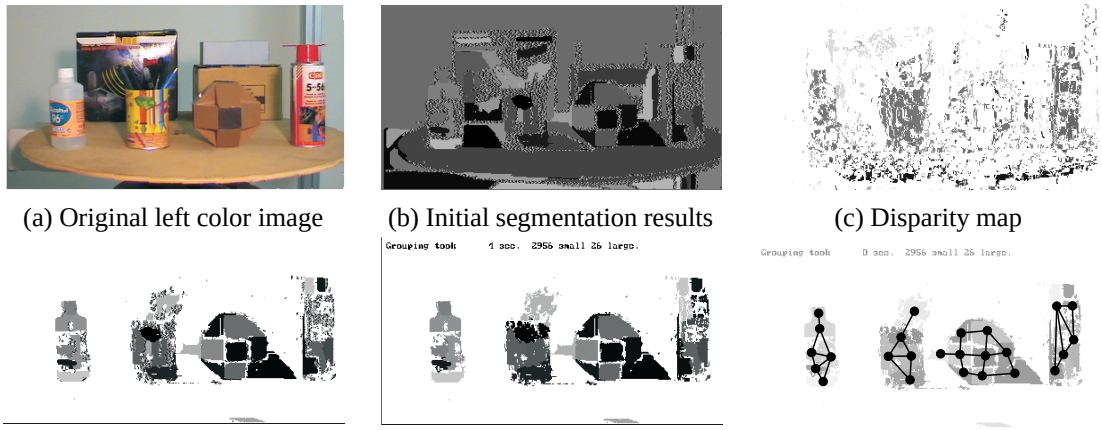
To solve for subsegmentation, in each region Ω_j the set of points $r_j = \left\{(x_i, y_i, z_i) | (u_i^d, v_i^d) \in \Omega_j\right\}$ is extracted from the depth map, and for each point $(u_i^s, v_i^s) \in \Omega_j$ a new point (x_i, y_i, z_i) is added to r_j, where the corresponding $d_i(u_i^s, v_i^s)$ is obtained from the average disparity from the points in the window $(u_i^s - 1 \ldots u_i^s + 1, v_i^s - 1 \ldots v_i^s + 1)$ with entry in r_j. The process is repeated iteratively until all pixels in Ω_j have a corresponding entry in r_j. This is, until all pixels in Ω_j have been assigned a depth estimate.

Smooth surface segments S_l^j are recursively generated by starting at any point $\vec{\mathbf{p}}_i \in r_j$ and growing outwards while meeting the following two criteria for the neighbouring points $\vec{\mathbf{p}}_i$ and $\vec{\mathbf{p}}_k$:

$$\|\vec{\mathbf{p}}_i - \vec{\mathbf{p}}_k\| > t_J \quad \text{Jump Edge Criterion} \quad (2)$$

$$\frac{\cos^{-1}(\hat{\mathbf{n}}_i^T \hat{\mathbf{n}}_k)}{\|\vec{\mathbf{p}}_i - \vec{\mathbf{p}}_k\|} > t_U \quad \text{Curvature Criterion} \quad (3)$$

The normals $\hat{\mathbf{n}}_i$ are computed minimizing the error of fitting a local planar patch in the vicinity of $\vec{\mathbf{p}}_i$ [4]. The



(a) Original left color image (b) Initial segmentation results (c) Disparity map

(d) Initial pruning (e) Data fusion results (f) Graphs due for recognition

Figure 1. Data fusion steps

region $R_j \in \mathbf{\Omega}$ is then replaced in $\mathbf{R}$ by the segments in $\mathbf{S}^j$. A new division of $\mathbf{R}$ into $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ is necessary.

For the case of oversegmentation we first generate a set of initial classes $\omega_l = \{(u_i^d, v_i^d, d_i) | (u_i^d, v_i^d) \in \Omega_l\}$, and the sets of points to be classified $\gamma_m = \{(u_i^d, v_i^d, d_i) | (u_i^d, v_i^d) \in \Gamma_m\}$.

The task at hand is to associate each Γ_m to its parent region Ω_l based on their spatial proximity. This classification constitutes the merging of smooth and detail into spatially coherent entities. Consider one detail region Γ_m. We must compute the distance from the points $(u_i^d, v_i^d, d_i) \in \gamma_m$ to the classes in ω. It is clear that the distribution of sample points in the classes in ω does not follow any typical probability distribution, but in those areas where the image is rich in detail, the samples resemble a uniform distribution if projected on the image plane. The minimum distance between points being the pixel width and length. This observation suggests the use of a parametric distance measure for classification. The normalized distance from point $(u_i^d, v_i^d, d_i) \in \gamma_m$ to class ω_l is

$$d_{il} = \mathrm{tr}\left(\Sigma_l^{-1}\left((u_i^d, v_i^d, d_i) - (\bar{u}_i^d, \bar{v}_i^d, \bar{d}_i)\right)^T \left((u_i^d, v_i^d, d_i) - (\bar{u}_i^d, \bar{v}_i^d, \bar{d}_i)\right)\right) \quad (4)$$

where $(\bar{u}_i^d, \bar{v}_i^d, \bar{d}_i)$ is the mean vector of class ω_l, and Σ_l the covariance matrix. The votes $V_i = \min_l(d_{il})$ are accumulated for each $(u_i^d, v_i^d, d_i) \in \gamma_m$, and the region in $\mathbf{\Omega}$ associated to the class ω_l with most votes is considered the parent region for Γ_m.

The result is a new set of regions $\mathbf{\Omega}'$ where each element $\Omega'_j = \{\Omega_j \cup \Gamma_m \ldots \Gamma_n\}$, represents a region in the scene where color continuity and depth continuity are merged to constitute spatially coherent entities. A set of characteristics can be measured on each of these regions, such as position, normal orientation, curvature, level of detail, area, compactness, etc. The immediate step to follow from these results will be an attempt to learn and recognize these groups of segments as objects.

The time complexity of the depth from stereo module is $O(dk^2n)$, where k is the width of the kernel window, d is the maximum expected disparity, and n is the number of pixels in the image. If the edges in G are sorted in linear time, the segmentation module is bounded by $O(n)$.

The time required to compute equation 4 is bounded by $O(m^2/a^2)$, where m is the number of points in the disparity map associated to regions in $\mathbf{\Omega}$, and $a = |\mathbf{\Omega}|$. Given that the disparity map is dense in the perimeter of Ω_j and negligible inside smooth regions, $m \approx \sum_a p_j k/2$, and from the compactness constraint $p_j^2 < t_C|R_j|$. The overall cost of the oversegmentation part of the algorithm is bounded by

$$O(m^2/a) \approx O(a\bar{p}_j^2 k^2) < O(ak^2 t_C |\Omega_j|) < O(k^2 t_C n). \quad (5)$$

The time complexity of our algorithm is linear with respect to the number of pixels in the image, and is asymptotically comparable to that of the individual data acquisition modules.

4. Experiments

Several test cases were performed, and an example that best shows the advantages and drawbacks of our algorithm is presented. The original left color image from the stereo pair is shown in Figure 1(a) (its b/w version may appear in the conference proceedings). Both the left and right images were rectified to meet the parallel epipolar constraint. The results of the color segmentation algorithm applied to the rectified left image are shown in Figure 1(b). Note that labels assigned to each segment do not resemble the intensity values in the original image, as would be expected from a color based segmentation algorithm. This was done only

to ease visual identification of large segments from small or highly compact ones. In this figure, the textured regions represent the small segments that need to be classified as belonging to the nearby objects.

Figure 1(c) shows the disparity map obtained from the stereo module. It is most clear from this image how depth information is dense at highly detailed areas in the scene, whereas smooth regions are poorly represented. Although the left-to-right right-to-left constraint could have been relaxed from equality to similarity when creating the depth map, this was not implemented; letting the segmentation module overcome the weaknesses of any typical depth-from-stereo module.

An initial pruning of the segmented image is done based purely on the mean disparity value of each segment. We eliminated from our three-dimensional region of interest those segments that fall too close or too far from the camera by computing their mean disparity. Also those segments with very low points-in-depth-map to segmented-area ratio were discarded, as they do not contain enough disparity information to accurately estimate their depth, and are not suitable for later attempts at object characterization. Figure 1(d) shows how the boxes behind the four objects of interest are virtually eliminated, as well as the table and the vertical bar in the back.

Results from the data fusion algorithm are shown in Figure 1(e). The image shows our four objects of interest easily identifiable. These segment groups and their attributes can be used to characterize the objects they represent. Figure 1(f) shows a set of graphs representing the hypothesized objects that are due for recognition.

5. Conclusion

In this paper, we have shown how the shortcomings of two individual low-level processing modules can be overcome in an integrated environment. The inherent variability of the data formats is tackled by exploiting their individual characteristics. While color-based segmentation methods are robust in smooth regions and tend to fail in areas where detail is prominent, the opposite is true for a depth from stereo module. We have provided the necessary framework to exploit this situation by relabeling those areas where segmentation fails based on depth cues, both in sub- and oversegmentation situations. A qualitative example on how our algorithm behaves is presented for a complicated scene including different textures as well as overlapping objects with very similar color reflectance. It is of major relevance that the time complexity of our algorithm is linear with respect to the number of pixels in the image, and does not exceed that of the individual acquisition modules.

The results of our algorithm are expressed as graphs with the nodes representing spatially coherent regions in the scene where smoothness and detail is blended together, and the edges represent spatial connectivity between regions. Evidently, the next step is to train a system with these segmentation results and to attempt recognition of the same objects at different positions and orientations and varying illumination conditions.

Acknowledgements

The authors would like to thank Jaume Vergés-Llahí for providing the initial segmented images. This work was partially supported by CICYT project TAP98-0473, CONACYT and CETYS.

References

[1] W. Austin and A. Wallace. Object location by parallel pose clustering. *Comput. Vision and Image Understanding*, 72(3):304–327, Dec. 1998.

[2] A. Blake, A. Zisserman, and G. Knowels. Surface description from stereo and shading. In B. Horn and M. Brooks, editors, *Shape from Shading*, Artificial Intelligence Series, chapter 2, pages 29–52. The MIT Press, Cambridge, Aug. 1989.

[3] H. Bulthoff and H. Mallot. Integration of depth modules: Stereo and shading. *J. Optical Society of America*, 5:1749–1758, Oct. 1988.

[4] C. Chen and A. Kak. Robot vision system for recognizing objects in low-order polynomial time. *IEEE Trans. on Syst. Man and Cybernetics*, 18(6):1535–1536, Nov. 1989.

[5] J. Cryer, P.-S. Tsai, and M. Shah. Integration of shape from shading and stereo. *Pattern Recog.*, 28(7):1033–1043, July 1995.

[6] P. Felzenszwalb and D. Huttenlocher. Image segmentation using local variation. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recog.*, pages 98–104, Santa Barbara, June 1998.

[7] Y. Hel-Or and M. Werman. Pose estimation by fusing noisy data of different dimensions. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 17(2):195–201, Feb. 1995.

[8] S. Lacroix, P. Grandjean, and M. Ghallab. Perception planning for a multi-sensory interpretation machine. In *Proc. IEEE Intl. Conf. on Robotics and Automation*, volume 2, pages 1818–1824, Nice, May 1992.

[9] S. Nadabar and A. Jain. Fusion of range and intensity images on a connection machine (cm-2). *Pattern Recog.*, 28(1):11–26, Jan. 1995.

[10] S. Pankanti and A. Jain. Integrating vision modules: Stereo, shading, grouping, and line labeling. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 17(8):831–842, Sept. 1995.

[11] S. Sarkar and K. Boyer. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 15(3):256–276, Mar. 1993.

[12] J. Vergés-Llahí and A. Sanfeliu. Color image segmentation solving hard constraints using a graph partitioning greedy algorithm. submitted to IAPR Intl. Conf. on Pattern Recognition, Sept. 2000.