Universität Hamburg

Technical Report FBI-HH-M-345/10

eTRIMS Scene Interpretation Datasets

Johannes Hartz
Patrick Koopmann
Arne Kreutzmann
Kasim Terzić

{hartz | koopmann | terzic}@
informatik.uni-hamburg.de

{mail}@arne-kreutzmann.de

November 15, 2010

Abstract

We describe two image datasets for learning and evaluating interpretations of man-made scenes. The datasets consist of (A) 110 and (B) 200 fully annotated images from the building façade domain. Dataset A emphasises prominent object structures, and has been manually double-checked. Dataset B puts more emphasis on single object recognition on a larger set of classes, and has not been double checked. For both datasets, we define an object partonomy as a set of related labelled polygons, which are provided in an XML format based on the MIT LabelMe database. The datasets can be used as ground truth for training and evaluating object detection and classification algorithms, structure detection and classification algorithms as well as for evaluating complete interpretations of structured scenes, which makes use of the partonomy and taxonomy of objects.

1 Introduction

In the field of Computer Vision, there is a growing interest in using high-level knowledge for interpreting scenes from a wide range of domains. This involves vision tasks which go beyond single-object detection to provide an explanation of the observed scene. These tasks include recognising structure and relationships between objects in the scene and inferring missing and occluded parts. Typically, scene interpretation attempts to explain all objects in a scene, and use contextual knowledge in terms of compositional hierarchy and spatial relations to improve classification of ambiguous detections. While there are a number of benchmarks available for evaluating object detection and categorisation, there are very few datasets catering specifically to object structures and scene interpretation.

The two datasets presented in this report consists of images from the façade domain. The main difference between them is that dataset A emphasises pictures with prominent object structures. The annotation of dataset A includes the *complete* tag, which denotes if an object structure has been completely visible in the image (to avoid distortion on the rectified images). It has been manually double checked, and features a total of 12 structure and 26 object classes. Dataset B focuses more on single object recognition tasks, with a total of 200 images and 30 different classes. It has not been double-checked.

The domain of building facades is not only relevant due to the increasing interest in 3D city modelling, landmark recognition and autonomous navigation, but also presents challenges in the fields of object and structure recognition:

- the appearance of many object classes is extremely variable in terms of shape, colour and texture,

- objects of different classes are often visually similar, e.g. *doors* and *windows*,

- structures like *window-array* and *upper-floor* are very similar in terms of their spatial arrangement

- instances of structures like *entrance* or *balcony* are very variable in the composition and arrangement of their parts

Since the domain of building façades is a highly structured one, higher-level knowledge and image structure can be used to improve the performance of detection and interpretation algorithms, thus measuring the contribution of top-down processing steps on the overall scene understanding.

2 Annotation

For each image a full annotation of objects and object structures is given, which together form a compositional hierarchy. All objects in the image are marked with a polygon and assigned a class label. Object structures are additionally described by a set description of their parts (i.e. the objects it is constituted of), on the next level of the composition. Transitive compositional relations are not annotated. The annotations are provided as bounding polygons which closely follow the objects' contours, but are not accurate at a pixel-level. The polygons are closed (the starting and ending points are the same). For the images in dataset A, the annotation features the *complete* tag, which denotes if an object structure is visible in the image completely, or if it is clipped off at the image border.

An XML format is used for representing the annotations. The format is an extension of the well-known MIT LabelMe [8] format, for which a number of tools are available, and which is easy to parse. The annotations can be viewed and modified using the Annotation Tool developed at University of Bonn [5]. The XML format is discussed in detail in Section 5.

All of the images used in the datasets are *rectified*, to make it easier to exploit the horizontal and vertical alignments specific to the domain.

The datasets consist of objects from 30 different classes, as shown in Table 4. The classes *Canopy, Car, Chimney, Door, Ground, Pavement, Person, Railing, Road, Sign, Sky, Stairs, Vegetation* and *Window* represent simple objects without parts, and are referred to as *primitives.* They are annotated as bounding polygons with a class label. The remaining classes represent formations of other objects, which may or may not have a typical appearance on their own, and are referred to as *aggregates.* For example, a *Balcony* can contain doors, windows and railings, a *Window Array* is a row of windows, and an *Upper Floor* is a horizontal alignment of windows and balconies above the ground level. Such aggregate classes are annotated using bounding polygons, a class label, and a list of their parts (expressed by IDs which are unique in the database). The objects are annotated to the extent that they are visible, excluding the occluded parts or parts that are located outside the image, with a few exceptions detailed below.

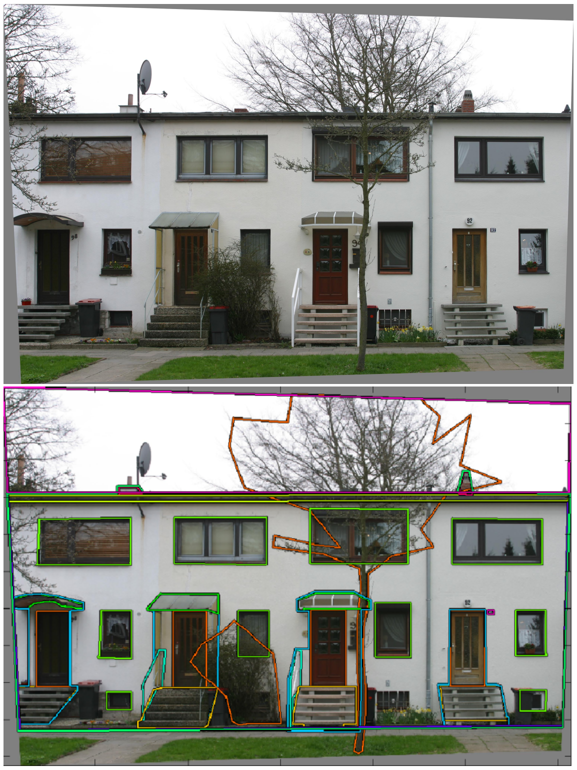
The datasets include a number of classes that are difficult to classify without knowledge about the scene, such as *Ground*, *Apartment Building* and *Office Building*. These classes are not always completely visible due to occlusion, but are semantically significant, and can be inferred from the primitives by a reasoning system. Higher-order aggregates like these are annotated completely, including the parts that are occluded by other objects (e.g. cars or vegetation). The reasoning behind this exception is that an interpretation system should be able to infer the location and position of a building and its bottom part even if it is not directly observed, and that the ground truth should contain this information. The class *Other Building* refers to buildings which do not fit into one of the other building models, such as garages or sheds.

3 Dataset A

Dataset A contains 110 images of buildings from various european cities including: Basel (Switzerland), Berlin (Germany), Bonn (Germany), Hamburg (Germany), Heidelberg (Germany), Karlsruhe (Germany), Munich (Germany), Prague (Czech Republic) and several cities from the United Kingdom. In these images 10 structure classes are annotated, as well as 16 primitive object classes. These structure classes feature 1632 object structures, of which 1154 are *complete*. A total of 5312 objects are annotated in the images. Some statistics for the annotations of images of Dataset A are presented in Tables 1 and 3.

Table 1: Structure class labels in Dataset A

`Entrance`	`Balcony`	`Dormer`	`Window-Array`
`Ground-Floor`	`Upper-Floor`	`Roof`	`Family-Home`
`Apartment-Building`	`Office-Building`		



```
single family home
  ...
single family home
  ...
single family home
  roof
    chimney
  facade
    upper floor
      window
    ground floor
      entrance
        door
        railing
        stairs
        canopy
        railing
    window
    window
single family home
  ...
vegetation
vegetation
sky
ground
```

Figure 1: Example image with the corresponding annotation shown below. The *has-part* relationships are shown on the right, with the partonomy indicated by indentation levels. Objects with the same scope and indentation level are siblings.

Table 2: All class labels in Dataset A

`Apartment-Building`	`Balcony`	`Canopy`	`Chimney`
`Cornice`	`Door`	`Dormer`	`Entrance`
`Gate`	`Ground`	`Ground-Floor`	`Office-Building`
`Pavement`	`Person`	`Railing`	`Road`
`Roof`	`Sign`	`Family-Home`	`Sky`
`Stairs`	`Transportation-Object`	`Upper-Floor`	`Vegetation`
`Window`	`Window-Array`		

Table 3: Relative frequencies of structure classes in dataset A

Entrance	0.11
Balcony	0.19
Window-Array-X	0.20
Dormer	0.01
Ground-Floor	0.12
Upper-Floor	0.20
Roof	0.04
Apartment-Building	0.08
Office-Building	0.01
Single-Family-Home	0.05

4 Dataset B

Dataset B contains 200 images of buildings from various european cities including: Basel (Switzerland), Berlin, Bonn, Hamburg, Heidelberg, Karlsruhe, Munich (Germany), Prague (Czech Republic) and some cities from the United Kingdom. There are 30 classes annotated, of which 10 are object structure classes. There are a total of 10311 objects annotated in the images of Dataset B (see Table 4).

5 XML Format

The annotations represent a compositional hierarchy with has-part relationships. Each object is also described by a bounding polygon. The annotations

Table 4: Frequency of all classes in Dataset B

Type	**Count**	
Apartment Building	328	2.94 %
Balcony	380	3.41 %
Canopy	56	0.50 %
Chimney	66	0.59 %
Cornice	14	0.13 %
Door	591	5.30 %
Dormer	85	0.76 %
Entrance	163	1.46 %
Facade	423	3.79 %
Gate	11	0.10 %
Ground	152	1.36 %
Ground Floor	369	3.31 %
Office Building	19	0.17 %
Other Building	42	0.38 %
Pavement	132	1.18 %
Person	28	0.25 %
Railing	430	3.86 %
Road	88	0.79 %
Roof	408	3.66 %
Scene	200	1.79 %
Sign	106	0.95 %
Single Family Home	58	0.52 %
Sky	211	1.89 %
Stairs	46	0.41 %
Transportation Object	176	1.58 %
Upper Floor	988	8.86 %
Vegetation	394	3.53 %
Wall	6	0.05 %
Window	4689	42.06 %
Window Array	489	4.39 %

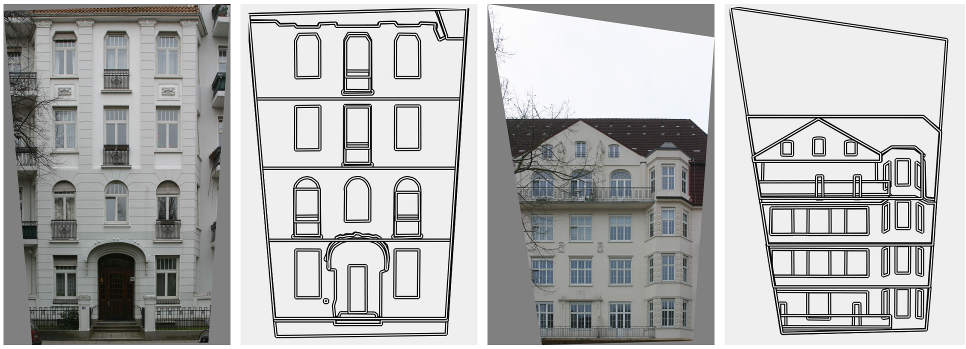


Figure 2: Two examples from the datasets. The image on the left is followed by the corresponding annotation on the right. The grey regions are the result of the image rectification process.

are described in an XML format which is an extension of the LabelMe XML format. The LabelMe XML format also defines objects as bounding polygons, but does not include partonomical relations.

All XML tags and the structure of an XML file are shown in Table 5, where ($\vdots$) indicates that multiple entities may be present. Additional tags are present to maintain compatibility with the Annotation Tool used to produce them [5], and do not contain important information. The indents represent *has-part* relationships, while tags on the same level, in the same scope, are siblings. A partial tree structure of the compositional hierarchy of an image is shown in Figure 1. The new tags compared to the LabelMe XML format are:

- `imageWidth` and `imageHeight`, which contain the size of the original image,
- `scale`, which is an estimate of how many pixels represent a centimeter at the facade level,
- `rectified`, which indicates whether the image has been rectified, and `transformationMatrix`, which contains the matrix used for the rectification
- `annotatedClasses`, which contains a list of classes that are annotated in the image, each described by a `className` tag

- `objectID`, which is an object identifier string unique in the whole database, and
- `objectParts`, which is a comma-separated list of the objectIDs of all the parts belonging to an object.

5.1 Folder Structure

The datasets contain two folders: **images** and **annotations**. This folder structure is compatible with the University of Bonn's Annotation Tool and consistent with the LabelMe convention. The images are given as JPGs with the resolution varying between 515 × 328 and 4064 × 3456 pixels. The annotations are in the described XML format.

6 Comparison to Other Datasets

The presented datasets are similar to the eTRIMS 8-class and 4-class dataset [4]. The main differences are that our datasets features more complex and varied scenes, and that the partonomy is explicitly annotated. Since our datasets aim to test complete interpretations and not pixel labelling, we use polygons instead of pixel masks. A pixel can therefore belong to a number of objects at different levels of the partonomical hierarchy.

In the eTRIMS Interpretation datasets, images contain full annotation, meaning that all important objects in the scene are annotated. This is in contrast to common annotated datasets used by the Computer Vision community, such as the TU Darmstadt Database [6], UIUC Car Database [1], the VOC Challenge datasets [2], and the Caltech [3], MIT-CSAIL [9] and TU Graz [7] datasets, where only a small number of the objects in the scene are annotated. The full annotation of objects and the partonomy makes our datasets suitable as ground truth for learning complex scene models and evaluating complete scene interpretations.

Table 5: XML File Structure

`annotation`	
`filename`	the filename of the corresponding image file
`folder`	the folder name of the corresponding image file
`sourceAnnotationXML`	the version string of the annotation tool
`rectified`	1 := *image is rectified* and 0:= *image is not rectified*
`imageWidth`	the width of the corresponding image
`imageHeight`	the height of the corresponding image
`transformationMatrix`	the transformation matrix used for the rectification
`annotatedClasses`	a list of all annotated classes
`className`	one element of the class list
⋮	
`scale`	estimate of the size of *pixel per cm*
`object`	an object in the scene, this can be an aggregate or a primitive
`name`	the type of the object, e.g. `window`
`objectID`	an ID for the object (unique in the image)
`date`	the date of the annotation
`sourceAnnotation`	the person who annotated the image
`polygon`	the polygon that describes the object
`pt`	a point in the polygon
`x`	the x coordinate of the point
`y`	the y coordinate of the point
⋮	
`objectParts`	a list of `objectID`s of the parts that belong to this object

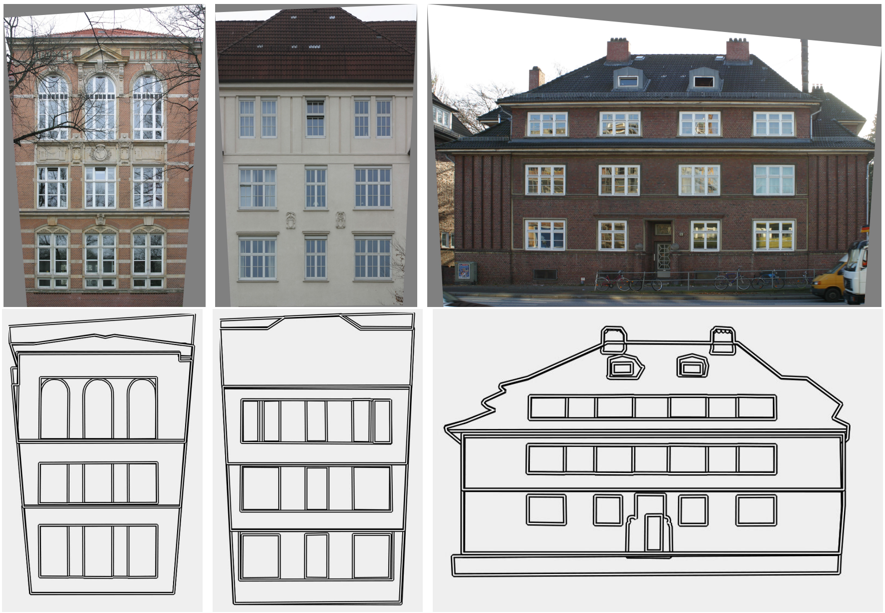


Figure 3: Some examples from the datasets. The images are shown in the top row, the corresponding annotations in the bottom row.

Acknowledgment

This research has been supported by the European Community under the grant IST 027113, eTRIMS - eTraining for Interpreting Images of Man-Made Scenes. We would also like to thank the members of the eTRIMS Consortium, namely: Jan Čech, Martin Drauschke, Ondřej Drbohlav, Daniel Heesch, Lothar Hotz, Mohammad Jahangiri, Radim Šára, Hanns-F. Schuster, Jan Šochman, Robby T. Tan and Susanne Wenzel, who contributed to the database by collecting data, providing annotations and supervising students. We thank all members of the eTRIMS consortium for valuable discussions and feedback.

References

[1] S. Agarwal, A. Awan, and D. Roth. UIUC image database for car detection. http://l2r.cs.uiuc.edu/ cogcomp/Data/Car/. [Online; accessed 20-Jan-2010].

[2] M. Everingham. The VOC 2006 database. http://pascallin.ecs.soton.ac.uk/challenges/VOC/databases.html. [Online; accessed 20-Jan-2010].

[3] R. Fergus and P. Perona. The Caltech database. http://www.robots.ox.ac.uk/ vgg/data3.html. [Online; accessed 20-Jan-2010].

[4] F. Korč and W. Förstner. eTRIMS Image Database for interpreting images of man-made scenes. Technical Report TR-IGG-P-2009-01, April 2009.

[5] F. Korč and D. Schneider. Annotation tool. Technical Report TR-IGG-P-2007-01, June 2007.

[6] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004.

[7] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):416–431, 2006.

[8] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. Technical report, Tech. Rep. MIT-CSAIL-TR-2005-056, Massachusetts Institute of Technology, 2005.

[9] A. Torralba, K. P. Murphy, and W. T. Freeman. The MIT-CSAIL database of objects and scenes. http://web.mit.edu/torralba/www/database.html. [Online; accessed 20-Jan-2010].