Pairwise Similarities for Scene Segmentation combining Color and Depth data

F. Bergamasco, A. Albarelli and A. Torsello
Università Ca' Foscari Venezia
www.dais.unive.it

M. Favaro and P. Zanuttigh
Università di Padova
www.unipd.it

Abstract

The advent of cheap consumer level depth-aware cameras and the steady advances with dense stereo algorithms urge the exploitation of combined photometric and geometric information to attain a more robust scene understanding. To this end, segmentation is a fundamental task, since it can be used to feed with meaningfully grouped data the following steps in a more complex pipeline. Color segmentation has been explored thoroughly in the image processing literature, as much as geometric-based clustering has been widely adopted with 3D data. We introduce a novel approach that mixes both features to overcome the ambiguity that arises when using only one kind of information. This idea has already appeared in recent techniques, however they often work by combining color and depth data in a common Euclidean space. By contrast, we avoid any embedding by virtue of a game-theoretic clustering schema that leverages on specially crafted pairwise similarities.

1. Introduction

The goal of segmentation is to isolate each single object that appears in an image. This can be deemed to be a somewhat ill-posed problem since it is often difficult to state at a semantic level if an object should be split in two segments or kept as a whole. This kind of ambiguity can be observed when humans are asked to manually build a ground-truth for evaluation purposes, as different individuals will provide very different segmentations for the same image. This is illustrated in Fig. 1 where an image is shown along two manual segmentation selected from the dataset [2]. In the first segmentation the operator separated the body of the lizard from the spots on the skin. This is arguably the same kind of output that would be obtained using one of the many color-based segmentation techniques [5, 7, 6]. However, in most applications that perform model-based classification, handling the whole lizard as a single object would be desiderable. In the second ground-truth this over segmentation does not happen, nevertheless the foreground and background leafs have been wrongly merged as their depth level has not been accounted for. While this kind information is not available within standard digital images, it can be obtained by adopting a proper 3D reconstruction method like Dense Stereo [12], Structured Light [13] or even dedicated hardware [8]. Given the increasing popularity of such techniques it is not surprising that a number of algorithms that make use of depth information to boost the segmentation performance have appeared. Harville and Robinson [9] use stereo data to augment local appearance features extracted from the images. In [3] Bleiweiss and Werman use Mean Shift over a 6D vector that fuses color and depth data obtained with a ZCam and weighted with respect to their estimated reliability. Wallenberg at al. [14] generate a probabilistic representation of RGB color and depth derivatives, with the goal of overcoming the incompatibilities between the two measures due to their different physical units. The same problem is addressed by Dal Mutto et al. [11], that propose to cast the color component in the CIELab space and to append the 3D position of each pixel through a weighting constant that is learned experimentally. Most of these techniques have in common that the segmentation happens over an extended Euclidean space built on a base that combines two sources of information that are disjoint from both a physical and a semantical point of view. In this paper we propose a different approach, where pairwise similarities are defined between macropixels and a game-theoretic approach is employed to make them play in an evolutionary game until stable segmentation emerges.

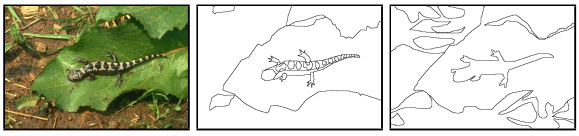

Figure 1. Two manual segmentations

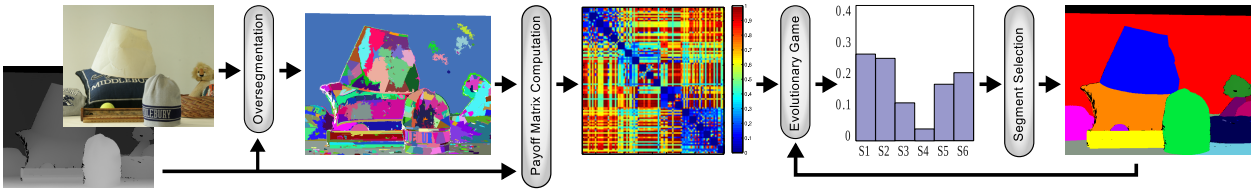

Figure 2. Illustration of the proposed pipeline (see text for details)

2. Segmentation through disjoint cues

The proposed pipeline is made up of three main tasks depicted in Fig. 2. The first step is an oversegmentation that will build a large number of macropixels. A measure of compatibility is then computed between each pair of macropixels, thus materializing a square similarity matrix. Finally, the obtained matrix is used as the payoff matrix of a non-cooperative game that will be played between macropixels in order to isolate segments as clusters of mutually compatible parts.

2.1. Macropixel extraction

Since the pipeline works by assembling macropixels it is very important to obtain a proper oversegmention of the objects in the scene. This means that (hopefully) no part should span over different objects. For this reason the process starts with a color-based segmentation using Felzenszwalb and Huttenlocher's technique [7], subsequently, for each segment, mean and variance are computed with respect to the depth of the pixels. Each time variance exceeds a threshold σ_t k-means with $k = 2$ is applied recursively to obtain two splitted sub-segments.

2.2. Distances and pairwise similarities

To compute a similarity between macropixels we first need to define distances with respect to both color and 3D information. The color distance is simply defined as the distance on the u/v plane of the average of the chroma components of the elements of the marcropixel (luminance is avoided to attain illumination invariance). That is, given macropixels m_1 and m_2 respectively with average chroma coordinates $[u_1 v_1]^T$ and $[u_2 v_2]^T$, their color distance is:

$$d_c(m_1, m_2) = \sqrt{(u_2 - u_1)^2 + (v_2 - v_1)^2} \quad (1)$$

While the just defined color distance is based on an Euclidean measure, the depth-aware distance is not. Specifically, we designed a modified Dijkstra algorithm where the most convenient step is not the one that shortens the total trip to the destination, but the one that performs the smaller plunge along the z axis (assuming x and y axis aligned with the image plane). The algorithm has been applied between the centroids of each pair of macropixels on a 4-connected graph built over the original pixels and a path that contains the sequence of 3D jumps with the smaller possible maximum drop has been recorded. Within this scenario we can expect that if two segment belong to the same uninterrupted surface the larger jump in the path will be small since there will always be a way around abrupt discontinuities. This of course could not be the shortest Euclidean route, albeit optimal according to our measure. This can be observed in Fig. 3 where the route connecting the green and blue macropixels goes through the baby to minimize the maximum jump. More formally, if p is the sequence of 3D displacements connecting macropixels m_1 and m_2 we first find the maximum drop:

$$[\Delta x_m \Delta y_m \Delta z_m]^T = \underset{[\Delta x \Delta y \Delta z] \in p}{\arg\max} \; |\Delta z| \quad (2)$$

And then we calculate the distance between m_1 and m_2 as a function of the angle along the drop:

$$d_z(m_1, m_2) = \frac{2}{\pi} asin\left(\frac{|\Delta z_m|}{\|[\Delta x_m \Delta y_m \Delta z_m]^T\|}\right) \quad (3)$$

The rationale of Eq. 3 is that some kind of normalization must be applied to the drop along the z axis to avoid a bias toward foreground objects. Since angles are not affected by the distance from the image plane this is a reasonable choice.

Given distances d_c and d_z we are able to define a similarity between macropixels. In a sense, our goal is to merge macropixels that have been produced by the same phenomenon, that is the same surface that should exhibit a reasonable level of contiguity and color consistency. To this end, we can deem d_c and d_z as deviations from an unknown average for the two independent

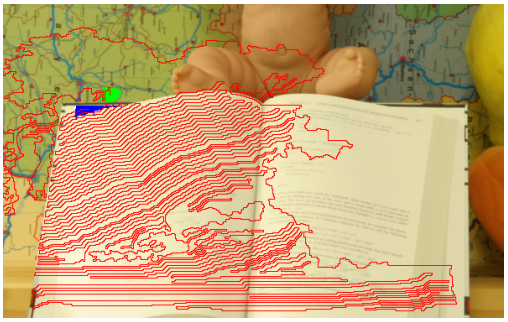

Figure 3. Behavior of the modified Dijkstra

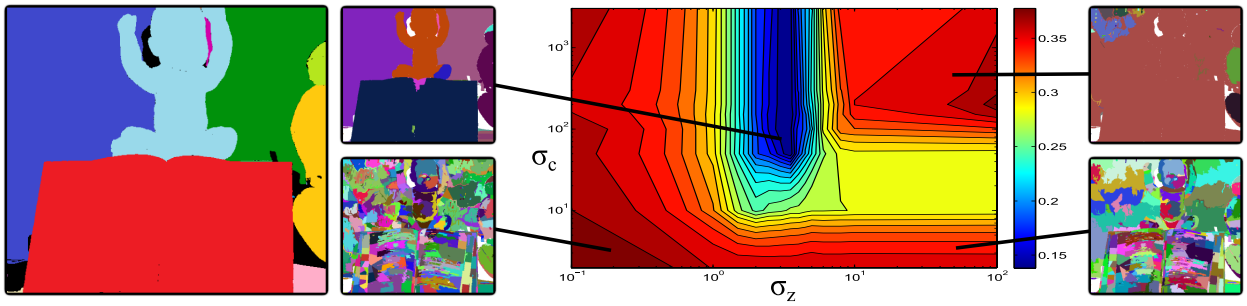

Figure 4. Exploration of the space of parameters σ_c and σ_z.

aspects of the phenomenon and thus we model the similarity between m_1 and m_2 as the density of a mixture of two orthogonal zero-mean Gaussians:

$$\pi(m_1, m_2) = e^{-\frac{1}{2}\left(\frac{d_c(m_1,m_2)^2}{\sigma_c} + \frac{d_z(m_1,m_2)^2}{\sigma_z}\right)} \quad (4)$$

Where σ_c and σ_z are two parameters that represent respectively the expected variance of color and depth information. This similarity defines a payoff matrix $\Pi = (\pi_{ij} = \pi(m_i, m_j))$ between all the macropixels.

2.3. Game-theoretic clustering

Following [1] we use the principles of Evolutionary Game Theory [15] to cluster macropixels into larger objects according to the mutual payoff defined with Eq. 4. Each macropixel is modeled as a strategy. When strategies i and j are played one against the other both the players obtain the same payoff (i.e. gain) π_{ij}. The amount of population that plays each strategy at a given time is expressed through the probability distribution $\mathbf{x} = (x_1, \ldots, x_n)^T$ (called *mixed strategy*) with $\mathbf{x} \in \Delta^n = \{\mathbf{x} \in \mathbb{R}^n : \forall i\ x_i \geq 0,\ \sum_{i=1}^n x_i = 1\}$. The *support* $\sigma(\mathbf{x})$ of a mixed strategy $\mathbf{x}$ is defined as the set of elements chosen with non-zero probability: $\sigma(\mathbf{x}) = \{i \in O \mid x_i > 0\}$. A mixed strategy $\mathbf{x}$ is said to be a *Nash equilibrium* if it is the best reply to itself, i.e. $\forall \mathbf{y} \in \Delta,\ \mathbf{x}^T\Pi\mathbf{x} \geq \mathbf{y}^T\Pi\mathbf{x}$. Finally, $\mathbf{x}$ is called an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and $\forall \mathbf{y} \in \Delta \quad \mathbf{x}^T\Pi\mathbf{x} = \mathbf{y}^T\Pi\mathbf{x} \Rightarrow \mathbf{x}^T\Pi\mathbf{y} > \mathbf{y}^T\Pi\mathbf{y}$. This condition guarantees that any deviation from the stable strategies does not pay. The search for a stable state is performed by simulating the evolution of a natural selection process. Under very loose conditions, any dynamics that respect the payoffs is guaranteed to converge to Nash equilibria [15] and (hopefully) to ESS's. We chose to use the replicator dynamics, governed by the following equation

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t)\frac{(\Pi\mathbf{x}(t))_i}{\mathbf{x}(t)^T\Pi\mathbf{x}(t)} \quad (5)$$

where $\mathbf{x}_i$ is the i-th element of the population and Π the payoff matrix. Once the population has reached a local maximum, all the non-extincted pure strategies (i.e., $\sigma(\mathbf{x})$) can be considered selected by the game and thus the associated macropixels can be merged into a single segment. After each game the selected macropixels are removed from the population and the selection process is iterated until all the segments have been merged.

3. Experimental evaluation

All the following experiments have been performed using the Middlebury Stereo Dataset [10], which consists of a large set of stereo pairs associated with a ground-truth disparity (obtained with a structured light scanner). A total of 7 subjects were selected to be used in our tests and, for each subject, a manual segmentation was performed in order to separate the objects found in the scene. Macropixel have been produced by using Felzenszwalb and Huttenlocher's segmentation with parameters $\sigma = 0.5$, $k = 500$ and $min = 20$ while the value of σ_t has been set to 3 for all the images.

With the first set of experiments we evaluate the effect of parameters σ_c and σ_z. The quality of the result is assessed using the Hamming distance (see [4]) with respect to the manual segmentation. In the middle of Fig. 4 we show a plot of the segmentation quality over the σ_c/σ_z plane. Two interesting observation can be done. First, there exists a single large optimal region in witch the two parameters lead to a small Hamming distance. Second, once one parameter is optimal and the other is greater than a certain threshold, the Hamming distance remains almost stable thus simplifying the parameter tuning process. In Fig. 4 we also show the qualitative effect of σ_c and σ_z. When the two parameters are both too low, the evolutionary game becomes very selective hence producing a lot of erroneous clusters. On the other hand, if they are too high the process is unable to separate each region. In general, if σ_c is low with respect of σ_z, the process care more about color distance (bottom right figure) and vice versa. When σ_c and σ_z are chosen from the optimal region, the resulting segmentation is almost identical to the ground-truth.

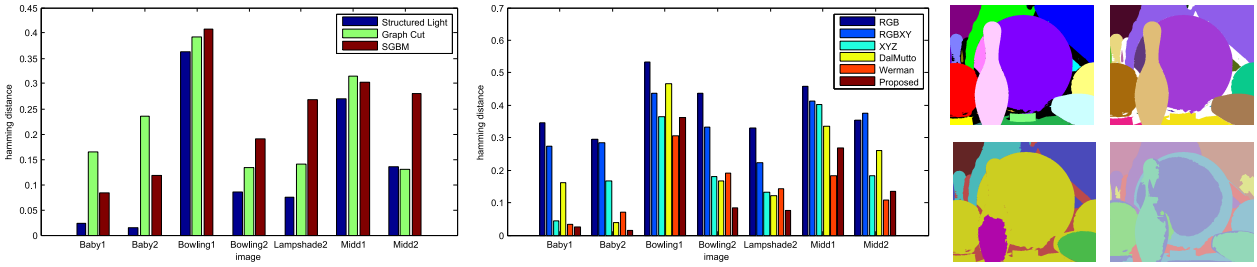

Figure 5. Comparisons with other methods and influence of the 3D information source.

The aim of the second experimental batch is to evaluate the impact of replacing the ground-truth 3D data with depth maps produced with dense stereo and to compare the results obtained by our method with other techniques. In the first histogram of Fig. 5 we compare the quality obtained for different scenes using the optimal values of σ_c and σ_z and three different sources of 3D data (groud-truth, Graph-Cut Stereo and SGBM Stereo). It can be seen that there are almost half cases in witch Graph-Cut lead to a better segmentation, and the residual in witch SGBM is better. This is due to the fact that, in scenes containing large untextured areas, a global method like GC is able to better discriminate depth regions. On the other hand, in scenes with lots of textures, the depth map produced may be smoother than local methods, hence hindering the segmentation. In the second histogram we compare our results with three k-means based segmentations (using only color, only spatial and a union of color and spatial data) and two methods recently introduced in [11] and [3]. As expected k-means, albeit supplied with the ground-truth number of segments, tends to performs badly with respect to the other methods. K-means with spatial data performs better because depth information is more descriptive to define object boundaries for this dataset. However, the ability to properly catch the interplay between color and spatial data is crucial in some cases like Baby2 and Midd1. Overall, our approach obtains very good results in most situations. Finally, in the right part of Fig. 5 some qualitative results are shown, respectively ground-truth (top-left), the proposed method (top-right), Werman (bottom-left) and DalMutto (bottom-right).

4. Conclusions

We proposed a segmentation technique that exploits both color and depth data by casting the clustering problem into a non-coperative game defined through pairwise similarities over an initial set of macropixels. Our approach has shown to be quite tolerant with respect to the choice of parameters and behaves better than other recent methods that perform the segmentation in a combined Euclidean space.

References

[1] A. Albarelli, S. R. Bulò, A. Torsello, and M. Pelillo. Matching as a non-cooperative game. In *ICCV: IEEE Intl. Conf. on Comp. Vis.* IEEE Computer Society, 2009.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 33(5), 2011.

[3] A. Bleiweiss and M. Werman. Fusing time-of-flight depth and color for real-time segmentation and tracking. In *DAGM Workshop: Dyn3D'09*, pages 58–69, 2009.

[4] X. Chen, A. Golovinskiy, and T. Funkhouser. A benchmark for 3D mesh segmentation. *SIGGRAPH*, 2009.

[5] D. Comaniciu, P. Meer, and S. Member. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:603–619, 2002.

[6] T. Cour, F. Benezit, and J. Shi. Spectral Segmentation with Multiscale Graph Decomposition. pages 1124–1131. IEEE Computer Society, 2005.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. Journal of Computer Vision*, 59(2):167–181, 2004.

[8] V. Gulshan, V. Lempitsky, and A. Zisserman. Humanising grabcut: Learning to segment humans using the kinect. In *IEEE Workshop on Consumer Depth Cameras for Computer Vision, ICCV*, 2011.

[9] M. Harville and I. N. Robinson. Fusion of local appearance with stereo depth for object tracking. *CVPR Workshops: IEEE Computer Society Conf. on Comput.Vis. and Pat. Rec.*, pages 1–8, 2008.

[10] H. Hirschmuller and D. Scharstein. Evaluation of Cost Functions for Stereo Matching. *CVPR: IEEE Comp. Society Conf. on Comput.Vis. and Pat. Rec.*, 1:1–8, 2007.

[11] C. D. Mutto, P. Zanuttigh, G. M. Cortelazzo, and S. Mattoccia. Scene segmentation assisted by stereo vision. In *3DIMPVT'11*, pages 57–64, 2011.

[12] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. Journal of Computer Vision*, 47:7–42, 2002.

[13] D. Scharstein and R. Szeliski. High-Accuracy Stereo Depth Maps Using Structured Light. *CVPR: IEEE Computer Society Conf. on Comput.Vis. and Pat. Rec.*, 1:195–202, June 2003.

[14] M. Wallenberg, M. Felsberg, P.-E. Forssén, and B. Dellen. Channel coding for joint colour and depth segmentation. In *Pattern Recognition*, volume 6835, pages 306–315. Springer Berlin / Heidelberg, 2011.

[15] J. Weibull. *Evolutionary Game Theory*. MIT, 1995.