Geo-semantic Segmentation

Shervin Ardeshir[1], Kofi Malcolm Collins-Sibley[2] Mubarak Shah[1],
[1]Center for Research in Computer Vision, University of Central Florida. [2]Northeastern University.

Segmentation of an in image into coherent and semantically meaningful regions is a fundamental problem in computer vision. Many methods employing image content have been proposed in the literature during the last few decades. In particular, semantic segmentation of images taken from urban areas has been studied in the past few years [1, 2, 3, 4]. However, to the best of our knowledge, the problem of segmenting images into exact geo-referenced semantic labels has yet to be studied. Given the availability of a plethora of geographical information, geo-semantic segmentation could be a relevant topic of research in computer vision. In particular, GIS databases containing information about the exact boundaries (2D footprints in long-lat domain) of semantic regions such as buildings, roads, etc. and are available for many urban areas, making it a suitable source of information for performing geo-semantic segmentation. In this work, we propose a method to leverage the information extracted from GIS, to perform geo-semantic segmentation of the image content, and simultaneously refine the misalignment of the projections. First, the image is segmented into a set of initial super-pixels. Also, camera matrix is formed using the provided camera parameters such as sensor size and focal length, alongside with the GPS location of the camera using which the image was captured. Using the camera matrix, 3D points in the GIS (which belong to regions of interest, in our case buildings and streets) are projected onto the 2D image plane. In many cases, due to the inaccuracies in the meta-data and mosty due to the GPS location (EXIF Tag), the projections are far from perfect and usually not aligned with the objects of interest. This misalignment is more intense for regions which are at larger distances from the camera and/or have smaller projection sizes in the field of view of the camera. We show that these misaligned projections turn into misleading priors for the semantic segmentation step. Therefore, we evaluate each of the projections in terms of reliability, and weight their contribution before using them in the segmentation process. Once the projections are weighted, we get some priors for the initial super-pixels, and combining them with the spatial-visual consistency of super-pixels we obtain some initial semantic segmentation results. Given the fact that each semantic segment is computed using a projection, the reliability score of the projections can be propagated to the semantic segments. As a result, semantic segments corresponding to the most reliable projections (which usually belong to large buildings covering considerable proportion of the content of the image) will get significant weights. Using the projected GIS segments, and their corresponding semantic segments in the content of the image, the misalignment of the projections could be computed with respect to their location in the image content. Our experiments indicate that the aforementioned alignment can be used to obtain a new set of updated projections, with less misalignment. In other words, the results of the segmentation can be used for refining the projections and vice versa. This iterative process continues until our defined convergence criteria is met. We evaluated the authenticity of this refinement procedure by observing the alignment of the projections with the annotated content of the image.

The block diagram of the proposed method is shown in figure 2 . Given a GPS-tagged image and the GIS database containing the outlines of the buildings, we project the points on the outlines and a set of points representing the streets to the 2D image plane in order to obtain some priors about the semantic segments present in the image. In addition, we perform some initial super-pixel segmentation of the image using a typical segmentation method. Finally, we construct a graph over the super-pixels and use the GIS-obtained priors, and the pairwise visual similarity among the superpixels for labeling the super-pixels with at most one of the geo-semantic labels which are supposed to be visible in the image. The labeling is done in an iterative manner and at each iteration, the quality of projections improve in addition to the segmentation accuracy.

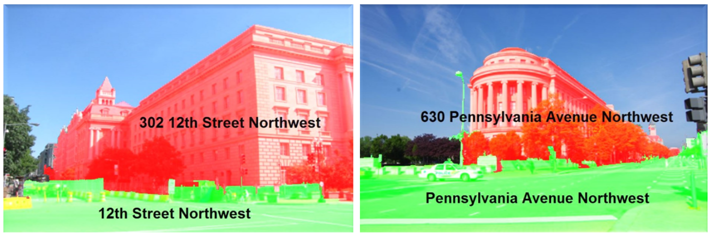
302 12th Street Northwest
12th Street Northwest
630 Pennsylvania Avenue Northwest
Pennsylvania Avenue Northwest

Figure 1: Labeling semantic entities in the image such as buildings (red) and streets (green) alongside with their geo-semantic labels (addresses).

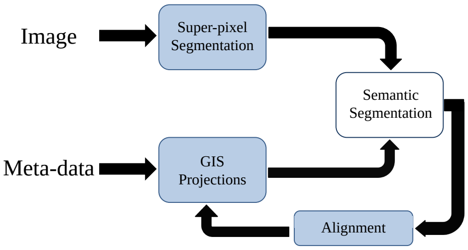

Figure 2: The block diagram of the proposed method. GIS projections are used for labeling the super-pixels, and the resulting semantic segments are being used for updating the quality of the priors in an iterative manner.

Geo-Semantic Segmentation using GIS Projections We perform parallel random walks on a fully connected graph, which can be solved in linear time using closed form solution:

$$X_{k+1} = (1-\alpha)(I-\alpha S)^{-1} X_k. \tag{1}$$

Here X_k and X_{k+1} are $m \times n$ matrices whose elements x_{ij} capture the probability of super-pixel j belonging to semantic segment i. $S_{m \times m}$ captures the pairwise similarity among the super-pixels and $0 < \alpha < 1$ is a constant weighting contribution of the unary and binary scores.

Alignment and Updating Projections Given a set of projections and their corresponding semantic segments, we aim to find a global transformation, mapping the projections to their corresponding semantic segments. We perform that alignment using a weighted least square, giving high weights to reliable projection-segmentation pairs. The transformation is found and applied to the projections using the following:

$$\mathbf{g_{k+1}} = [((\mathbf{WQ})^T(\mathbf{WQ}))^{-1}\mathbf{Q}^T\mathbf{WY}]\,\mathbf{g_k}. \tag{2}$$

Here $\mathbf{W}$ is a diagonal matrix containing the squared roots of the weights of the corresponding pairs, $\mathbf{Q}$ and $\mathbf{Y}$ contain the points extracted from the projections and segmentations respectively. Also, $\mathbf{g}_k$ and $\mathbf{g}_{k+1}$ are the projections at iteration k and $k+1$.

[1] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *Computer Vision–ECCV 2008*, pages 44–57. Springer, 2008.

[2] D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up., 2005.

[3] Olivier Teboul, Loic Simon, Panagiotis Koutsourakis, and Nikos Paragios. Segmentation of building facades using procedural shape priors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3105–3112. IEEE, 2010.

[4] Jianxiong Xiao, Tian Fang, Peng Zhao, Maxime Lhuillier, and Long Quan. Image-based street-side city modeling. *ACM Transactions on Graphics (TOG)*, 28(5):114, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation web page.