Leveraging flickr images for object detection

Elisavet Chatzilari

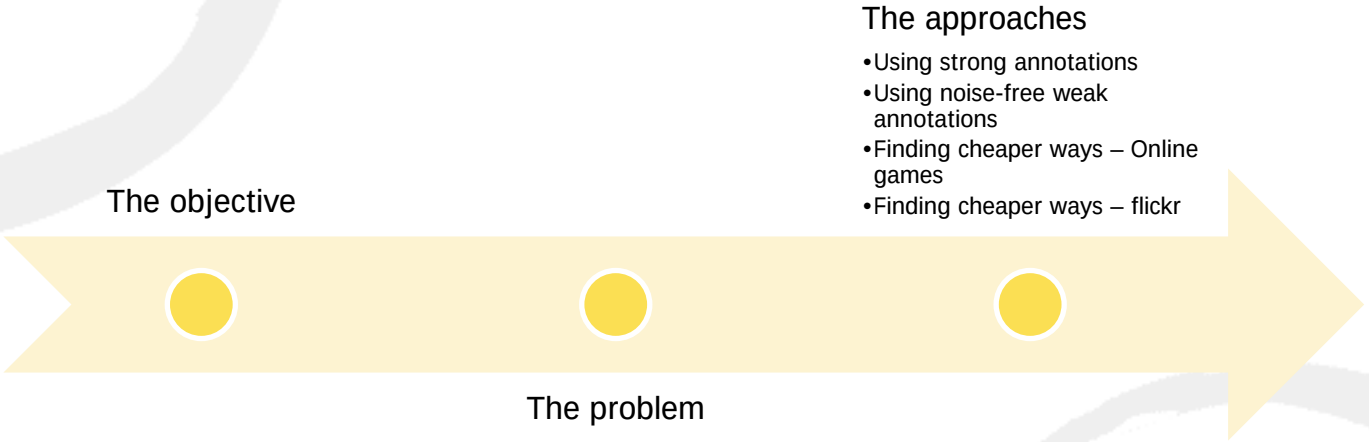
Spiros Nikolopoulos

Yiannis Kompatsiaris

Outline

- Introduction to object detection
- Our proposal
- Experiments
- Current research

Introduction to object detection



The objective

Given an unseen image, the objective is to automatically identify and localize the present visual objects in a scalable and effortless way

The problem

Machine learning algorithms were extensively used to issue the object recognition problem:

Simulate the functionality of the human visual system to recognize visual objects by using a number of samples to train a model for a semantic concept.

The efficient estimation of object detection model parameters mainly depends on two factors:

Quality of the training examples (manual annotation)

Quantity of the training examples (large corpus)

Such samples are very expensive to obtain raising scalability problems

The approaches

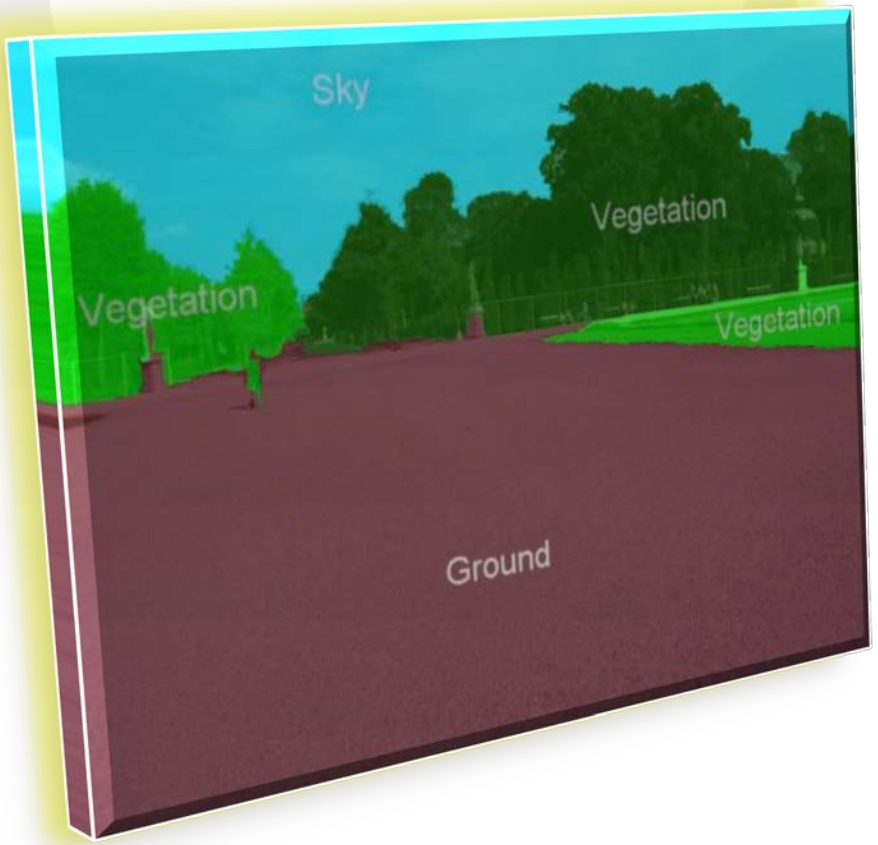
Annotation type of the training images

- Strong annotations
- Weak annotations – noise free
- Rough annotations – noisy

Algorithms

- Support Vector Machines
- Bayesian Networks
- Random Forests
- Probabilistic Latent Semantic Analysis
- **...**

Strong Annotations



Training models of object classes using strong annotations (manually annotated images at region or pixel level e.g. MSRC dataset)

- Model parameters are learned more efficiently
- It is practically impossible to have strong annotations for many concepts

Approaches

- learn the conditional distribution of the class labels given an image, using a Conditional Random Field (CRF) model

J. Shotton, J. M. Winn, C. Rother, A. Criminisi, TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: ECCV (1), 2006, pp. 1–15.

Noise-free weak annotations

Clouds;

Sea;

Sun;

Tree;

Training models of object classes using noise-free weak annotations (manually annotated images at image level e.g. Corel dataset)

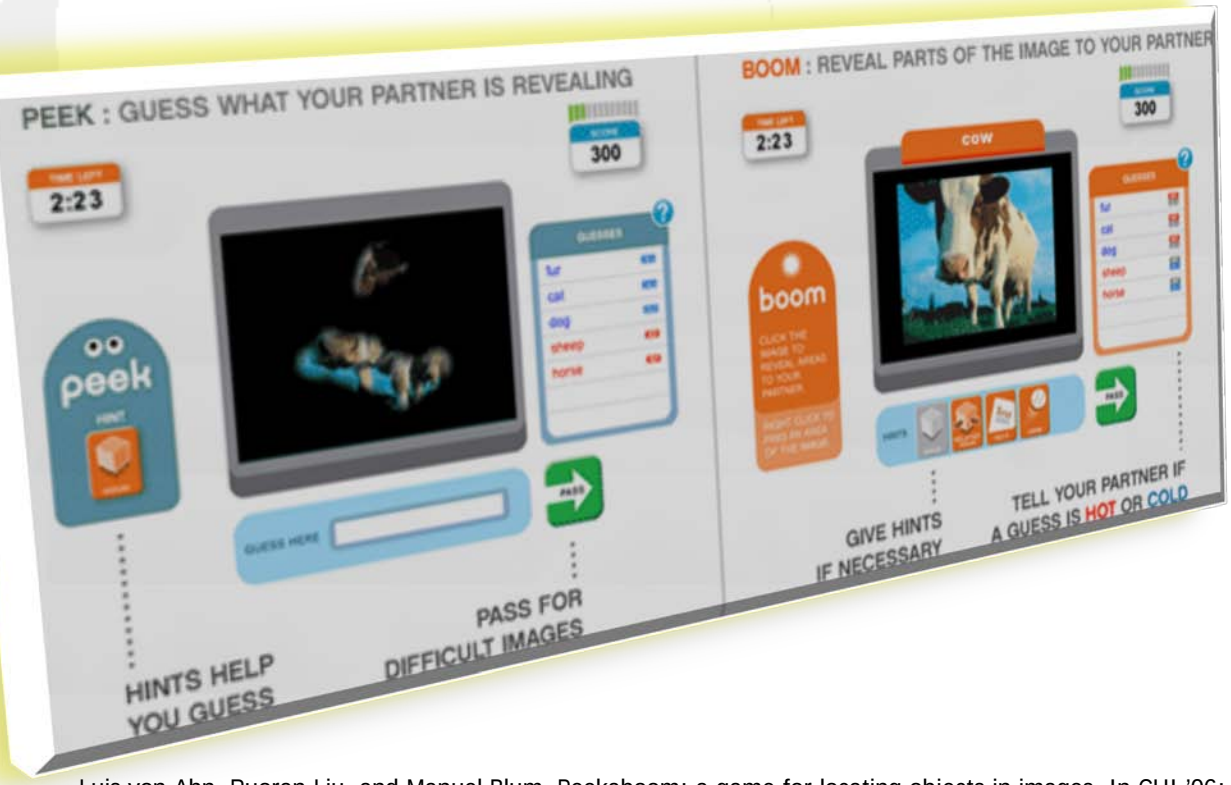
- Cheaper to obtain
- Manual annotation, even at image level is a time-consuming and laborious task limiting the scalability of frameworks depending on such annotations

Approaches

- Combine aspect models (PLSA) with spatial models (MRF) [1]
- Learn correspondences between regions and labels using EM [2]

[1] J. J. Verbeek, B. Triggs, Region classification with markov field aspect models, in: CVPR, 2007.
[2] P. Duygulu, K. Barnard, J. F. G. de Freitas, D. A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in: ECCV (4), 2002, pp. 97–112.

Finding cheaper ways – online annotation games?



Online annotation games

- Effortless and scalable learning

Approaches

- The task of annotation was presented as a game
- People play the game for entertainment
- Collecting valuable metadata is a side effect

Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 55–64, New York, NY, USA, 2006. ACM.

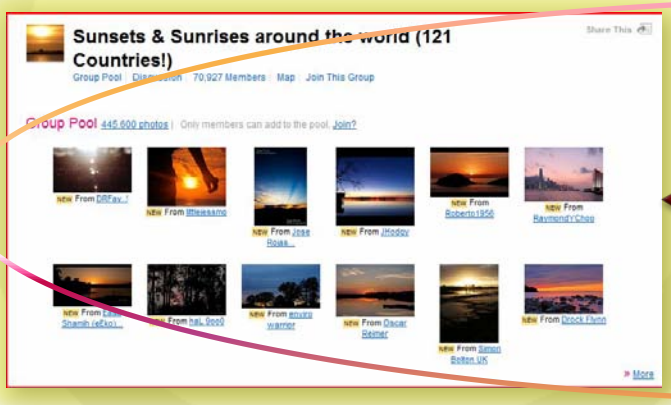
The big bang of social sites

The excessive use of web 2.0 applications resulted to mass user-generated content (text, images, video)

flickr is populated daily with thousands of images with associated information (tags, geo-tags, notes etc)

Can we leverage effectively the unlimited and “cheap” social content in order to train object detectors?

flickr – what it has to offer



flickr Groups

Tags

Blue • sky • propeller • Blue sky • Clouds • plane • photography • Photo • Image • Imagery • Nikon • Lightroom • Photoshop • Adobe • Cuba Gallery • presets • Light • Tone • amazing • best • awesome • color • colour • full color • color grading • art • cool • background • square • square format • style • Nikon Camera • texture • lighting • composition • artistic • digital • Tips • Tricks • Techniques • pictures

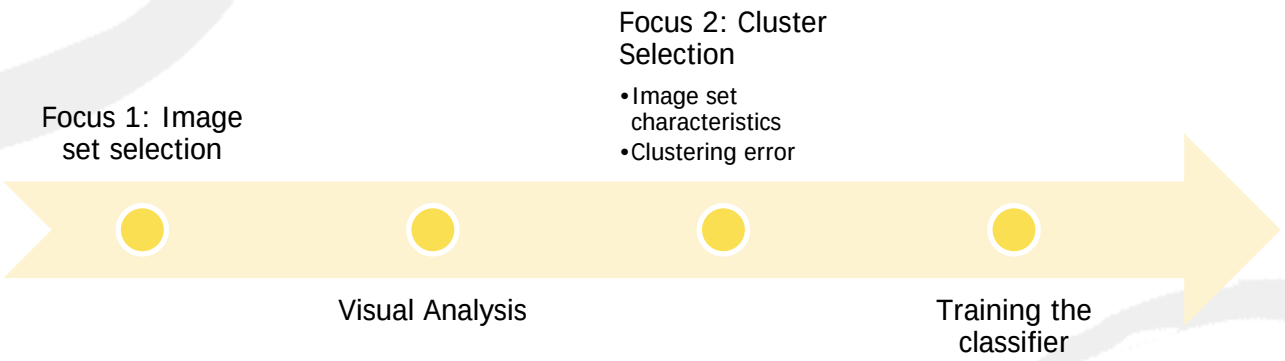
Images + tags

flickr

flickr notes

Geo-tagged images

Our proposal



E. Chatzilari, S. Nikolopoulos, I. Patras and I. Kompatsiaris, " Enhancing Computer Vision using the Collective Intelligence of Social Media" in book: "New Directions in Web Data Management 1", Springer, Series: "Studies in Computational Intelligence", Editors: Athena Vakali and Lakhmi Jain, Publishing: February 5, 2011

Our proposal

Proper image set selection: images/tags emphasize visually/textually on the same semantic concept

Visual representation of the concept

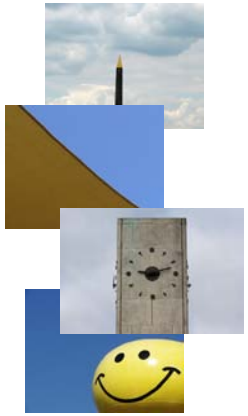
Object detection model

Textual representation of the concept

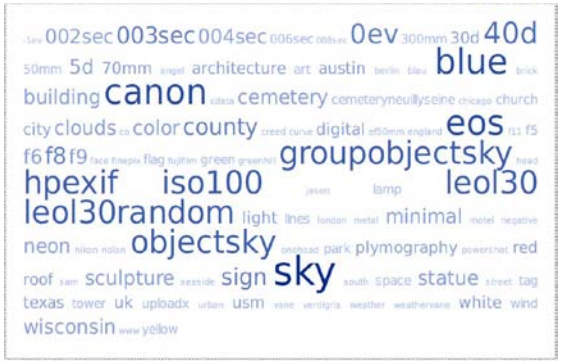
Sky

Flickr images

Image Group

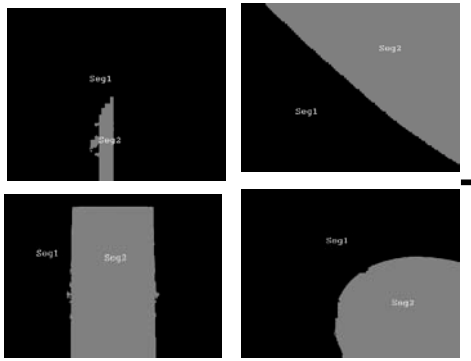


Tag cloud of all images in the Image Group

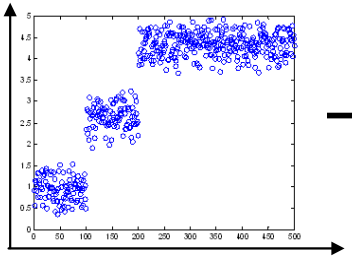


- Each word corresponds to a tag-“term”
- Sky is the most populated tag-“term” in this group

Segmentation

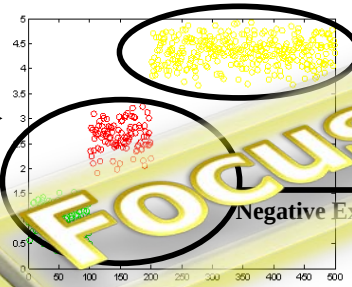


Visual Feature Extraction



Visual Features Space

Clustering



Positive Examples

FOCUS 2

Negative Examples

Supervised Learning

Object Detection Model for Sky

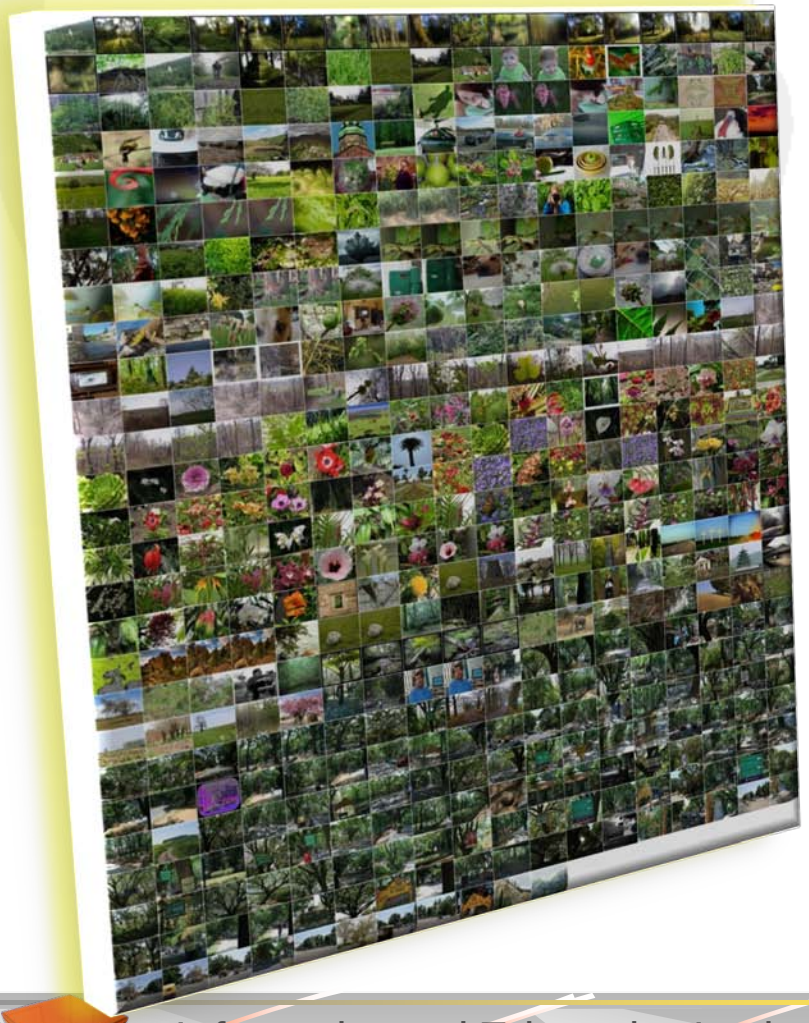
- Each group of colored points corresponds to a visual-“term”
- The group of points colored in yellow is the most populated visual-“term”

Focus 1: image set selection

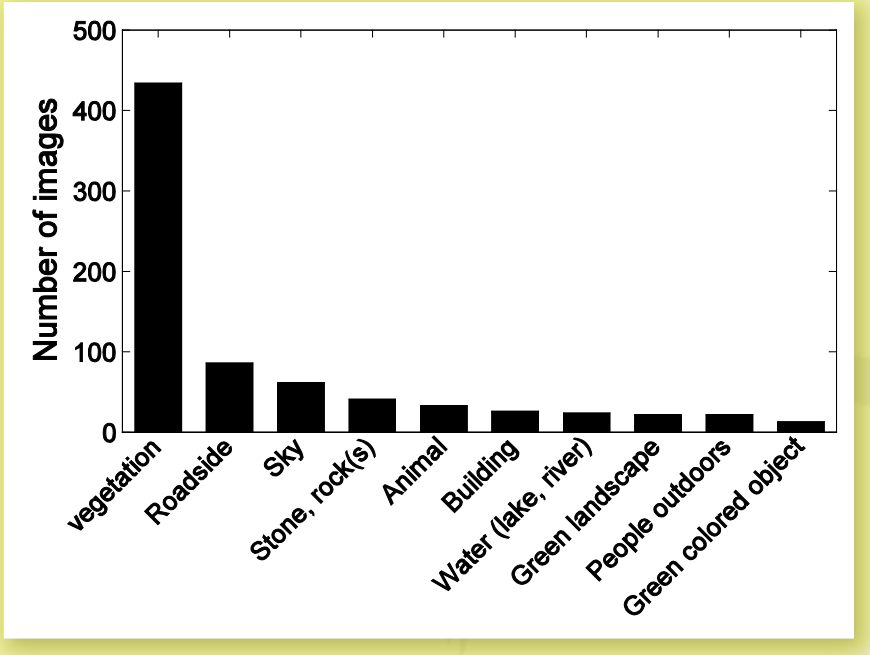
Collect a set of images the majority of which depict the object of interest

- **SEMSOC**, tag based clustering that provides semantically coherent image groups which emphasize on a semantic concept. The textual representation of this concept is given by the most populated tag of the image cluster.
  - *E. Giannakidou, I. Kompatsiaris, A. Vakali, Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems, in: ICSC, 2008, pp. 128–135.*
- **flickr groups**, virtual places hosted in flickr that allow social users to share content on the grounds of a certain topic

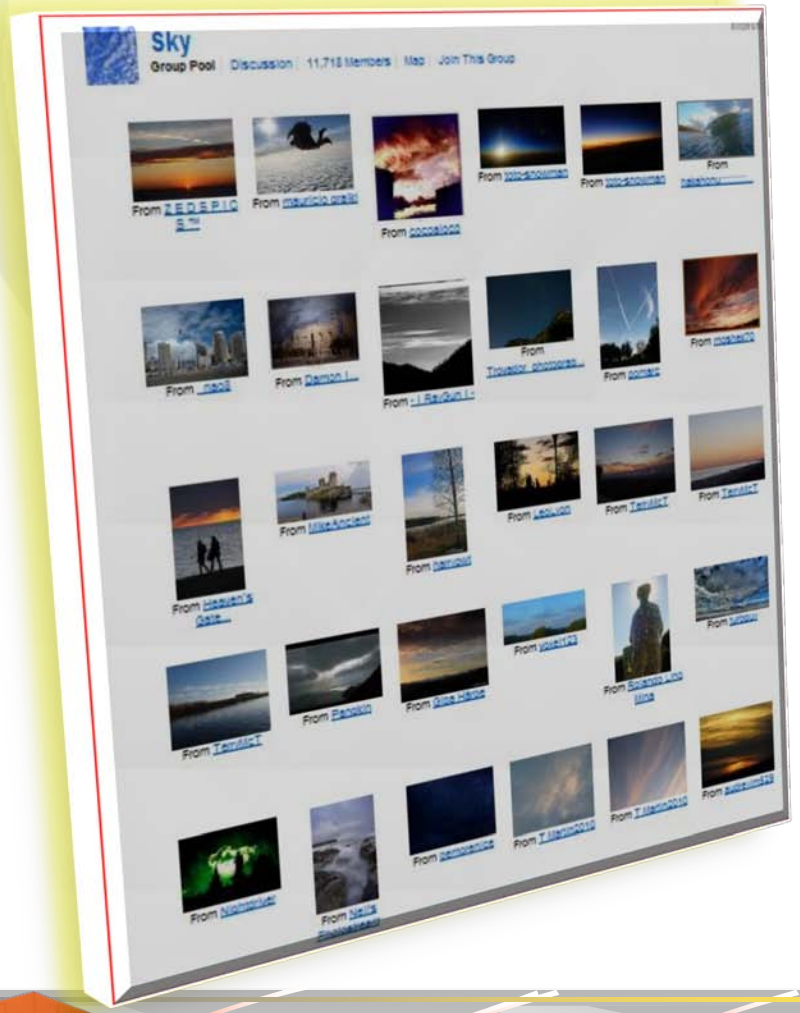
SEMSOC output example (vegetation)



Distribution of objects based on their frequency rank



flickr group example (Sky)



The textual representation of the semantic concept is extracted by the group title.

Visual Analysis

Segmentation

- K-means with connectivity constraint (KMCC) [1]

Visual Descriptors

- Bag-of-words approach by vector quantizing all SIFT descriptors found in each segmented region. [2]

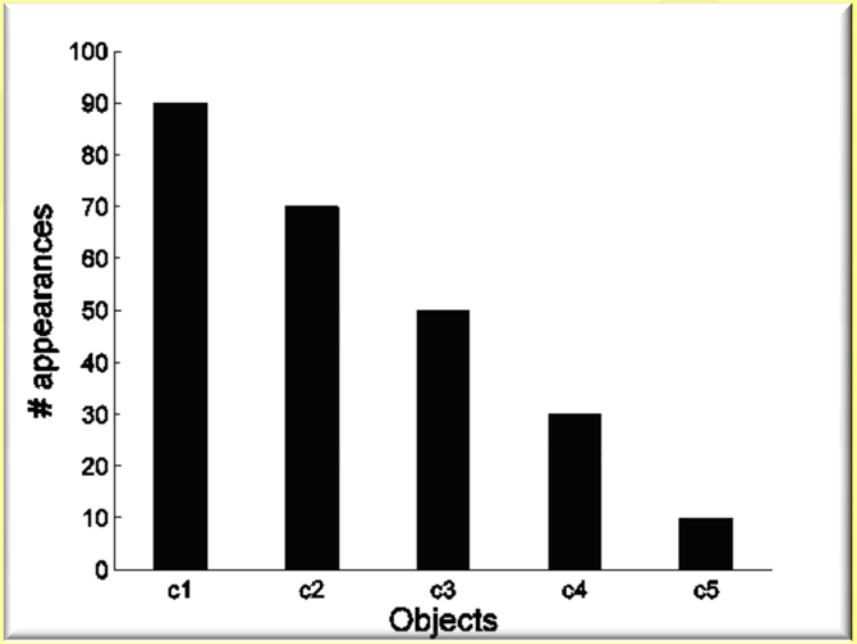
Region clustering

- Affinity propagation Clustering [3]

[1] *V. Mezaris, I. Kompatsiaris, M. G. Strintzis,* Still image segmentation tools for object-based multimedia applications, IJPRAI 18 (4) (2004) 701–725.
[2] *D. G. Lowe*, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[3] *B. J. Frey, D. Dueck,* Clustering by passing messages between data points, Science 315 (2007) 972–976.

Focus 2: Cluster selection - Imageset characteristics

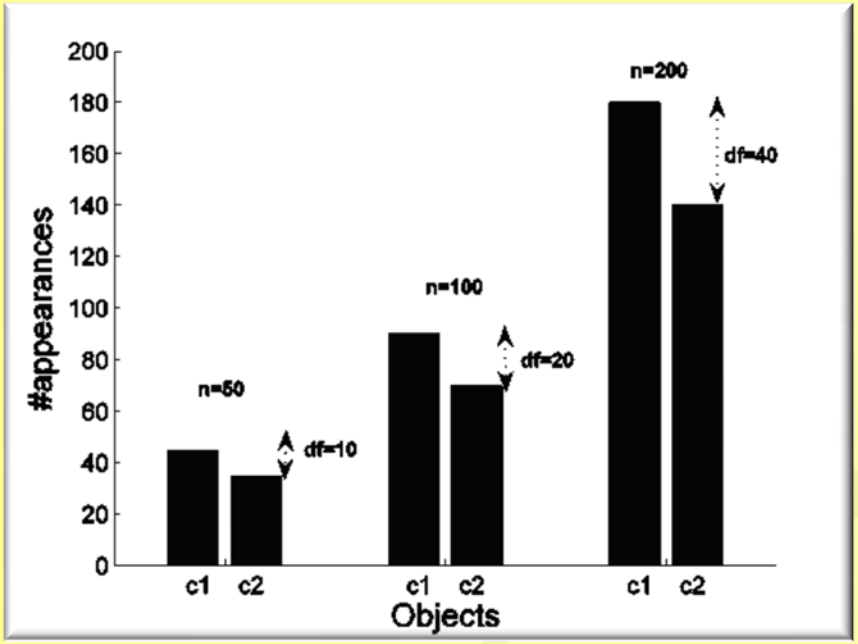
Distribution of objects in the dataset (synthetic data)



The images emphasize on a single concept (c1)

- **The frequency of the visual object related to c1 will be higher than any other object.**
- **Another concept semantically associated with the selected will appear more than others (e.g. sea, sand).**

Gap between 1st and 2nd most frequent objects at dataset size n=50, 100 and 200 (synthetic data)

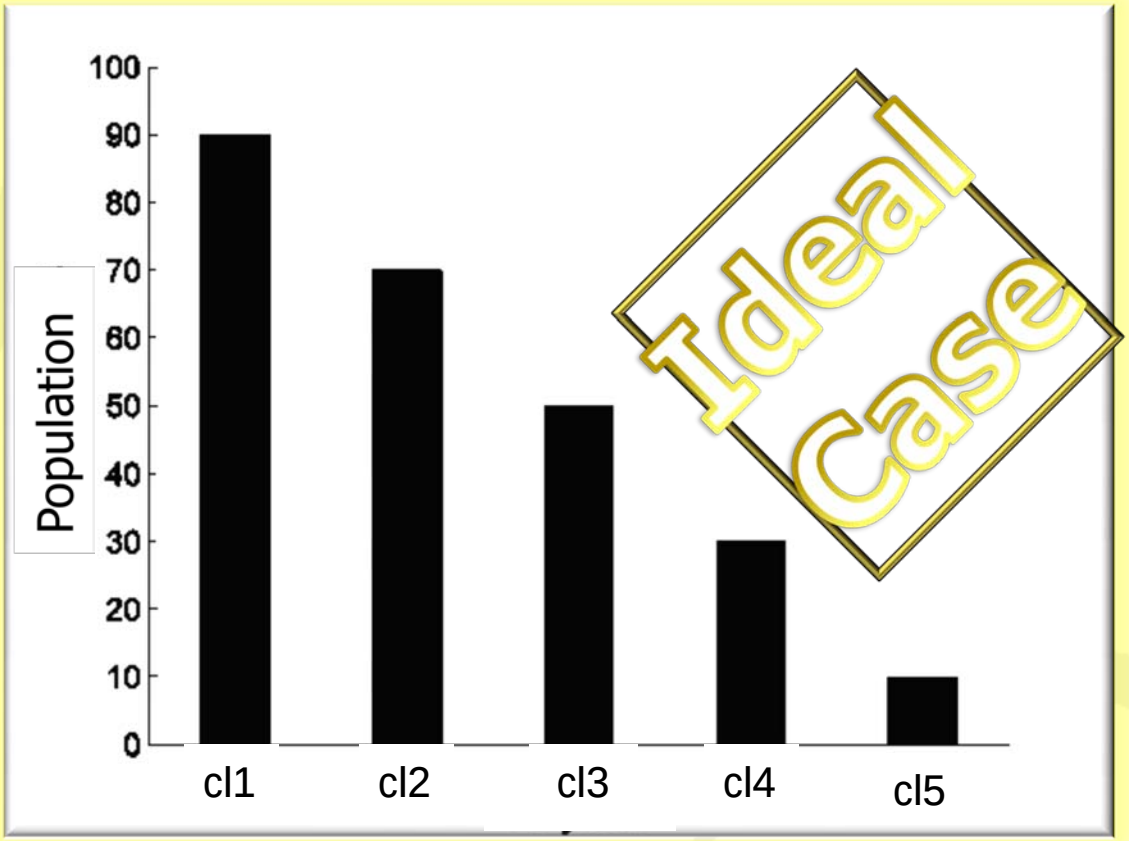


As the dataset size (n) increases

- **The population of c1 will increase with a higher rate than any other object**
- **The gap between the frequency of the 1st and 2nd most populated objects will increase**

Focus 2: Cluster selection – Ideal case

Distribution of clusters' population if visual analysis algorithms performed ideally



Perfect case

- The distribution of objects within the image set and the distribution of the population of the clusters coincide:
- The most populated cluster contains all regions depicting the most frequently appearing object.

Real case

- Examine how a possible solution deviates from the perfect solution
- How the dataset size is connected to the success of our choice (the most populated cluster containing the most frequently appearing object)

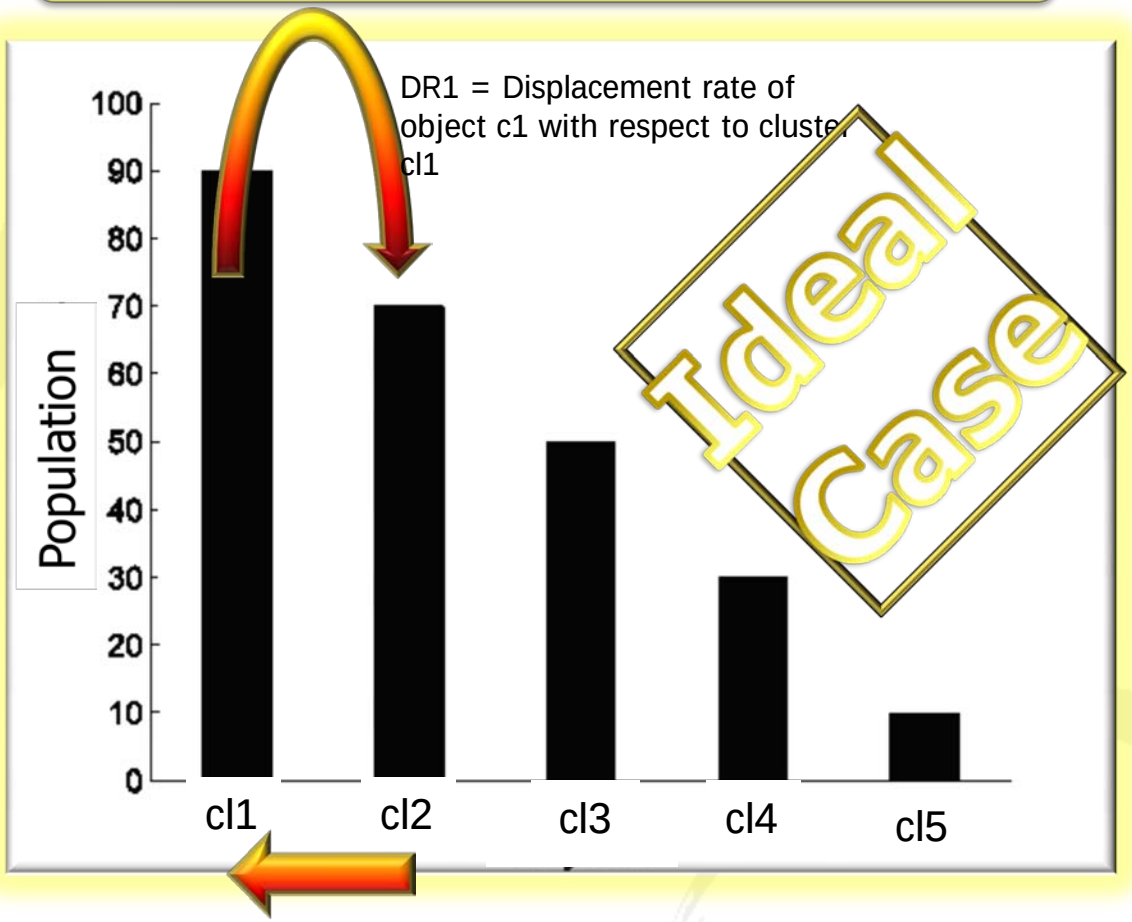
Focus 2: Cluster selection - Clustering error

We view the error introduced by all visual analysis algorithms as a cluster-to-object assignment error ($error_{cl\text{-}obj}$)

The cluster corresponding to the second most frequently appearing object (cl2) is more likely to become more populated than the cluster corresponding to the first most frequently appearing object(cl1).

When the dataset size increases, the gap between the frequency of c1 and c2 widens, allowing more error to be introduced and still select the appropriate cluster.

Distribution of clusters' population if visual analysis algorithms performed ideally



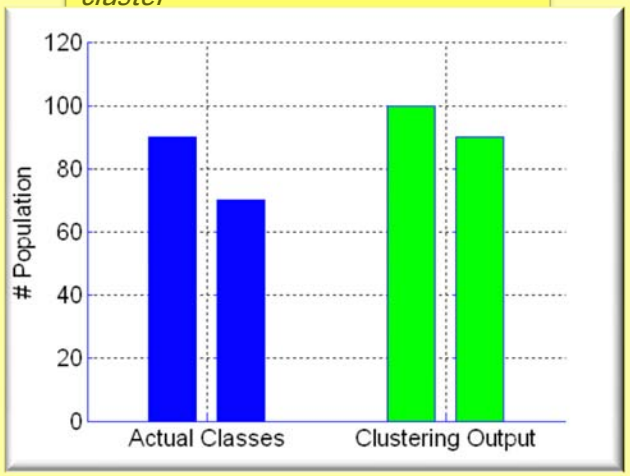
As cluster-to-object error Increases …

Perfect Case

error = 0

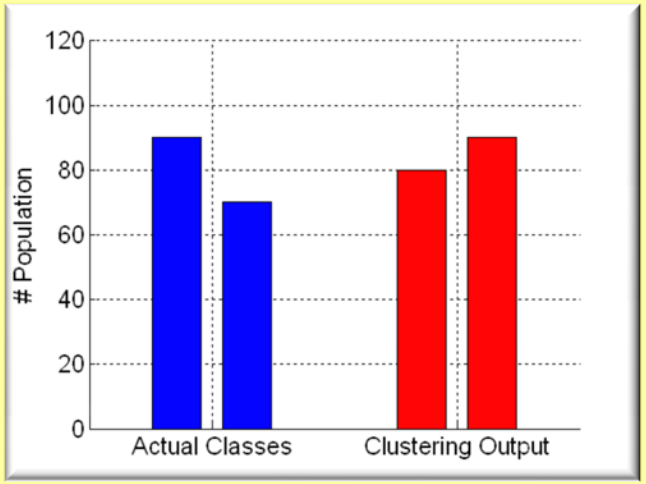
Most populated still the correct cluster

error > 0



Situation is reversed

error >> 0



The correct cluster is by far missed

error >>> 0

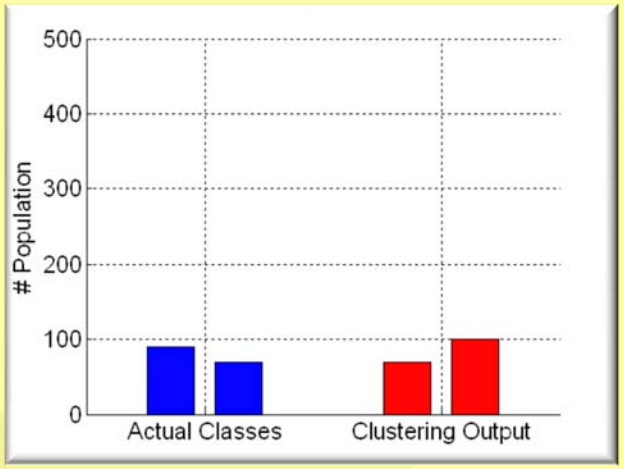
FIXED DATASET SIZE

As Dataset Size Increases …

FIXED CLUSTERING ERROR

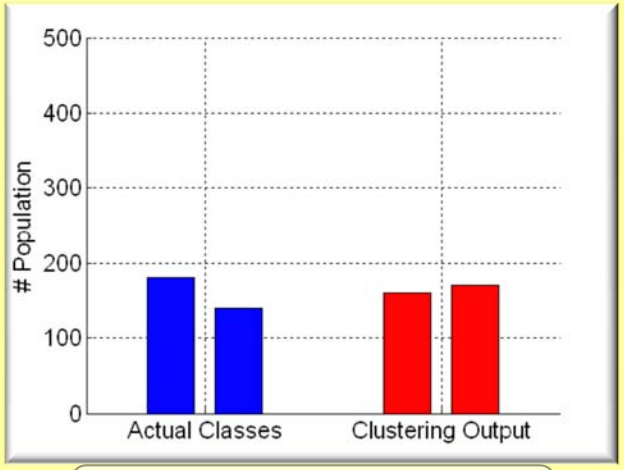
Problematic Case

N = α



The gap is shortened

N > α

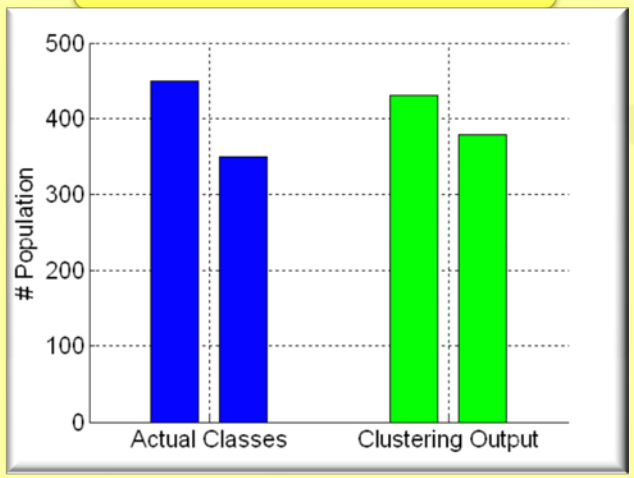


Situation is reversed

N >> α

The correct cluster is easily identified

N >>> α

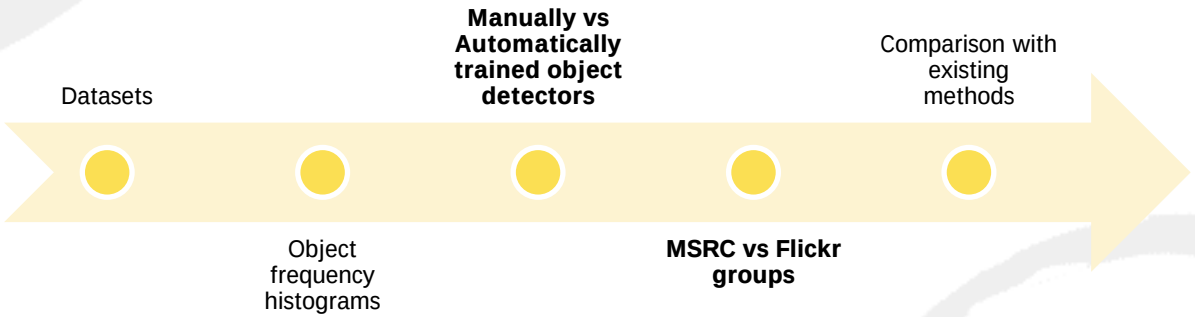


Step 4: Train models

Support Vector Machines were chosen for training the object detection models

- Select the regions belonging to the most populated visual cluster to be the positive examples
- Negative examples are chosen randomly from the remaining dataset

Experiments



Datasets

Name	Source	Annotation type	No. Of Images	objects	Selection approach
Flickr3k	flickr	Weak – images + tags	3000	cityscape, sea-side, mountain, roadside, landscape, sport-side	SEMSOC
Flickr10k	flickr	Weak – images + tags	10000	Jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis	SEMSOC
Flickr Groups	Flickr groups	Weak – flickr groups	500 images per concept	sky, sea, person, vegetation and the 21 MSRC objects	Group title
Seaside	internal	Strong	536	sky, sea, person, sand, rock, boat, vegetation	Keyword based
MSRC	MSRC	Strong	591	aeroplane, bicycle, bird, boat, body, book, cat, chair, cow, dog, face, flower, road, sheep, sing, water, car, grass, tree, building, sky	Keyword based

Evaluation

Object frequency histograms for different sizes of the dataset

Compare the performance of automatically trained object detectors from flickr images to the ones generated by strongly annotated images

Compare with existing methods

Object frequency histograms - setup

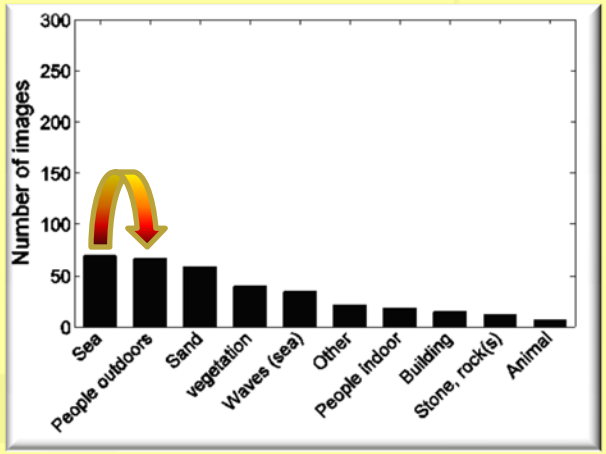
Flickr 3k

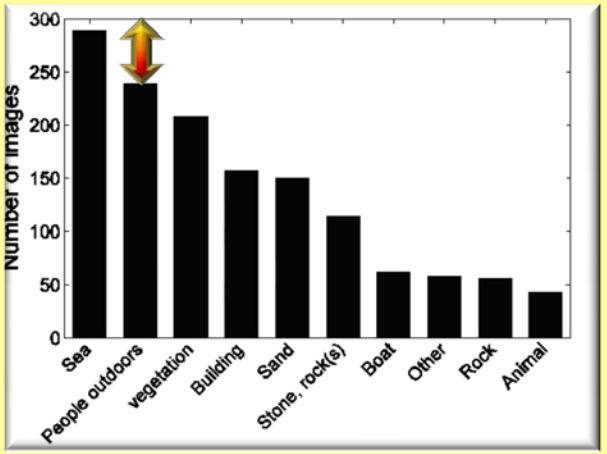
Flickr 10k

SEMSOC

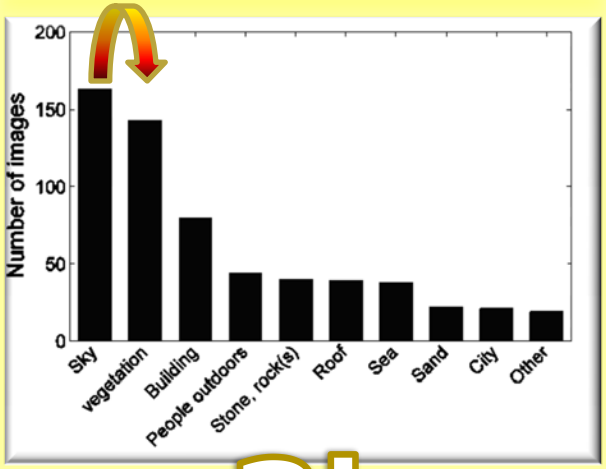
Object frequency histograms

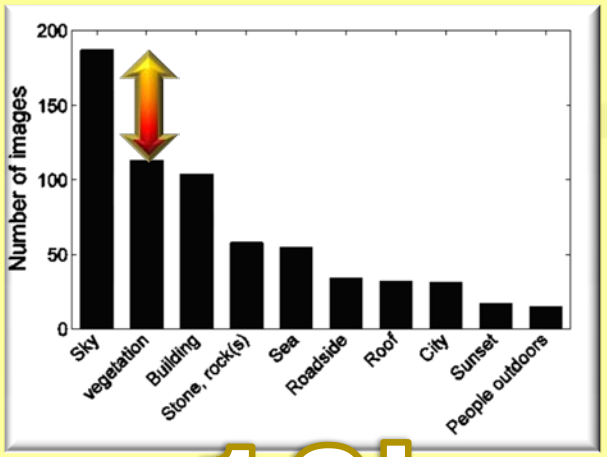
Sea





Sky



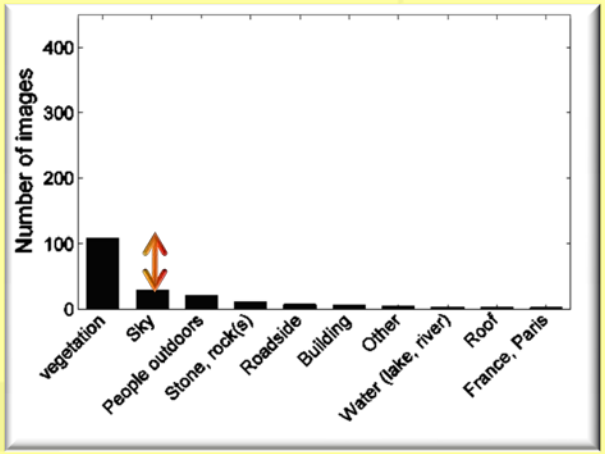


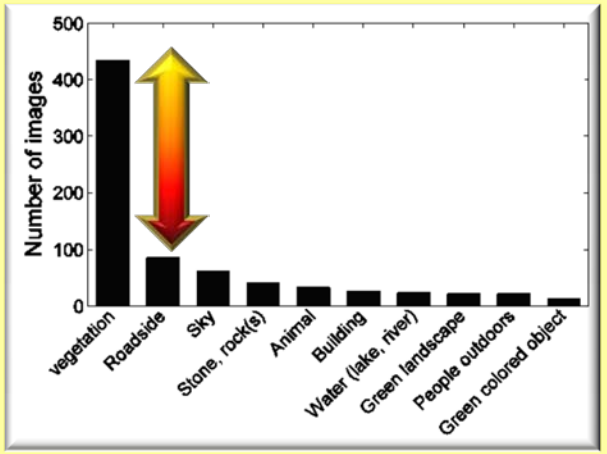
3k

10k

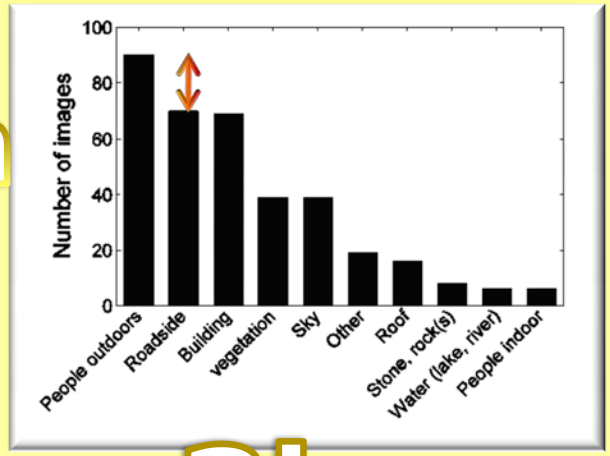
Object frequency histograms

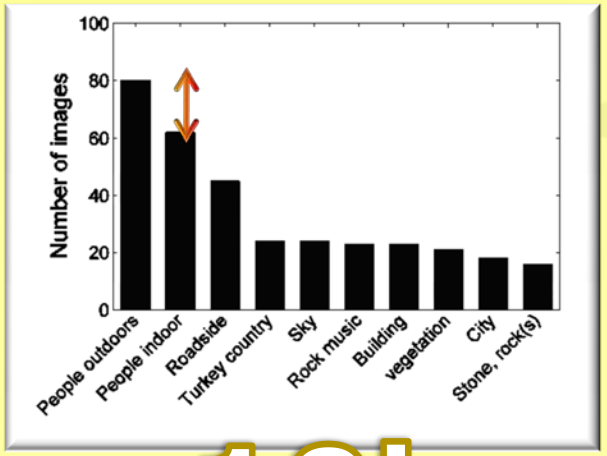
Vegetation





Person





3k

10k

Manually vs Automatically trained object detectors - setup

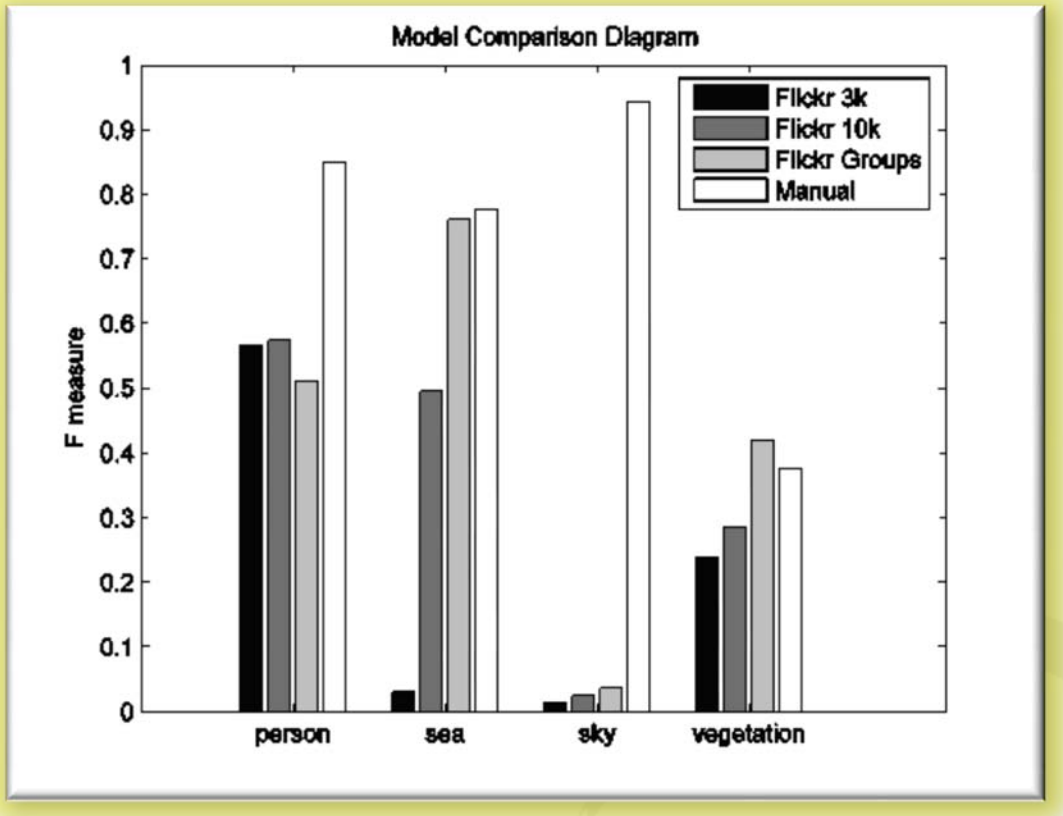
Dataset	Flickr 3k	Flickr 10k	Flickr groups	Seaside
Selection algorithm	SEMSOC	SEMSOC	Title	Keyword based
Framework	Proposed approach	Proposed approach	Proposed approach	Keyword based selection

Manually vs Automatically trained object detectors

Performance lower than manually trained detectors

Consistent performance improvement as the dataset size increases

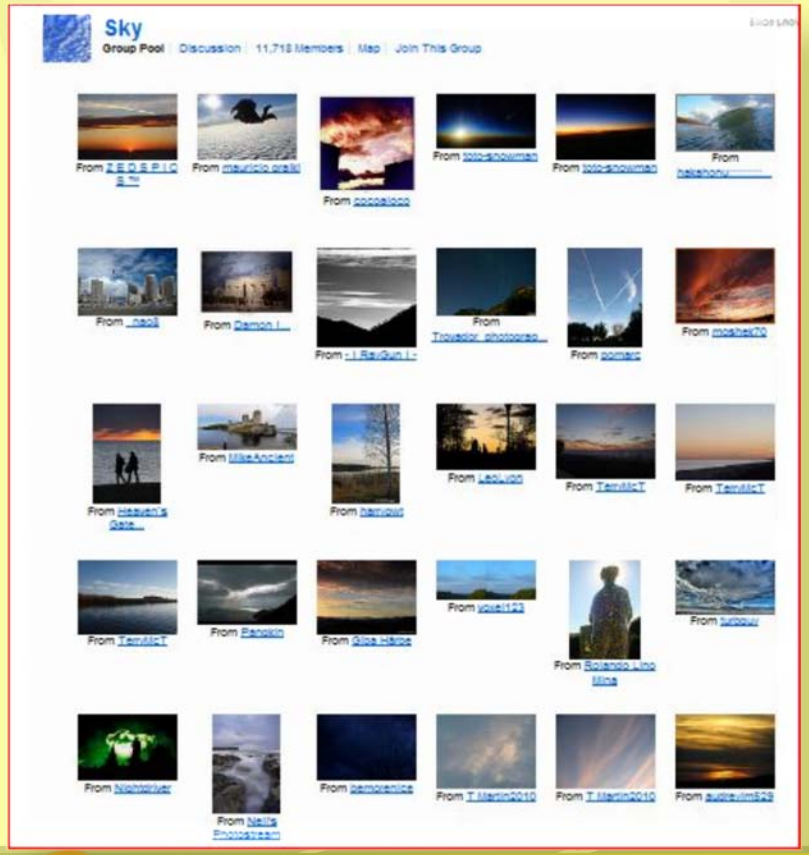
- Sea: The increase of the dataset size allows us to choose the appropriate cluster
- Sky: The dataset size needs to increase more, so that the most populated cluster to become the one containing the sky regions.



MSRC vs Flickr groups - setup

Flickr Group for object sky

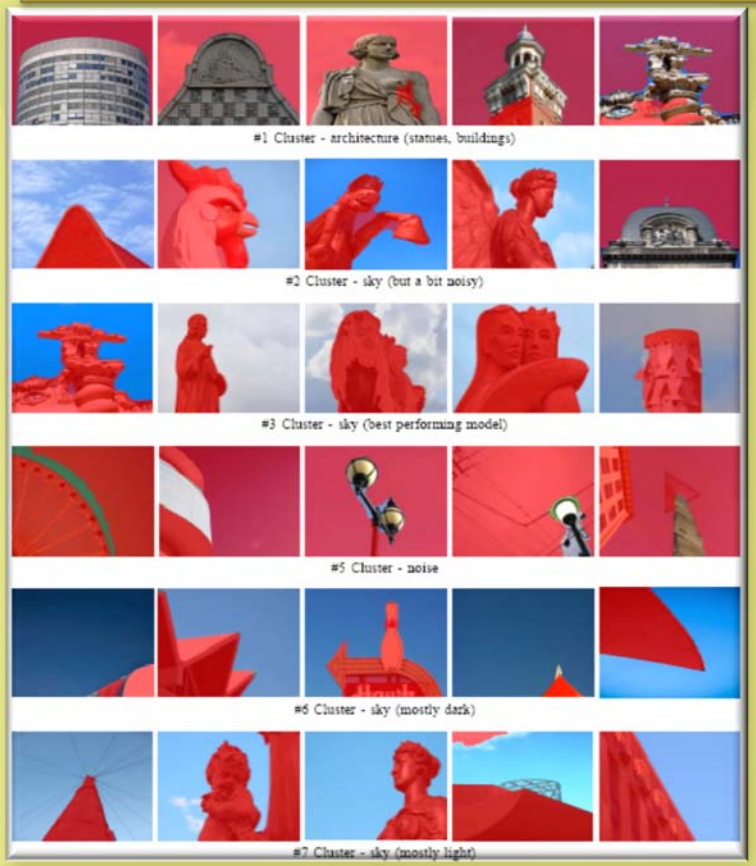
Clusters of regions



MSRC vs Flickr groups - setup

Clusters of regions

Select each cluster as positive each time and train classifiers



Train SVM classifier

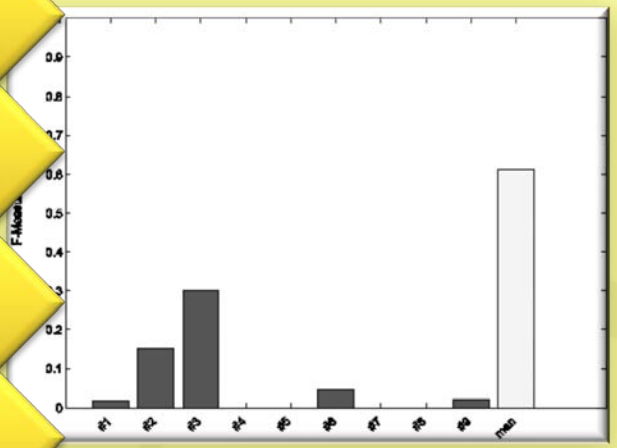
Train SVM classifier

Train SVM classifier

Train SVM classifier

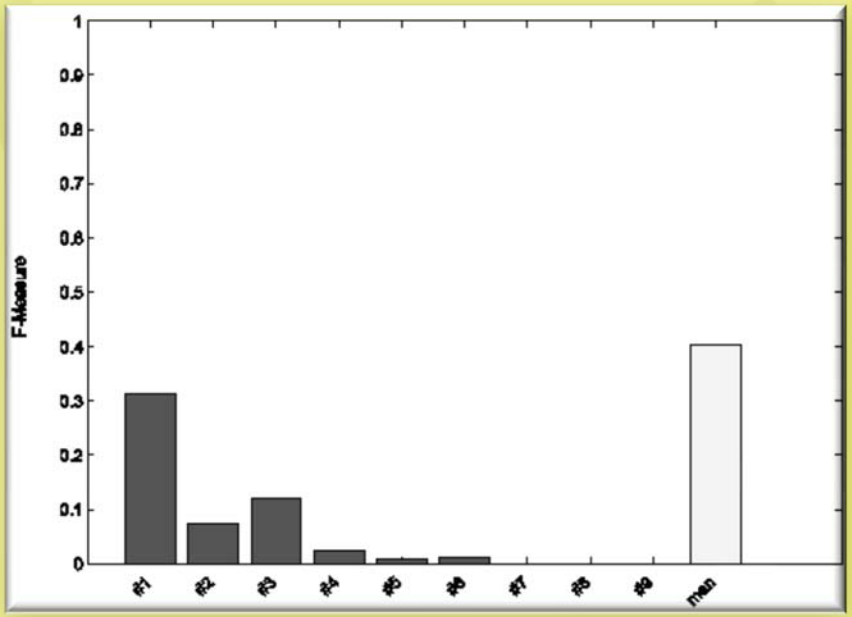
Train SVM classifier

Train SVM classifier



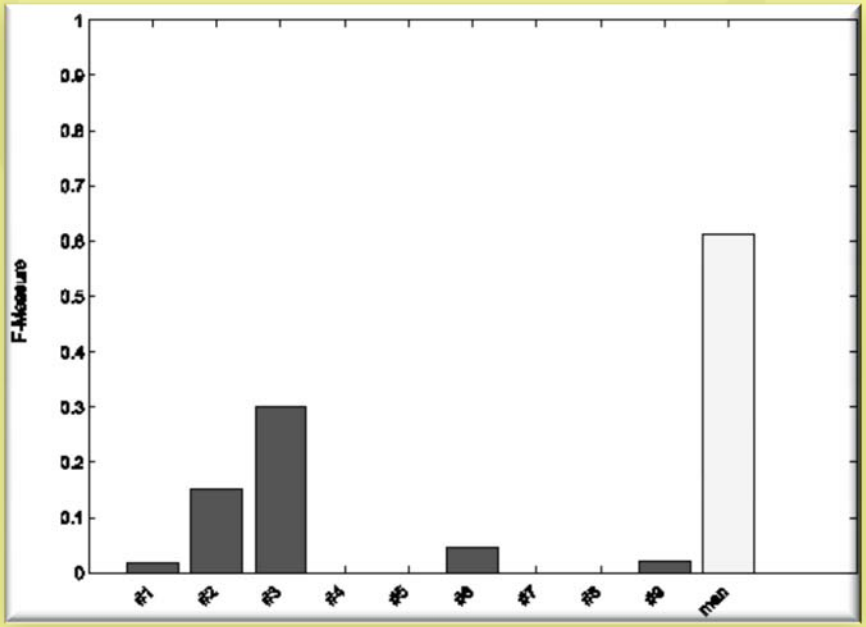
Experimental Results - MSRC vs Flickr groups

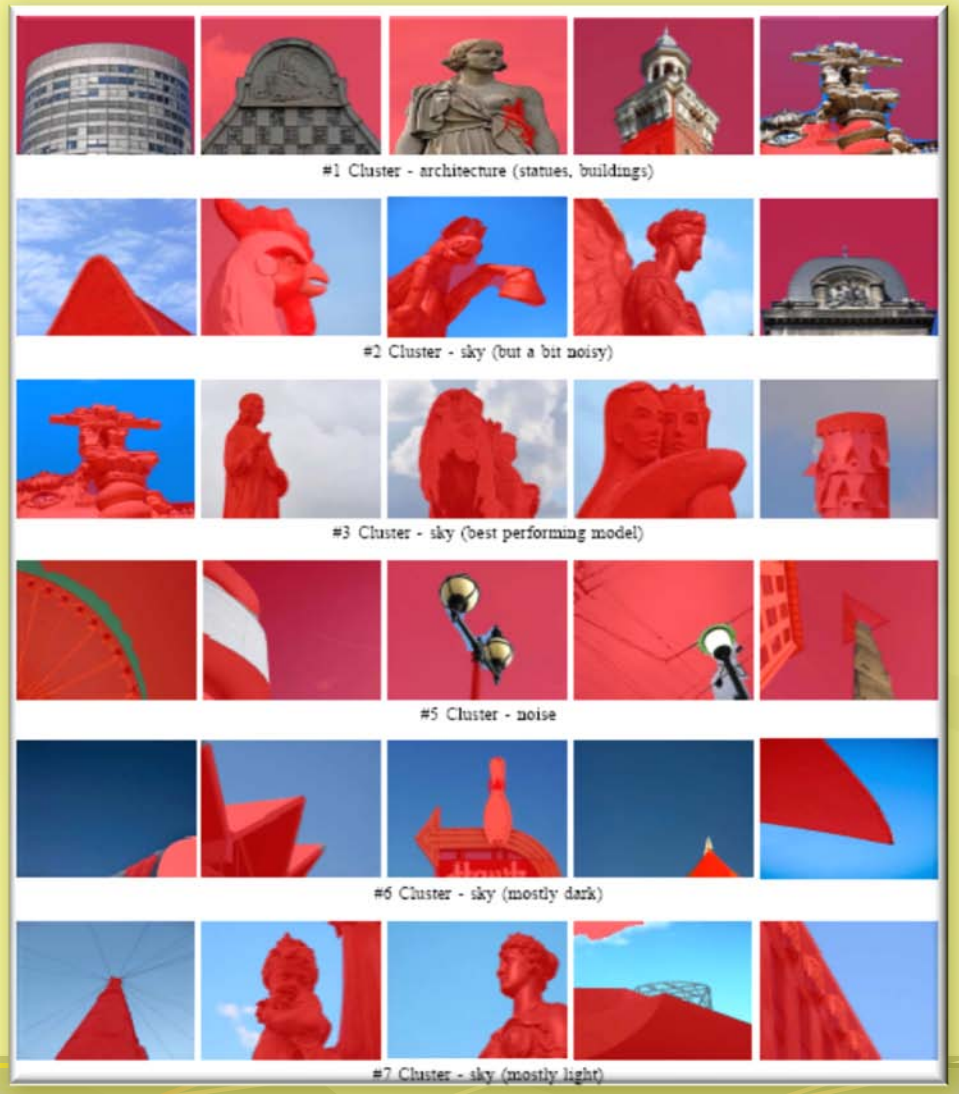
Tree object



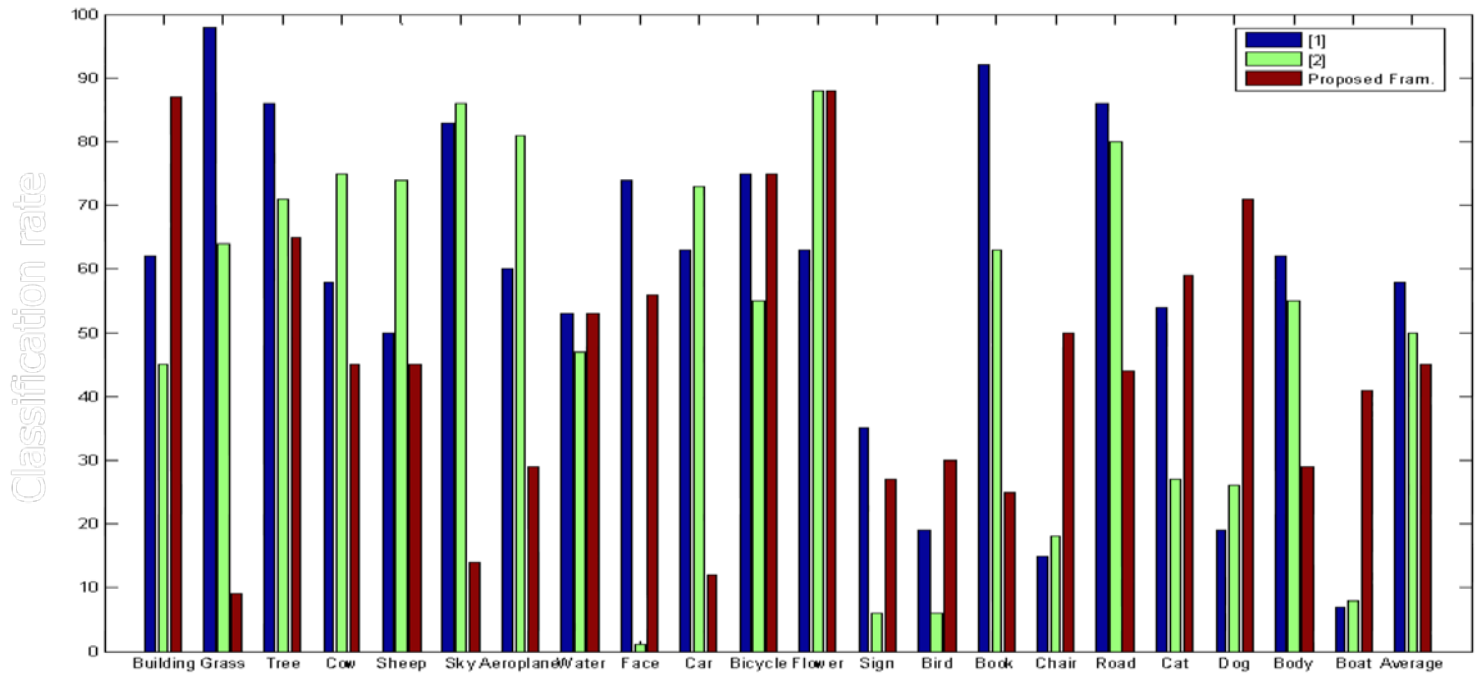
Experimental Results - MSRC vs Flickr groups

Sky object





Comparison with existing methods



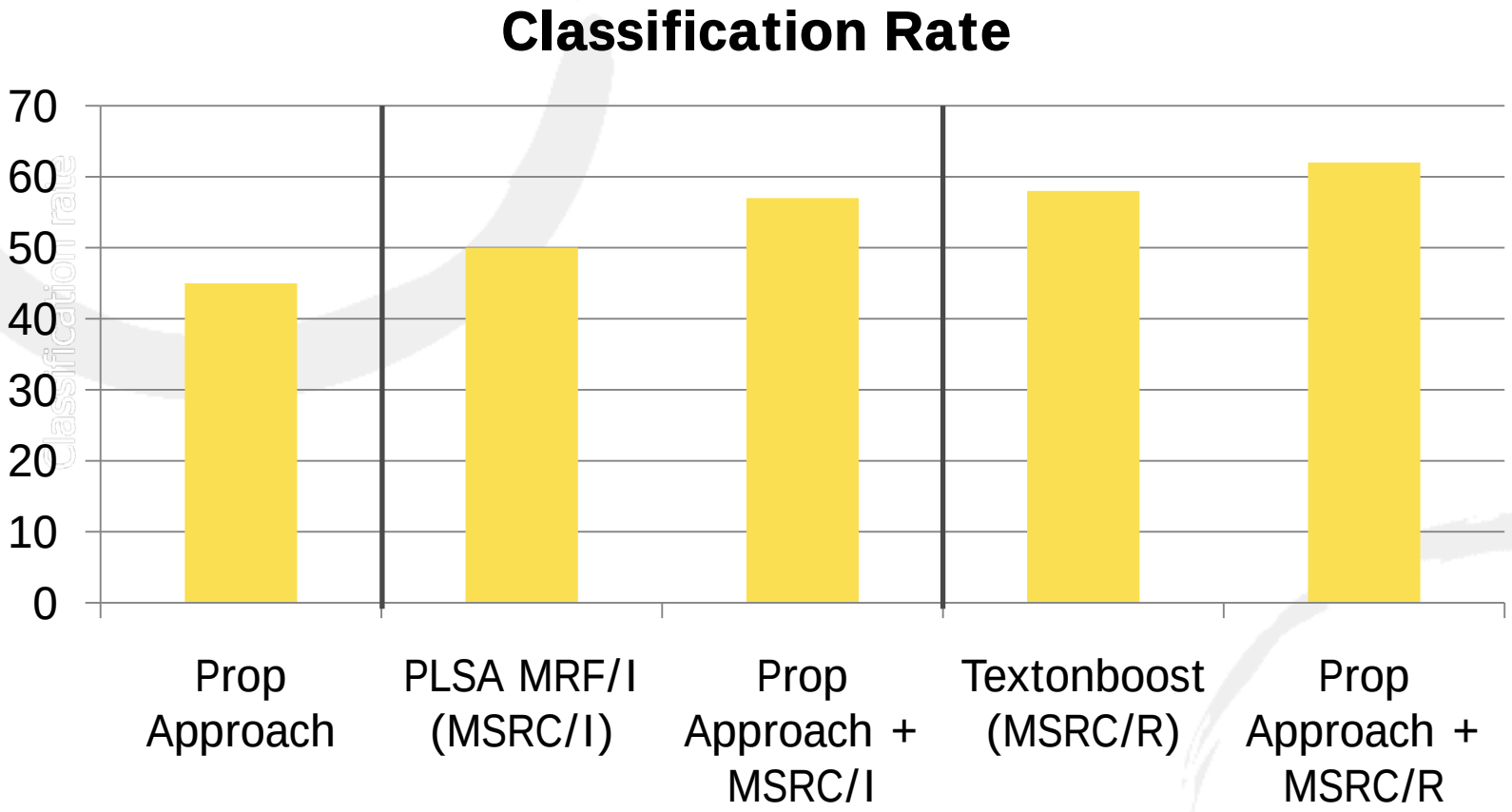
Textonboost (uses strong annotations) outperforms the other two methods comparing the average classification rate among all concepts.

Our method yields the best performance in 9 out of 21 cases, compared to 7 out of 21 for the PLSA-MRF/I and 8 out of 21 for the Textonboost (in three cases *Water, Flower, Bicycle* the classification rates are identical for two different methods).

[1] *J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in ECCV (1), 2006, pp. 1–15.*

[2] *J. J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in CVPR, 2007*

Comparison with existing methods



[1] *J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in ECCV (1), 2006, pp. 1–15.*

[2] *J. J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in CVPR, 2007*

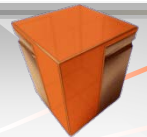
Current research

Preliminary results

Current research

Employ semi-supervised techniques

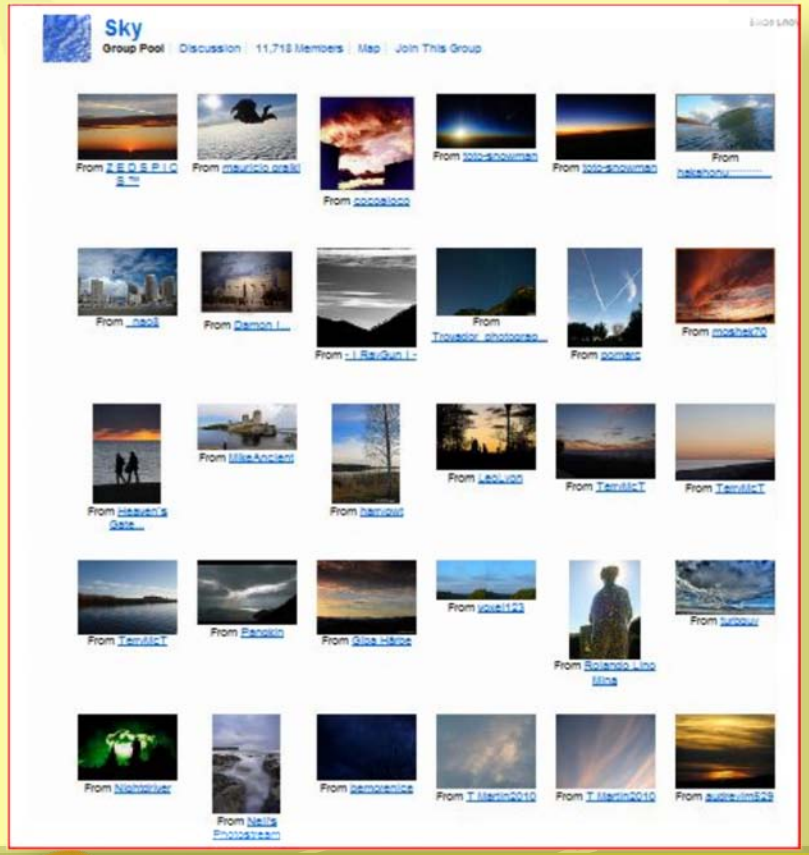
- Use a small portion of labeled data in the clustering and/or cluster selection process.
- Select and merge the proper region clusters using the labeled data as guides



Semi-Supervised selection- setup

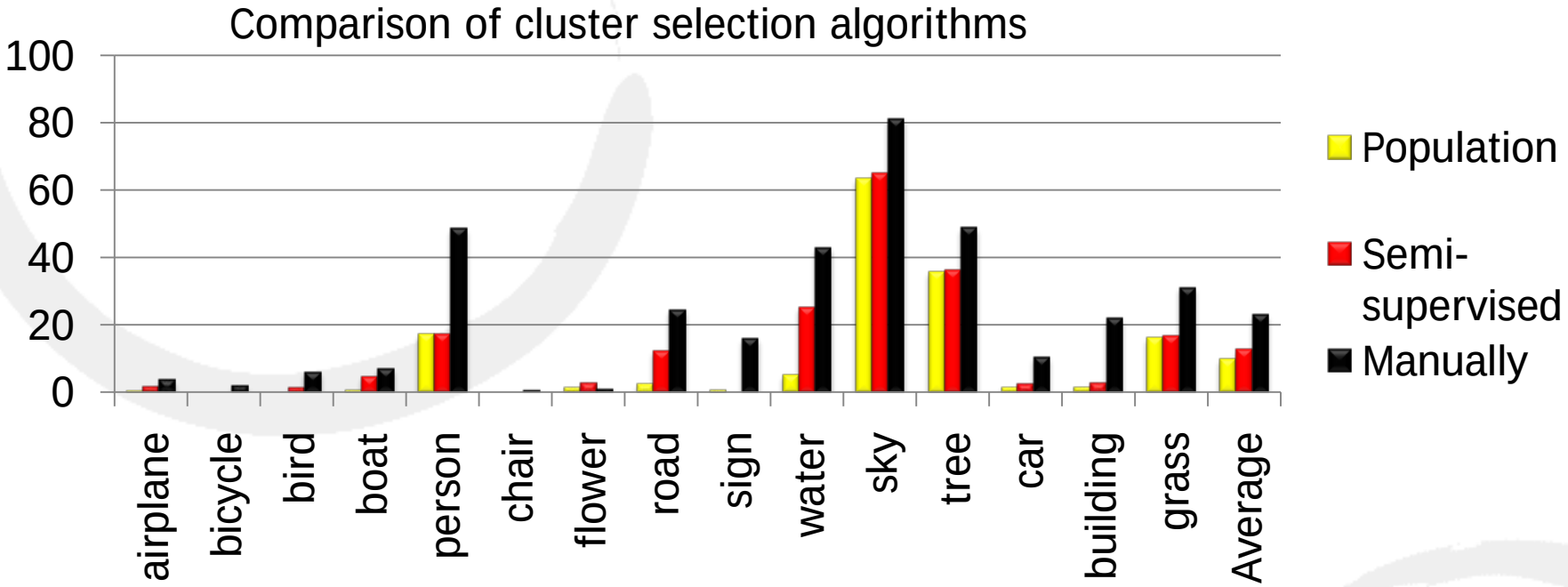
Flickr Group for object sky

Clusters of regions



Semi-Supervised selection- setup

Results – Semi-supervised cluster selection



- Dataset – SAIAPR TC-12 dataset (imageCLEF – 20k images)
- SAIAPR TC-12 dataset is split into 3 subsets (train 14k images, validation 2k images, test 4k images)
- Community detection clustering [1]
- Each cluster, formed from the flickr groups dataset, was picked and the trained model generated by the regions contained in it was evaluated on the validation set
- The best performing clusters were merged and re-evaluated
- The best performing merge of clusters was chosen to generate the final model

[1] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali. A graphbased clustering scheme for identifying related tags in folksonomies. In *DaWaK '10.*

Questions?

Thank you!