Introduction to a Large-Scale General Purpose Ground Truth Database: Methodology, Annotation Tool and Benchmarks

Benjamin Yao, Xiong Yang, and Song-Chun Zhu

Lotus Hill Institute of Computer Vision and Information Sciences
EZhou City, HuBei Province, P.R. China
{st.zyyao,xyang.lhi}@gmail.com,
sczhu@stat.ucla.edu
http://www.lotushill.org

Abstract. This paper presents a large scale general purpose image database with human annotated ground truth. Firstly, an all-in-all labeling framework is proposed to group visual knowledge of three levels: scene level (global geometric description), object level (segmentation, sketch representation, hierarchical decomposition), and low-mid level (2.1D layered representation, object boundary attributes, curve completion, etc.). Much of this data has not appeared in previous databases. In addition, *And-Or Graph* is used to organize visual elements to facilitate top-down labeling. An annotation tool is developed to realize and integrate all tasks. With this tool, we've been able to create a database consisting of more than 636,748 annotated images and video frames. Lastly, the data is organized into 13 common subsets to serve as benchmarks for diverse evaluation endeavors.

Keywords: Ground truth Annotation, Image database, Benchmark, Sketch representation, Top-down/Bottom-up Labeling.

1 Introduction

The importance of having an image database containing *ground truth* annotations parsed by humans for a wide variety of images is widely recognized by the machine vision community. The goal of our project is to build up a publicly accessible annotated image database with over 1,000,000 of images and more than 200 categories of objects. Because manual annotation of millions of images is too time-consuming a task for every vision lab to do independently, we hope to compile this centralized database to serve the community's diverse training and evaluation endeavors.

The challenges are many fold, however. First, there is no standard or handy tools for general purpose annotation. We need to find answers to questions like "what to label" and "how to represent common visual knowledge", so that we can develop a suitable labeling tool. Secondly, with the potential scale of the database

A.L. Yuille et al. (Eds.): EMMCVPR 2007, LNCS 4679, pp. 169–183, 2007.

in mind, we would like to devise top-down algorithms to guide and speed up annotation, yet it is a non-trivial task to organize and abstract visual knowledge from labeled images for this purpose. In addition, to make the database general enough to be used for different evaluation tasks, we need to build up benchmarks for a variety of visual patterns, thus we need to define equivalent distances over different spaces.

In this paper, we present our efforts in confronting these challenges, and show examples of data from our database. By consulting with several vision groups, we have gathered a consensus on the commonly desired information for labeling in three levels:

- *Scene Level*: Global geometry information, scene category (indoor/outdoor), events and activities;
- *Object Level*: Hierarchical decomposition, object segmentation, sketching and semantic annotation;
- *Low-middle Level*: Contours types (object boundary, surface norm change or albedo change), Amodal completions, Layered representation (2.1-D), etc.

According to the requirements listed above, we developed a novel annotation tool, which integrates several functional modules designed for specific task(s). We show that by properly combining these functions, the tool can perform customized annotation tasks blending all kinds of information. Moreover, the tool is associated with an *And-Or Graph knowledgebase*[4], which organizes and summarizes labeled visual knowledge in a universal way.

To the best of our knowledge, there has not been much previous work on building a large scale general purpose database. However, there are many special purpose databases publicly available, which provide us with some valuable insight. Here we only list those most related to our database:

- **LabelMe** database of MIT-CSAIL [Torralba et al.[11]]**.** This is the most similar dataset to ours. Images in this database are of natural images and cluttered scenes and contain objects under multiple views/poses. Its current limitation is that only the rough boundary of the object is annotated, as opposed to fine segmentation or hierarchical decomposition.
- **CalTech 101 and 256** [Fei-Fei, Griffinet et al.[6,7]]**.** These two datasets provide a great number of diverse object classes. Its limitations are that the objects are not positioned in real scenes, are centered in the image, and have a limited number of viewpoints.
- **The Berkeley Segmentation Dataset** [Martin and Fowlkes[9]]**.** This dataset is a pioneer effort on large scale ground truth annotation on general natural images. It makes a valuable contribution to error control and benchmark, but it is limited in regards to scale and content.
- **UA (Arizona) localized semantics dataset** [Barnard et al.[1]]**.** This dataset provides a good semantic annotation standard based on the data of Berkeley dataset.

2 Methodology: Representation and Organization of Labeling Information

2.1 Region Segmentation and Semantic Annotation

Segmentation is the foundation of image annotation. Common annotation task requires *object level* segmentation. For example, when we see the *water* in Figure 1, we tend to interpret it as a single *thing*, even though it is actually composed of several disconnected image areas. Therefore, in our representation framework, there are two levels of data. At the lower level, we define a *region* as an image area with closed boundary. At the higher level, we define an entity represent for object named *PO* (as short for "Physical Object"). A *PO* can represent any meaningful entity in an image, such as a scene, an object, an object part, a texture area or a text block. In Figure 1, all the regions of the road are aggregated into a single *PO* (masked with a same color). *PO* is a core element in our framework. It is composed of several lower level elements including regions and sketch graph (section 2.2). It also corresponds to a node in the parsing graph (section 2.3).

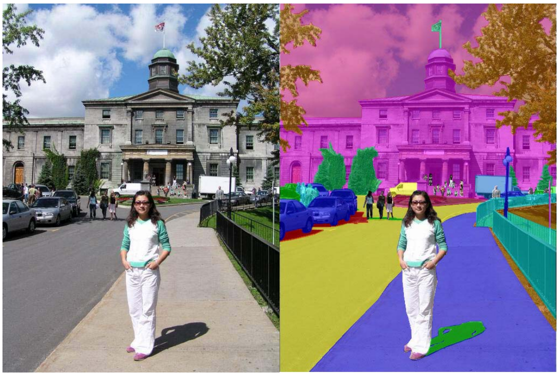

Fig. 1. This figure shows the result of object level segmentation. Sub-figure on the left is the original image. Sub-figure on the right is the segmentation mask. The *region* of an object (e.g. the road surface) may be composed of several disconnected image areas.

It is also worth mentioning that our annotation tool enables users to do "fine" segmentation and thus can output accurate object boundary, in comparison to coarse outlines provided by other datasets. This is especially important when the objects are small and in great numbers. A typical example is the annotation task of aerial images. As shown in Figure 2, the average number of *PO*s in aerial images is over one hundred. Such task can hardly be accomplished without fine segmentation.

Furthermore, to make the naming convention general and accurate, we use WordNet [10] as the reference. The following rules are adopted when a semantic

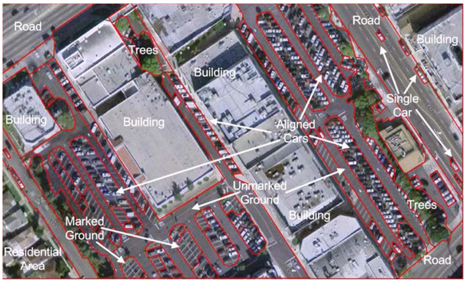

Fig. 2. This figure illustrates the segmentation and annotation results of a typical aerial image (a parking area)

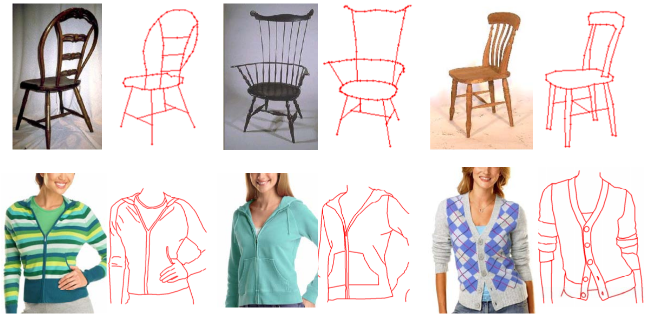

Fig. 3. Sketch graph representation of object: The sketch graph can capture most essential perceptual information of an object, like the structural information of chair or the folds, sewing lines, albedo/lighting changes of cloth

name is chosen for a *PO*: (1) Words of object name correspond to their WordNet definition. (2) The sense in WordNet (if multiple) should be mentioned as word[*i*], where *i* is the sense number in WordNet. (3) Synonym or description of same object in a different way should be given as additional entry(s) (e.g. *grass, ground*). (4) For parts of an object, add the object name as prefix (e.g. *horse:head*). (5) Descriptive words can be added to provide further information(e.g. *[white]human*, *[frontal view]car*).

2.2 Sketch Graph Representation

Since segmentation only provides the outer boundary of an object, we adopt a sketch graph representation like the Primal Sketch model [8] to record structural features inside the object boundary. The sketch is composed of a set of strokes (long curves) aligned through landmarks. It is not required to have a closed boundary (see the small figures on the right side of each object in Figure 3). This figure demonstrates that the sketch graphs can capture essential perceptual information of the object, such as the structural information of the chair and the folds, sewing lines, albedo/lighting changes of cloth. We believe that it is valuable to record sketch information in addition to the object boundary. Another example is showed in Figure 6.

Sketch graph representation is further augmented to include low-middle level vision elements by adding attributes to each curve (*attributed curves*) . As illustrated in Figure 4, the curves of object are classified by different colors to represent for *occlusion*, *surface normal change*, *lighting/albedo change* respectively. Besides, model/amodel completions and 2.1D layered representation are clearly defined with attributed curves (illusory contours, Figure 5).

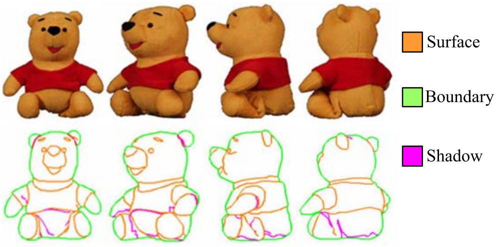

Fig. 4. Attributed curves. "Winnie Pooh" is labeled with three types of curve attributes: *Surface*: curves generated by surface norm change; *Boundary*: curves on object boundary; *Shadow*: curves generated by shadow.

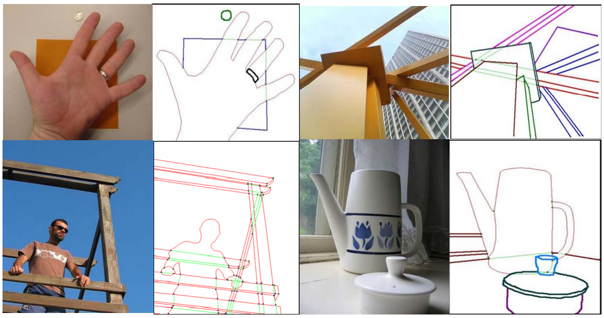

Fig. 5. Illusory contours for model/amodel completion and 2.1D layered representation. This figure shows how to label 2.1D layered representation and curve completions in a 2D image. The green line stand for the illusory(occluded) contours.

Figure 6 uses a high resolution sketch to do artistic rendering. We can see that the sketch graph image on the right welly depicts the appearance, clothing, and even expression of the little girl in the left image, with little distortion. Therefore, with such kind of *sketch*, people can easily do image rendering by filling in color patterns with all kinds of artistic styles. Moreover, the sketch graph is also a flexible and expressive tool to localize image features. As shown in Figure 11, we use the landmarks of sketch graph to represent the body pose and limb directions of human.

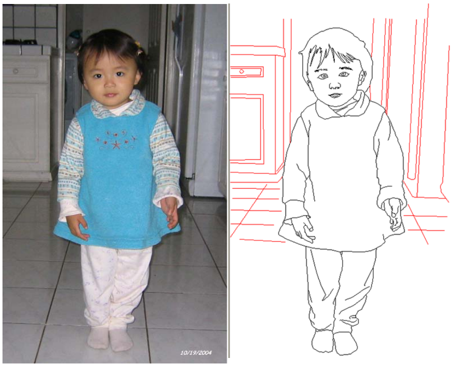

Fig. 6. Artistic rendering. *Sketch* captures the appearance, clothing, and expression of the child in the image with little distortion. Based on this, people can easily do image rendering with all kinds of artistic styles.

2.3 Hierarchical Decomposition and Parsing Graph

Compositionality is a common phenomenon in natural images. As shown in Figure 7(a), the image is composed of objects (car and background), while the car is composed of several parts and, if the resolution of the image were higher, we could further decompose the car body into sub-parts. Therefore it is natural to organize the contents of image in a hierarchical style. In our annotation framework, we adopt a tree structure (parsing tree) to record the hierarchical decomposition of objects (similar to the *Image Parsing* concept of Zhu et al. [12]). The image is decomposed hierarchically from scene to object then to parts and so on. This process terminates when the resolution of leaf nodes is too low to decompose further. Nodes of the parsing tree are the *PO*s mentioned in 2.1.

By adding horizontal connections between parsing tree nodes, we further augment the parsing tree into a parsing graph. The horizontal links between nodes represent the relationship between object/parts. For example, the dashed lines in Figure 7(b) stand for supporting and occluding relationships between objects in the image.

Another example of a parsing graph is shown in Figure 8. In this figure, the bike is labeled at two different scales. At the low resolution, the bike is labeled as a whole. At the high resolution, it is further decomposed into parts with details. This process represents the scaling phenomena in vision.

2.4 And-Or Graph Knowledgebase and Bottom-Up/Top-Down Labeling Procedure

To build up an annotated image database with millions of images, it is important to abstract and organize visual knowledge from labeled images. It is

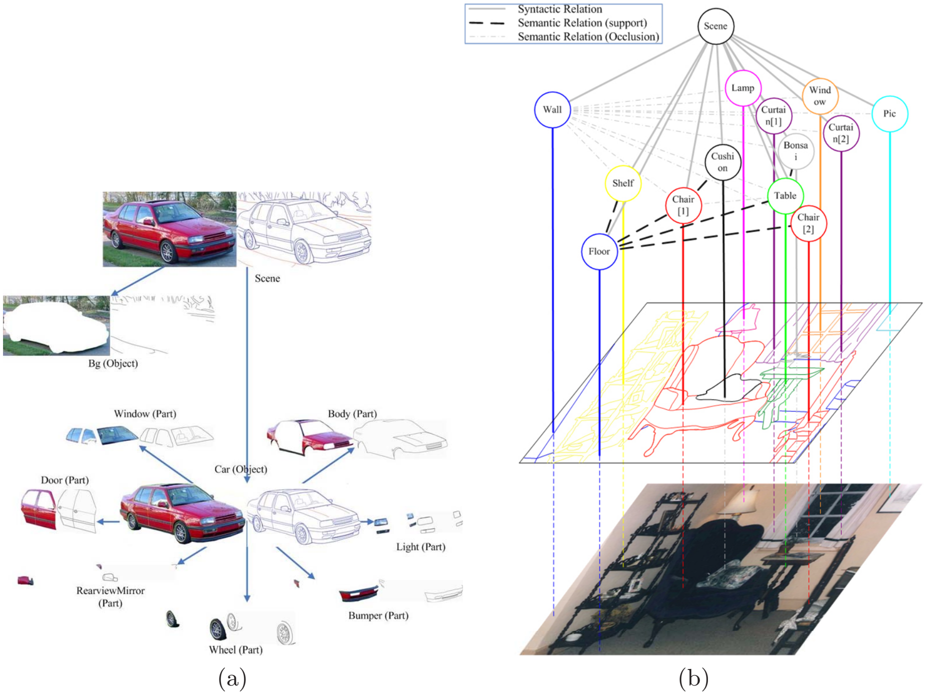

Fig. 7. (a) A parsing tree: The image is hierarchically decomposed from scene to object then to parts like a tree. The node of the parsing tree is *PO* (section 2.1), which includes both region and sketch graph. (b) A parsing graph. Solid lines represent hierarchical decomposition. Dashed lines represent spatial relationships (supporting and occluding in this figure).

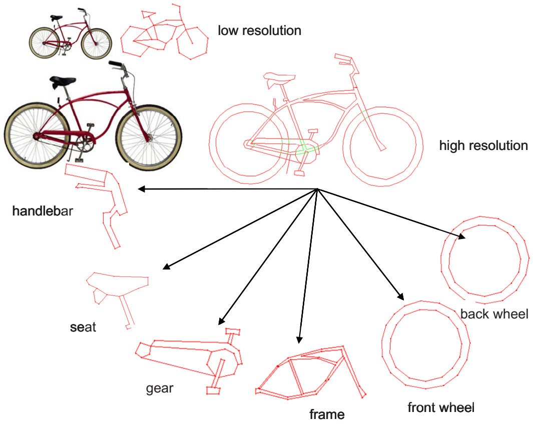

Fig. 8. Multiple scale/resolution annotation: The bike in the figure is labeled at two different scales. At low resolution, the bike is represented by a single sketch graph. At high resolution, it is further decomposed into parts.

also desirable to use previously extracted information from labeled images to train top-down algorithms to guide and speed up later annotations. To this end, we adopt the And-Or Graph concept brought up by Chen et al. [4] to organize labeling information in our database. The And-Or Graph is a uniform representation of visual knowledge, which combines Stochastic Context Free Grammar (hierarchical decomposition) and Markov Random Field (horizontal relationship). As mentioned previously, *Parsing Graph* integrates segmentation and sketch graph in a hierarchial structure with spatial relationships. Therefore, each parsing graph can be regarded as an instance of the And-Or graph in a real image. The And-Or Graph is an abstraction of all labeled information and hence can be compiled into an *And-Or Graph knowledgebase*. The indexing information used by the *And-Or Graph knowledgebase* is sketch graph representation of POs, which can be thought as prototypes or templates.

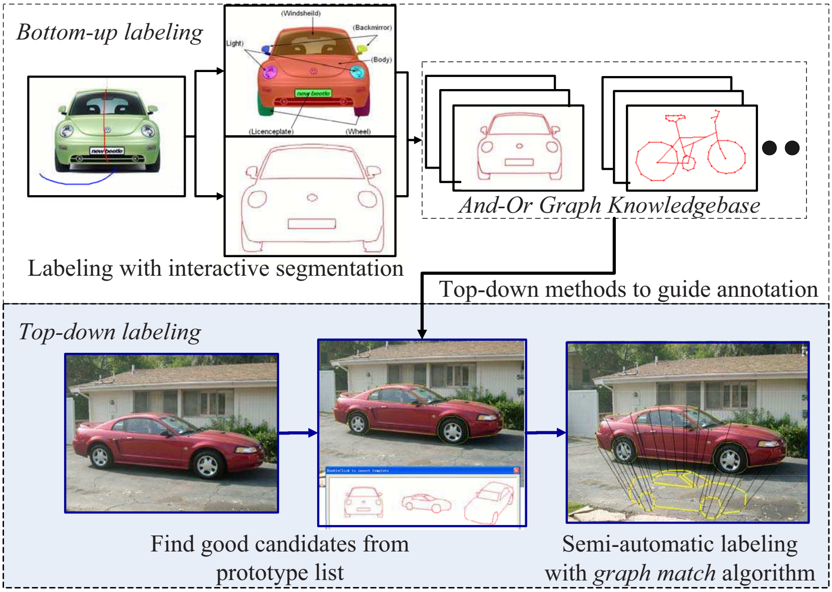
Bottom-up labeling
(Windsheild)
(Backmirror)
Light)
(Body)
(Licenceplate)
(Wheel)
And-Or Graph Knowledgebase
Labeling with interactive segmentation
Top-down methods to guide annotation
Top-down labeling
Find good candidates from prototype list
Semi-automatic labeling with *graph match* algorithm

Fig. 9. Bottom-up/top-down labeling process: In the bottom-up process, labeled information of objects is summarized into the And-Or graph knowledge base and stored as prototypes of different object categories. In the top-down labeling process, these templates are utilized to guide the annotation process.

As shown in Figure 9, we devise a bottom-up/top-down labeling procedure with the *And-Or Graph Knowledgebase*. As shown in the upper part of Figure 9, when a new category or novel instance is input, the object is labeled manually or with interactive method such as GraphCut[3]. The graph representation is then stored into the *And-Or Graph Knowledgebase* as templates. This process is called bottom-up labeling. When there are sufficient templates recorded to cover the inter-class variety of an object category, the templates can be utilized in a downward direction. First, good candidates are automatically selected from the template pool in the *And-Or Graph Knowledgebase*. After the best template is selected (manually or automatically), match algorithm based on sketch graph representation such as *Shape Context*[2] or *Graph Match*[1] is used to fit the template onto object. Thus, the labeling procedure is speeded up dramatically. The top-down labeling process[2] is shown in the bottom part of Figure 9. Through this bottom-up/top-down labeling procedure, visual knowledge is accumulated towards the final goal of automated image parsing and labeling.

2.5 Global 3D Geometry Information

Global 3D geometric information is very important for scene understanding. Our annotation tool provides a module to label Global coordinate frame and perspective projection parameters in 2D image. As shown in Figure 10, the pairs of yellow lines in the figure are perspective parallel lines in real world. *Vanishing points* can be computed by a group of parallel lines. The *horizon line* is easily derived by connecting two *vanishing points*. The *ground plane* can be created using two pairs of parallel lines, which is same for the vertical planes.

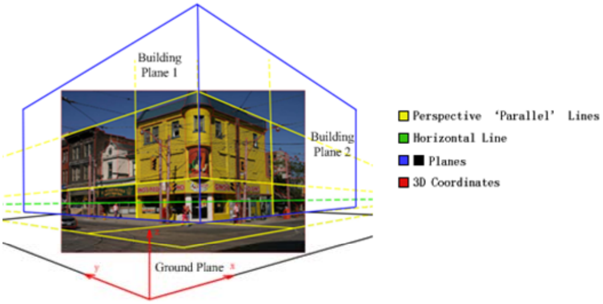

Fig. 10. This figure illustrate the labeling results of global geometry information. Yellow lines are *Perspective 'Parallel' Lines*. *Vanishing Points* can be computed by the intersection of parallel lines. Green line is *Horizon Line*. Black frame is ground plane, blue frame stands for vertical planes. Red axis is global coordinate.

3 Annotation Tool: Integrating Functional Modules

In previous section, we present the representation and organization methodology of labeling information with examples from our database. In our annotation tool, we realize the functions with seven modules:

---

[1] Discussed in a companion paper in CVPR07.

[2] The automatic algorithms exploited in top-down labeling process is detailed discussed in a another companion paper submitted to ICCV07.

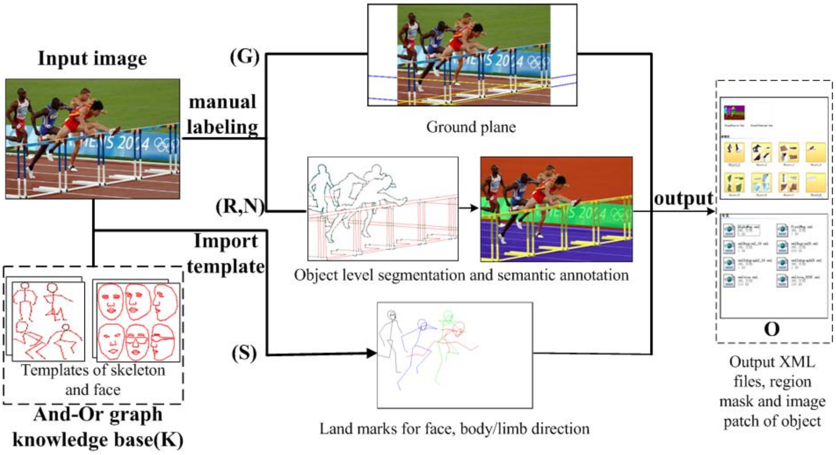

Fig. 11. Integration of functional modules. The task is to label images for sports activities. Four aspects of information are required: 1) ground plane and horizon line 2) segmentation of objects and semantic labels 3) layer labels for foreground and background 4) body direction and faces of athletes. To finish this task we integrate three modules: 1) geometry and 3D scene label module(G) for marking out ground plane on the image 2) region segmentation and annotation module(R) for labeling segmentation of objects, semantic and layer labels 3) And-Or Graph knowledge base assistant label module(S) for marking the body direction and human faces. At last, we output the labeling results.

1. Region segmentation *(R)*.
2. Sketching and graph representation *(S)*.
3. Hierarchical decomposition using *Parsing Graph* representation *(P)*.
4. Semantic annotation *(N)*.
5. Attributed Curves *(A)*.
6. Geometry and 3D information labeling *(G)*.
7. *And-Or Graph Knowledgebase (K)*.

In this section, we study the case of a practical image annotation task to illustrate how we integrate several functional modules to perform a specific annotation task. The task is to create an annotation subset for sports activities. It requires visual knowledge of following aspects:

1. Ground plane and horizon line of image
2. Segmentation of objects and semantic labels
3. Layer labels for foreground and background
4. Body direction and faces of athletes.

It is obvious that requirement (1) is a global coordinate frame and perspective projection parameters labeling issue; Requirement (2) and (3) can be reduced to region segmentation and semantic annotation. Since there are many templates of face and body skeleton in our *And-Or Graph knowledgebase*, we can fulfil requirement (4) by a top-down procedure introduced in section 2.4.

As shown in Figure 11, we integrate five modules to finish this task. First, we use the 3D Geometry module to label ground plane and horizon line on the image(G). Second, we use the segmentation and annotation to perform object level segmentation (R,N). Third, we use templates of human body skeleton and faces from And-Or Graph knowledge base to mark the body direction and faces(S,K). Integrating these five modules, we derive a customized annotation procedure.

4 Database Statistics, Subsets and Benchmarks

There are 3,927,130 *PO*s, 636,748 images and video frames in our database at present, and the number is growing everyday. As illustrated in Figure 14, widespread images have been annotated.

To serve the community's dire needs for *dataset* for training and evaluation, we have organized 13 common subsets from our database to serve as benchmarks. Figure 12 illustrate the typical image collection of these subsets. Table 1 illustrates more detailed statistics of these subsets, including image number, class number, visual knowledge included (functional modules involved), etc. These subsets are:

1. *Common scene classification:* A subset for general scene classification (Row1 in Figure 12, Row1 in Table 1).

   classes: Images are categorized into 14 classes including: *bathroom, bedroom, cityview, corridor, hall, harbor, highway, kitchen, living room, office, parking, rural, seashore, street.*

   labeling information: 3D Geometric description of scene, Object-level Segmentation (objects included in a scene, such as sky, tree, pedestrians, cars), Semantic Annotations, Parsing graph is used to perform scene decomposition and record occluding relation between objects.
2. *Activity and Events classification:* A subset for activity and event classification (Row2 in Figure 12, Row2 in Table 1).

   classes: Images are categorized into 16 classes including: *dinning, harvest, lecture, meeting, shopping, badminton, bocce, croquet, high jump, hurdles, iceskate, polo, rowing, snowboarding, RockClimbing, sailing.*

   labeling information: Similar with the common scene classification subset. Since judgement of events and activities highly related with human in the scene, special annotations of human body are added (for specific, the face, body and limb directions, the case in Section 3 is an example).
3. *Aerial images:* A subset for aerial image segmentation (Row3 in Figure 12, Row3 in Table 1).

   classes: Images are grouped into 10 classes: *airport, business, harbor, industry, intersection, mountain, orchard, parking, residential, school.*

   labeling information: Segmentation of main objects in a scene, such as building roof, parking area, single car and road surface.

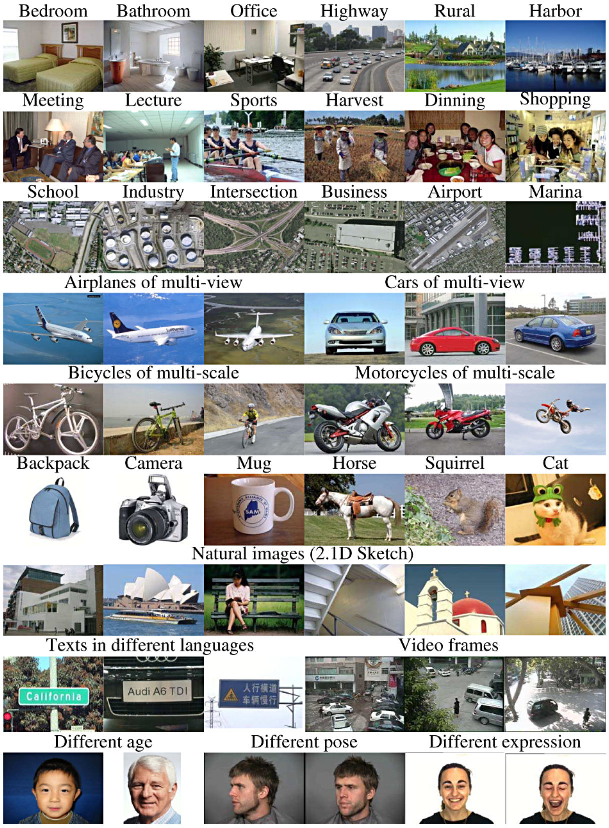

Fig. 12. Exemplary images of 13 subsets for benchmarks. Subset1: scene classification (Row 1); Subset2: events and activity (Row 2); Subset3: aerial images (Row 3); Subset4-6: 20 categories of popular objects, which are in multiple views, scales and resolutions (Row 4-5); Subset7-9: Generic objects of 200 categories (Row 6); Subset10: Human faces with different age, pose, expression and etc. (Row 9); Subset11: Surveillance video frames (Row 8); Subset12: Text (Row 8); Subset13: Natural images for 2.1D layered representation.

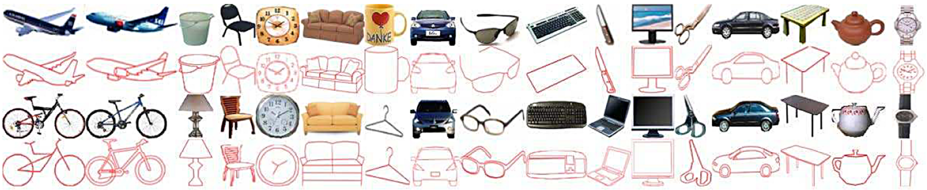

Fig. 13. Popular object subsets: 20 categories of common objects are selected and labeled with multiple views, scales and resolution. The first and third lines of this figure are original image patches. The second and fourth lines of this figure are sketch representation of objects.

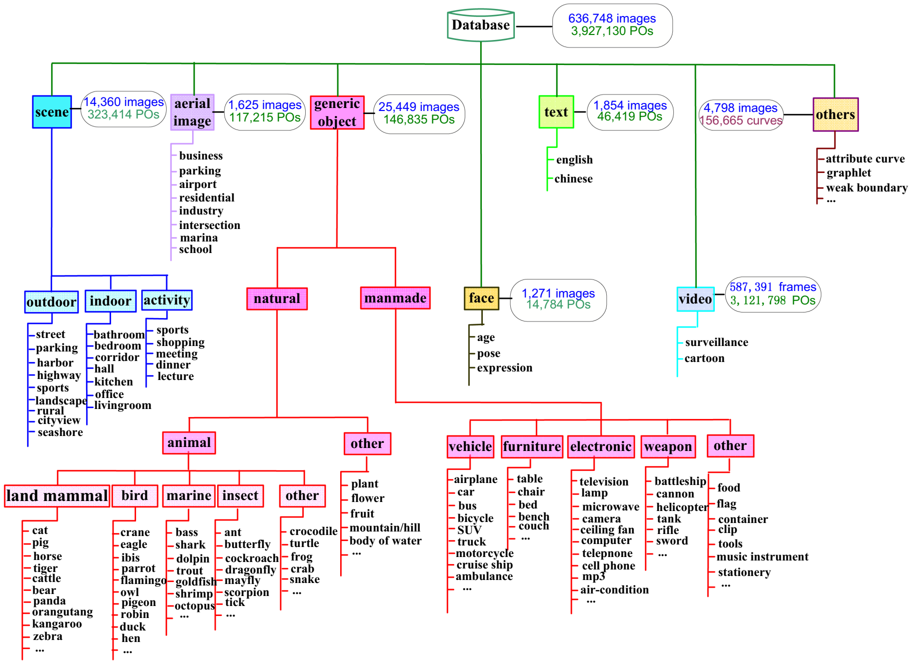

Fig. 14. This tree list is a comprehensive inventory of our dataset. From root node to leaf nodes, the entire set is decomposed into subsets and categories hierarchically. Terminal nodes without boxes are corresponding to the most detailed categories. The numbers in arc angle boxes are the statistics relatively. The *PO* in the figure means Physical Object mentioned in section 2.1.

4. *Popular objects:* Three subsets for object categorization, object bounding box localization and object outline detection. These objects are putted in multiple views and resolutions. (Row 4,5 in Figure 12, Row 4-6 in Table 1).

   classes: airplane, bicycle, bucket, chair, clock, couch, cup, frontcar, glasses, hanger, keyboard, knife, lamp, laptop, monitor, motorcycle, sidecar, table, teapot, watch.

   labeling information: Both segmentation and sketch representation of object are labeled. Objects are labeled under two or three resolutions. At the high resolution, object is decomposed hierarchically into parts and sub-parts.

   Figure 13 shows the segmentation and sketch representation of both high and low resolution of this subset.

5. *Generic object:* Three subsets for object categorization, bounding box localization and outline detection. (Row 6 in Figure 12, Row 7-9 in Table 1).

   labeling information: Similar with popular objects, except that only one resolution and single view is labeled. Because many objects are rarely seen, thus it is very hard to collect enough images for different views and resolutions.

6. *Face:* A subset for human face categorization. (Row 9 in Figure 12, Row 10 in Table 1).

   class: Four classes differ in facial expression, lighting condition, age and pose.

   labeling information: Landmarks of sketch graph are used to record the feature points of human faces. These landmarks are compatible with the Active Appearance Model (AAM)[5].
7. *Video clips:* A subset for video surveillance task. (Row 8 in Figure 12, Row 11 in Table 1).

   labeling information: Both segmentation and sketch representation for foreground objects. Segmentation for background areas.
8. *Text:* A subset for text recognition tasks. Two kinds of languages are included: English and Chinese (Row 8 in Figure 12, Row 12 in Table 1).

   labeling information: Segmentation of letter(character), hierarchical decomposition from *text block* to *lines* to *words* until *letters or characters*.
9. *2.1D layered representation:* A subset with natural images for general 2.1D segmentation tasks (Row 7 in Figure 12, Row 13 in Table 1).

   labeling information: Segmentation and sketch representation, occluding relation between objects are recorded.

Table 1. Detailed Statistics of Subsets, functional module abbreviations see sec. 3

Subsets		*Class Num*	Functional Modules							*Image Num*
			R	*S*	*P*	*K*	*N*	*A*	*G*	
Scene Classification	Common scene	14	∨		∨		∨		∨	9637
	Activity	16	∨	∨	∨	∨	∨		∨	4723
	Aerial	10	∨		∨	∨	∨			1625
Popular Object	Categorization	20					∨			9585
	Bounding Box	20	∨				∨			9585
	Outline	20	∨	∨	∨	∨	∨	∨		9585
Generic Object	Categorization	200					∨			15864
	Bounding Box	200	∨				∨			15864
	Outline	200	∨	∨		∨	∨	∨		15864
Face		4	∨	∨	∨	∨	∨			1271
Video		1		∨	∨		∨			587391
Text		2	∨	∨	∨	∨	∨			1854
2.1D Sketch		1		∨				∨		1446

5 Conclusions and Future Works

In this paper, we present a new large-scale general purpose ground truth image database. We bring up the representation and organization methodology of generally desired labeling information. We also demonstrate that, by properly combining the functional modules of our annotation tool, one can perform annotation tasks blending all kinds of desired information. Besides, a bottom-up/top-down labeling framework is proposed using the And-Or Graph knowledgebase to speed up labeling process. Lastly, thirteen subsets of labeled data are organized to serve as standard *Benchmarks*. Further investigations are needed on the automatic algorithms related with bottom-up/top-down labeling procedure to realize the long term goal of semi-automatic labeling and automatic labeling. Besides, to set up benchmarks for image understanding (rather than simple classification), further investigation are needed on defining equivalent distance over diverse visual spaces.

Acknowledgements. The project is supported by: National 863 project, Contact No. 2006AA01Z121; National Natural Science Foundation of China, Contact No. 60672162; National Natural Science Foundation of China, Contact No. 60673198; National 863 project, contact No. 2006AA01Z339.

References

1. Barnard, K., Fan, Q., et al.: Evaluation of localized semantics: Data, methodology, and experiments. University of Arizona, Computing Science, Technical Report,TR-05-08. (September 2005)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Recognition and Machine Intelligence, 509–522 (April 2002)
3. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 1222–1239 (2001)
4. Chen, H., Xu, Z.J., Zhu, S.: Composite templates for cloth modeling and sketching. In: CVPR'2006, pp. 943–950 (2006)
5. Cootes, T.F., Taylor, C.J.: Active appearance models. In: ECCV'1998 (1998)
6. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Recognition and Machine Intelligence, pp. 594–611 (April 2006)
7. Griffin, G., Holub, A., Perona, P.: The caltech 256. Caltech Technical Report
8. Guo, C., Zhu, S., Wu, Y.: Primal sketch: Integrating texture and structure. Computer Vision and Image Understanding (2006)
9. Martin, D., Fowlkes, C., et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV'2001, p. 416 (2001)
10. Miller, F.C., Tengi, R., Wakefield, P., et al.: Wordnet - a lexical database for english (1990)
11. Russel, B.C., Torralba, A., Murphy, K.P.: Labelme: a database and web-based tool for image annotation, M.I.T., C.S. and A.I. Lab Techinical Report, MIT-CSAIL-TR-2005-056 (September 2005)
12. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.-C.: Image parsing: Unifying segmentation, detection and recognition. Int'l. J. of Computer Vision, Marr Prize Issue (2005)