Can Similar Scenes help Surface Layout Estimation?

Santosh K. Divvala, Alexei A. Efros, Martial Hebert
Robotics Institute, Carnegie Mellon University.
{santosh,efros,hebert}@cs.cmu.edu

Abstract

We describe a preliminary investigation of utilising large amounts of unlabelled image data to help in the estimation of rough scene layout. We take the single-view geometry estimation system of Hoiem et al [3] as the baseline and see if it is possible to improve its performance by considering a set of similar scenes gathered from the web. The two complimentary approaches being considered are 1) improving surface classification by using average geometry estimated from the matches, and 2) improving surface segmentation by injecting segments generated from the average of the matched images. The system is evaluated using the labelled 300-image dataset of Hoiem et al. and shows promising results.

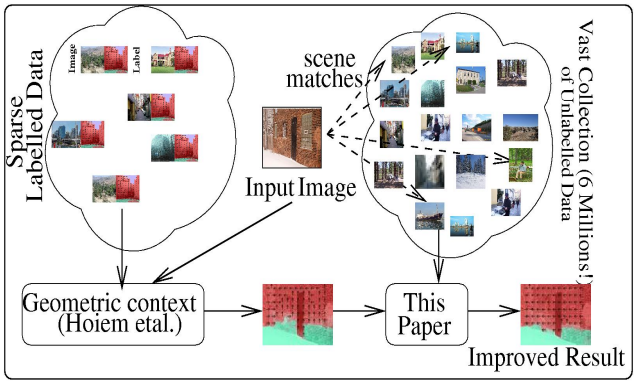


Figure 1. In this paper we show that Geometric Context [3] performance can be improved by using a set of scene matches drawn from a large unlabelled image collection

1. Introduction

Reasoning about a scene from a photograph is an inherently ambiguous task. This is because a single image in itself does not carry enough information to disambiguate the world that it is depicting. Of course, humans have no problems understanding photographs because of all the prior visual experience they can bring to bear on the task. How can we help computers do the same?

After a lull of some 30 years, there is a resurgence of interest in single view approaches to scene understanding. Good progress has been made on a number of fronts, including contextual priming [10], depth estimation [6, 8], surface layout estimation [3] and classification [5], among others. One common ingredient in all these approaches is that they pose the problem in terms of supervised learning, using training datasets of labelled input/output image pairs.

However, labelled data is hard to come by, especially in large enough quantities that are likely to be necessary for modelling the full richness of our visual world. At the same time, the popularity of Internet photo-sharing websites such as Flickr means that there are now enormous amounts of freely available but largely *unlabelled* data, literally billions of images! Some approaches suggested using this data via keyword searches, as a way of providing weak text-based annotations. But while this works really well for specific (and well-known) geographic locations [9] or specific objects [12], it fails to find more generic scene types (e.g. "alleyway").

Very recently, a new class of approaches have emerged which successfully use huge amounts of completely unlabelled visual data. Hays and Efros [2] demonstrate that it is possible to find semantically similar scene matches in a dataset of 2 million Flickr images and use that information to fill holes in images. Torralba *et al*. [11] show that using a dataset of 80 million 32×32 images, it is possible to perform operations such as gray-image colorization and detecting image orientation without any labels, just from the power of the data itself. But we believe that these application are only scratching the surface of what is possible with very-large-scale, unlabelled image databases.

1.1. Overview

In this paper, we will describe a preliminary investigation of utilising large amounts of unlabelled image data to help in the estimation of rough scene layout (See Fig. 1). We will take the single-view geometry estimation system of Hoiem *et al*. [3] as the baseline and see if it is possible to improve its performance by considering a set of similar scenes gathered from the web. Given an input image, the system

attempts to segment it into meaningful surfaces and to classify each surface into a set of *geometric classes* which correspond to rough surface orientations. The system provides two classifiers: the main one, deciding between "ground", "vertical", and "sky" classes, and the sub-classifier which predicts the orientation of the "vertical" surface into planar "left", "center", and "right", and non-planar "porous" and "solid" classes. The performance of the main classifier as reported in [3] is already at the level close to the labelling noise and is unlikely to be improved further. The performance of the sub-classifier however is still rather weak, and that is where we will concentrate our efforts here.

We propose to improve the performance of the system of Hoiem *et al*. by considering not only the actual input image to the system, but also a set of similar scenes found in a large image dataset of 6 million images collected from Flickr. Following the approach of Hays and Efros [2], we use the scene gist descriptor of Oliva and Torralba [6] to query a large image dataset to find a set of 100-200 nearest neighbour matches. Our goal is to use these scene matches to help improve the algorithm's understanding of the input image at hand.

Of course, the matches are not labelled, so it is not immediately obvious that any improvement can be achieved since little new information is being added to the process. However, in this paper we argue that even without labels, similar matching images could be of help for a difficult single-image task. In the case of the Geometric Context system of Hoiem *et al*., we consider two complimentary approaches for improving the performance:

Improving surface classification Consider a surface such as a side of a building, but with an unusual or confusing texture which prevents the geometric context algorithm from estimating the surface orientation correctly. Now further consider that a set of similar scenes is available where the geometric arrangement of buildings is very much like in the original image, but the texturing of the surfaces is different. In this case, the geometric context output for the matches might produce better results than for the original image. Now, if we can figure out a way of "injecting" this new knowledge into the system, we can improve the performance of surface classification (See Fig. 2).

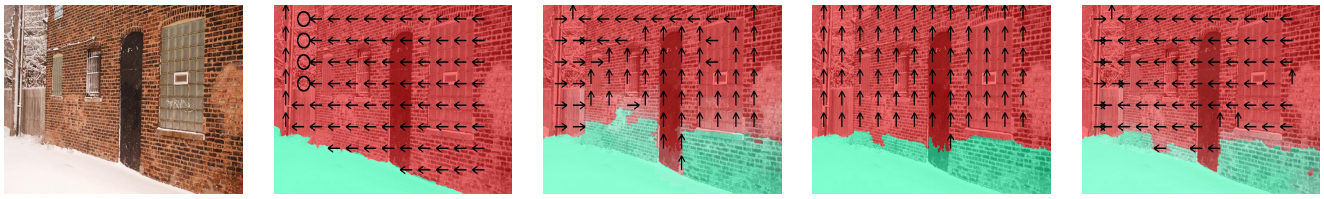
Improving surface segmentation Another problem with the geometric context approach is that it often does not find a suitable segmentation for each surface, which leads to poor surface classification performance. Indeed, Hoiem *et al*. have shown that their classification performance improves a great deal if a "perfect" ground-truth segmentation is used in their system. This suggests a potential benefit to try and improve segmentation. Again, consider an input image where the boundary between two surfaces is not well-defined in the image and thus is missed by all the segmentations of Hoiem *et al*.'s multiple segmentation algorithm. Now consider the set of scene matches which, hopefully, depict geometrically similar scenes, but perhaps with different failure modes. Therefore, one can hope that on average, the two surfaces in question would be visually more dissimilar and easier to segment out. Again, if we could utilise the knowledge provided by the set of matching scenes, and "inject" it into the segmentation algorithm, we might be able to improve the performance of the system (See Fig. 4).

In the rest of this paper, we will describe our approach for operationalizing these two ideas and thereby present results of our investigation on the images from the Geometric Context dataset [3] with scene matches retrieved from *Flickr*. In the following, we use the words 'scene matches', 'nearest neighbours' and 'similar scene images' interchangeably.

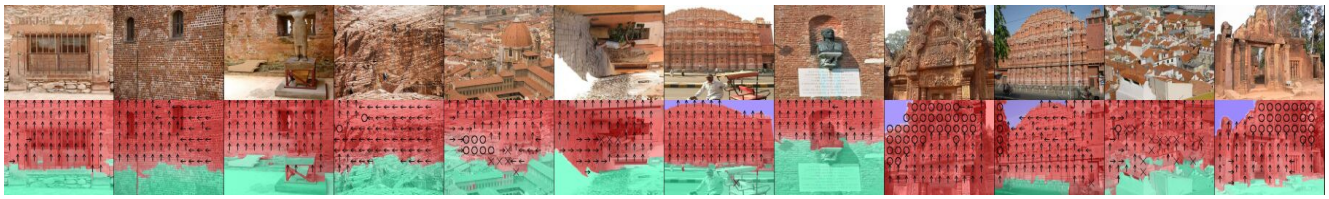
2. Improving Surface Classification

Our basic motivation is to use information from huge amounts of unlabelled but similar images to aid the surface classification of an input image. As mentioned before, we retrieve a set of similar scene images for an input image from the vast collection of digital images on Flickr using the method of [2]. The nearest neighbours are based on the GIST feature descriptor plus color. The top 200 of the retrieved nearest neighbours are used in our algorithm. Given these neighbour images, we apply the Geometric Context (GC) algorithm to obtain their coarse image geometry. The GC algorithm proposed by Hoiem *et al*. [3] employs a multiple segmentation framework to build larger segments by merging simple super-pixel [1] elements of an image. By using the features extracted from these larger segments in a classification framework, the mapping between scene features and their corresponding rough surface layout is learnt. The algorithm outputs a 'confidence' vector V_s for every super-pixel s in the image (which indicates the probability of the super-pixel element belonging to either of the seven geometric classes) and the max component of this vector is chosen as the most probable label.

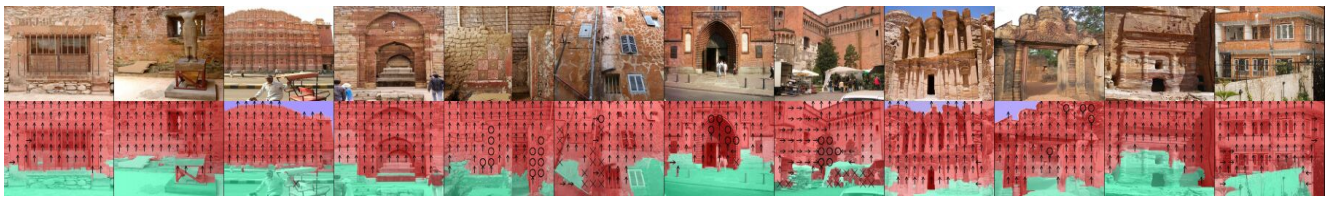
The nearest neighbours retrieved using [2] are similar to the input image with respect to their overall scene appearance (based on GIST). As no geometrical cues are utilized in the matching process, some of the retrieved matches could be random and may not have any relation to the input image in terms of their surface layout. In order to filter the matches, we perform a clustering step. It must be emphasised that in [2] scene matches were filtered by user interaction, while in [7] the problem was resolved by clustering on the ground-truth labels associated with the matches. As we do not have access to the labels, we perform a mean-shift clustering of the images by using the GC output confidences as the features for clustering. More precisely, the



(a) Input image (b) Ground truth (c) Hoiem *et al*. (d) Average scene prior (e) Our result



(f) Top 12 scene matches retrieved using GIST (above) and their Geometric Context outputs (below)

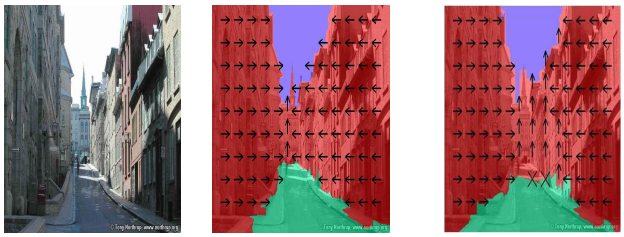


(g) Top 12 scene matches obtained after clustering (above) and their Geometric Context outputs (below)

Figure 2. Using average scene prior to aid Geometric Context (GC): Given an input image, the scene matches retrieved using [2] (second row) are clustered based on their GC output confidences to obtain a relevant set of matches (third row). The clustering helps in filtering out poor scene matches (images 4,5,6,11 in the second row). Using the clustered matches, the *average* scene prior is computed by marginalizing over their GC confidences and is used as a feature for the input image (only the max component is displayed in figure). This has helped in improving the surface layout result of the 'brick wall' scene.

confidence vectors V_s obtained for every pixel in the ith neighbour image are vectorized to obtain a single large confidence vector V_i' and is used as the feature vector to cluster the N neighbours. At the end of the clustering process, we only consider the images belonging to the largest cluster and ignore the rest. The clustering step is quite helpful in obtaining a refined set of matches with similar rough surface layouts amongst them (as observed in Fig.2(g)). Moreover we observed a quantitative improvement in the sub-classifier classifier accuracy (of around 1.5%) when using the average scene prior from clustered neighbours (accuracy was 63.2%) as against the scene prior from all the neighbours (accuracy was 61.7%).

Given the clustered scene matches, we can now get an average geometric scene prior for the input image by marginalizing the GC output confidences over the clustered scene matches. More precisely, given a super-pixel element s of the input image, the GC confidence vectors V_s of the corresponding pixel elements across the neighbour images are averaged to yield a prior confidence vector P_s for the super-pixel. As the matches possess similar surface layout, by marginalizing their output confidences, we could obtain a good guess of the most probable geometric class label of a super-pixel element in the input image. In Fig. 3, we see that this average prior by itself performs well in classifying

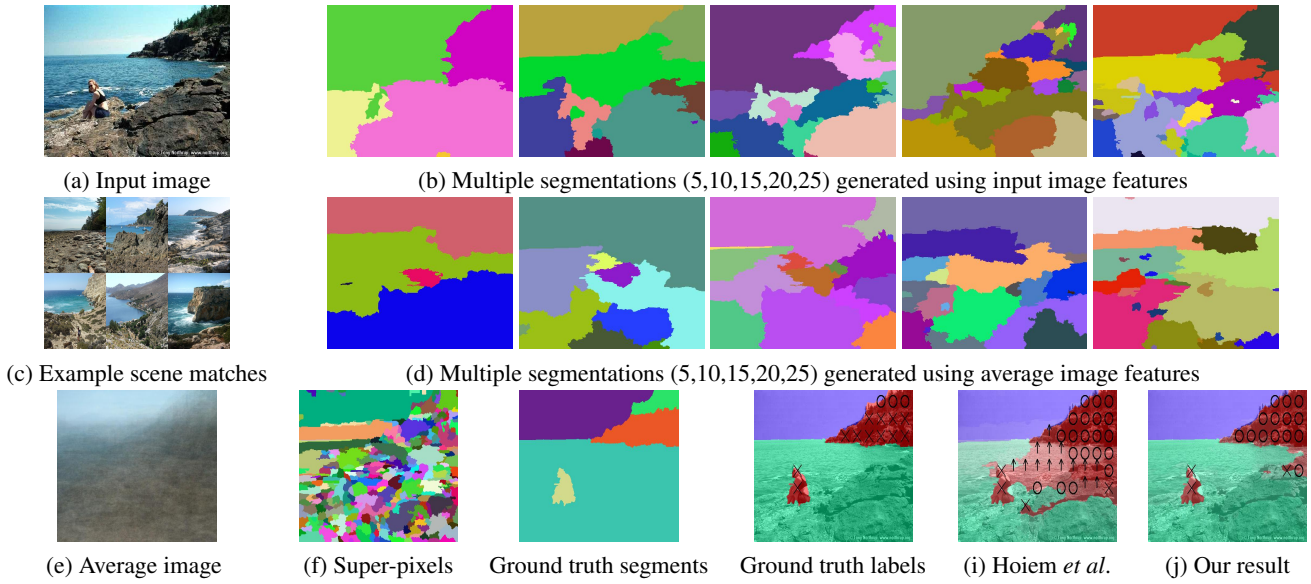


(a) Input image (b) Ground truth (c) Average scene prior

Figure 3. The average scene prior performs surprisingly well in estimating the surface layout of an input image.

the input image. Interestingly (as we will see in Section. 4), the classification accuracy obtained just by using this prior information (without actually looking at the image) is comparable to the baseline result of Hoiem *et al*.

The average geometric scene prior acts as a good cue for scene classification of the input image. Thus we use it as a feature in conjunction with the the original set of features employed by Hoiem *et al*. in their algorithm (*i.e.,* colour, texture, location and perspectivity) and retrain the classifiers. Had this prior information contained discriminative characteristics about the input image, it should have aided in the better classification of the super-pixels. We indeed observed that the classifiers had used these features during



(a) Input image (b) Multiple segmentations (5,10,15,20,25) generated using input image features

(c) Example scene matches (d) Multiple segmentations (5,10,15,20,25) generated using average image features

(e) Average image (f) Super-pixels Ground truth segments Ground truth labels (i) Hoiem *et al.* (j) Our result

Figure 4. Similar scenes aid in proposing better segmentations for an image: Given an input image, features from its scene matches (retrieved using [2]) are extracted and marginalised to obtain an average scene feature (which could be visualised as the feature corresponding to the average image). The multiple segmentations generated using this average feature set are characterised based on the overall scene geometry and thus could be helpful in the classification of the input image. Observe the improvement in the surface layout result of the 'shore' scene.

the classification process which signifies the prior as a useful cue.

Fig. 2 graphically illustrates the overall approach. Given the input image of a 'brick wall' scene, the top 12 (of the 200) nearest neighbour matches retrieved by [2] are displayed along with their Geometric Context outputs. Some of the top matches (*e.g.,* images 4,5,6,11) fail to possess similar surface layout as the original image. However by clustering the scene matches based on their GC output confidences, we gather a cluster of matches with similar surface layouts. Using these clustered neighbours, average scene prior is computed by marginalizing over them and is used as a feature in the classification algorithm. Observe that the estimated average scene geometry correctly predicts (with few errors) the entire block of the brick wall as a single large *vertical* class segment. This helps in an improved estimation of the perspectivity cues that leads to its correct sub-classification as belonging to the planar *left* class.

3. Improving Surface Segmentation

Having good spatial support for regions in the input image significantly helps in improving segmentation and classification [3, 4]. Unfortunately, this task is not easy and it is often a challenge to produce good segmentations (corresponding to different homogeneous regions in an image) by only using the information from the image. Here we study the possibility of utilising the information from the retrieved scene matches to aid the surface segmentation task.

In the previous section, we had considered the average scene geometry for improving the classification result. It is also possible to consider the average of the neighbour images themselves in hope that they can improve the segmentation. However, we did not find the direct pixel-wise average of the scene matches to be useful (see Fig. 4(e)). Instead, given the retrieved scene matches, we marginalise the low-level scene features *i.e.,* colour, texture, location (that were already computed while applying the Geometric Context on them) to obtain an average in the feature space.

The average feature vector is a useful cue for forming better segments of the input image. To see this, consider the case where the input image consists of a 'ground' region with varying texture and colour characteristics. In such a case, using the super-pixel features exclusively from the input image would fail to merge these regions together to form a holistic ground segment. However, by retrieving the similar scene matches and computing the average scene feature, the overall structure of the scene (unaffected by artifacts such as shadows etc) could be captured. Thus the averaging yields a high-level descriptor for the super-pixel elements of the input image based on the global scene organisation. Using this descriptor potentially helps in forming improved segmentations that better convey the overall scene structure and in turn offer good spatial support.

Fig. 4 illustrates our idea. Given the 'shore' scene image, the features corresponding to its similar scene matches are

used to obtain an average scene feature (as it is difficult to visualise the feature vectors corresponding to scene matches and the average, we instead display the images). Using these features, we generate the multiple segmentations as done in Hoiem *et al*. These multiple segmentations provide improved spatial support for different homogeneous regions and when used along with the segmentations generated using input image features leads to an improved surface layout estimation for the scene. Hence the scene matches can be used for improving the segmentation of the input image, in addition to improving the classification accuracy.

4. Experimental Results

The evaluation of the proposed hypothesis was performed on the 300 images of the Geometric Context dataset [3]. The top 200 scene matches to each image were retrieved using the method of [2] from the collection of six million images in *Flickr*. By applying the Geometric Context algorithm on these images and marginalizing their output confidences, we obtain an average scene prior (as described in Section. 2). In Fig. 5, we analyse the quality of this scene prior estimated with a varying number of neighbour images. The figure displays the main-classifier and vertical sub-classifier classifier accuracies obtained by employing only the scene prior and without using any input image features. It is interesting to note that by just using the scene prior, one could achieve a modest classification result (*i.e.,* the accuracies drop only by around 5-8% compared to the baseline result of Hoiem *et al*.). Also the peak is obtained when around top 40 nearest neighbours are used.

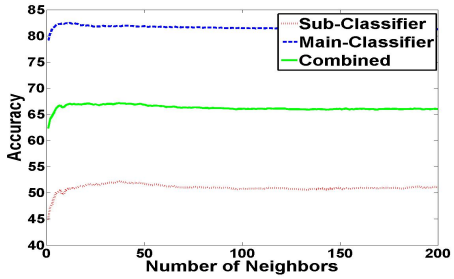


Figure 5. Accuracies obtained by employing only the average scene prior computed from a varying number of scene matches. By using only scene matches (and no information from the input image), we could achieve an accuracy of 82.5% for the main-classifier and 52.3% for the sub-classifier.

In Fig. 6, we demonstrate few results of using the average scene prior as a useful cue for improved surface classification. The figure illustrates few of the examples where the surface layout result was improved by inclusion of the scene prior. Notice that the clustering process helps in gathering similar scenes out of the nearest neighbour image matches and the marginalization results in providing the overall spatial layout of the scene.

In Fig. 7, we display few results of the improved segmentations. For every input image, the features from the retrieved scene matches are marginalised to obtain an average feature vector. This feature vector is used in the segmentation process to yield better spatial support for the homogeneous regions within the image. Notice that the improved spatial support has helped in obtaining better Geometric context result.

We supplement the above qualitative analysis of the approach by a quantitative performance analysis. In Table. 4, the classification accuracies obtained for the main-class and the vertical sub-class classifiers are displayed for the various cases. As mentioned in Section. 1.1, we mainly focus on the improvement of the sub-classifier classifier. Notice that using the augmented features set (*i.e.*, with the inclusion of average scene prior from clustered neighbours) helps in improving the sub-classifier classifier accuracy by around 3%, while using the average scene features yields an improvement around 1.5%. The improvements are statistically significant ($p < 0.05$). However, by combining both sources of information, we do not see much additive increase in the accuracies suggesting that the two sources of information are not completely independent.

Table 1. Classification Accuracies

	Sub-Class	Main-Class
1. Baseline (by Hoiem *et al*.)	**60.5%**	87.2%
2. Using only average scene prior (without using input image features)	**52.3%**	82.5%
3. Augmenting feature set with the average scene prior	**63.2%**	87.3%
4. Using segmentations from the average scene features	**61.5%**	87.7%
5. Combining both 3 & 4 above	**63.6%**	87.9%

5. Discussion

In this work, we have analysed the utility of large amounts of unlabelled image data to aid the process of surface layout estimation. By using information from similar scene matches to an input image, the Geometric context [3] result was improved. The improvement in accuracy is modest (from 60.5% to 63.6%). But it should be noted that the maximum possible sub-classifier classifier accuracy when tested on ground-truth segmentations using a classifier trained on multiple segmentations is 65.5%. (If the classifier is trained and tested both on ground-truth segmentations, then it acheives 71.5% [3] but this is not realistic). Thus the problem of achieving improved performance is quite challenging (as we are close to the optimal

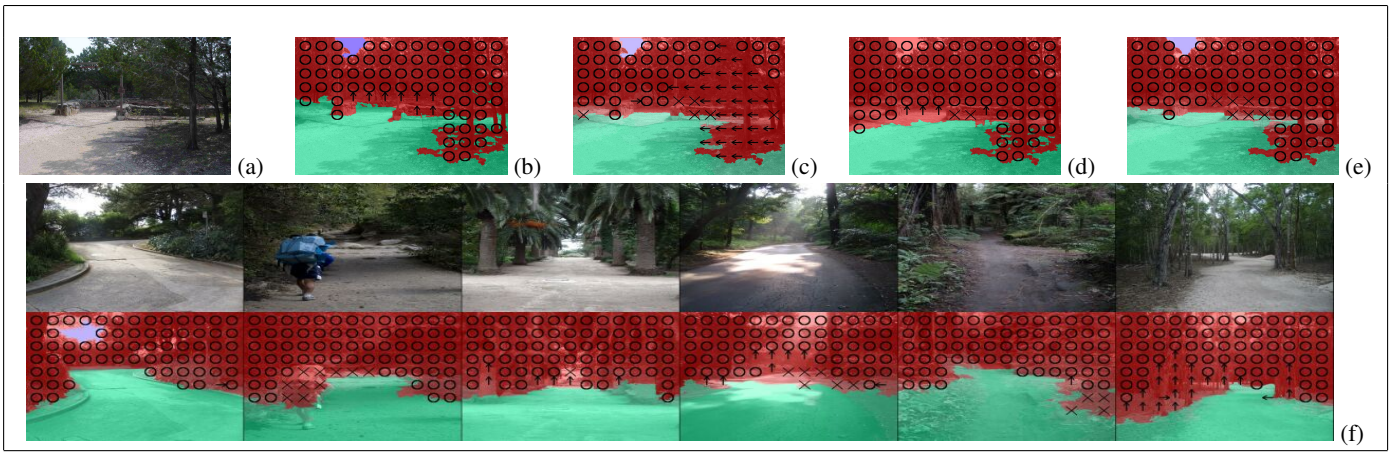


Figure 6. Using average scene prior as an input image feature helps in improving the Geometric Context (GC) result. (a) Input image (b) Ground truth labelling (c) Result from Hoiem *et al*. [3] (d) Average scene prior (e) Our result (f) Top few scene matches and their corresponding Geometric Context outputs. In this example, the classification of the tree is corrected upon inclusion of the scene prior.

operating point). Also further evaluating the qualitative results, we observe that for many example images where good scene matches were acquired, the improvement in the result is substantial. While for many other, the approach of Hoiem *et al*. already performs well and thus there is little scope for improvement. For the remaining cases (*i.e.,* extremely complicated scenes), the retrieved scene matches were poor and had not aided the classification. Nevertheless the most interesting observation is that even without using any input image features, one can achieve a good classification accuracy by just marginalizing the scene matches. This suggests that there exists an inherent consistancy amongst real world scenes that is captured by these unlabelled scene matches, thus providing a global constraint over any given test image.

Our preliminary investigation has shown the potential for using unlabelled data along with labelled data to achieve improved performance. Our future plan is to study better methods for injecting the scene prior (such as a mixture of experts framework) to further effectively use this information. Also our proposed approach could act as a potential tool for gathering more data for training the classifiers in a co-training based framework.

Acknowledgments We thank Derek Hoiem for providing the code and data used in [3], and James Hays for generating the scene matches from his *Flickr* dataset. This research was supported in part by the National Science Foundation under Grant IIS0745636.

References

[1] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 25, 2004.

[2] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007.

[3] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1), 2007.

[4] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. *British Machine Vision Conference*, 2007.

[5] V. Nedovic, A. W. Smeulders, A. G. Redert, and Jan-Mark. Depth information by stage classification. *ICCV*, 2007.

[6] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. Journal of Computer Vision*, 42(3):145–175, 2001.

[7] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[8] A. Saxena, S. Chung, and A. Ng. Depth reconstruction from a single still image. *International Journal of Computer Vision*, 74(1), 2007.

[9] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics (SIGGRAPH)*, 25(3), 2006.

[10] A. Torralba. Contextual priming for object detection. *Int. Journal of Computer Vision*, 53(2):169–191, 2003.

[11] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical Report MIT-CSAIL-TR-2007-024, MIT CSAIL, 2007.

[12] A. Torralba and A. Oliva. Depth estimation from image structure. *PAMI*, 24(9), 2002.

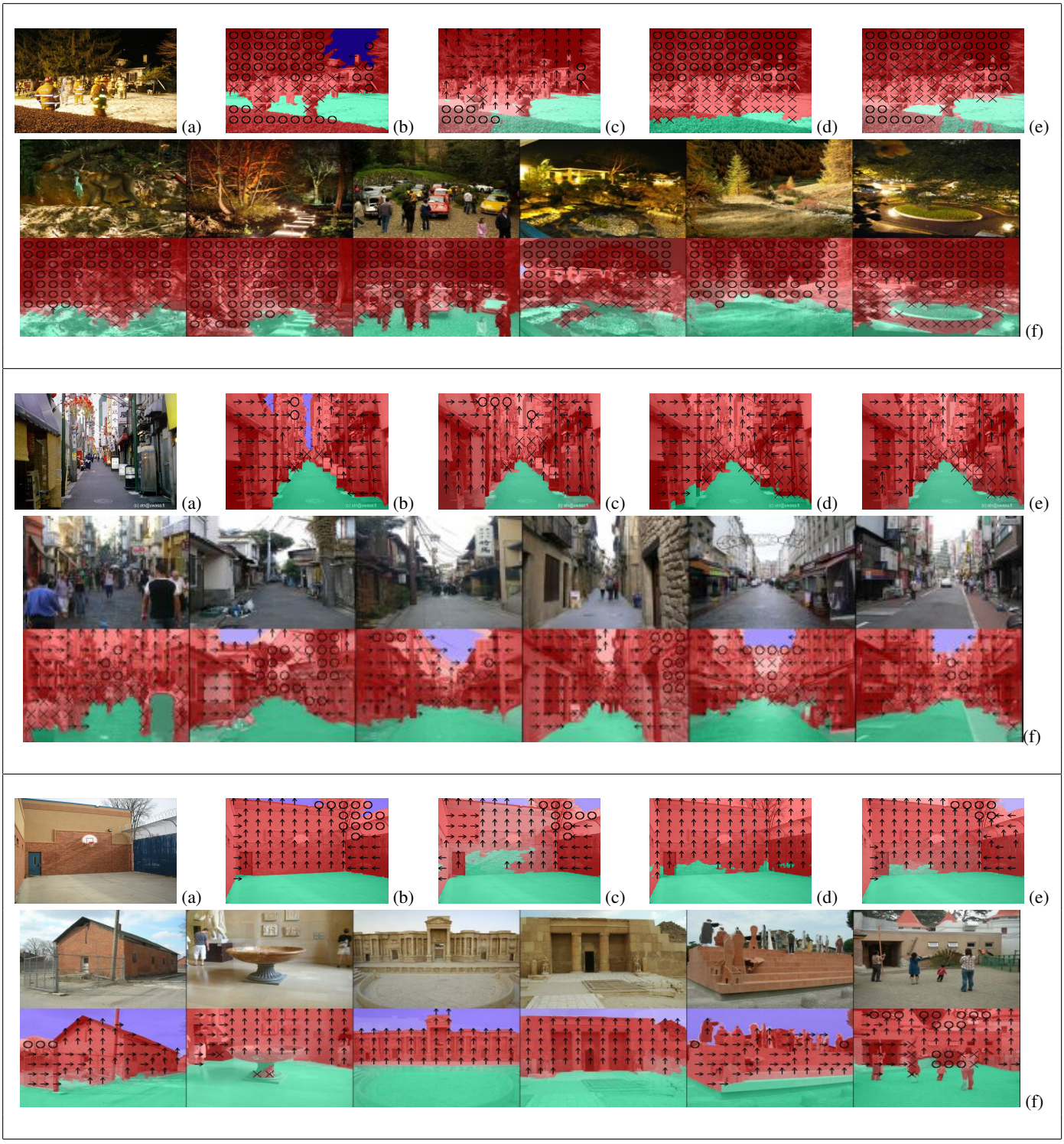


Figure 6(contd). Using average scene prior as an input image feature helps in improving the Geometric Context (GC) result. (a) Input image (b) Ground truth labelling (c) Result from Hoiem *et al*. [3] (d) Average scene prior (e) Our result (f) Top few scene matches and their corresponding Geometric Context outputs. The classification of the tree, side of the alley and the wall in examples 1, 2 and 3 respectively is improved upon inclusion of the scene prior.

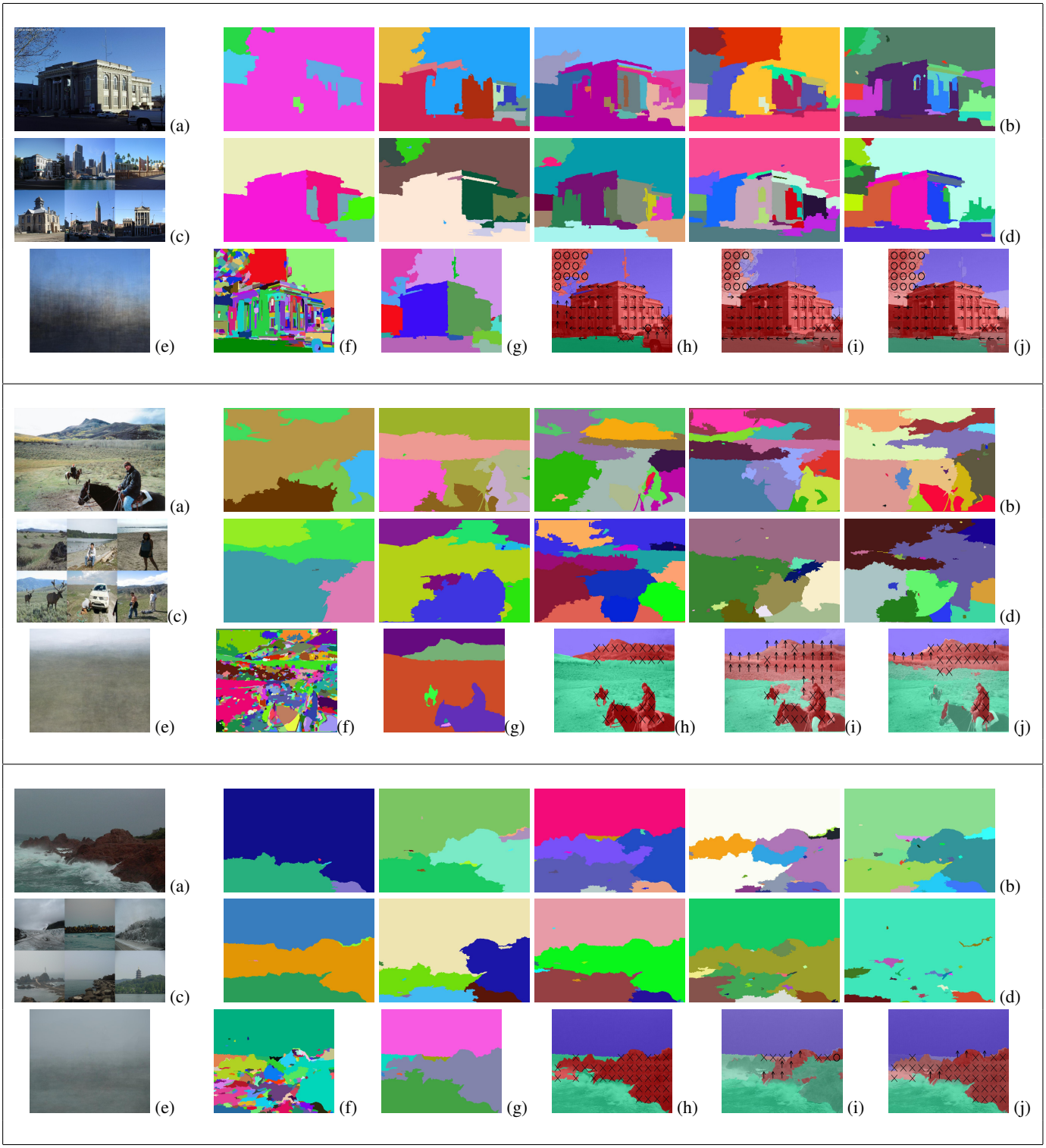


Figure 7. Similar scenes can aid the segmentation of an image. (a) Input image (b) 5 multiple segmentations (5,10,15,20,25 segments) generated using input image features (c) Example scene matches (d) 5 multiple segmentations (5,10,15,20,25 segments) using average scene features (e) Average neighbour image (f) Super-pixeled image (g) Ground-truth segmentation (h) Ground-truth labelling (i) Hoiem *et al*. [3] (j) Our result. The multiple segmentations generated using the average image features characterise the overall scene geometry and are helpful in estimating an improved surface layout for the input image.