Label Transfer Exploiting Three-dimensional Structure for Semantic Segmentation

Valeria Garro
DI, University of Verona
Strada Le Grazie 15, 37134
Verona (Italy)
valeria.garro@univr.it

Andrea Fusiello*
DIEGM, University of Udine
Via Delle Scienze 206, 33100
Udine (Italy)
andrea.fusiello@uniud.it

Silvio Savarese
EECS, University of Michigan
1301 Beal Avenue, Ann Arbor,
MI 48109-2122, US
silvio@eecs.umich.edu

ABSTRACT

This paper deals with the problem of computing a semantic segmentation of an image via label transfer from an already labeled image set. In particular it proposes a method that takes advantage of sparse 3D structure to infer the category of superpixel in the novel image. The label assignment is computed by a Markov random field that has the superpixels of the image as nodes. The data term combines labeling proposals from the appearance of the superpixel and from the 3D structure, while the pairwise term incorporates spatial context, both in the image and in 3D space. Exploratory results indicate that 3D structure, albeit sparse, improves the process of label transfer.

Categories and Subject Descriptors

I.4.6 [**Image Processing And Computer Vision**]: Segmentation—*Pixel classification*

Keywords

Image Parsing, Segmentation, Labeling, Image Understanding, Markov Random Fields, Structure From Motion

1. INTRODUCTION

Semantic image segmentation, also known as image parsing (labeling image regions with their categories)xs, has been recently tackled as a *label transfer* approach, which reduces the inference problem for a new image to the problem of matching it to an existing set of labeled images. Whereas this matching is usually accomplished by exploiting local similarity between images [12, 18, 16], in this paper we investigate the leverage of sparse 3D information coming from Structure from Motion to improve the transfer. What motivates our work is that we expect 3D structure to be a powerful cue for matching, while its computation is becoming a "commodity" thanks to the availability of efficient and robust implementations [15, 8, 6]. In particular, in this work we concentrate on urban environment application, that has received much attention both in the segmentation [12, 16, 17] and in the Structure from Motion literature [2].

[**2D approaches**] Previous work on annotation transfer come from the field of object recognition, where some existing techniques have been modified in order to deal specifically with urban outdoor scenarios. To overcome the massive dataset dimension in [12] the authors present a method that involves first a retrieval step on a large database of annotated images using a modified version of SIFT flow [11]. The subsequent steps are applied only on a small subset of images presenting similar structures alignment. This system integrates multiple cues in a Markov random field framework to segment and recognize the query image. Even if the algorithm gives accurate results, the matching time for a pair of images with very small resolution (256×256) is around 30 seconds, which is still not suitable for real-time applications.

Zhang et al. in [18] proposed a similar procedure that involves retrieving multiple image sets of similar annotated images each of which can cover all semantic categories of the query image. Then the dense correspondences between these sets of images and the query are computed at superpixel level to increase the computational efficiency. The matching scheme, called KNN-MRF, does not rely only on the single feature appearance of each superpixel but it considers also the spatial relation between pairs of neighbor superpixels, modeling it as a Markov random field. After establishing the correspondences, a pruning procedure is applied in order to remove the semantic incorrect matches using a trained classification model.

Another interesting work has been proposed in [16]. The authors present a non-parametric scalable approach to image labeling at a superpixel level. For each superpixel of the dataset images, a large number of disparate features have been extracted describing its appearance. Unlike [18] it requires no training, it also combines semantic (e.g. sky, road, building) and geometric (e.g. vertical, horizontal) labeling under the assumption that each semantic label is associated with a unique geometric label.

Recently [7] proposed a solution that leverages on a graph of dense overlapping patch correspondences across large datasets of images avoiding the needs for exhaustive pairwise comparisons.

*Part of this work has been carried out while a.f. was with the University of Verona

Mirage 2013 Berlin, Germany

[3D approaches] An early attempt of label transfer involving 3D points coming from a Structure from Motion reconstruction can be found in [14]. In that work a user's annotation in one image, in the form of a rectangle with text, is transferred to the other images of the Structure from Motion dataset simply by retrieving the 3D points whose 2D projections are inside the specified rectangle and then reprojecting them on the other images. The authors apply an additional check involving visibility and scale information that avoids incorrect reprojections, but the label transfer is still rough, penalized by tight visibility constraints and it is restricted to the set of images processed by the Structure from Motion algorithm.

We propose a method that falls in between these two different approaches exploiting the effectiveness given by the additional 3D data to increase the precision of image approaches. To the best of our knowledge our work is the first attempt to combine appearance-based label transfer with 3D structure.

The rest of the paper is organized as follows. The next section gives an overview of the algorithm; Section 2.2 describes the pre-computation step of superpixel and feature extraction and Section 2.3 explains two different techniques for retrieving a subset of images in order to limit the main computation to a more specific and reliable dataset. In Section 2.4 we will describe in detail how we map our problem in a Markov random field framework; Section 3 describes the experiments and conclusions are drawn in Section 4.

2. METHOD

2.1 Overview

Our method integrates information from the appearance of an image, captured in a urban environment, with the three-dimensional structure of the scene in order to perform the label transfer. In a nutshell, the query image is first segmented into superpixels, then both information sources – namely, image appearance and 3D structure – propose a labeling for each superpixel; these proposals are combined and then relaxed with a Markov Random Field (MRF) to include spatial context.

It is assumed that a set of images D is available, capturing a specific urban environment, together with a sparse cloud of 3D points S representing the structure of the scene depicted in D, as it is commonly produced by a Structure from Motion pipeline. Moreover, a subset of these images $A \subset D$ is already labeled, i.e. , it has been segmented into small regular and coherent regions, called *superpixels*, and each superpixel has been assigned to a specific semantic class. Additional available data, obtained by Structure from Motion, is the camera pose estimation (the orientation and the 3D position of the camera center) of each image belonging to D.

The goal of our work is to label the images belonging to the set $D \setminus A$ or another external image not included in D, captured in the same scene environment. In such case, the procedure should rely on a localization system, as described in [9, 5], that efficiently computes the external camera parameters (orientation and camera center) of the image with respect to the existing reconstruction. From this point forward, let us call I_q, query image, the one that is going to be labeled.

2.2 Superpixels extraction

As a pre-processing step, computed only once, the labeled images in A are segmented into superpixels using the Simple Linear Iterative Clustering algorithm (SLIC) [1], which starts with a regular grid of cluster centers and then locally clusters pixels in the combined five-dimensional color and image plane space. The extracted superpixels are characterized by a compact and nearly uniform shape, as shown in Figure
1.

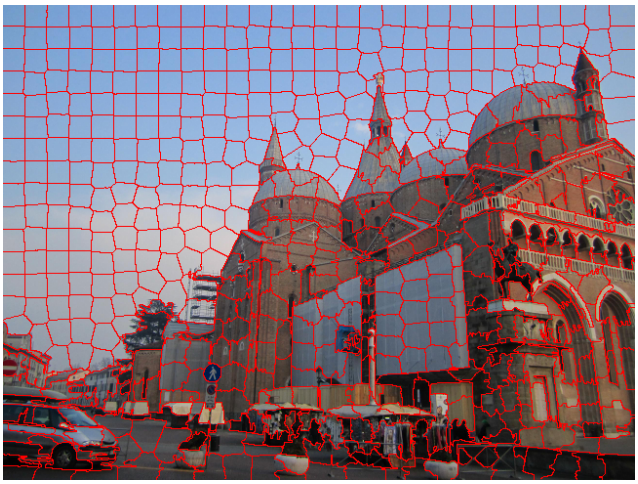

Figure 1: Example of superpixel extraction.

For each superpixel the algorithm extracts three features related to its location on the image and its color and texture appearance:

- the top height; linked to the relative "location" of the superpixels in the image (e.g. in urban scenarios "sky" superpixels will more likely appear on the top of the image and "ground" superpixel on the bottom);
- the color histogram in the log-chromaticity space correlated to the color appearance of the superpixel;
- an histogram of occurrences of the set of SURF feature descriptors already extracted from the images during the Structure from Motion procedure that are related to keypoints inside the superpixels.

The number of features is much smaller than in [16], in which 4 global and 13 superpixel features have been extracted. The main reasons are that:

i) we do not need to compute global features, mainly related to the image retrieval step, because the algorithm relies only on the 3D structure for the computation of the "working reference subset" of most similar images. Section 2.3 describes in detail the procedure.

ii) due to the superpixels' property to have a uniform shape there is no need to describe in detail both shape and location properties, only one feature, in our case the top height, can sufficiently characterize the superpixels.

2.3 Active set

Dealing with a potentially large input dataset of images D taken in a huge urban environment like district areas or entire cities [4], we need an efficient method to extract the effective subset of images $A_q \subset A$ that capture the same scene portion as I_q (which we call the "active" set). We describe now the two possible scenarios already mentioned in Section 2.1.

If $I_q \in D$ we can define A_q exploiting the 3D points visibility, since I_q shares the visibility of a common portion of the 3D points cloud. Let us call $S_q \subset S$ the set of 3D points visible from the query image I_q; we reproject the 3D points in S_q onto the images belonging to A and we define A_q as the subset of A in which the 3D projection gives valid values (i.e., at least a subset of S_q is inside the cameras' view frustum of A_q).

If $I_q \notin D$ the image should be initially localized as already mentioned in Section 2.1. The first stage of a typical localization system consists on a retrieval procedure of most similar images D_q that is done using image information applying a bag-of-word approach [13] and can be refined involving also a 3D structure constraint [5]. A_q is then defined as the intersection $D_q \cap A$.

2.4 Label assignment

The label assignment of I_q is computed by a MRF that has the superpixels of I_q as nodes. An edge is set between two nodes if the corresponding superpixels are adjacent. The set of superpixels $\{s_i\}$ is indicated with $\mathcal{V}$ while $\mathcal{E}$ represents the set of edges between neighbor nodes $\{(s_i, s_j)\}$

2.4.1 Unary Term

The unary term of the MRF encodes labeling proposals coming from the actual data, both in terms of appearance of superpixels and 3D structure.

Appearance based.

For each superpixel s_i of I_q we assign a score at each label l based on on image appearance, as in [16].

The probability $P\left(f_i^k|l\right)$ of observing f_i^k, the k-th feature of superpixel s_i, given the label l can be estimated empirically from the frequency of features from the given label in the neighborhood of f_i^k (in feature space). Specifically, let N_i^k denote the set of all superpixels in A_q whose k-th feature distance from f_i^k is below a fixed threshold. Then:

$$P_a\left(f_i^k|l\right) = \frac{1 + n\left(l, N_i^k\right)}{1 + n\left(l, A_q\right)} \tag{1}$$

where $n(l, Z)$ denotes the number of superpixels in set Z associated to label l.

Making the custom assumption that the extracted features are independent of each other given the label, the likelihood ratio for label l and superpixel s_i is:

$$\Lambda_a(l, s_i) = \frac{P(s_i|l)}{P(s_i|\bar{l})} = \prod_k \frac{P\left(f_i^k|l\right)}{P\left(f_i^k|\bar{l}\right)} \tag{2}$$

where $\bar{l}$ is the set of all labels excluding l and $P_a(f_i^k|\bar{l})$ is computed likewise.

Structure based.

In order to exploit the additional information given by the 3D structure, we establish a relationship ϕ that maps superpixels from of I_q to superpixels in A_q via the 3D structure (details in the following). Thanks to ϕ we can compute $\Pi(s_i)$, the set of points of S_q that are projected onto superpixel s_i. Points in $\Pi(s_i)$ carry forward the label of the associated superpixel in A_q, so every point in $\Pi(s_i)$ casts a vote for the label of s_i; the probability of s_i having label l (according to the structure) is estimated by count as:

$$P_s(l|s_i) = \frac{1 + n(l, \Pi(s_i))}{1 + n(*, \Pi(s_i))} \tag{3}$$

where $n(*, \Pi(s_i))$ is simply the cardinality of $\Pi(s_i)$.

We combine the two terms $\Lambda_a(l, s_i)$ and $P_s(l|s_i)$ after converting to an energy to obtain the unary (data) term:

$$E_{\mathrm{d}}(s_i) = -\log \Lambda_a(s_i, l) - \alpha \log P_s(s_i|l). \tag{4}$$

The α coefficient modulates the relative influence of appearance and geometry, and was set to $\alpha = 0.8$ in our experiments.

It remains to discuss the method by which we establish a relationship between a superpixel in A_q and a superpixel of I_q via the 3D structure. The first step (performed once and for all) is to compute the visibility of 3D points with respect to the superpixels in A_q. This is done with ray-casting in a discretized volumetric space, where a cell (or voxel) is deemed occupied if more than a given number of points are inside. Marching along the ray from the 3D point $\mathbf{x}$ to the center of the superpixel s (back-to-front), s is added to the visibility set of $\mathbf{x}$ if all the cells on the path are free (see Figure 2). Then for each superpixel s we compute $\Pi(s)$, the set of 3D points that projects onto s, according to the visibility.

The same procedure is applied to the superpixel in I_q, and in this way a map ϕ from a superpixel $s \in I_q$ to a set (possibly empty) of superpixels of A_q is established:

$$\phi(s) = \{s_i \in A_q \mid \Pi(s_i) \cap \Pi(s) \neq \emptyset\} \tag{5}$$

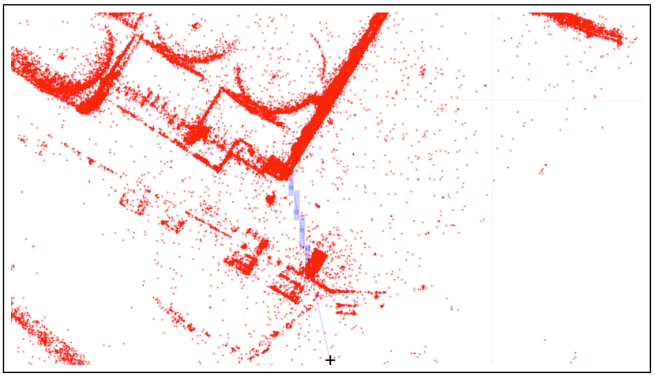

Figure 2: Illustration of the ray casting procedure. The path from the corner of the church to the camera position (cross) is obstructed by the statue, so the ray tracing (light-blue squares) stops there.

This geometric way of handling occlusions fails if the occlusor have not been reconstructed (no 3D points are attached to it). In order to cope with this problem we also perform a photoconsistency check between superpixels that have been associated via ϕ. In other words, superpixel s

must be photoconsistent (i.e. , have similar appearance) with the superpixels in $\phi(s)$.

The photoconsistency of two superpixels $s_i \in I_q$ and $s_j \in \phi(s_i)$ is evaluated as the chi-square distance between two-dimensional histograms of the log-chromaticity representation of the superpixels. In formula:

$$\chi^2(h_i, h_j) = \frac{1}{2} \sum_m \frac{(h_i(m) - h_j(m))^2}{h_i(m) + h_j(m)} \quad (6)$$

where h_i is the color histogram of s_i transformed in the log-chromaticity space and normalized. The log-chromaticity [3] transformation of an RGB image is defined as:

$$\{R, G, B\} \rightarrow \{log(R/G), log(B/G)\}. \quad (7)$$

After transforming in log-chromaticity space, two patches representing the same surface under different illumination (in color and intensity) are related by a linear transform that we factor out by subtracting the mean and dividing by the standard deviation of the log-chromaticity values. After such normalization, two patches representing the same surface should have identical values.

2.4.2 *Pairwise Term*

The pairwise term of the MRF enforces the spatial contextual constraint on the label assignment. It differs from the classical constant Potts penalty because the cost is modulated by a term which is inversely proportional to distance in 3D space, the rationale being that little or no cost is paid for assigning different labels to superpixels that are associated to 3D points that are far apart in space:

$$\phi(l_a, l_b) = \beta w_d \delta\left(l_a \neq l_b\right). \quad (8)$$

The distance term w_d is defined as:

$$w_d(s_i, s_j) = \begin{cases} e^{(-\mu \|p_i - p_j\|)} & \text{if } \Pi(s_i) \neq \emptyset \wedge \Pi(s_j) \neq \emptyset \\ 1 & \text{otherwise} \end{cases} \quad (9)$$

where p_i is the centroid of $\Pi(s_i)$, and μ is a rate of decay that depends on the scale of the 3D structure. In our experiments we set β and μ respectively to 10 and 4. The resulting smoothness term is:

$$E_S = \sum_{(s_i, s_j) \in E} \phi_{s_i, s_j} \quad (10)$$

2.4.3 *Inference*

The total energy cost function is the sum of the unary terms (Eq. 4) and pairwise terms (Eq. 10):

$$E(s) = \sum_{s_i \in \mathcal{V}} \underbrace{(-\log \Lambda_a(s_i, l) - \alpha \log P_s(s_i|l).)}_{\text{unary term}} + \underbrace{+ \sum_{(s_i, s_j) \in \mathcal{E}} \phi_{s_i, s_j}}_{\text{pairwise term}}$$

In order to minimize this submodular function we decide to apply the tree-reweighted belief propagation algorithm [10], as its implementation lends itself well to code parallelization (in CUDA as well) in order to reach high efficiency performance. However, this implementation requires a regular connectivity. While at the onset of SLIC every superpixel has exactly four neighbors, during the evolution of the clusters in principle this connectivity may vary, although in practice it is mainly preserved. Therefore we decided to *force* the connectivity to remain unchanged.

3. EVALUATION

Since no existing benchmark database are available that meet our requirements, we created a brand new labeled dataset of 50 images with the related 3D points cloud reconstruction obtained with "Samantha", a Structure from Motion algorithm [6] whose implementation is available online[1].

Figure 3: Top: sample images of *Piazza del Santo* from the labelled dataset. Bottom: a view of the 3D structure.

The depicted scene is a public square (*Piazza del Santo*, Padova, IT) in front of a church (*Basilica di S. Antonio*), with a horse statue occluding the background in some particular views[2]. This particular scenario has been chosen in order to fully evaluate the 3D data potential.

The cardinality of S, the set of 3D points, in this dataset is about 70k, a view of the 3D structure can be seen at the bottom of Figure 3, while some sample images are shown on the top. The original image resolution, used in the Structure from Motion algorithm, is 2592×1944 pixels but for the labeling procedure the images have been scaled by a factor of four in order to speed up the computation and to dampen

[1] samantha.3dflow.net

[2] Dataset available from www.diegm.uniud.it/fusiello/demo/3lt/.

the number of superpixels extracted. From each low resolution image (648 × 486) an average of 600 superpixels have been computed.

The ground truth data have been obtained by manually labeling each image of the dataset (in this case $D = A$), assigning to each superpixel one of the following labels: "base statue", "statue's pedestal", "statue's horse", "scaffolding cover", "woodboard", "church", "church's cupola" , "ground", "building" and "sky".

Figure 4 shows an example of ground truth labeling, the portions of the image in grayscale are associated with no label.

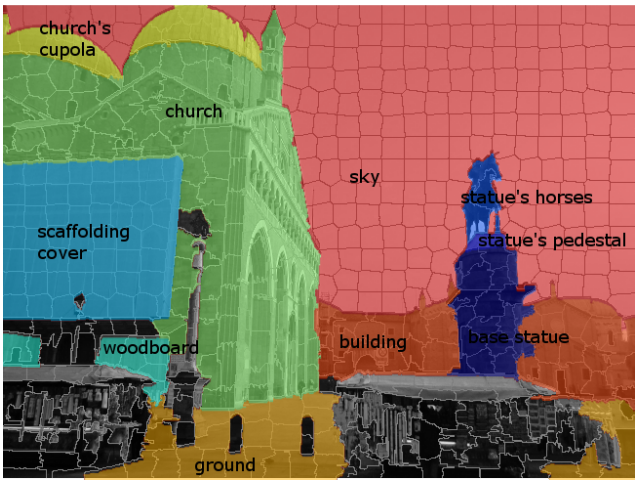

Figure 4: Example of hand-made labeling.

We tested our system with a leave-one-out approach. One at a time each image has been removed from the dataset. As $I_q \in D$ the experiment falls in the first of the two retrieval scenarios explained in Section 2.3, we reproject into $A \setminus I_q$ the 3D points visible from I_q in order to retrieve a subset of at most 10 most similar images A_q that will be used for the labeling procedure. In this way, in case of a dataset of more considerable dimension, we do not have to employ the feature data of the entire dataset during the nearest neighbour extraction, preserving the system efficiency.

We compared our proposed method to a partial implementation that ignores the 3D data contributions and is similar in spirit to [16]. The performance of our system is presented in terms of per-pixel classification rate, our proposed method achieves 82.6% whereas the restricted version exploiting only 2D data reaches 78.1%.

As shown in Figure 5, due to the lack of uniformity in the dataset label distributions we report also the results in terms of per-pixel classification rate for each label, see Figure 6. The average per-class rate is 59.4% for the complete version of the proposed method and 49.7% for the 2D restricted version.

The addition of the 3D data in the MRF clearly improves the labeling results, particularly in urban scenarios where some labels could be difficult to discern, like different buildings with similar colors and textures appearance (e.g. "base statue" and "church"). This is clearly displayed also in Figure 7 that shows a qualitative example of two labeling results related to the first two images of Figure 3.

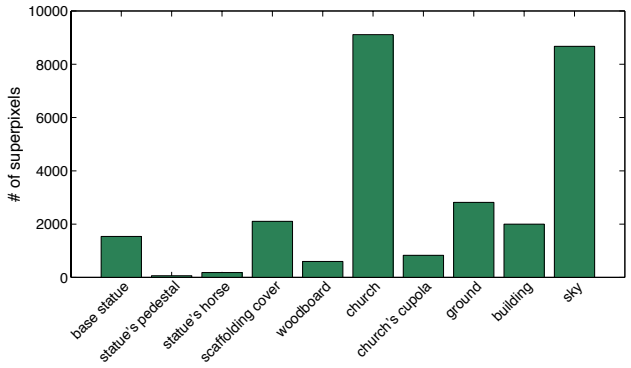

Figure 5: Superpixels label frequencies in the dataset.

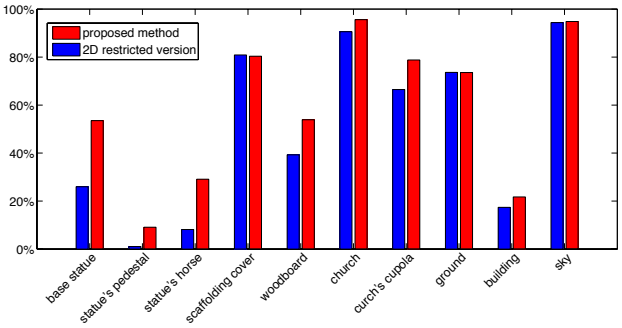

Figure 6: Per-class classification rate.

On the other hand, there are other labels for which little or no improvement is obtained, e.g. "ground" , "sky" and "scaffolding cover". These labels are associated to textureless region that have not been reconstructed by the Structure from Motion pipeline, because no SURF keypoints have been detected there, as shown in Figure 3 (bottom).

The time cost of labeling a query image is summarized in Table 1 focusing the attention to each steps of the proposed system.

Table 1: Average execution time

Superpixel extraction	2 sec
Feature extraction	23 sec
Visibility check	25 sec
Label inference	1 sec

The current version of the proposed system is implemented in MATLAB with no use of parallelization and optimized data structure; the computational bottlenecks are the Feature Extraction and the Visibility check stages that could be speed up exploiting the processing power of a graphic card.

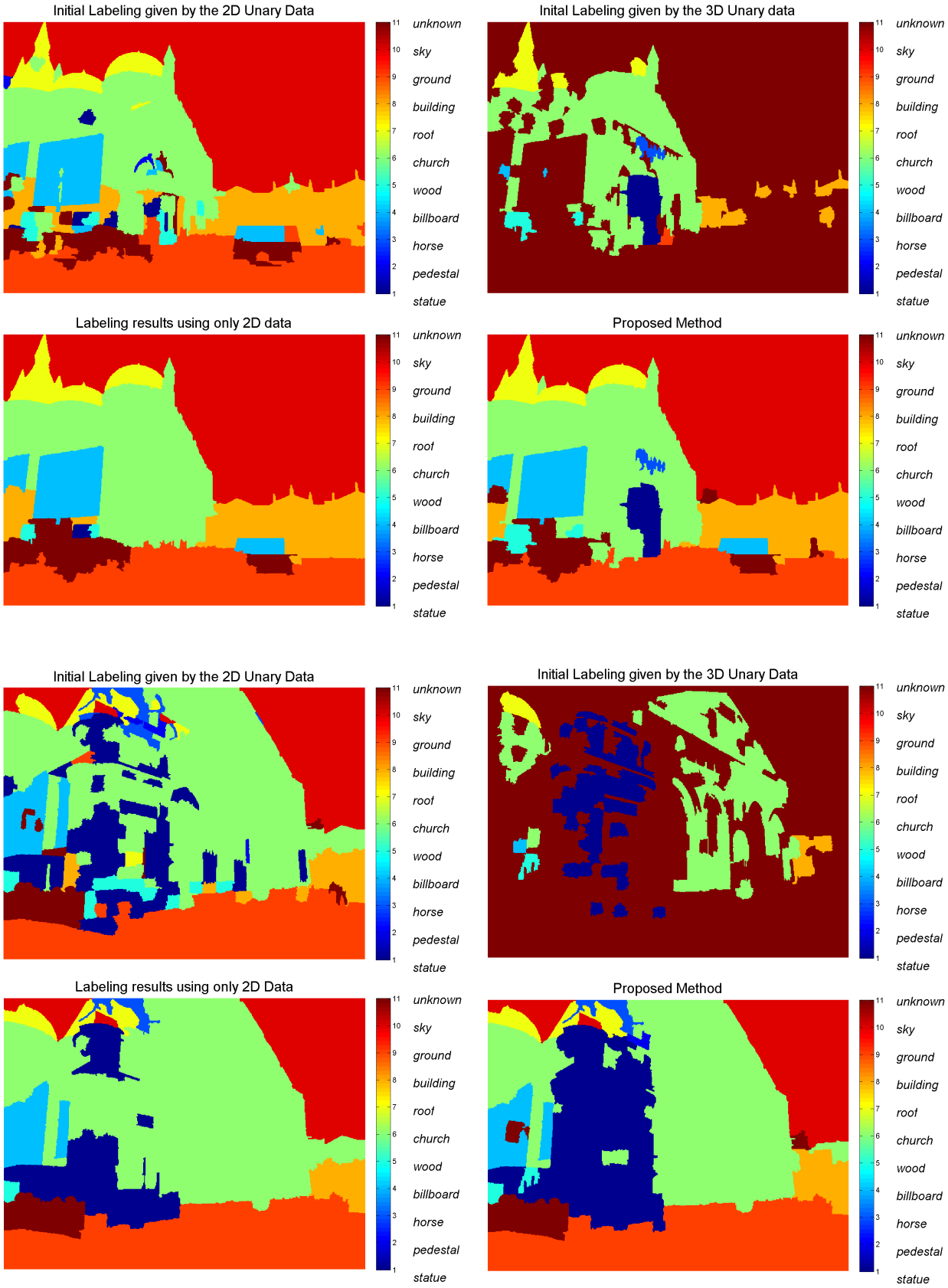

Figure 7: Two examples of labeling results. The amelioration brought by 3D structure can be appreciated by comparing the left and right images of the second and third rows.

4. DISCUSSION

This paper has investigated the leveraging of 3D structure in the process of label transfer. The method that has been presented takes advantage of i) already labeled images, and ii) sparse 3D structure to infer the category of superpixel in a novel image. Due to the laboriousness of creating annotated datasets with 3D structure, experiments that have been reported are only exploratory, but they clearly indicate that the 3D structure, even if sparse, can improve the process of label transfer.

Future work will be aimed at probing the method on larger and different datasets, keeping focused on scalability.

5. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274 – 2282, 2012.

[2] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proceedings of the European Conference on Computer Vision*, 2008.

[3] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. Removing shadows from images. In *Proceedings of the European Conference on Computer Vision*, pages 823–836, 2002.

[4] J. Frahm, P. Fite Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Proceedings of the European Conference on Computer Vision*, pages IV: 368–381, 2010.

[5] V. Garro, M. Galassi, and A. Fusiello. Wide area camera localization. In *Proceedings of the International Conference on Image Analysis and Processing*. Springer. To appear.

[6] R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] S. Gould and Y. Zhang. Patchmatchgraph: Building a graph of dense patch correspondences for label transfer. In *Proceedings of the European Conference on Computer Vision*, 2012.

[8] A. Irschara, C. Zach, and H. Bischof. Towards wiki-based dense city modeling. In *Proceedings of the International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[9] A. Irschara, C. Zach, J. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 2599–2606, 2009.

[10] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, October 2006.

[11] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *Proceedings of the 10th European Conference on Computer Vision*, pages 28–42, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] C. Liu, J. Yuen, and A. B. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1972–1979, 2009.

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct. 2003.

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[15] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, November 2008.

[16] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Proceedings of the 11th European conference on Computer vision*, pages 352–365, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Proceedings of the International Conference on Computer Vision*, 2009.

[18] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *Proceedings of the 11th European conference on Computer vision*, pages 561–574, Berlin, Heidelberg, 2010. Springer-Verlag.