Basic level scene understanding: from labels to structure and beyond

Basic Level Scene Understanding: From Labels to Structure and Beyond

Jianxiong Xiao Bryan C. Russell[+] James Hays[*] Krista A. Ehinger Aude Oliva Antonio Torralba
Massachusetts Institute of Technology [+]University of Washington [*]Brown University

Abstract

An early goal of computer vision was to build a system that could automatically understand a 3D scene just by looking. This requires not only the ability to extract 3D information from image information alone, but also to handle the large variety of different environments that comprise our visual world. This paper summarizes our recent efforts toward these goals. First, we describe the SUN database, which is a collection of annotated images spanning 908 different scene categories. This database allows us to systematically study the space of possible everyday scenes and to establish a benchmark for scene and object recognition. We also explore ways of coping with the variety of viewpoints within these scenes. For this, we have introduced a database of 360°panoramic images for many of the scene categories in the SUN database and have explored viewpoint recognition within the environments. Finally, we describe steps toward a unified 3D parsing of everyday scenes: (i) the ability to localize geometric primitives in images, such as cuboids and cylinders, which often comprise many everyday objects, and (ii) an integrated system to extract the 3D structure of the scene and objects depicted in an image.

Keywords: SUN, basic level scene understanding, scene viewpoint recognition, scene detection, 3D context model

Links: DL PDF

1 Introduction

The ability to understand a 3D scene depicted in a static 2D image goes to the very heart of the computer vision problem. By "scene" we mean a place in which a human can act within or navigate. What does it mean to *understand a scene*? There is no universal answer as it heavily depends on the task involved, and this seemingly simple question hides a lot of complexity.

The dominant view in the current computer vision literature is to name the scene and objects present in an image. However, this level of understanding is rather superficial and limits scene understanding applications. If we can reason about important semantic properties and structures of scenes, it will enable richer applications. Furthermore, the danger of working on an over-simplified task is that it may distract us from exploiting the natural structures of the problem, which may be important for the ultimate solution. Scaling up to real world scenes will require working with databases that exhaustively span our visual experience. Moreover, with more diverse data 3D reasoning becomes necessary to cope with the wide variety of viewpoints for a given scene.

http://sundatabase.mit.edu/

Figure 1: *List of 908 scene categories in our SUN database – the most exahustive scene dataset to date. The size of each category name is proportional to the number of images belonging to the category.*

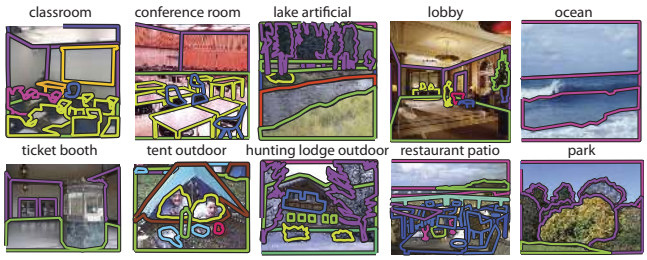

Figure 2: *Examples from 18,507 fully annotated images in SUN.*

Our ultimate goal of research is to pass the **Turing test for scene understanding**: Given an image depicting a static scene, a human judge will ask a human or a machine questions about the picture. If the judge cannot reliably tell the machine from the human, the machine is said to have passed the test. More specifically, this means that the machine must be able to convert the image into a data structure that represents all knowledge extracted from the picture. As the question can be diverse (e.g. how many cracks are on a wall of the picture?), seeking such a generic representation is a challenging task at the current level of research.

Therefore, we want to define a set of goals that are suitable for the current state of research in computer vision that are not too simplistic nor challenging, and also to produce a natural representation of the scene. Based on these considerations, we define the task of scene understanding as predicting the scene category, the 3D enclosure of the space, and all the objects in the images. For each object, we want to know its category and 3D bounding box, as well as its 3D orientation relative to the scene. As an image is a viewer-centric observation of the space, we also want to recover the camera parameters, such as observer viewpoint and field of view. We call this task **basic level scene understanding**, with analogy to basic level in cognitive categorization [Rosch 1978]. We believe that this basic level scene representation may be directly relevant to privileged sensory-motor affordances. It also has practical applications for providing sufficient information for simple interaction with the scene, such as navigation and object manipulation. In this paper we discuss several aspects of this basic level scene representation.

Figure 3: *Examples of 7,971 chairs that were manually annotated in 2,173 images of our SUN database.*

2 What are the categories for scenes?

One of the fundamental tasks of basic level scene understanding is to be able to classify a natural image into a limited number of semantic categories. What are the scene categories? From a human-centric perspective, the categories should capture the richness and diversity of environments that make up our daily experiences. Although the visual world is continuous, most environmental scenes are visual entities that can be organized in functional and semantic groups. A given scene or place may allow for specific actions, such as eating in a restaurant, drinking in a pub, reading in a library, or sleeping in a bedroom. To capture this diversity, we have constructed a quasi-exhaustive taxonomy and dataset representing the diversity of visual scene categories that can be encountered in the world. We have used WordNet [Fellbaum 1998], an electronic dictionary of the English language containing more than 100,000 words, and have manually selected all of the terms that describe scenes, places, and environments (any concrete noun that could reasonably complete the phrase "I am in a *place*", or "Let's go to the *place*"). This has yielded 908 scene categories, which are illustrated in Figure 1. Once we have a list of scenes, the next task is to collect images belonging to each scene category. Since one of our goals is to create a large collection of images with variability in visual appearance, we have collected images available on the Internet using various image search engines for each scene category term. Then, a group of reliable human participants manually pruned the images that did not correspond to the definition of the scene category. We have found that within the same scene category, some exemplars are more typical than others and category membership is naturally graded [Ehinger et al. 2011]. Despite this, we are able to consistently collect 131,072 images to date. We refer to this dataset as the SUN (Scene UNderstanding) database [Xiao et al. 2010]. To provide data for research and natural statistics of objects in scenes, we have also labeled objects in a large portion of the image collection with polygon outlines and object category names [Barriuso and Torralba 2012]. To date, there are 249,522 manually segmented objects for the 3,819 object categories labeled. Example images are shown in Figures 2 and 3. [Patterson and Hays 2012] also labels scene attribute for our image collection, and [Satkin et al. 2012] aligned computer graphics model on 500 SUN pictures.

3 Seeing Scenes within Scenes

3.1 Scene Detection

Imagine that you are walking down a street. A scene classification system will tell you that you are on the street and allow you to localize people, cars, etc. However, there are additional detection tasks that lie in between objects and scenes. For instance, we want to detect restaurant terraces, markets, or parking lots. These concepts also define localized regions, but they lack the structure of objects (i.e. a collection of parts in a stable geometric arrangement) and they are more organized than textures. We refer to these scenes within scenes as "sub-scenes" to distinguish them from global scene labels. A single image might contain multiple scenes, where a scene is a bounded region of the environment that has a distinct functionality with respect to the rest. For instance, an image can be composed of storefronts, a restaurant terrace, and a park. The objects and the actions that happen within subscenes have to be interpreted in the framework created by each local scene, and they might be only weakly related to the global scene that encompasses them. Just as people can move continuously between scene categories (eg. "office" into "corridor" and "street" into "storefront"), it is frequently the case that a single photograph depicts multiple scene types at different scales and locations within the image (see Figures 4 and 5). As scenes are more flexible than objects, it is unclear what the appropriate representation is in order to detect them in complex images. In our current investigation, we use our scene classification framework [Xiao et al. 2010] to directly classify image crops into sub-scene categories. We refer to this task as "scene detection". Figure 6 shows prediction results of the scene detector we have built.

Figure 4: *These scenes of a beach, a village, and a river are all from a single image, shown in the image of Figure 5. They have totally different semantic meanings and functions.*

Figure 5: *There are many complementary levels of image understanding. One can understand images on a continuum from the global scene level (left) to the local object level (right).*

Figure 6: *Example scene detection results for the "harbor" detector. Green box is correct detection and red box is mistake. The text above each image is the ground truth annotation obtained from Amazon Mechanical Turk voting task.*

Figure 7: *Example images for 3 categories out of the 80 scene categories in our high resolution SUN360 panorama database.*

3.2 Scene Viewpoint Recognition

Besides there being sub-scenes inside the view, there may be other scenes at different views of the same observer location that carries crucial semantic meaning for function and usage. For instance, a theater has a distinct distribution of objects – a stage on one side and seats on the other – that defines unique views in different orientations. Just as observers will choose a view of a television that allows them to see the screen, observers in a theater will sit facing the stage when watching a show. In [Xiao et al. 2012a], we introduce the problem of scene viewpoint recognition, the goal of which is to classify the type of place shown in a photo, and also to recognize the observer's viewpoint within that category of place. We have constructed a database of 360° panoramic images organized into 80 place categories, as shown in Figure 7. For each category, our algorithm automatically aligns the panoramas to build a full-view representation of the surrounding place. We also study the symmetry properties and canonical viewpoint of each place category. At test time, given a photo of a scene, the model can recognize the place category, produce a compass-like indication of the observer's most likely viewpoint within that place, and use this information to extrapolate beyond the available view by filling in the probable visual layout that would appear beyond the boundary of the photo.

4 3D Structure of Scenes

Although an image is a 2D array, we live in a 3D world, where scenes have volume, affordances, and can be spatially arranged where one object can be occluded by another. The ability to reason about these 3D properties would be of benefit for tasks such as navigation and object manipulation.

4.1 Geometric Primitive Recognition

For many objects, such as boxes, soda cans, and balls, their 3D shape can be entirely expressed by a simple geometric primitive, such as a cube, cylinder, or sphere. For other objects, their 3D shape may include one or more of these geometric primitives. The ability to detect these primitives and to recover their parameters would allow at least a partial 3D description for many depicted objects. Our desired output is not simply an indication of the presence of a geometric primitive and its 2D bounding box on the image. Instead, in [Xiao et al. 2012b], we build a 3D object detector to recover a parameterization of the object's 3D shape, along with the camera parameters. Figure 8 shows some examples of our geometric primitive detector.

4.2 Unified 3D Scene Parsing

Our goal is to realize a fully integrated system for basic level scene understanding that produces a full 3D parse result for a single image

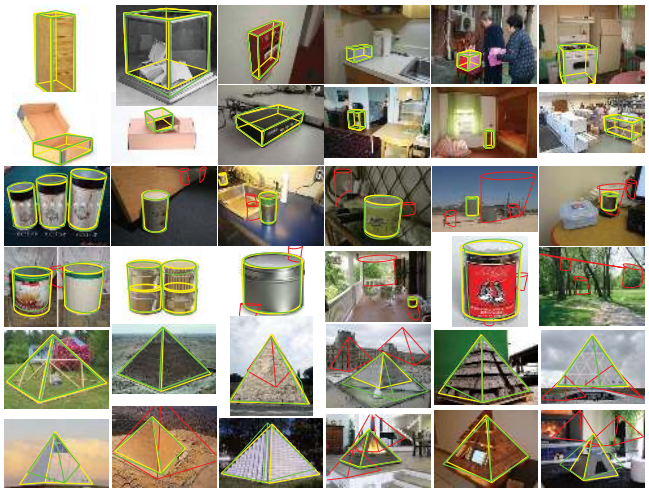

Figure 8: *Parameterized 3D interpretation of geometric primitives in scenes. Yellow outlines are ground truth annotations, green outlines are correct detections by our algorithm, and red outlines are incorrect detections.*

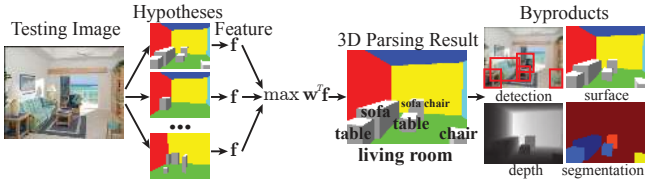

Figure 9: *Unified 3D scene parsing for basic level scene understanding. For each image, we generate a pool of possible hypotheses and encode image features and contextual rules for each hypothesis in a feature vector* $\mathbf{f}$ *(Figure 10). We choose the hypothesis that maximizes the objective function* $\mathbf{w}^T\mathbf{f}$ *as the result of 3D scene parsing. As by-products of our 3D parsing result we obtain information that have traditionally been considered in isolation, such as the 2D location of objects, their depth and 3D surface orientation.*

depicting a 3D scene. The output 3D parse result would be a 3D scene with objects annotated. This is depicted in Figure 9 for a living room scene. As by-products of our 3D parsing result we can obtain information that have traditionally been considered in isolation, such as the 2D bounding box of objects, their depth, and their 3D surface orientation.

The key idea of the algorithm is to generate a pool of possible outputs as hypotheses and pick the best one. We define a list of parsing rules and use structural SVM [Joachims et al. 2009] to learn the relative importance of these rules from the training data. For each hypothesis, we extract a vector about how well each rule is satisfied by this hypothesis, and score the hypothesis based on the importance of the rules. More specifically, given an image $\mathbf{x}$, we aim to predict a structured representation $\mathbf{y}$ for the 3D parsing result using a linear prediction rule: $\arg\max_{\mathbf{y}\in\mathcal{Y}}\mathbf{w}^T\mathbf{f}(\mathbf{x},\mathbf{y})$, where $\mathcal{Y}$ is the hypothesis space of all possible 3D parsing results for $\mathbf{x}$. The label $\mathbf{y}$ is a data structure of the 3D scene parsing result, which includes the scene category, camera parameters, space boundaries, and objects [1]. We encode image evidence and contextual constraints into the feature vector $\mathbf{f}(\mathbf{x},\mathbf{y})$. Therefore, a good scene parsing result

[1]The objects in the scene are assumed to be grounded on the floor or stacked on another object. Each object is represented as an object category, a 3D bounding box, including its center location, size, and yaw angle, with the assumption that the vertical axis of the bounding box is parallel with the gravity direction.

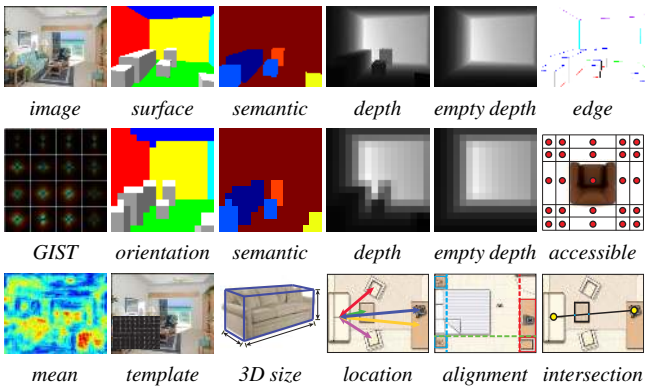

Figure 10: *Illustration of various rules we design to describe both the image evidence and context compatibility, All these rules are encoded in the structural SVM feature function* $\mathbf{f}(\mathbf{x}, \mathbf{y})$.

$\mathbf{y}$ not only explains the image evidence well, but also satisfies the contextual constraints. The parsing rules are illustrated in Figure 10 (for their definitions, please refer to the supplementary materials).

To learn the weights $\mathbf{w}$ from a set of input-output pairs $((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N))$ obtained from manual annotation, we seek to optimize the following convex problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n, \tag{1}$$

such that $\mathbf{w}^T\mathbf{f}(\mathbf{x}_n, \mathbf{y}_n) - \mathbf{w}^T\mathbf{f}(\mathbf{x}_n, \hat{\mathbf{y}}) \geq \triangle(\mathbf{y}_n, \hat{\mathbf{y}}) - \xi_n$, for all n and all possible output structures $\hat{\mathbf{y}} \in \mathcal{Y}_n$ in the hypothesis space. $\triangle(\mathbf{y}_n, \hat{\mathbf{y}})$ is the loss function controlling the margin between the correct label $\mathbf{y}_n$ and the prediction $\hat{\mathbf{y}}$. One of the major differences between a structural SVM and a standard SVM is that $\mathbf{f}(\mathbf{x}, \mathbf{y})$ depends not only on $\mathbf{x}$, but also on $\mathbf{y}$. This allows us to encode both the image evidence and contextual relations in a uniform manner. Moreover, the SVM discriminatively learns the relative importance of features and relations based on training data.

In Figure 11, we see that our algorithm produces encouraging results. In the first column, although our model fails to recognize the drawers, it manages to recognize a chair that is not labeled in the ground truth due to labeling error. In column 5, row 3 of Figure 11 we see that our model mistakes the white carpet in front of the sofa as a table, which is a typical configuration for a living room.

5 Conclusion

We have proposed to define basic level scene understanding as a tractable research goal, and have summarized our recent effort to probe the state of the art of several domains and questions related to visual scene understanding. Current and future investigations are concerned with applications of the work to domains such as image-based modeling [Xiao and Furukawa 2012; Xiao et al. 2008; Xiao et al. 2009; Xiao and Quan 2009], viewpoint extrapolation [Xiao et al. 2012a], and assessment of subjective visual scene properties [Khosla* et al. 2012a; Khosla et al. 2012b; Isola et al. 2011].

Acknowledgements

This work is funded by NSF grant (1016862) to A.O, Google research awards to A.O and A.T, ONR MURI N000141010933 and NSF Career Award (0747120) to A.T. J.X. is supported by Google U.S./Canada Ph.D. Fellowship in Computer Vision.

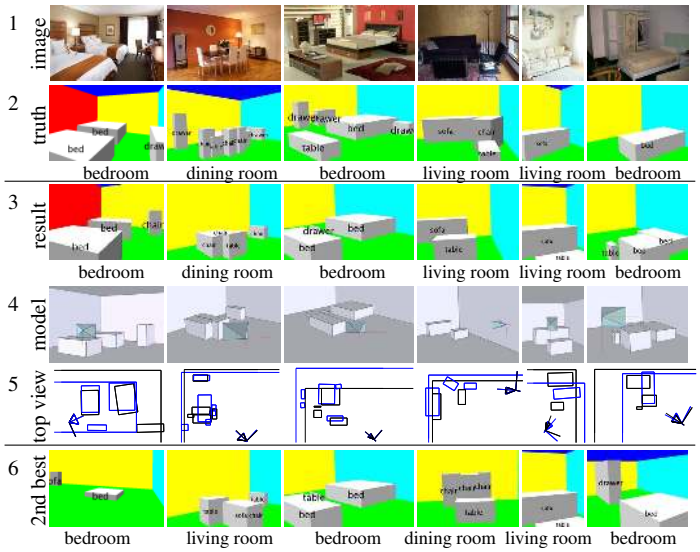

Figure 11: *Example results of our algorithm. The 1st row contains the input test image. The 2nd row contains the ground truth. The 3rd row contains a 2D rendering of our 3D parsing results. The 4th row is the rendering of the 3D mesh model and the camera frustum from a different viewpoint. The 5th row is the top view of the 3D parsing result: the* blue *lines are the ground* truth*, and the* black *lines are the* result*. The last row is the 2nd best prediction, which gives us an alternative parsing result.*

References

BARRIUSO, A., AND TORRALBA, A., 2012. Notes on image annotation.

EHINGER, K., XIAO, J., TORRALBA, A., AND OLIVA, A. 2011. Estimating scene typicality from human ratings and image features. In *CogSci*.

FELLBAUM, C. 1998. *Wordnet: An Electronic Lexical Database*. Bradford Books.

ISOLA, P., XIAO, J., TORRALBA, A., AND OLIVA, A. 2011. What makes an image memorable? In *CVPR*.

JOACHIMS, T., FINLEY, T., AND YU, C.-N. J. 2009. Cutting-plane training of structural svms. *Machine Learning*.

KHOSLA*, A., XIAO*, J., ISOLA, P., TORRALBA, A., AND OLIVA, A. 2012. Image memorability and visual inception. In *SIGGRAPH Asia*.

KHOSLA, A., XIAO, J., TORRALBA, A., AND OLIVA, A. 2012. Memorability of image regions. In *NIPS*.

PATTERSON, G., AND HAYS, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*.

ROSCH, E. 1978. *Principles of Categorization*. John Wiley & Sons Inc, 27–48.

SATKIN, S., LIN, J., AND HEBERT, M. 2012. Data-driven scene understanding from 3D models. In *BMVC*.

XIAO, J., AND FURUKAWA, Y. 2012. Reconstructing the world's museums. In *ECCV*.

XIAO, J., AND QUAN, L. 2009. Multiple view semantic segmentation for street view images. In *ICCV*.

XIAO, J., FANG, T., TAN, P., ZHAO, P., OFEK, E., AND QUAN, L. 2008. Image-based façade modeling. In *SIGGRAPH Asia*.

XIAO, J., FANG, T., ZHAO, P., LHUILLIER, M., AND QUAN, L. 2009. Image-based street-side city modeling. In *SIGGRAPH Asia*.

XIAO, J., HAYS, J., EHINGER, K., OLIVA, A., AND TORRALBA, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

XIAO, J., EHINGER, K., OLIVA, A., AND TORRALBA, A. 2012. Recognizing scene viewpoint using panoramic place representation. In *CVPR*.

XIAO, J., RUSSELL, B. C., AND TORRALBA, A. 2012. Localizing 3D cuboids in single-view images. In *NIPS*.