Abstract Scene Dataset v1.1

Last updated: May 1st, 2013

The contents of this file are described below. Any questions related to this dataset please direct to Larry Zitnick (larryz@microsoft.com) or Devi Parikh (parikh@vt.edu). If this dataset is used in any publications please reference:

C. L. Zitnick, and D. Parikh, Bringing Semantics Into Focus Using Visual Abstraction, In *CVPR*, 2013.

New for version 1.1:

Data associated with the following paper has been added to the dataset:

C. L. Zitnick, D. Parikh and L. Vanderwende, Learning the Visual Interpretation of Sentences, In *ICCV*, 2013.

This includes the 60k "simple" sentences, the tuples extracted from the sentences and the scene features. If this data is used in a paper please reference the *ICCV* paper above.

Contents 1.0:

RenderedScenes and RendereredSeedScenes: Directory containing the rendered scenes described in the paper referenced above. RenderedSeedScenes are the 1,002 original scenes used as seeds to generate 1,002 scene descriptions. RenderedScenes are the 10,020 scenes that are created by asking turkers to create scenes depicting the 1,002 written descriptions (10 scenes per description.) The scenes are named using "Scene<scene class>_<scene index>.png", where <scene class> has 1,002 values (0 to 1001) and <scene index> is the index over scenes belonging to the same scene class (ranges from 0 to 9 for the scenes in RenderedScenes.) All scenes generated from the same written description belong to the same class.

ClipArtScene.htm: Javascript code used by Amazon's Mechanical Turkers (AMT) to create scenes. The pngs used by the page are contained in the directory "Pngs." Please host the images on your own website, i.e. when running AMT experiments don't link to the images on Larry's website.

Sentences_1002.txt: Text file containing the 1,002 written scene descriptions (one per line.) For scenes from scene class *x*, their written description lies on line *x*+1 in the text file.

Scenes_10020.txt and SeedScenes_1002.txt: Text files containing the information needed to render the scenes in the directories RenderedScenes and RenderedSeedScenes. The files have the following format:

<Number of scenes>

<Scene Index> <Number of clip art pieces in scene>
<clip art name> <clip art type index> <clip art object index> <X position (pixels)> <Y position (pixels)> <Z position (0, 1, or 2)> <flip (0 = no flip, 1 = flip horizontally)>
…

The clip art type index has one of 8 values (type prefix):
0: sky object (s)
1: large objects (p)
2: boy in different poses and facial expressions (hb0)
3: girl in different poses and facial expressions (hb1)
4: animals (a)
5: clothing (hats and glasses) (c)
6: food (e)
7: toys (t)

The clip art object index has a different amount of values for each type. You may see the different types by looking at the clip art in the directory "Pngs". The image names have the format <type prefix>_<object index>s.png . Of specific interest are the objects indices for the boy and girl. They indicate the pose and expression of the children (35 possible values for combinations of 5 expressions and 7 poses.)

The <Z position> has three different values indicating the scale and depth of the clip art (0 is closer, 2 is further away.) The clip art is scaled by [1.0, 0.7, 0.49] for the three values [0, 1, 2]. When rendered, all objects at scale 2 are rendered before all objects at scale 1 and similarly for other combinations. The only exception is objects of the sky type are rendered first before all other objects, so for instance the sun can never be in front of a tree.

SceneRenderer.exe: Command line executable (Windows) for generating scenes using the following format:
SceneRenderer.exe <inputFile> <pngDirectory> <outputDirectory>

To generate the images in RenderedScenes, the following would be used:
SceneRenderer.exe Scenes_10020.txt Pngs RenderedScenes

Pngs: Directory containing the original clip art and background.

Word Features: Directory containing bag-of-words features for the written descriptions.

Words.txt: List of all words that occur at least 5 times in the written descriptions. The file has the following format <number of occurrences> <word>.

WordFeatures_355.txt: The bag-of-words features for each sentence. Each line indicates the words that appear in the written description. A value of 1 means the word appears and 0 otherwise. Each line indicates a new description. There are a total of 355 possible words, so the file contains 355x1002 values. The order of the words is the same as in Words.txt.

Visual Features: Directory containing the visual features for each scene. Each feature type has two files "*.txt" containing the feature values, and "*names.txt" containing descriptive names for each feature and one or two indices indicating on which object instances the feature is based (an illustration of the object indices can be viewed in ClipArtIndices.png.) The final number in the file's names indicates the dimensionality of the features. Please see the referenced paper for additional information on how the features are computed. The feature types are as follows:

10K_category_occurence_11: The occurrence of the 11 object categories.
10K_instance_occurence_58: The occurrence of the 58 object instances.
10K_category_co-occurrence_100_65: The co-occurrence of the object category pairs that occurred at least 100 times.
10K_instance_co-occurrence_100_377: The co-occurrence of the object instance pairs that occurred at least 100 times.
10K_category_Abs_GMM_44: The absolute location of the object categories.
10K_instance_Abs_GMM_232: The absolute location of the object instances.
10K_instance_Abs_depth_174: The absolute depth of the object instances.
10K_category_Rel_GMM_100_264: Relative location of object categories.
10K_instance_Rel_GMM_100_1508: Relative location of object instances.
10K_instance_Flip_Rel_GMM_100_3016: Relative location of object instances taking into account whether the reference object is facing left or right.
10K_instance_Rel_depth_100_1131: Relative depth of the object instances.
10K_instance_hand_116: Whether an object is near a child's hand.
10K_instance_head_116: Whether an object is near a child's head.
10K_person_24: Attributes of the children (pose and facial expression.)

Contents 1.1:

SimpleSentences: Additional Data associated with the paper "Learning the Visual Interpretation of Sentences", *ICCV* 2013.

SimpleSentences1_10020.txt: List of 30k sentences. There are 3 sentences per scene. Format is <Scene Index> <Sentence Index> <Sentence>.
SimpleSentences2_10020.txt: Second set of 30k sentences. There are 3 sentences per scene. Format is <Scene Index> <Sentence Index> <Sentence>.

SceneFeatures_v2_3124.txt: Set of scene features. Each scene has 58 x 3124 features corresponding to the 58 pieces of clipart and 3124 features per piece. Features for clipart not present in the scene are not stored (assume they are zero.) The format is:
<Number of scenes>
<Scene Index> <Number of clip art pieces in scene>
<clip art piece index 0> <3124 features>
…

SimpleSentences\Tuples: Tuple data extracted from the sentences.

TuplesText1_10020.txt and TuplesText2_10020.txt: List of extracted tuples. Format <Scene Index> <Sentence Index> <Primary object> <Relation> <Secondary object>.

TuplesIdx1_10020.txt and TuplesIdx2_10020.txt: List of extracted tuples. Format <Scene Index> <Sentence Index> <Primary Object Index> <Relation Index > <Secondary Object Index >.

ObjectList_10020.txt: List of objects. Format <Object Index> <Object Count> <Object Name>.

RelationList_10020.txt: List of relations. Format <Relation Index> < Relation Count> < Relation Name>.