Datasets and Dataset Creation

Visual Recognition and Search

Maysam Moussalem

Outline

- Importance of datasets
- Existing datasets
- Issues with current datasets
- New ways of acquiring large and diverse datasets
- LabelMe: a database and web-based tool
- Conclusion

Importance of datasets

- Datasets needed at all stages of object recognition
 - Learning visual models
 - Detecting and localizing instances of these models
 - Evaluating performance
- A good dataset must be
 - Very large
 - Very diverse
 - Well-annotated
- Drive research by providing common ground

Existing datasets

- Caltech 101
- Caltech 256
- PASCAL Visual Object Classes challenges
- Oxford buildings, flowers datasets
- CMU Face databases
- MIT Objects and Scenes
- Photo-tourism patches
- ...

Issues with current datasets...

- Unfortunately, most of these offer limited range of image variability!
 - Similar viewpoints and orientations
 - Sizes and image positions normalized
 - Little or no occlusion and background clutter
 - Often only one instance of object in image

• ...









A bit better... The Pascal 2006 average image (constructed by T. Malisiewicz)

Problems with existing datasets

- Some algorithms may exploit restrictions in datasets
 - E.g. those lacking scale, rotation invariance...
- Images are not challenging enough
 - More sophisticated algorithms might not show better results
 Results tend to converge around 100% accuracy

Outline

- Importance of datasets
- Existing datasets
- Issues with current datasets
- New ways of acquiring large and diverse datasets
- LabelMe: a database and web-based tool
- Conclusion

New ways of acquiring large and diverse datasets

- Web-based annotation tools
- Rely on collaborative effort of large population online
- Examples
- ESP
- Peekaboom
- LabelMe









- Online annotation tool
- Allows sharing of images and annotations
- Provides many functionalities
 - Drawing polygons
 - Querying images
 - Browsing the database



























