Object Discovery using CNN Features in Egocentric Videos

M. Bolaños, M. Garolera and P. Radeva
Lifelogging cameras passively acquire pictures 24-7-365. Producing data about the users' typical activities and habits.

SenseCam and Narrative wearable cameras
Lifelogging and Egocentric Vision

Lifelogging cameras passively acquire pictures 24-7-365. Producing data about the users' typical activities and habits.

Present a new field of research with a high potential:

- **Automatic nutrition diary.**
- **Memory aid** for mild cognitive impairment patients.
Focus of the Work

**Hypothesis:**
Individuals’ environment is “constructed” by sets of objects that characterize it.

**Goal:**
To develop automatic techniques for **object discovery to characterize the environment** of the person wearing the camera.

**Object Discovery ≠ Object Detection?**
Focus of the Work

**Hypothesis:**
Individuals’ environment is “constructed” by sets of objects that characterize it.

**Goal:**
To develop automatic techniques for object discovery to characterize the environment of the person wearing the camera.

Object Discovery ≠ Object Detection?
**Focus of the Work**

**Hypothesis:**
Individuals’ environment is “constructed” by sets of objects that characterize it.

**Goal:**
To develop automatic techniques for **object discovery to characterize the environment** of the person wearing the camera.

**What is the difference between Object Discovery and Object Detection?**
Baseline Object Discovery Proposal

Lee & Grauman’s “Learning the Easy Things First: Self-Paced Visual Category Discovery”

They propose to use:

a) As features for object candidates:

- LAB histograms: **colour** information.
- Pyramid HOG: **shape** information
- Spatial Pyramid Matching: **texture** information. Visual words technique based on SIFT.

b) Iterative clustering-based easy-first discovery.
Our Proposal

Discover iteratively and semi-supervisedly the most relevant classes from egocentric videos for a particular user, based on the following methodology:

- How can we apply a sampling on each image? **Objectness** object detector
- How can we represent real world objects? **CNN features**
- How can we apply a knowledge reuse? **Refill strategy**
- How can we discover new “concepts”? **Clustering**
Ego-Object Discovery Scheme (I)
Ego-Object Discovery Scheme (II)

Iterative Discovery

Easiest Objects Selection

Custering & Hard Instances Classification

Bag of Refill

Refilled Samples

Labeled Clusters

One-Class SVM

tvmonitor
Ego-Object Discovery Scheme (III)
Objectness detector

Object detection method for extracting a set of initial object candidates as an initial step for the object discovery.

Maximizes 3 complementary terms:

- **Multi-scale Saliency**

\[
MS(w, \theta_{MS}) = \sum_{p \in w} \frac{|\{p \in w | I_{MS}(p) \geq \theta_s\}|}{|w|}
\]

- **Color Contrast**

\[
CC(w, \theta_{CC}) = \chi^2(h(w), h(Surr(w, \theta_{CC})))
\]

- **Superpixels Straddling**

\[
SS(w, \theta_{SS}) = 1 - \sum_{s \in \mathcal{S}(\theta_{SS})} \frac{\min(\{|s \setminus w|, |s \cap w|\})}{|w|}
\]
We propose a set of features based on a pre-trained Convolutional Neural Network on ImageNet objects.
Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Preemptive DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about your genetic predisposition to musical aptitude of your unborn child?

Baxter: The Blue-Collar Robot

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make medical implants.

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people.

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss.

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket.

Ultra Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nano technology just might make it possible.

Big Data from Cheap Physics

Analyzing information from simple cell phones can provide surprising insights into people's behavior—and even help us understand the spread of diseases.

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical.

- visual
  - images
  - videos
- audio
  - speech
  - music
- text
  - natural language
- biology
  - Merck molecular activity
- physics?
What is a ConvNet?

- **Parameters sharing** and **pooling** take advantage of local coherence to learn invariant features.
- In its **simplest form**, a ConvNet is just a series of stages of the form:
  - convolution
  - bias
  - non-linearity (ReLu or sigmoid functions)
  - pooling
- Normalization layers (LCN) may be added.
ImageNet: 15 millions of images!!! Used to train the CNN.
Learned convolutional filters: Stage 1

9 patches with strongest activation

learned filters (7x7x96)

Strongest activations: Stage 2
Strongest activations: Stage 5
CNN Features

We propose a set of features based on a pre-trained Convolutional Neural Network on ImageNet objects.

After deleting the last layer, it can be used as a powerful feature extractor for real-world objects (output of dimension 4096).
We use an unsupervised clustering for finding groups of similar objects to label, based on three parts:

- **Agglomerative Ward Clustering:**
  - Minimum variance measure.
  - Clusters of similar size.
  
- **Silhouette Coefficient:**
  - Selects the best cluster for the user to label.
  
- **One-Class SVM:**
  - Trained on the last labeled cluster (discovers harder samples of the same class).
CNN Features

CNN improve significantly the cluster quality.
Refill strategy: useful for offering a knowledge reuse and an additional context from the objects. Consequently providing more robust clusters.
The selected easiest objects are complemented with a 20% of samples from the refill bag. They are equally divided between all the known classes except the "NoObject".

A SVM Filter is used to discard as many "No Object" as possible (i.e. FP of the Object detector).

The new (label, sample) pairs are added to the Bag of Refill. The new (label, sample) pairs are also added to the Bag of Refill.

**Ego-Object Discovery Scheme**

**Iteration: 2**

**INPUT:** photo/video sequence

- Extract Objectness Score
- ObjVSNoObj SVM filtering

**do**

- Get Easiest Objects
- Do Refill
- Get Selected Objects CNN Features
- Cluster
- Label Best Cluster
- Classify Harder Instances with OneClass SVM

**while(easyObjects)**

**OUTPUT:** object labels and locations

- 0.92 0.80 0.65
- 0.89 0.68 0.60
- 0.41 0.26
- 0.35 0.09
- 0.25 0.08
- 0.68
Used a pre-trained CNN provided by Hinton et al., trained on millions of ImageNet images.

The selected easiest objects are complemented with a 20% of labeled samples equally divided between all the known classes but the "No Object"s.

Before starting with the discovery iterations, 40% of all the object candidates are inserted into the "Previous knowledge" set.
1,000 images from a person's work day. 50,000 object candidates were extracted. To validate our method, we used the most frequent:
Tests Settings

True detection of an object candidate, if the Overlapping Score (OS) between the detected region and the Ground Truth is \( \geq 0.4 \):

\[
OS = \frac{|GT \cap \omega|}{|GT \cup \omega|}
\]

76% of the object detector samples are FPs! (“No Objects“)
Tests Comparison

Final F-Measure, Purity and Accuracy for each setting

F-Measure evolution for each different setting
Real Clusters Obtained (simplified)

hand

tvmonitor

cluster
hard samples
Extended Recent Work (I)

Egocentric Dataset of the University of Barcelona (EDUB)
- 4,912 images
- Acquired by 4 people in 8 different days
- 11,294 object labels (21 different classes).
Tested also on public datasets of intentional images.
Extended Recent Work (II)

Added **SVM Filter strategy for removing FPs** produced by the Object Detectors.

### Setting 5
- **CNN + Refill + Filter**

### Setting 3
- **CNN + Refill**

### Setting 1
- **Grauman**

<table>
<thead>
<tr>
<th>F-Measure</th>
<th>Setting 1</th>
<th>Setting 3 (ours)</th>
<th>Setting 5 (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>MSRC</em></td>
<td>0.121</td>
<td><strong>0.431</strong></td>
<td>0.410</td>
</tr>
<tr>
<td><em>PASCAL</em></td>
<td>0.002</td>
<td>0.145</td>
<td><strong>0.179</strong></td>
</tr>
<tr>
<td><em>EDUB</em></td>
<td>0.072</td>
<td><strong>0.285</strong></td>
<td>0.250</td>
</tr>
<tr>
<td><em>Average</em></td>
<td>0.065</td>
<td>0.287</td>
<td><strong>0.280</strong></td>
</tr>
</tbody>
</table>
Extended Recent Work (III)
Life-logging for MCI treatment

**Goal:** to develop tools for memory reinforcing of MCI and Alzheimer people.

To develop, for subjects with MCI, a program-based life-logging captured by a Wearable Camera recording specific autobiographical episodes for stimulating posteriorly episodic memory function known to be deficient in MCI.

To explore the association between changes in cognitive, functional and emotional outcomes.
Life-logging is not only useful for memory enhancement!

Towards lifestyle characterization
Object discovery algorithm that outperforms the state of the art and relies on:

- **Refill strategy** for providing knowledge reuse.
- **Features extraction** method using a pre-trained CNN.
- **SVM filtering** for providing a filter of FP candidates.
- Egocentric lifelogging dataset with object labels (**EDUB**).
- **Comparison on three datasets** EDUB, PASCAL and MSRC.

Future lines:

1. Discovering objects, scenes and people to characterize the user’s environment, adding also information provided by **event segmentation techniques**.
2. Algorithm code and dataset to be released soon.
Future Work

1. Propose an iterative scene and object complementary discovery for improving the context.

2. Make the method user-discriminative, being able to detect the objects that discriminate the persons environments from the egocentric data.

3. Algorithm code and dataset to be released soon.