Perception for robotics
Developmental approach

David Filliat
ENSTA ParisTech / INRIA FLOWERS

http://flowers.inria.fr
Robotics in social context

Need for learning, adaptation, human-robot interaction
Take inspiration from children

An old idea

Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education, one would obtain the adult brain [...] Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.

(Turing, 1950, “Computing Machinery and Intelligence”)
Behavioural and Cognitive Development in Human Infants

- How do developmental structures form?
- How do developmental structures impact the acquisition of novel skills?
Human cognitive development

Developmental robotics

Cognitive development in robots
Understanding the constraints guiding development to build robots

Families of developmental « forces »

Body morphology and growth :
- Morphology
- Self-organization of movement structures

Cognitive abstractions:
- Perceptual categories grounded in action
- Efficient learning in high-dimensions

Intrinsic motivation, active learning
- Autonomous collection of data
- Self-organization of developmental trajectories

Social learning, imitation
- Imitation of trajectories and goals
- Learning combinatorial motor primitives
Recognizing objects
Recognizing objects

Classical computer vision approaches

- **Structure** with parameters for prediction: \( f \)
- Sample database: \( x \rightarrow y \)
- Method to adapt parameters so that \( y = f(x) \) on training samples
- Generalization capability: \( y = f(x) \) on new samples
Representing images

Reduce representation size
  – Keep enough information
  – Reduce « noise »

Keypoint detection
  – Invariant to scale and orientation

SIFT, SURF, MSER
...
Representing images

Storing approximate appearance + position
– Bag of Words + Spatial Pyramid Matching
Representing images

Learning representation : deep learning
Recognizing objects

PASCAL - VOC challenge 2012
  – Segment objects
  – 20 categories
Recognizing objects

Very difficult task: VOC challenge 2012 - segmentation
Recognizing objects

Very difficult task: deep learning on imagenet

<table>
<thead>
<tr>
<th>Team name</th>
<th>Entry description</th>
<th>Description of outside data used</th>
<th>Number of object categories won</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet</td>
<td>Ensemble of detection models. Validation is 44.5% mAP</td>
<td>Pretraining on ILSVRC12 classification data.</td>
<td>142</td>
<td>0.439329</td>
</tr>
<tr>
<td>CUHK DeepID-Net</td>
<td>Combine multiple models described in the abstract without contextual modeling</td>
<td>ImageNet classification and localization data.</td>
<td>29</td>
<td>0.406659</td>
</tr>
<tr>
<td>Deep Insight</td>
<td>Combination of three detection models</td>
<td>Three CNNs from classification task are used for initialization.</td>
<td>27</td>
<td>0.404517</td>
</tr>
</tbody>
</table>

Surprising erroneous classification
Recognizing objects

Current limitations

– Still limited performances (vs human)
– Human-defined categories
– Need large image databases
  • 20 000 images / 50 000 objects
  • Collaborative labeling
– Learning is not incremental/continuous/active
Taking inspiration from children
Taking inspiration from children

Learn incrementally
  – No strong supervision (database)
  – Observe around: recognize / learn new
  – Interact with others

Perform experiences
  – Manipulate object
  – Create useful categories
    (self / other / objects)
  – Use language
Taking inspiration from children

Learning by observation

Learning by manipulation
What is an object?

Categorize shadows / objects / humans / myself

Recognize objects individually / their function

→ Proto-objects
Developmental learning of objects

1. Proto-objects

2. Learning appearance

3. Multi-view learning

4. Categorization

Caméra RGB-D

Joints

Robot model

Object model

Human model
1. Segmentation

Proto-objets

– Object-like unit of attention
– Detection through motion and appearance
– Segmented from depth data

RGB-D sensor

 Depth

Motion detection
KLT-tracking and clustering

Depth data
Depth contours

Proto-objects
2. Appearance Learning

Low level complementary features

Intermediate level local geometry

Bag of pairs recognition
2. Appearance Recognition

• Voting Method using mid-features
3. Multiple view learning

Add views to current model

– After recognition
– During tracking
Developmental learning of objects
Learning through manipulation

Objectives
- Improve segmentation
- Improve appearance models
- Discover physical properties
- ≠ sensori-motor theories

Requirements
- Capability to perform actions (knowledge of space)
- Recognize self / other / objects
4. Categorization

• Robot category
  – has high mutual information between the sensory data and proprioception

  \[
  MI(Lc; Ac) = Hc(Lc|Ac) - H(Lc)
  \]

  \[
  H(Lc) = - \sum_l p(l) \log(p(l)),
  \]

  \[
  Hc(Lc|Ac) = - \sum_a p(a) \sum_l p(l|a) \log(p(l|a)),
  \]

• Object category
  – static and independent on robot motors, when it is single,
  – can move, when it is connected to another entity

• Human category
  – independent on robot motors in all cases,
  – can move in all cases
4. Categorization

- Configuration of a robot arm
  - Localization of a physical entity
  - Connected entity?
    - \( M_{l_{c1}} < th_c \)
      - \( c_4 \) was identified?
        - \( p(c_4 = c_o) > th_c \)
          - robot category (\( c_o \))
          - unknown category (\( c_o \))
          - human category (\( c_h \))
          - object category (\( c_o \))
          - object + robot category (\( c_{o+r} \))
          - object + human category (\( c_{o+h} \))
    - \( M_{l_{c2}} < th_c \) & \( M_{l_{c1}} > th_c \)
      - human model
      - robot model
      - object model

- Human model
4. Categorization

For connected proto-objects
- Subtract robot views
- Update objects with remaining features
Developmental learning of objects
Developmental learning of objects

Object recognition rate

Learning through observation (blue color), TakeLiftFall manipulation (yellow color).
Improving skills autonomously and socially
Children always do something
Intrinsic motivation, curiosity and active learning

Hull (1943), White (1959)

Basic forms of motivations (e.g. food, reproduction...) can not account for the whole diversity of spontaneous exploratory behaviours of humans.

→ Intrinsic drive to reduce uncertainty, and to experience novelty
→ Interest = optimal difficulty = neither trivial nor too difficult

« Intelligent Adaptive Curiosity »

Errors in prediction in 4 activities

% of time spent in each activity

Autonomous creation of a developmental trajectory
Interacting with caregivers

Learning by observation
+ asking for specific object

Learning by manipulation
+ choosing an action
Active choice of objects, actions, actor
Curiosity-driven learning of objects

LEARNING TO RECOGNIZE OBJECTS THROUGH CURIOUSITY-DRIVEN MANIPULATION

Sao Mai Nguyen, Serena Ivaldi, Natalia Lyubova, Alain Droniou, Damien Gérardeaux-Viret, David Filliat, Vincent Padois, Olivier Sigaud, Pierre-Yves Oudeyer

David Filliat – ENSTA ParisTech / INRIA FLOWERS

(Nguyen et al.; ICDL-Epirob 2013; Ivaldi et al., IEEE TAMD 2013)
Curiosity-driven learning of objects

- Interact with human if useful
- Focus on difficult objects
Learning Visual Saliency
Visual Saliency

Object detection and recognition:
A major challenge in computer vision

- Most efficient algorithms
  - Offline
  - Supervised

- Developmental robotics constraints
  - Incremental
  - Autonomous
  - Active (exploration)

The IMAGENET challenge
Motivations

Human vision
  – Foveal/peripheral field of view
  – Active perception

Computer vision
  – Each pixel has the same resolution
  – Image is processed as a whole

Robotics vision
  – Computer vision algorithm are used
  – Action to improve perception
Motivations

Computer vision : Saliency map
- A common way to localize areas of interest in images
- Most available methods are static and not task-related

Human visual saliency
- Depends on people background (learned)
- Is most of the time task-related
Goals

A mechanism providing visual saliency maps
- Based on learning
- Task-related: Indoor object detection

Using developmental robotics mechanisms
- Incremental learning
- Guiding learning through action selection

Using foveal and peripheral vision alternance
- The fovea provides accurate and reliable data
- This data is used to learn saliency in the periphery
Proposed approach: General architecture

Exploitation mode

![Image](image1.png)

RGB-based feature extraction → Classifier → Saliency

Exploration mode

![Image](image2.png)

Foveal target selection → Intrinsic motivation criterion

Contextual view (RGB) features → Saliency learning (Classifier)

Saliency label → Foveal view (RGB-D) → Saliency discovery
Proposed approach: Salient elements discovery

- Salient elements are objects lying on plane surfaces
- Processing is too expensive to be processed on the whole frame
Proposed approach: Saliency learning

RGB-based feature vector

Local feature extraction

Local label extraction

“SALIENT”

“NOT SALIENT”

Classifier (Random forest) update
Proposed approach: Intrinsically motivated exploration

Select the fovea with an intrinsic motivation criterion

Uncertainty

- Select the foveal area based on saliency map fuziness

Novelty

- Select the foveal area where samples are far from the dataset
- Distance based on random forest sample proximity [2]
Experimental results: Comparison to state of the art


- Ground truth manually determined
- Compared with 4 methods

Visually

- Our method looks most like ground truth
- Provides rough idea of object shape
Experimental results: Comparison to state of the art
Experimental results: Comparison to state of the art

Saliency learning evolution

ROC curves on still images

ROC curves on video
Experimental results: Intrinsic motivation criteria

On a video sequence

- Outperform static method (BMS)
- Rnd: random, worst case
- Err: Learning is guided by an oracle (ideal case)
- Uncertainty: the most efficient criterion
- Some portions of the sequences are harder to learn
Application on a mobile robot

Segmentation

- Geometrical consideration
- Salient = Object lying on plane surface
- Accurate
- Slow and partial

Learning

- Features : RGB-based
- Labels : based on the segmentation result
Experimental results

State-of-the-art

Segmentation vs saliency
Exploration strategies

The robot moves in its environment
  – how to explore better than randomly?

Early strategies: Select samples in the input image
  – Based on novelty or uncertainty scores (ICDL)
  – No displacement strategies

New approach: IAC based
  – Divide the room to be explored into regions based on robot position and orientation
  – Robot moves to regions where saliency learning progress is the highest
IAC-based exploration
Results

Evolution of the error rate
Learning nouns and adjectives
Learning object names and colors

Challenges

– Associate each word to the right feature
– Cross-situational learning
– Benefit from natural behaviors of adult teachers
Learning object names and colors

Encoding objects and speech
  – Fixed size vector for each feature

Decompose data to discover “topics”
  – Explain data with few topics: \( V = \sum w_i h_i \)
Topic discovery

Find underlying topics that explain data
- Reduce data dimension
- Many existing techniques
- We use Non negative Matrix factorization

\[
\begin{align*}
S & \quad S \\
C & \quad C \\
T & \quad T \\
\end{align*}
\]

\[
S \times C \times T
\]

\[
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
1 & 1 \\
\end{bmatrix}
\]

Learn \[\geq 0\]
Topic discovery

Test on new objects
- Decompose using vision -> predict speech

New object

Predicted names

Fixed
Topic discovery

Learning during interaction with robot
- Training during 10 mins ($\approx$ 50 samples)
- Test several robot behaviors
Topic discovery

Topics -> Words + definition

- Good results if number of topics is known
Providing feedback to humans

Robot curiosity can guide humans
– Look at most uncertain objects
Providing feedback to humans

Was interaction pleasant?

<table>
<thead>
<tr>
<th></th>
<th>Curiosity</th>
<th>No Curiosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>No tests</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Error does not evolve as intuition would suggest

→ change learning approach
Summary

Recognizing objects is difficult
• Need large database
• Is not incremental
• Performances still limited

Take inspiration from children
• Use computer vision method
• Take inspiration from behaviour
• Gather samples incrementally
• Learn and categorize
• Actively get new samples
• Cross-situational learning