Interprétation d'images et robotique

École Nationale Supérieure de **Techniques Avancées**

David Filliat
Unité Informatique et Ingénierie des Systèmes
Equipe ENSTA – INRIA FLOWERS
`david.filliat@ensta-paristech.fr`

ENSTA ParisTech
université PARIS-SACLAY

Plan du cours

Introduction

- Robotique développementale
- Modèles de sacs de mots visuels

Cartographie - Localisation

- Application des sacs de mots visuels

Apprendre à interpréter des images

- Distinguer soi / non soi
- Modéliser soi/objets/humain

Apprendre à chercher des objets

- Apprentissage de saillance visuelle

Apprendre à éviter des obstacles

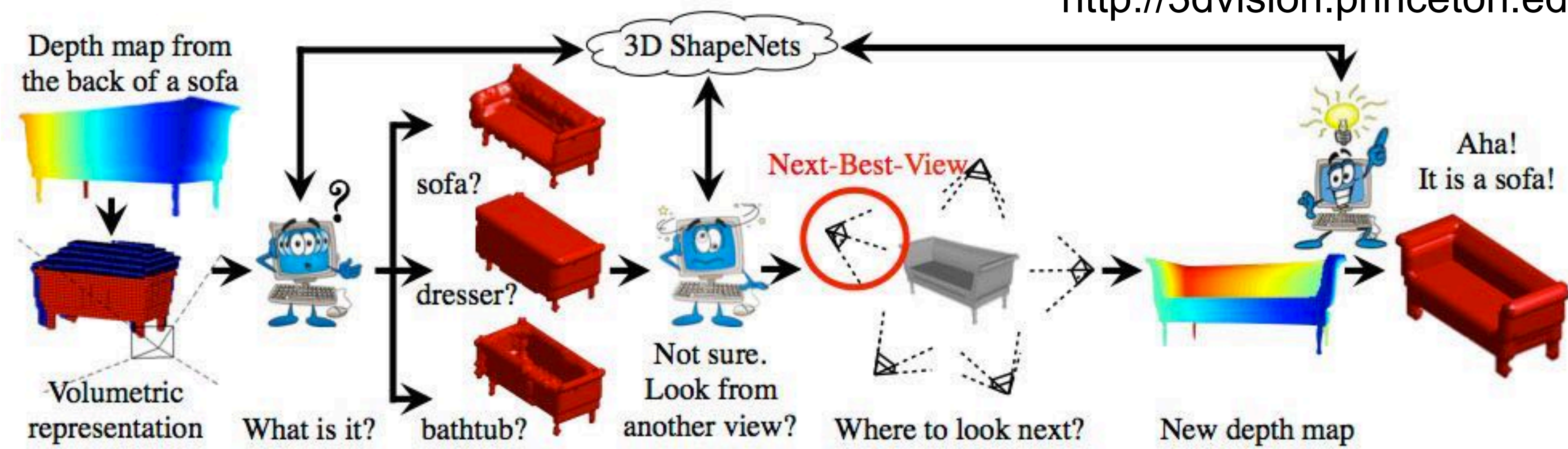
- Prédiction de profondeur en video monoculaire

Interprétation d'images et Robotique ?

Le robot peut agir

- Percevoir pour agir : asservissement visuel, cartographie, ...
- Agir pour percevoir : Choisir un point de vue pour simplifier la vision, vision active, ...
- Obtenir une information de supervision : caractéristiques invariantes, …

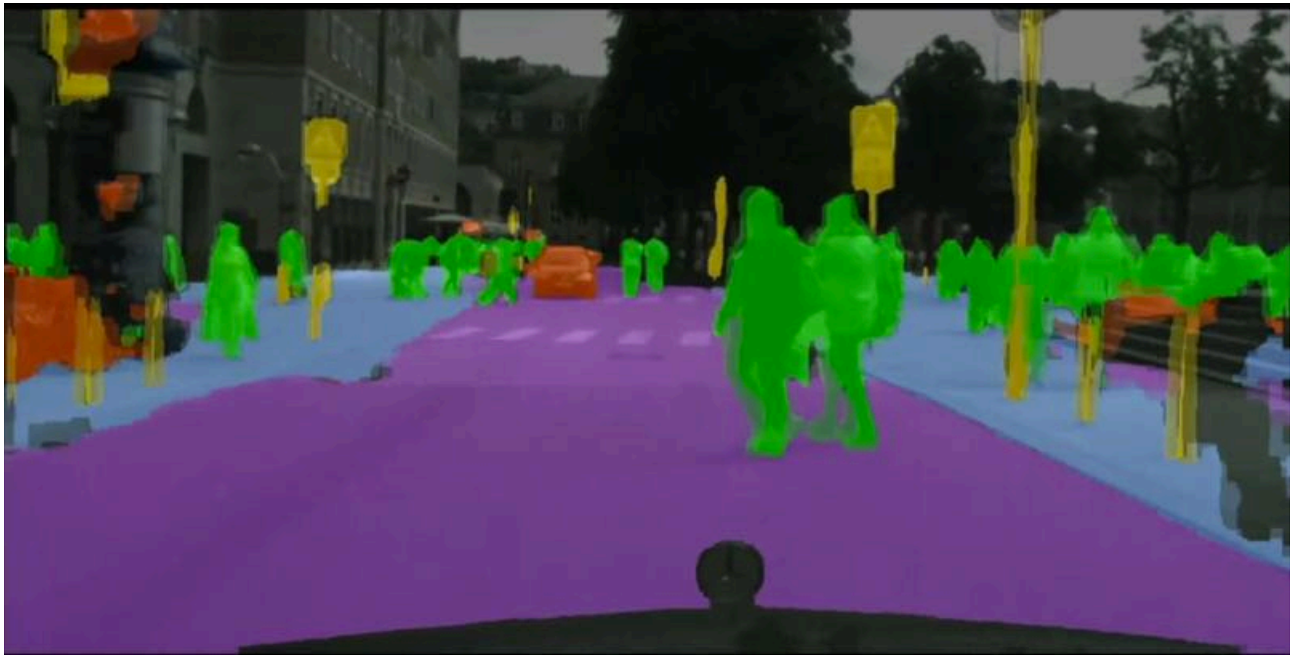
http://3dvision.princeton.edu/



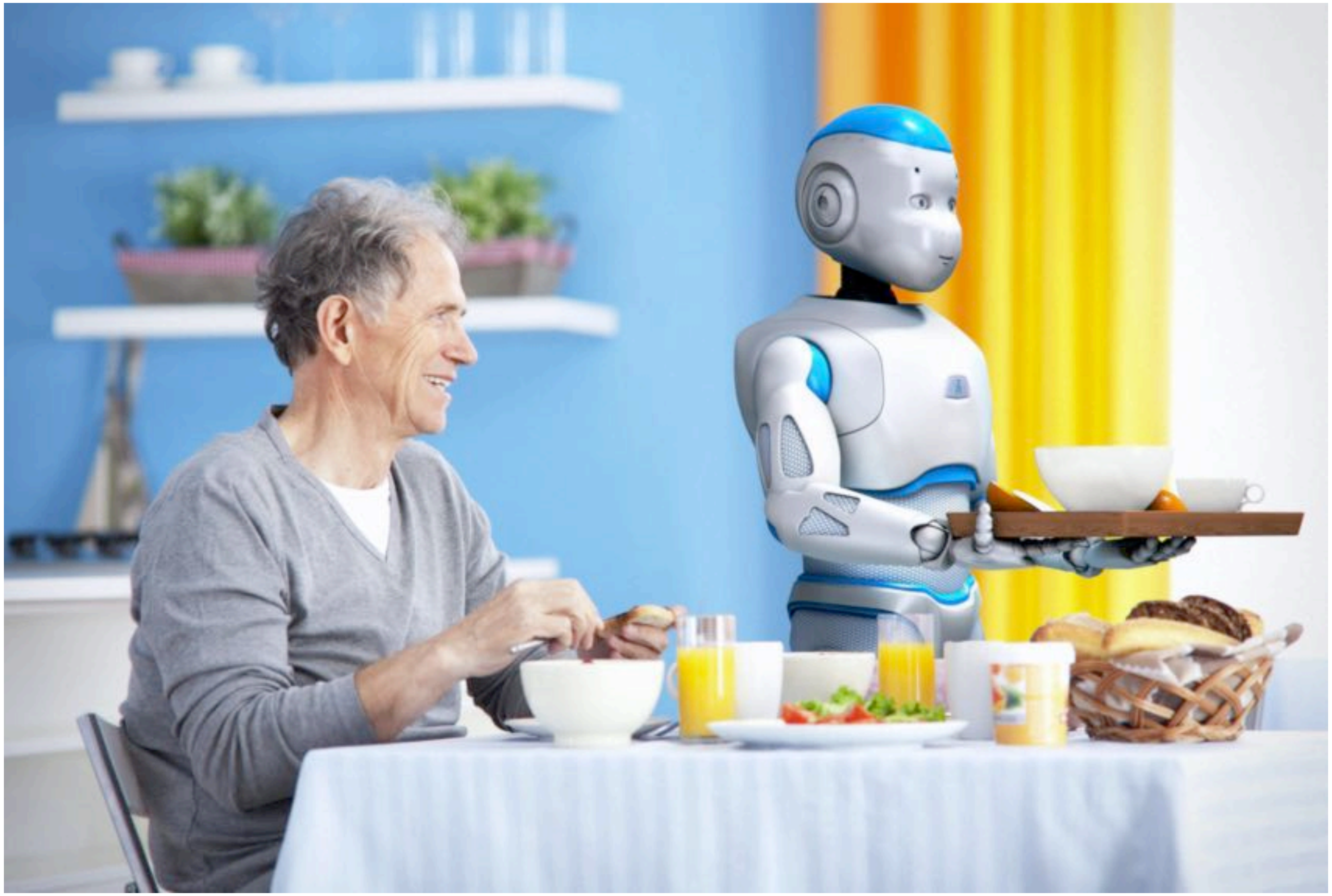
Interprétation d'images et Robotique ?

Interprétation d'image en robotique

- Nombreuses applications pour le robot ou sa mission
- Localisation, Guidage, Cartographie
- Reconnaissance d'objets, Recherche d'objets
- Robotique de service, véhicules intelligents, drones…



Robotique dans un contexte social



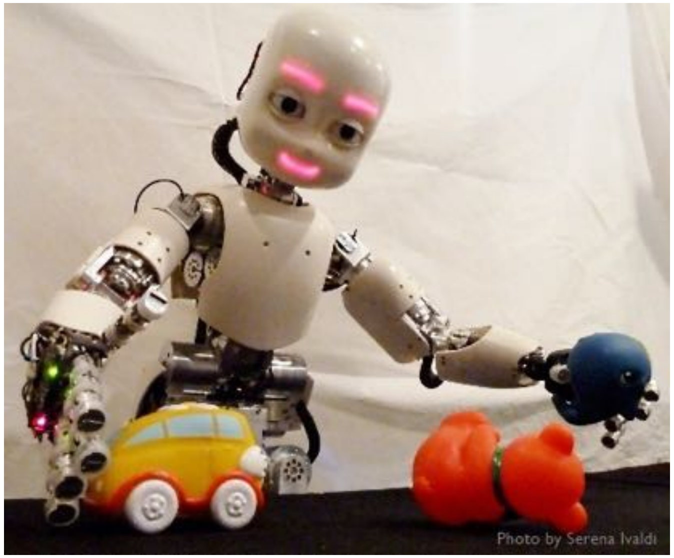
Besoin d'apprentissage, d'adaptation, d'interaction

S'inspirer des enfants

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)



Robotique développementale

Apprentissage de compétences sensori-motrices et sociales:

- **de manière autonome**
- **ouvert, sur le long terme**
- **dans le monde réel, physique et social**

➔ Validation expérimentale

Compréhension des mécanisme fondamentaux du développement

Application à la robotique d'assistance

Contraintes de développement

Intrinsic motivation, active learning

- ***Autonomous collection of data***
- Efficient learning
- Self-organization of developmental trajectories

Social learning, imitation

- Imitation of trajectories and goals
- Learning combinatorial motor primitives
- Optimal teaching

Cognitive abstraction

- ***Perceptual categories grounded in action***
- Active goal babbling, macro-actions, macro-states
- Efficient learning in high-dimensions

Body morphology and growth

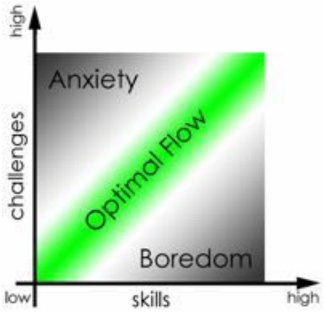
- Morphology
- Self-organization of movement structures
- Self-organization of maturational schedule

Motivations intrinsèques

Motivations intrinsèques

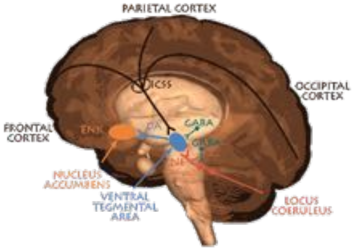
Mécanismes de l'exploration spontanée chez les enfants

Psychologie développementale



White (1959), Berlyne (1960), Csikszentmihalyi (1996)

Neurosciences



Dayan and Belleine (2002), Kakade and Dayan (2002), Horvitz (2000)

➔ Les **motivations intrinsèques** poussent les humains à explorer des activités de complexité/nouveauté/difficulté intermédiaire pour elle-même, grâce à un mécanisme de de régulation active de la croissance de la complexité

Motivations intrinsèques

Modèle « Intelligent Adaptive Curiosity »

Oudeyer P-Y, Kaplan , F. and Hafner, V. (2007)
Intrinsic Motivation Systems for Autonomous Mental Development,
IEEE Transactions on Evolutionary Computation, 11(2), pp. 265--286.

Motivations intrinsèques

Exemple de fonctionnement de la Curiosité Intelligente Adaptative (IAC)

Impossible d'afficher l'image. Votre ordinateur manque peut-être de mémoire pour ouvrir l'image ou l'image est endommagée. Redémarrez l'ordinateur, puis ouvrez à nouveau le fichier. Si le x rouge est toujours affiché, vous devrez peut-être supprimer l'image avant de la réinsérer.

Développement de la perception

Développement de la perception

Reconnaître

- Des visages
- Des objets
- Des catégories
- Des affordances
- Des lieux
- ….

Apprendre

- De nouveaux éléments
- Qu'est-ce qu'un objet ?

Reconnaître des objets

Exemples d’erreurs (VOC, 2009)

Impossible d'afficher l'image. Votre ordinateur manque peut-être de mémoire pour ouvrir l'image ou l'image est endommagée. Redémarrez l'ordinateur, puis ouvrez à nouveau le fichier. Si le x rouge est toujours affiché, vous devez peut-être supprimer l'image avant de la réinsérer.

Impossible d'afficher l'image. Votre ordinateur manque peut-être de mémoire pour ouvrir l'image ou l'image est endommagée. Redémarrez l'ordinateur, puis ouvrez à nouveau le fichier. Si le x rouge est toujours affiché, vous devez peut-être supprimer l'image avant de la réinsérer.

Adversarial examples in deep-learning (OpenAI, 2017)

Impossible d'afficher l'image. Votre ordinateur manque peut-être de mémoire pour ouvrir l'image ou l'image est endommagée. Redémarrez l'ordinateur, puis ouvrez à nouveau le fichier. Si le x rouge est toujours affiché, vous devez peut-être supprimer l'image avant de la réinsérer.

Reconnaître des objets

Limitations des méthodes supervisées

- Catégories définies a priori
- Besoin de bases d'exemples
  - ImageNet : Millions d'images / milliers d'objets
  - Annotation via le web
- Séparation apprentissage/utilisation

Alternative ?

- Apprentissage incrémental / en ligne
- Expérimentation / Supervision sociale
- Approche développementale

Approche développementale

Principes

- Inspiré des enfants
  - Données utilisées
  - Développement :

    Interaction sociale / identification de soi / expérimentations

- Apprentissage en-ligne, incrémental, non supervisé
  - Pas de bases de données
  - Pas d'objets pré-définis
  - Pas de détecteur spécialisé (peau, visages, markers)

Plan du cours

Introduction
- Robotique développementale
- **Modèles de sacs de mots visuels**

Cartographie - Localisation
- Application des sacs de mots visuels

Apprendre à interpréter des images
- Distinguer soi / non soi
- Modéliser soi/objets/humain

Apprendre à chercher des objets
- Apprentissage de saillance visuelle

Apprendre à éviter des obstacles
- Prédiction de profondeur en video monoculaire

Représentation des images

Représenter les images

- Réduire la taille des représentations
- Conserver l'information pertinente
- Diminuer le « bruit »

Détecteurs de points d'intérêt

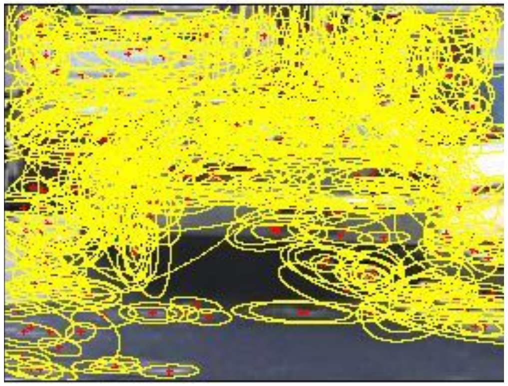
- Robuste aux changement d'échelle et l'orientation

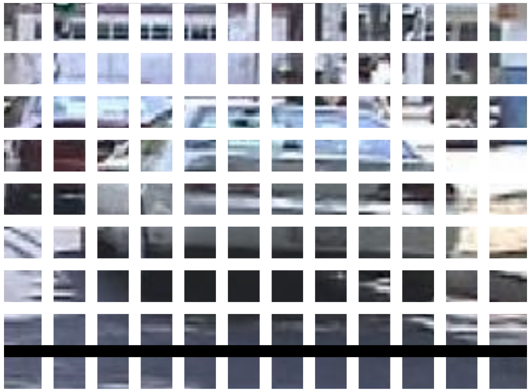
Descripteur

SIFT, SURF, MSER

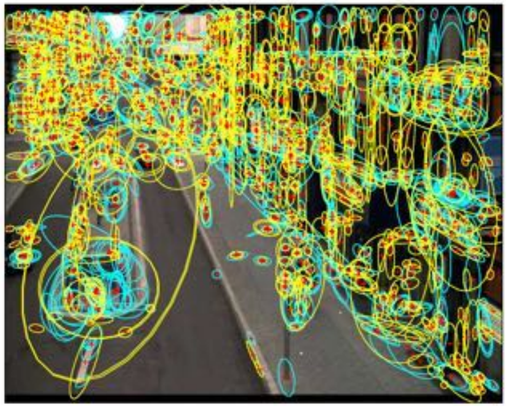
…

Echantillonage


Sparse, at interest points


Dense, uniformly


Randomly

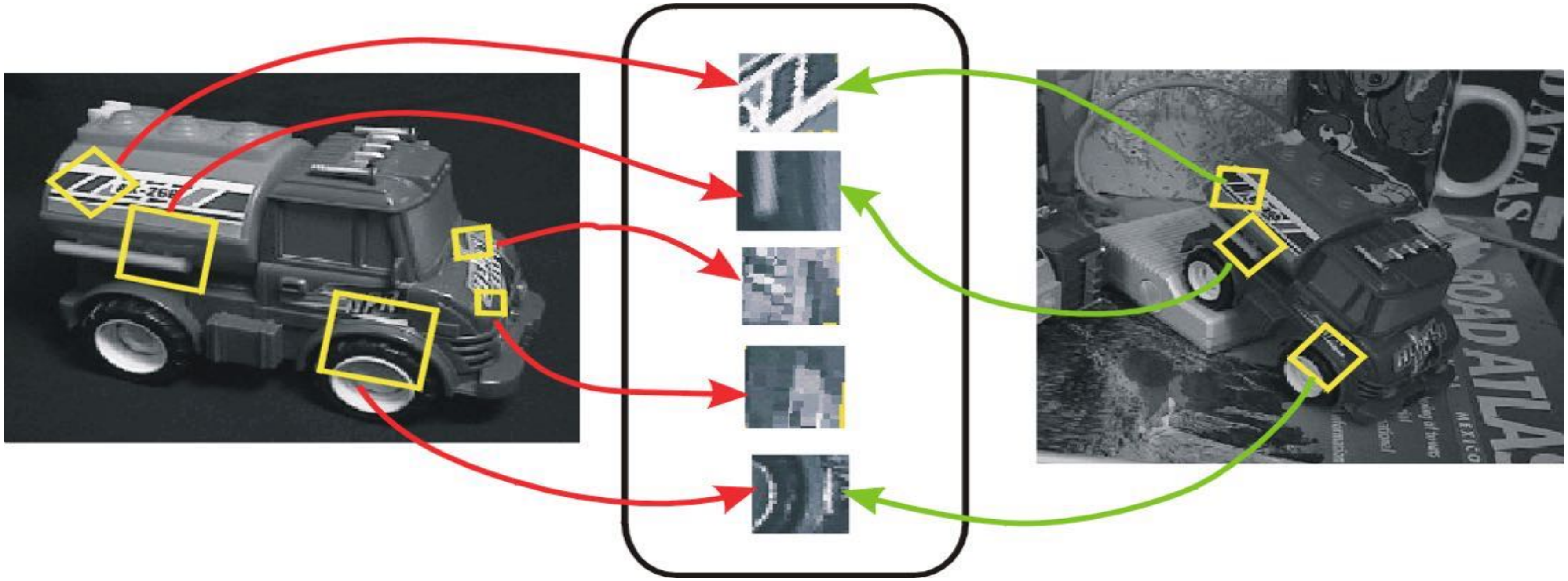

Multiple interest operators

- To find specific, textured objects, sparse sampling from interest points often more reliable.
- Multiple complementary interest operators offer more image coverage.
- For object categorization, dense sampling offers better coverage.

[See Nowak, Jurie & Triggs, ECCV 2006]

Indexation de caractéristiques locales

Possibilité de créer un index pour comparer des images ?



Slides de K. Grauman, B. Leibe

Index Inversé

Index

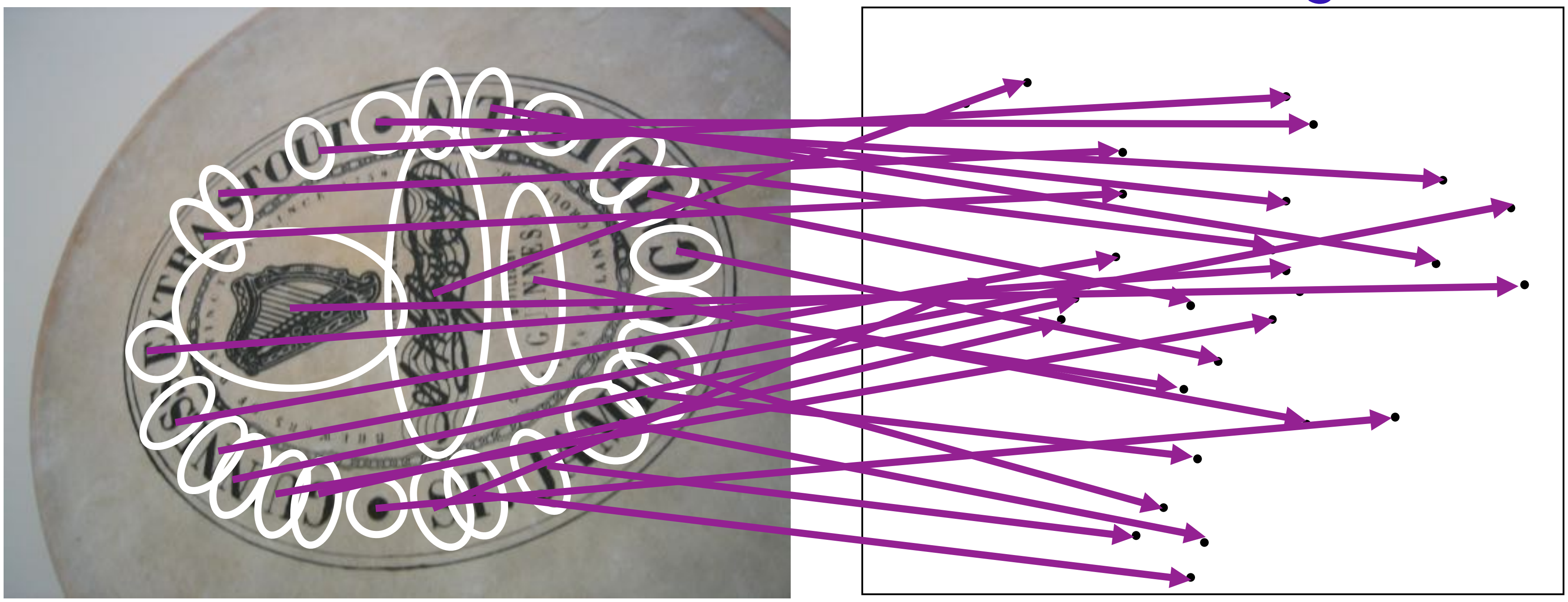
For text documents, an efficient way to find all *pages* on which a *word* occurs is to use an index…

We want to find all *images* in which a *feature* occurs.

To use this idea, we'll need to map our features to "visual words".

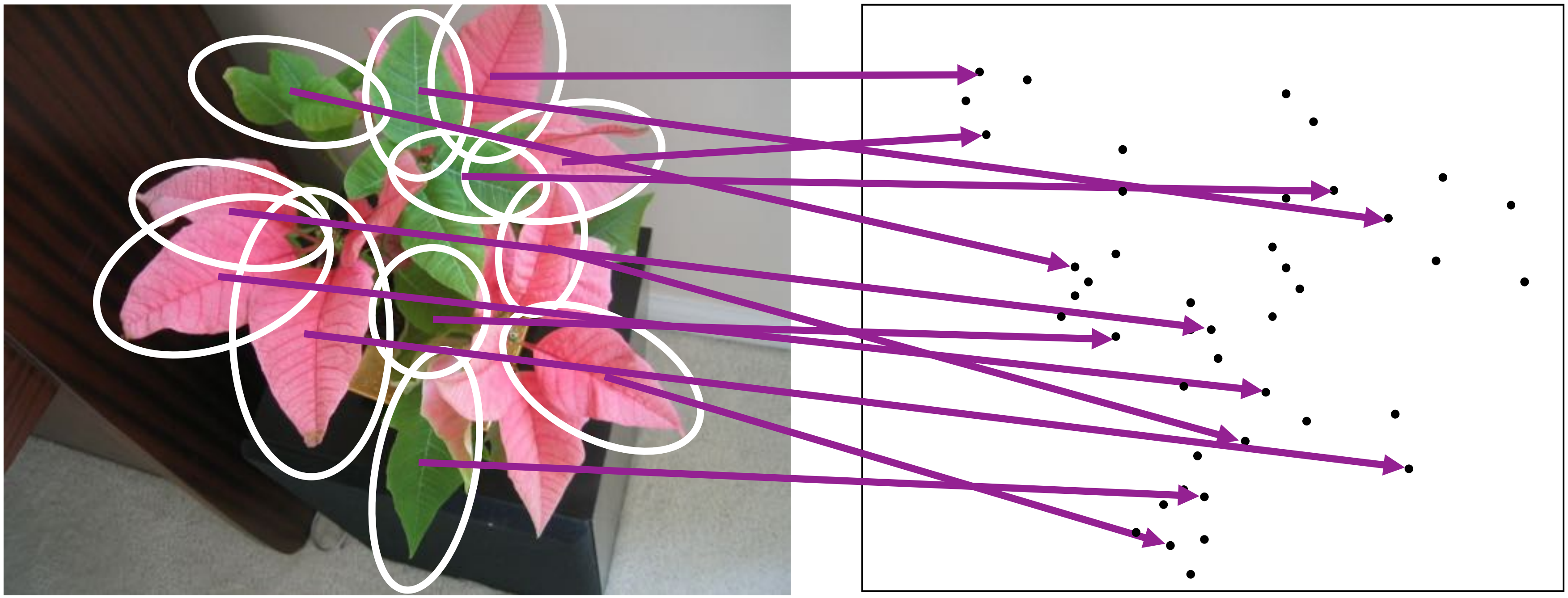
Mots visuels

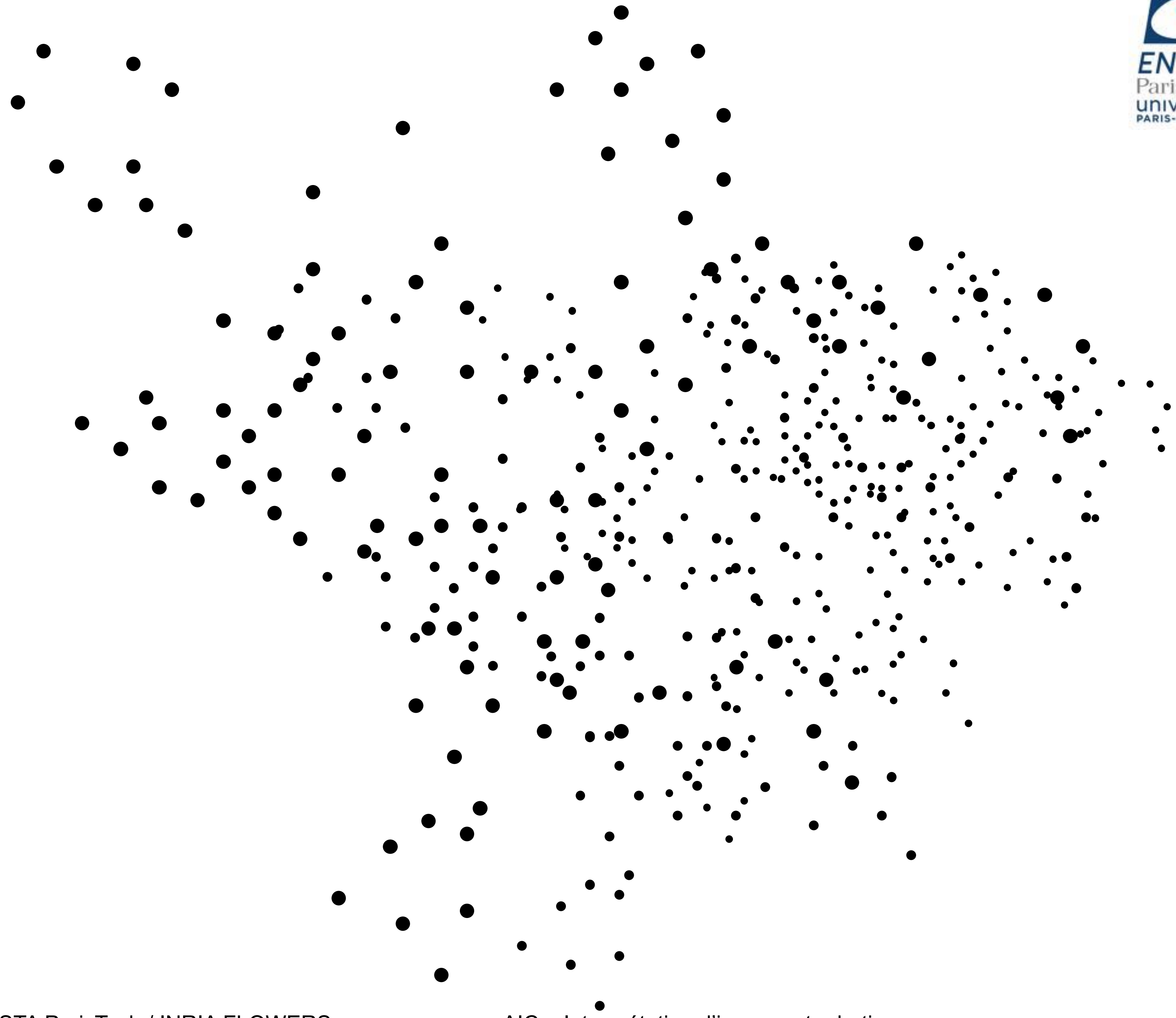
Extract some local features from a number of images …

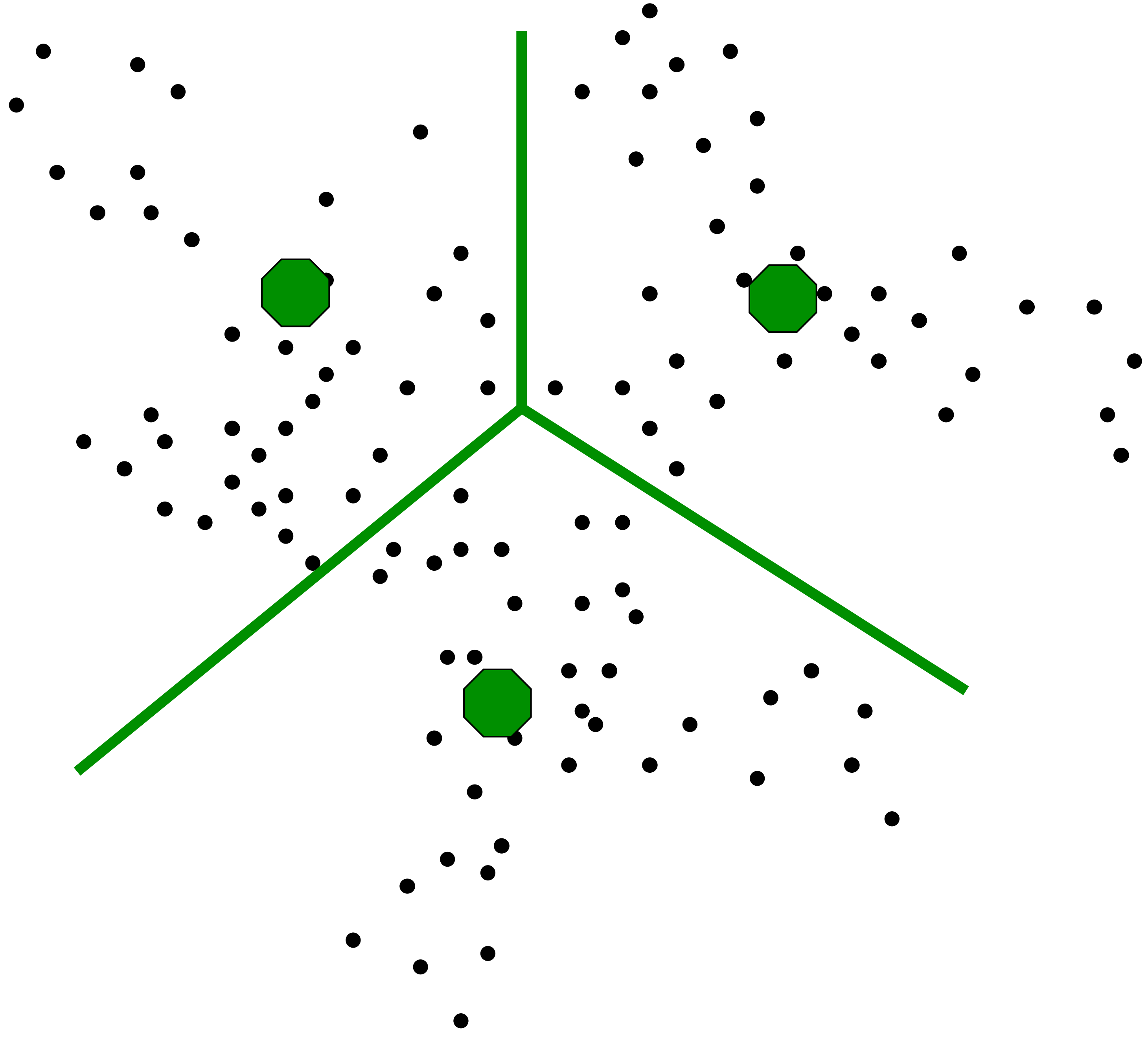


e.g., SIFT descriptor space: each point is 128-dimensional

Mots visuels

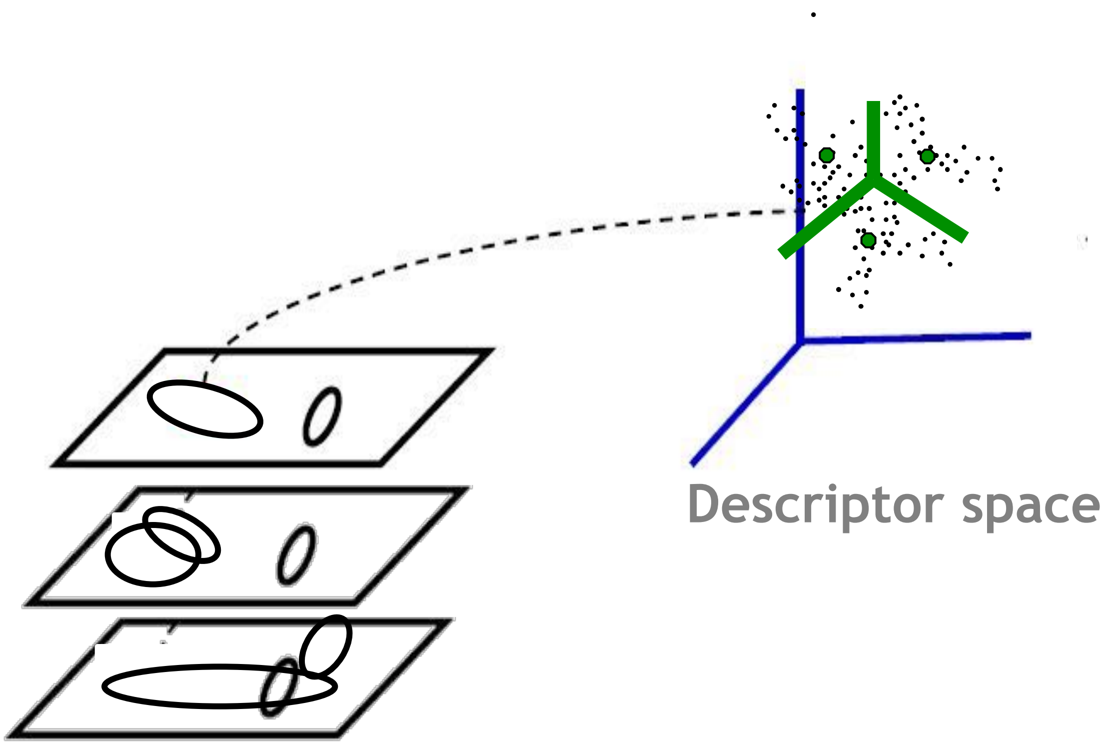






Mots visuels

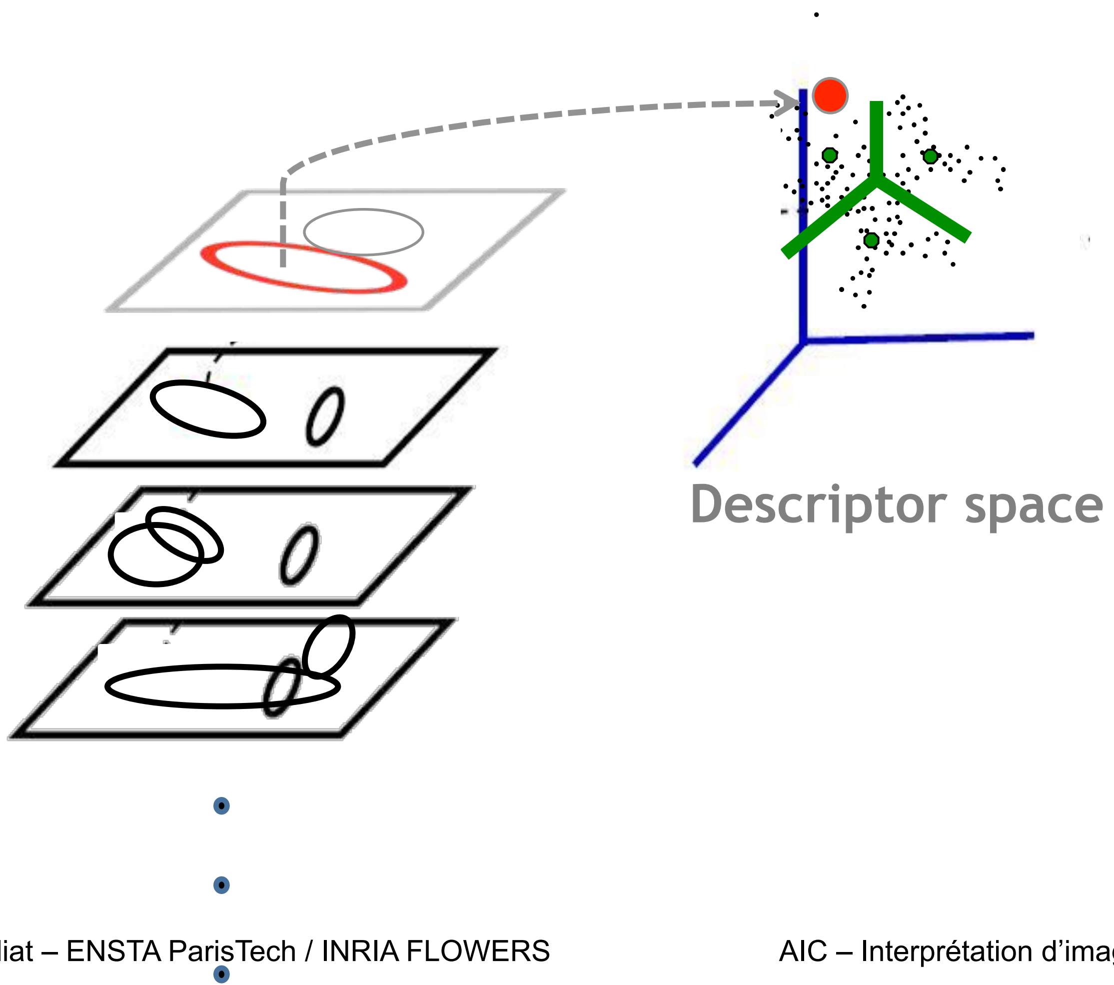
Map high-dimensional descriptors to tokens/words by quantizing the feature space



- Quantize via clustering, let cluster centers be the prototype "words"

Mots visuels

Map high-dimensional descriptors to tokens/words by quantizing the feature space



- Determine which word to assign to each new image region by finding the closest cluster center.

Mots visuels

Example: each group of patches belongs to the same visual word

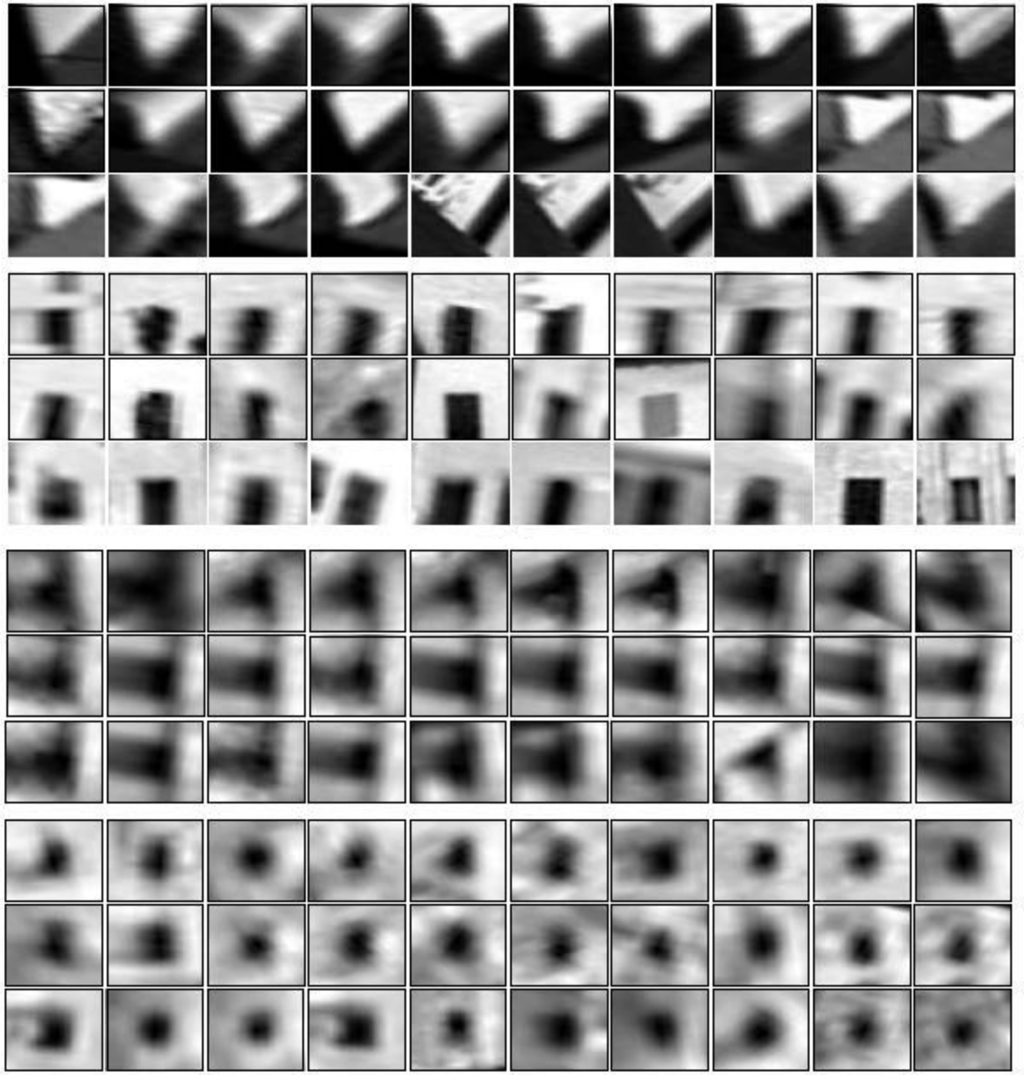


Figure from Sivic & Zisserman, ICCV 2003

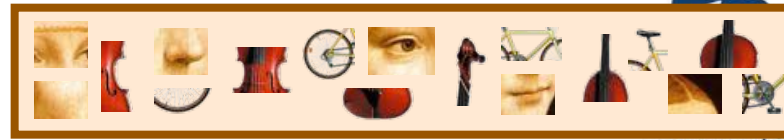
Index Inversé

frame #5

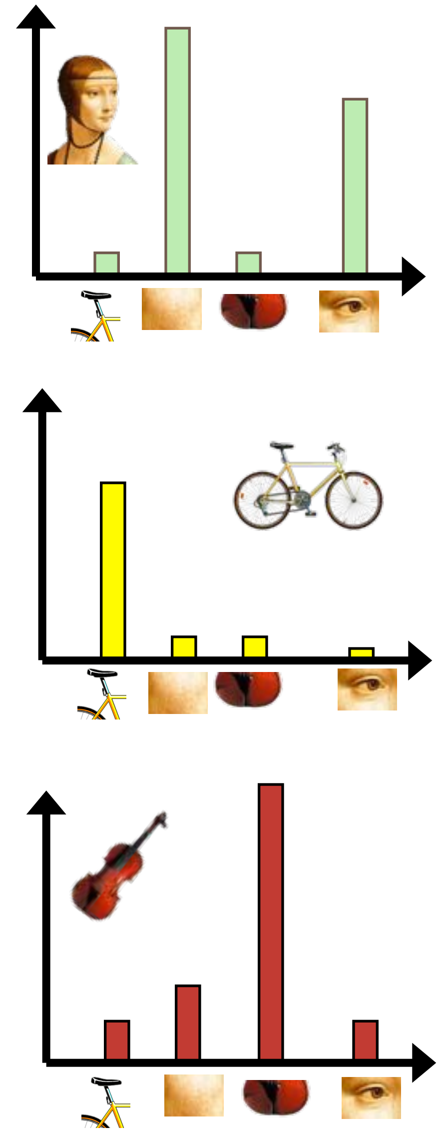
frame #10

Word number	List of image numbers
1	5,10, ...
2	10,...
...	...

Sac de Mots visuels



- Résumer l'image entière par sa distribution (histogramme) des occurrences de mots.
- Analogue à la représentation de sac de de mots couramment utilisée pour les documents.
- Représentation de taille fixe, indépendamment du nombre d'éléments



Méthodes de quantification

k-means (typical choice), agglomerative clustering, mean-shift,…

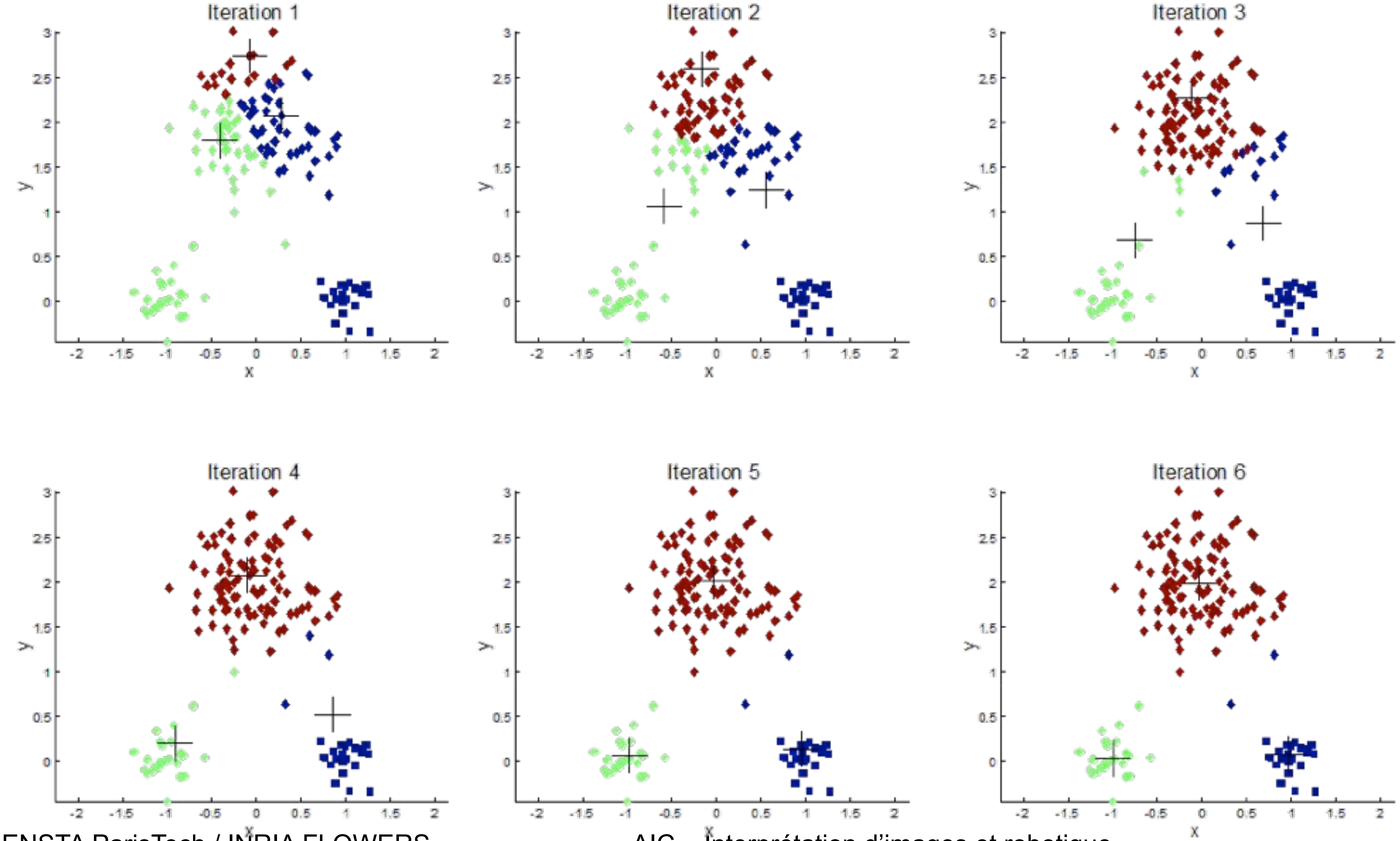
Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies

- Vocabulary tree [Nister & Stewenius, CVPR 2006]

K-means

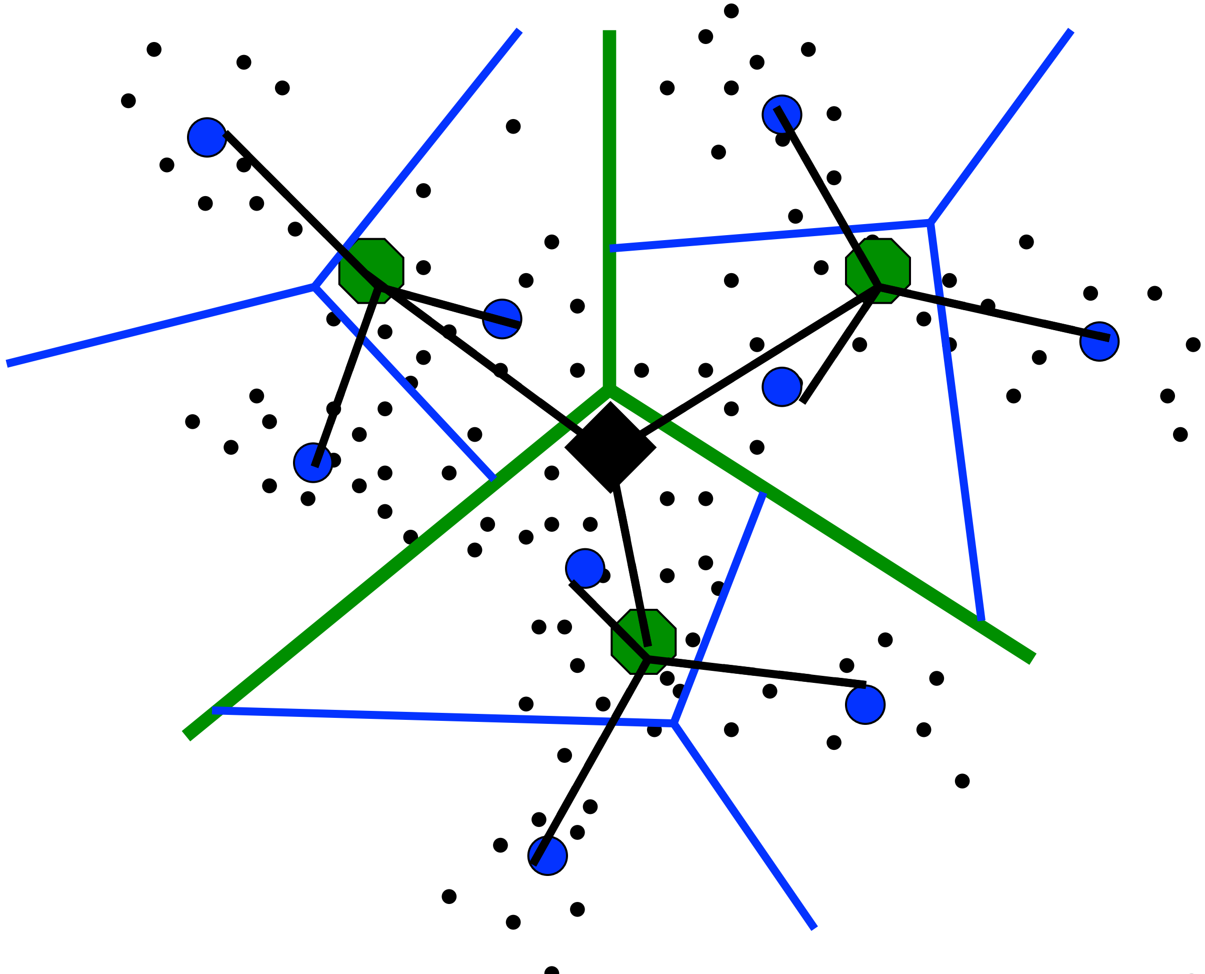
Initialize random centers

Loop : update center as mean of closest points



Vocabulary Tree

Tree construction:

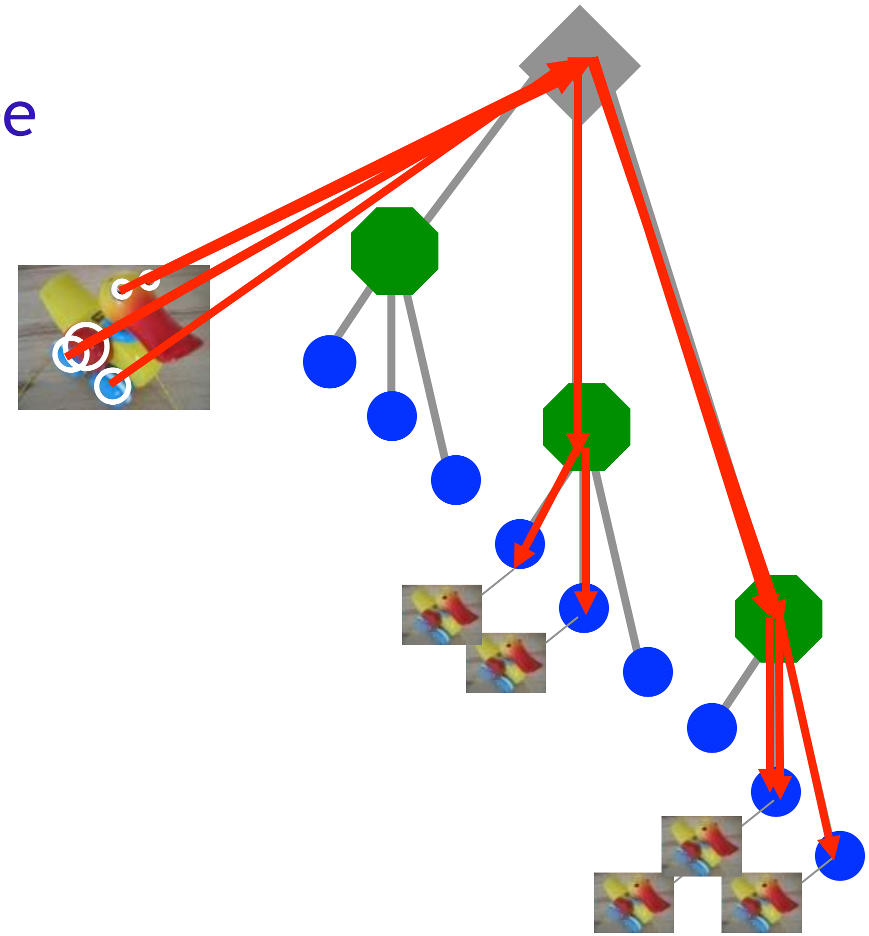


K. Grauman, B. Leibe

34

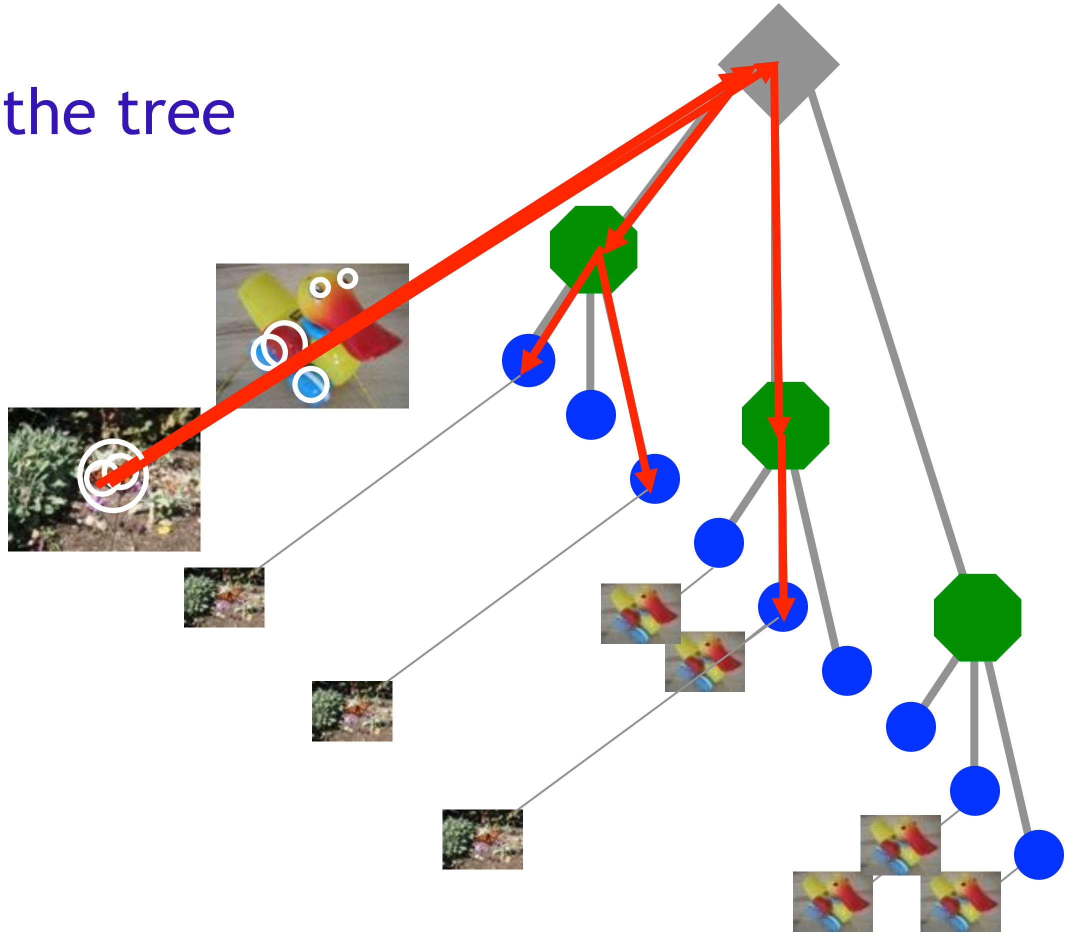
Vocabulary Tree

Training: Filling the tree



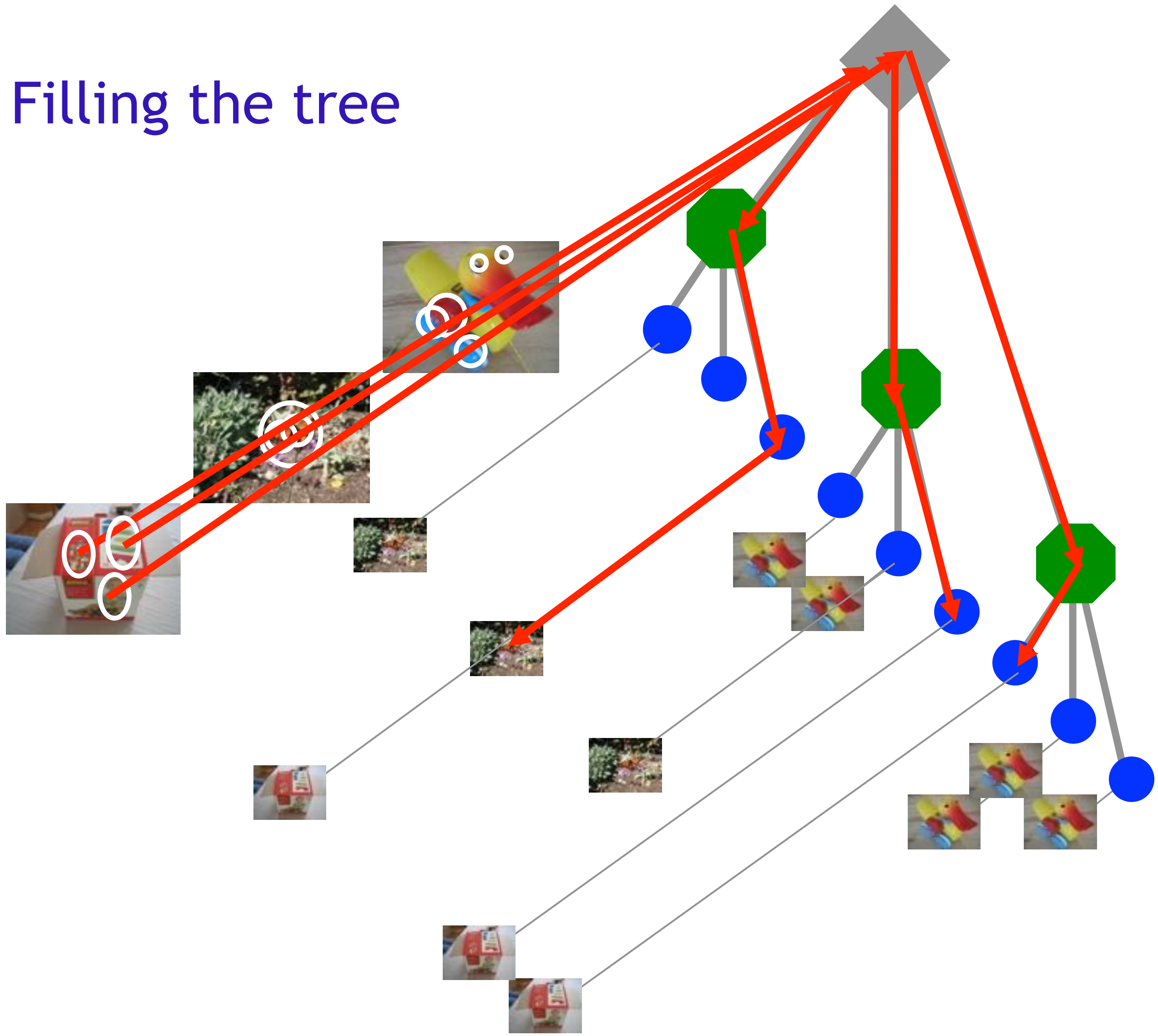
Vocabulary Tree

Training: Filling the tree



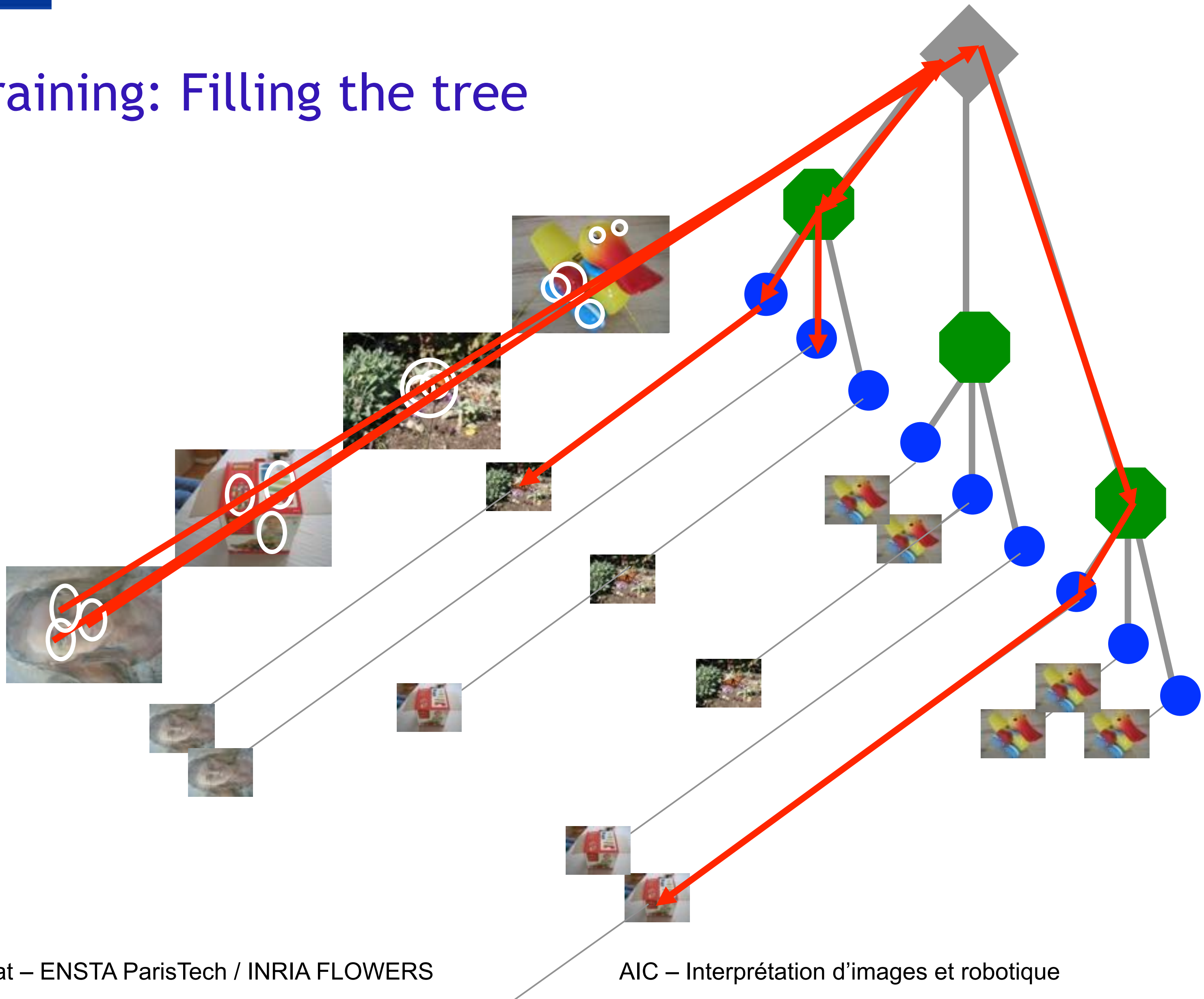
Vocabulary Tree

Training: Filling the tree



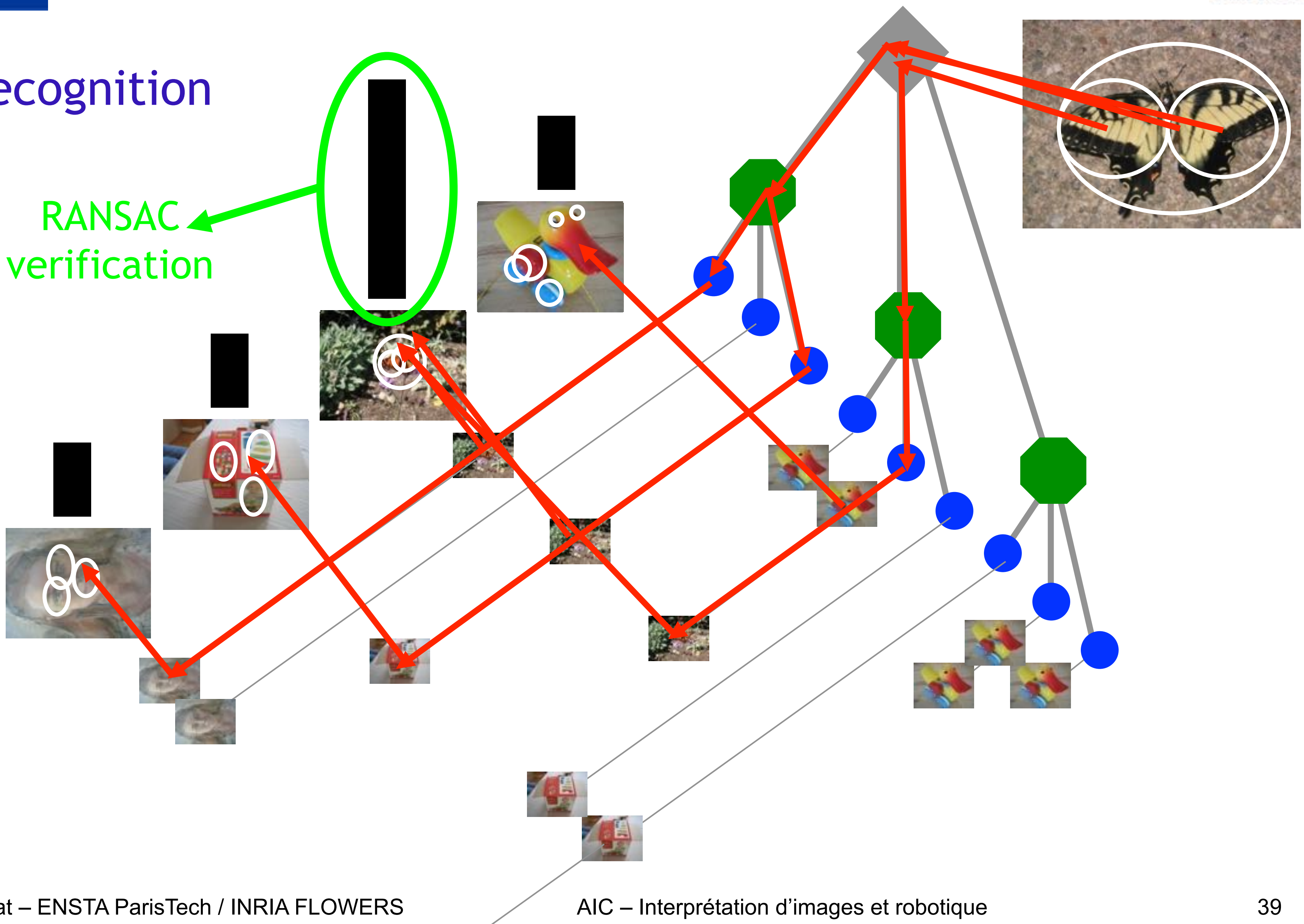
Vocabulary Tree

Training: Filling the tree



Vocabulary Tree

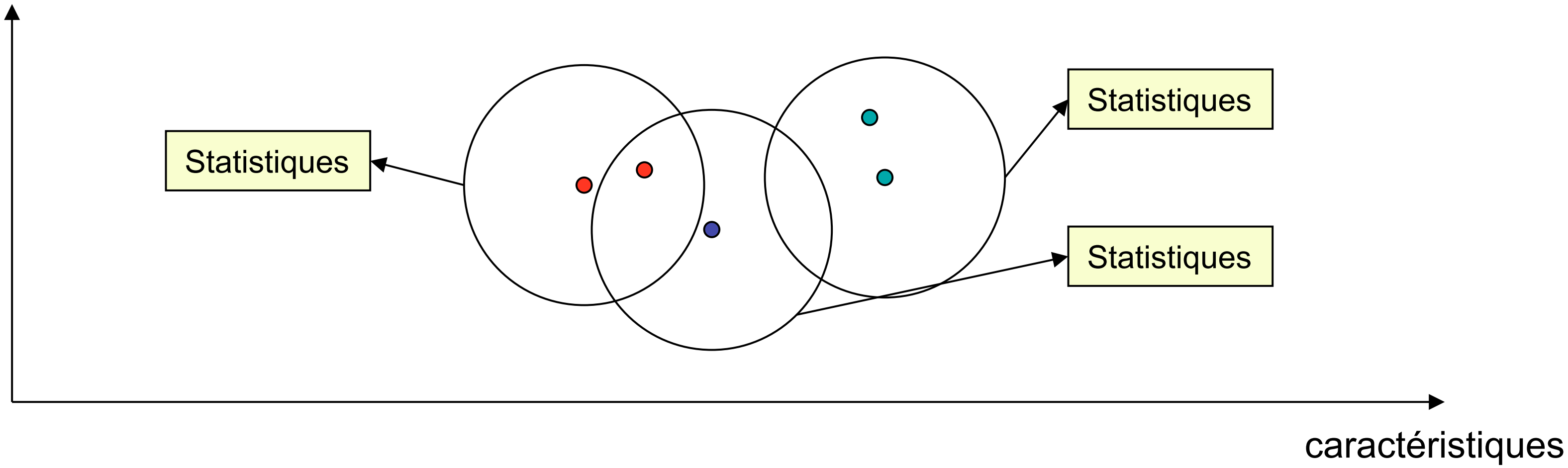
Recognition



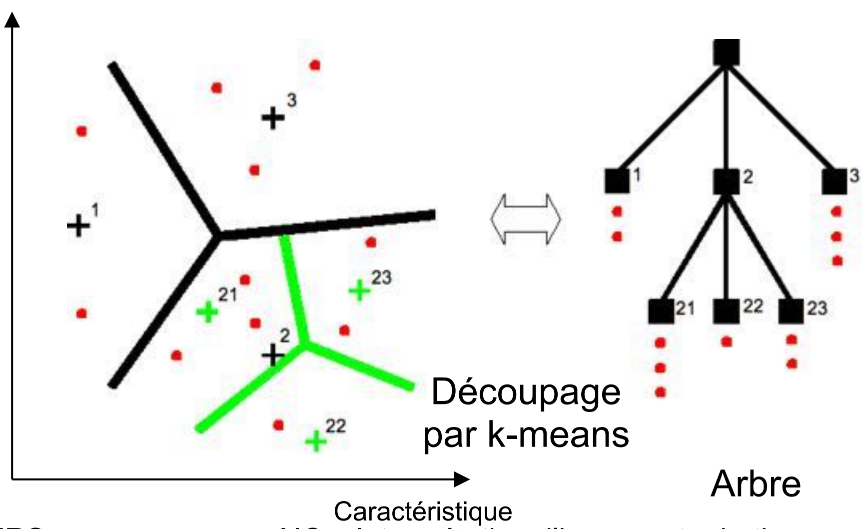
Utilisation en robotique : Adaptation incrémentale

Catégorisation incrémentale

– Dictionnaire : plus proche voisin incrémental



Recherche rapide



Plan du cours

Introduction
- Robotique développementale
- Modèles de sacs de mots visuels

Cartographie - Localisation
- **Application des sacs de mots visuels**

Apprendre à interpréter des images
- Distinguer soi / non soi
- Modéliser soi/objets/humain

Apprendre à chercher des objets
- Apprentissage de saillance visuelle

Apprendre à éviter des obstacles
- Prédiction de profondeur en video monoculaire

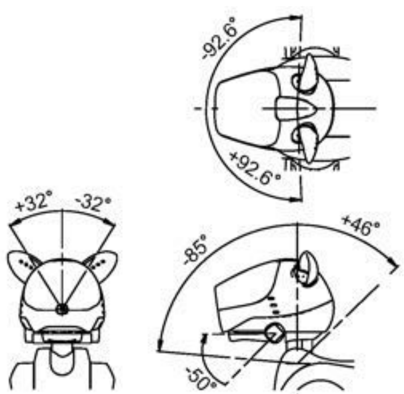
Localisation et cartographie qualitative

Navigation topologique

- Aibo arrive dans une nouvelle maison
- Apprend à reconnaître les pièces
- Va d'une pièce à l'autre

Approche

- Perception active
- Apprentissage par interaction discontinue avec l'utilisateur
  - ➔ Robustesse aux manipulations et à la qualité des images



Localisation et cartographie qualitative

Structure du problème

Des images appartiennent à plusieurs catégories

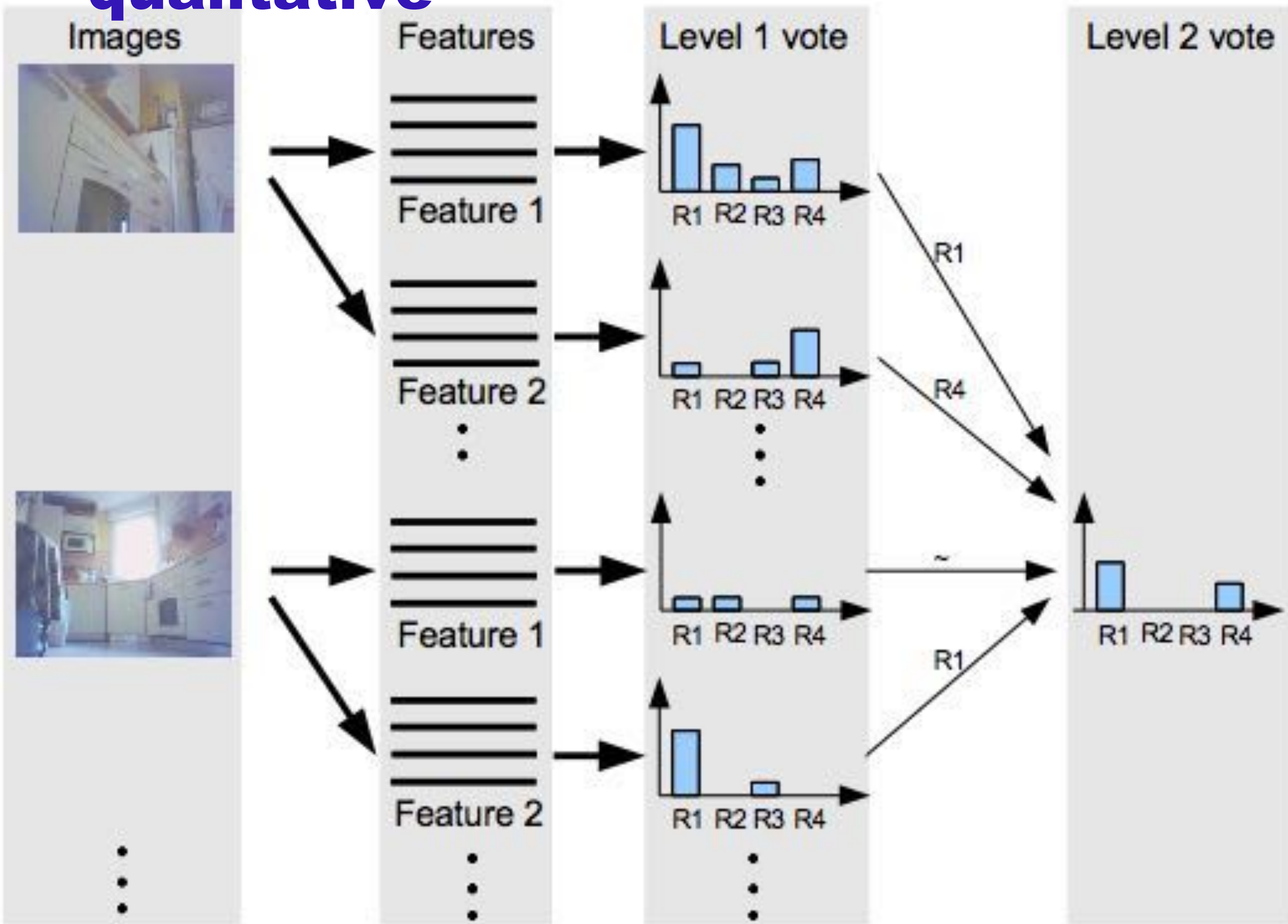
Toutes les images prise d'une position appartiennent à la même catégorie

Localisation Active

Prendre des images informatives

Prendre de nouvelles images jusqu'à confiance suffisante

Localisation et cartographie qualitative



Localisation et cartographie qualitative

Cartographie (apprentissage actif)

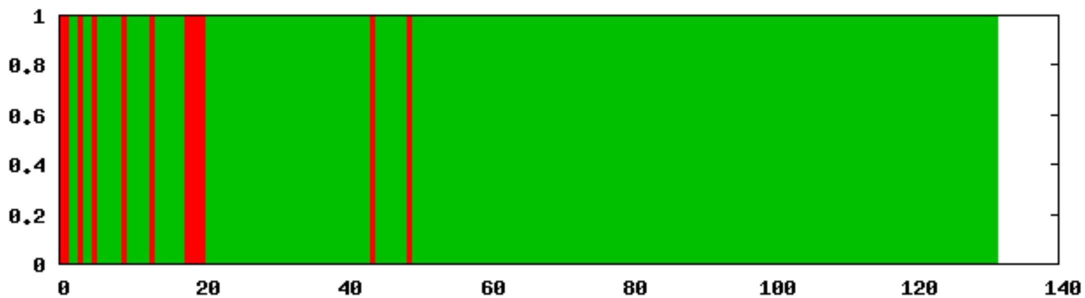
- Localiser le robot
- Si la localisation est erronée (info utilisateur)
- Demander la position correcte
- Apprendre avec les images utilisées pour la localisation

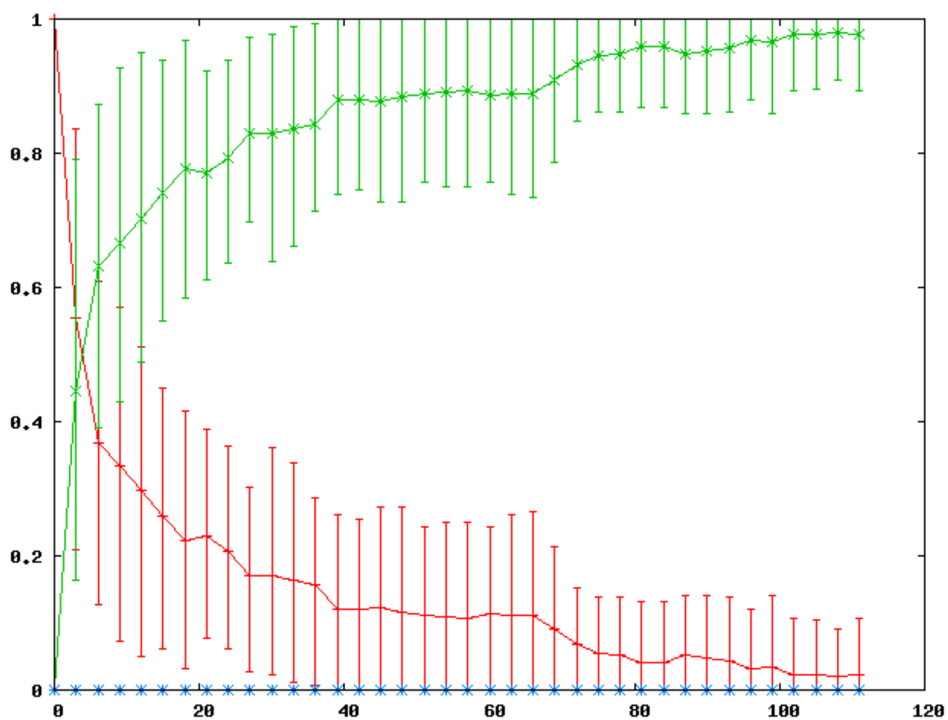
Apprentissage d'une image

- Pour chaque caractéristique :
  - Extraire les caractéristiques
  - Chercher les caractéristiques dans le dictionnaire
  - Si (inconnue) ajouter un mot
  - Mettre a jour les statistiques des mots trouvés

Localisation et cartographie qualitative

[FILLIAT07]





Localisation et cartographie qualitative


Barbara's office


Corridor


Elin's office


Kitchen

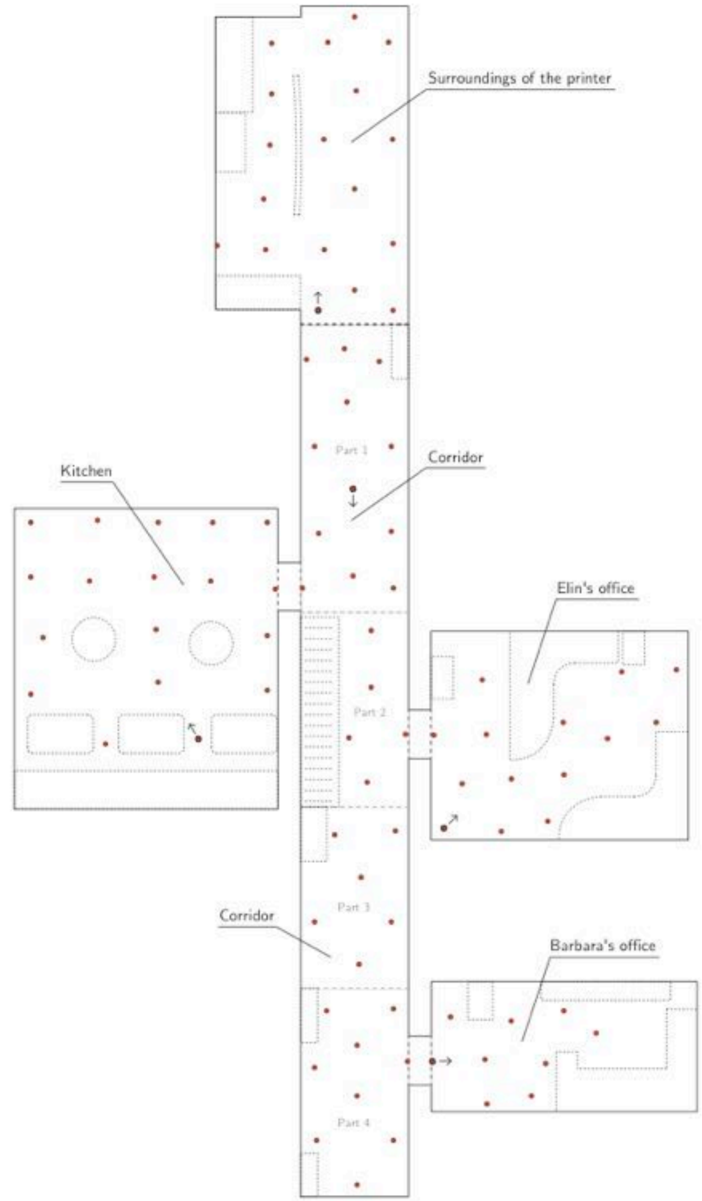

Surroundings of the printer

Barbara's office


Cloudy

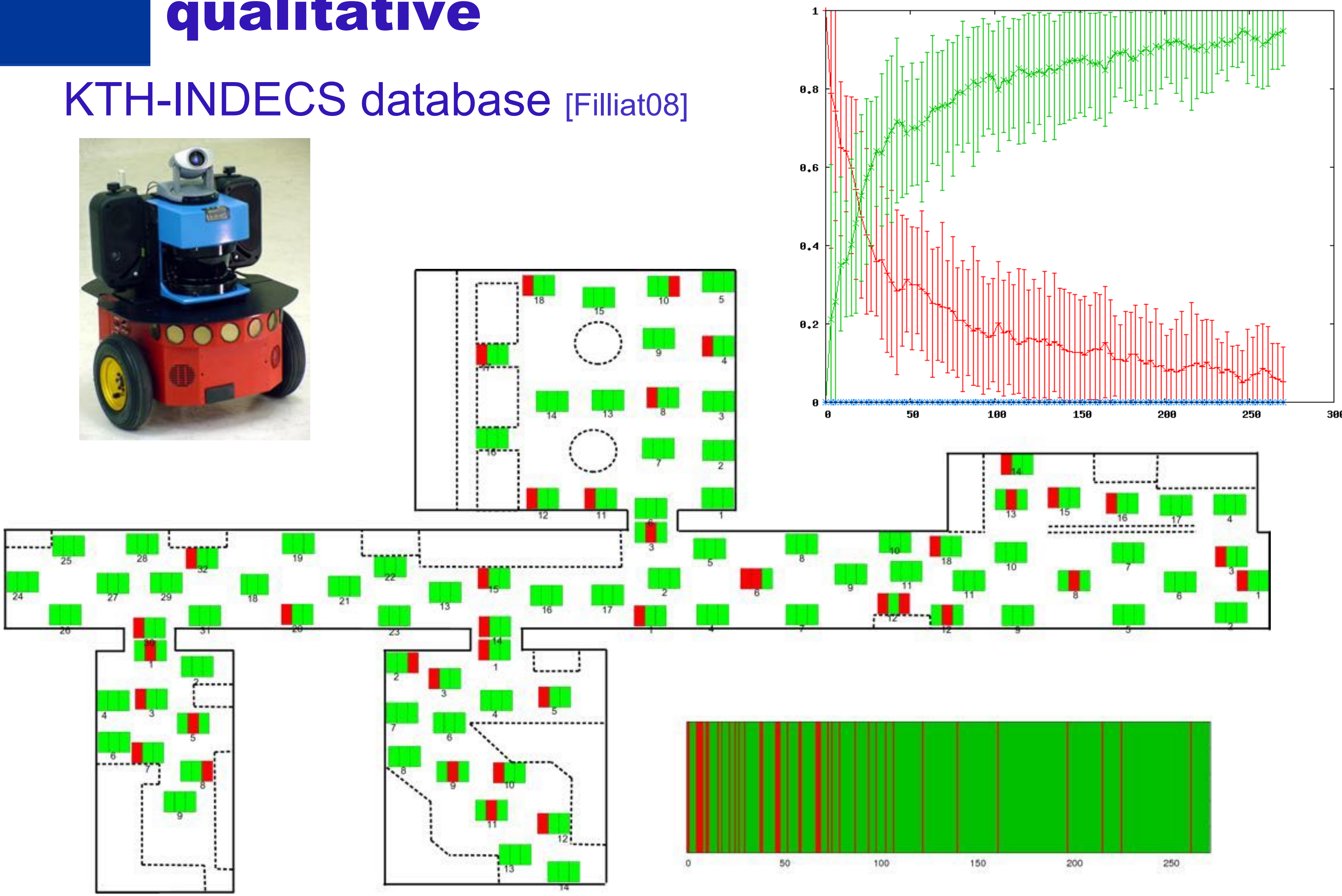

Night


Sunny



Localisation et cartographie qualitative

KTH-INDECS database [Filliat08]

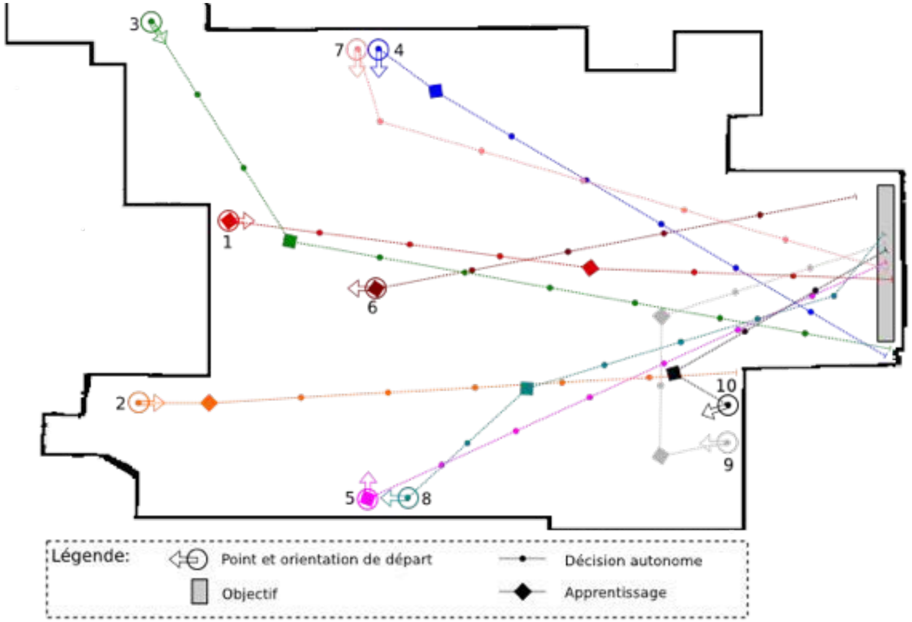


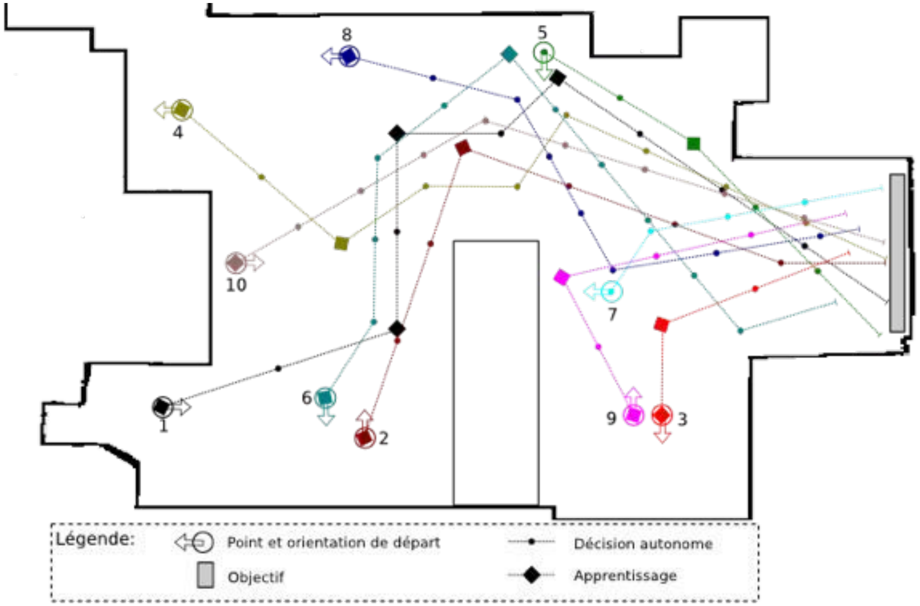
Guidage visuel

Apprendre et rejouer un chemin

- Approche identique (perception active / apprentissage)
- Prédire direction locale du but

Guidage visuel

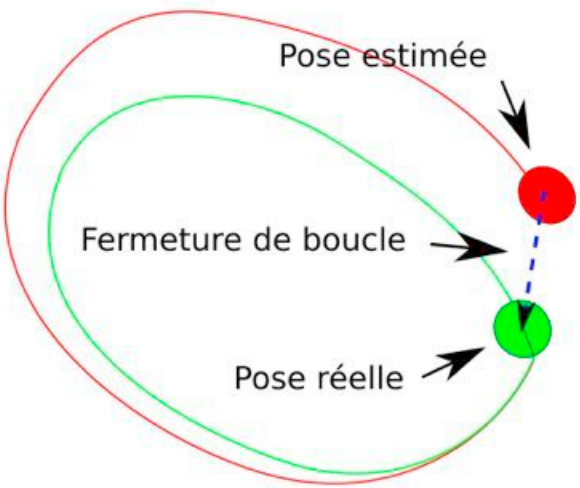




Détection de fermetures de boucles

Détection de fermeture de boucle par apparence

– Détecte si l'image courante
a déjà été vue
– Permet la correction de la
carte et la localisation (SLAM métrique)



Utilisation des sacs de mots incrémentaux pour la détection

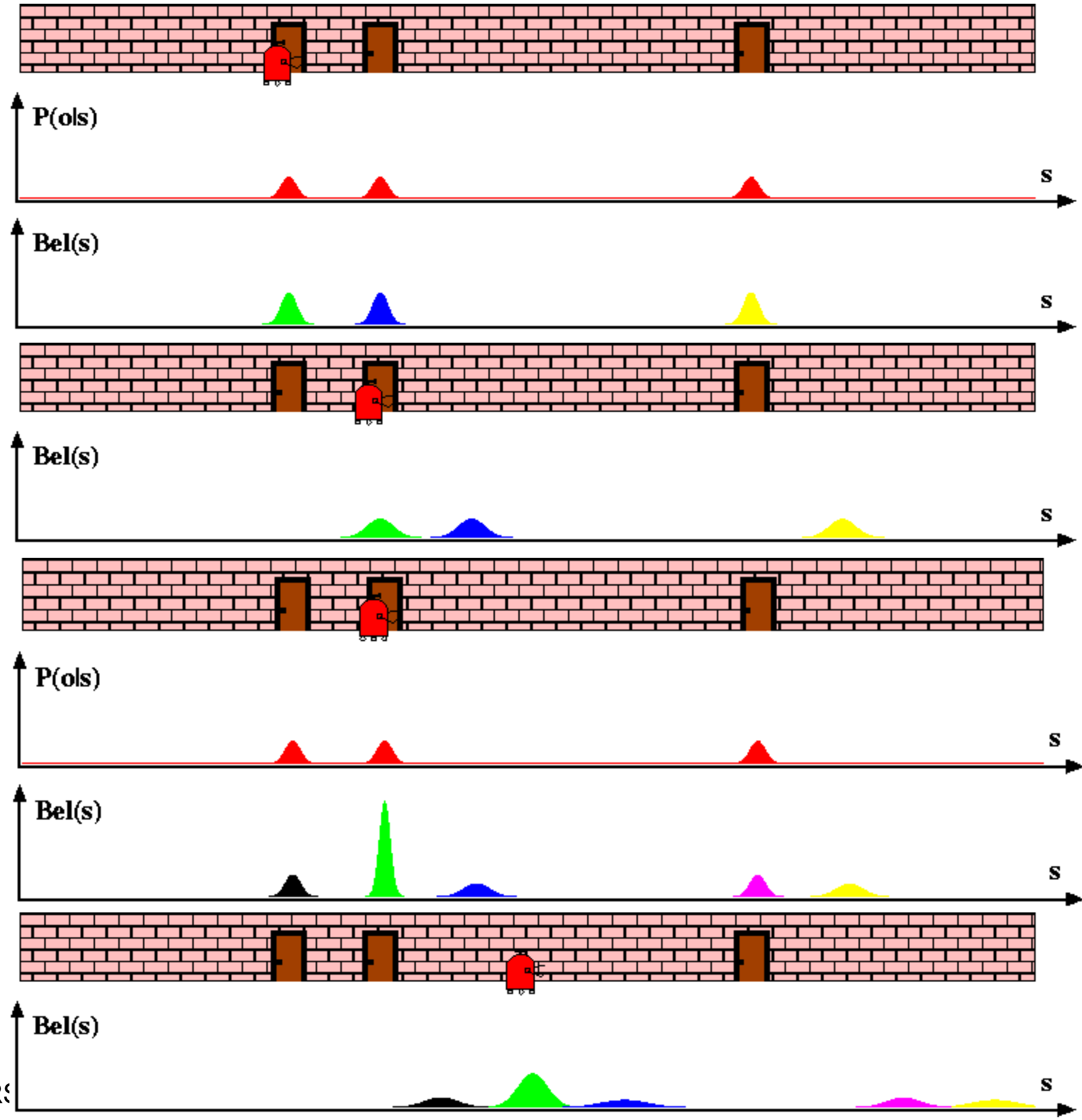
Exploitation du mouvement pour la reconnaissance

– Image seule ambigüe
– Mise en correspondance de séquences
– Utilisation d'un filtre bayésien

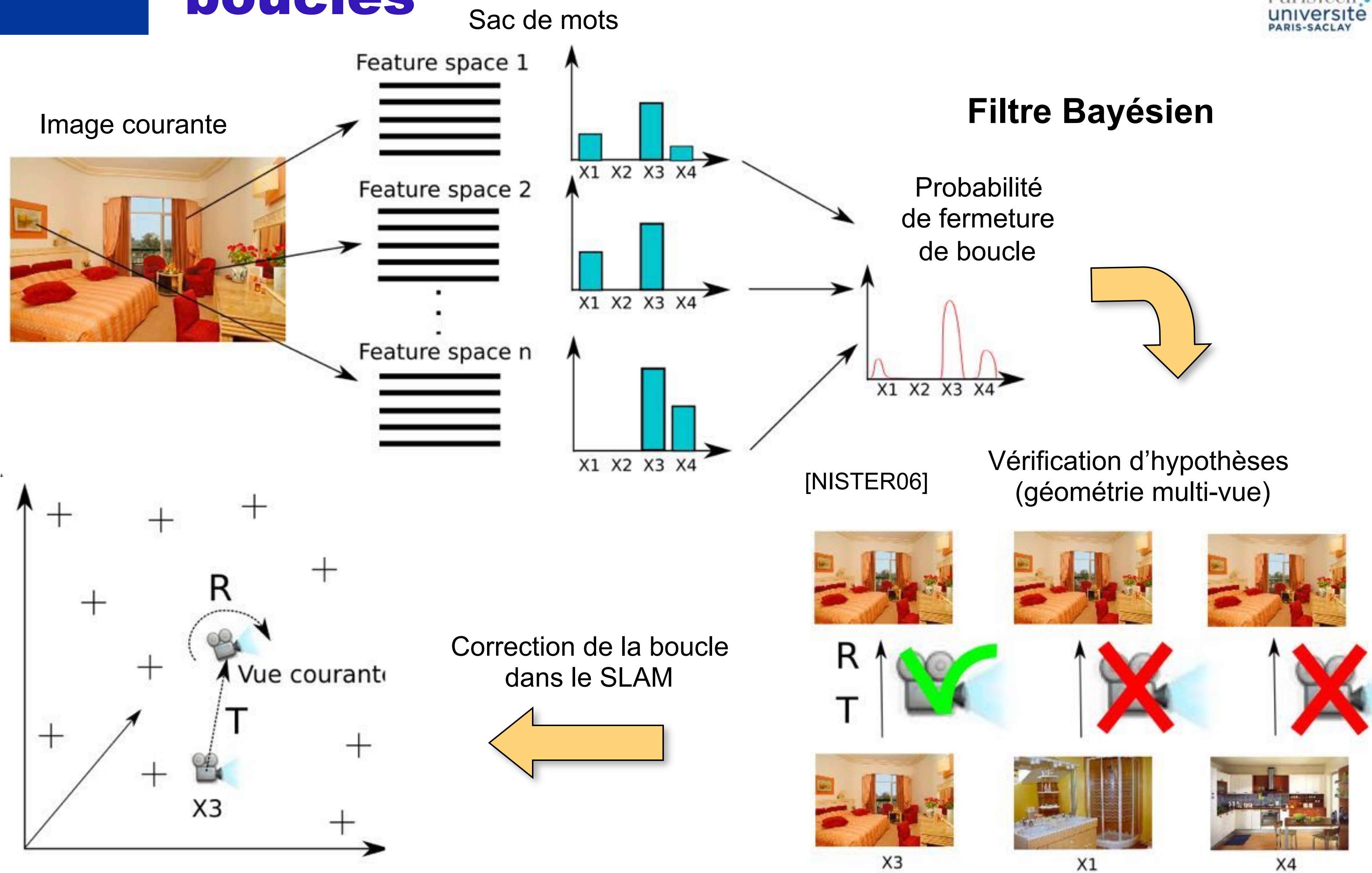
Détection de fermetures de boucles

Filtre Bayésien

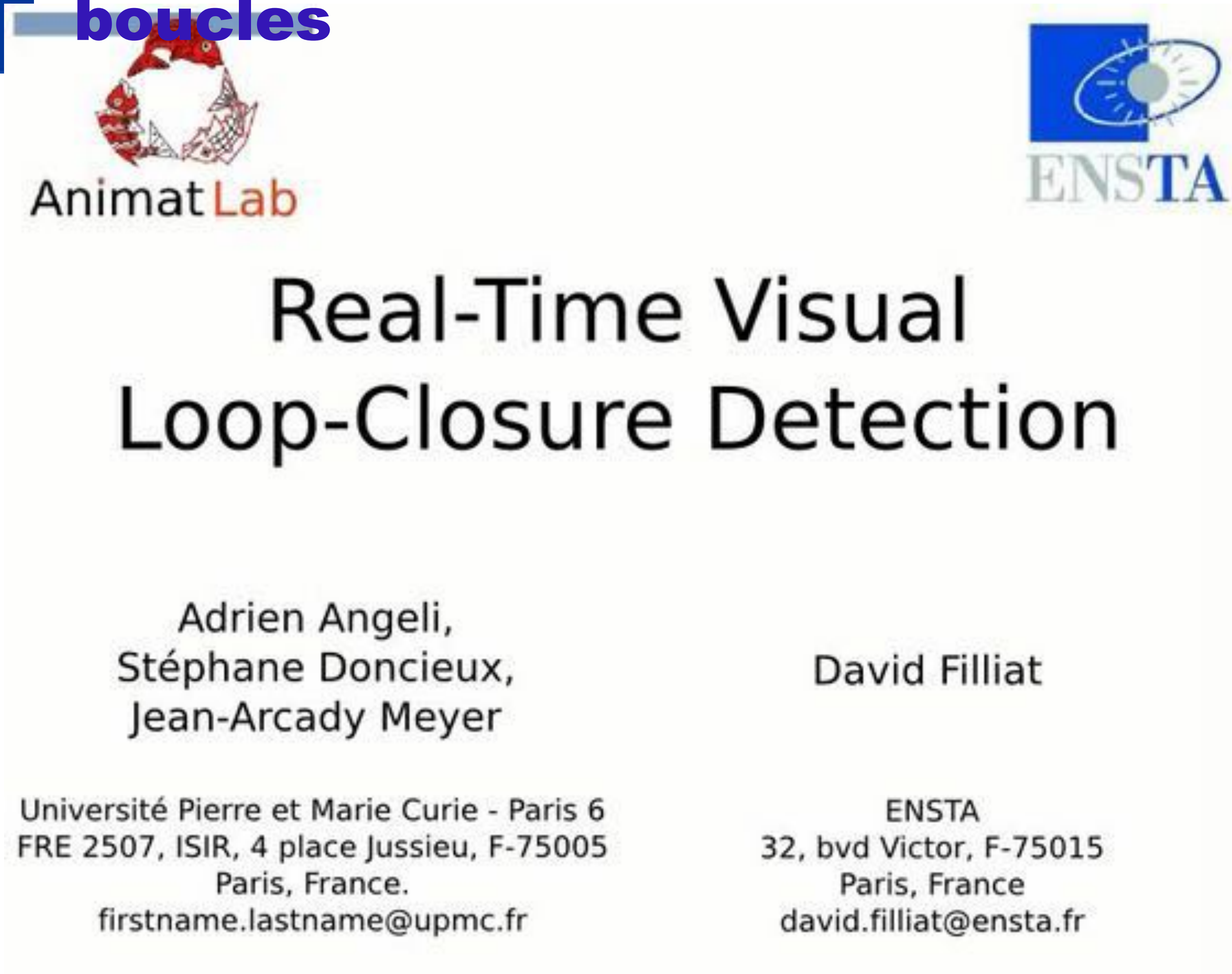
[Fox, Thrun et Burgard Probabilistic Robotics]



Détection de fermetures de boucles

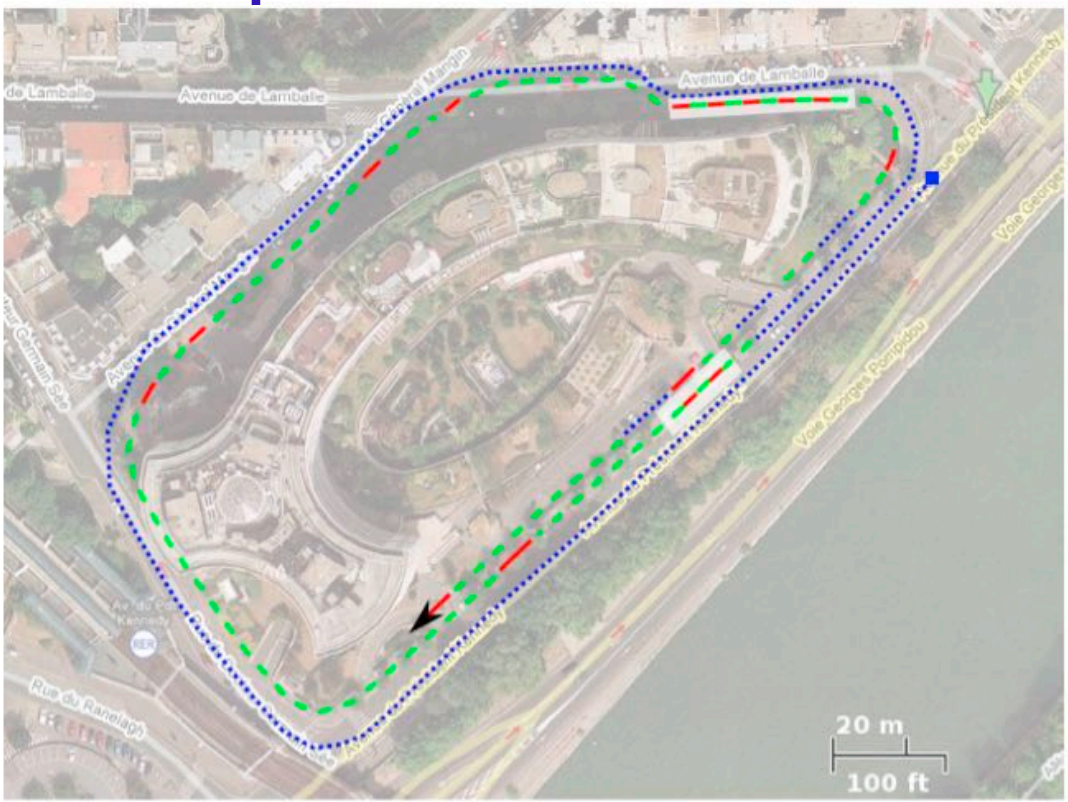


Détection de fermetures de boucles



Détection de fermetures de boucles

Exemple en extérieur



Plan du cours

Introduction
- Robotique développementale
- Modèles de sacs de mots visuels

Cartographie - Localisation
- Application des sacs de mots visuels

Apprendre à interpréter des images
- **Distinguer soi / non soi**
- **Modéliser soi/objets/humain**

Apprendre à chercher des objets
- Apprentissage de saillance visuelle

Apprendre à éviter des obstacles
- Prédiction de profondeur en video monoculaire

Approche développementale de la perception

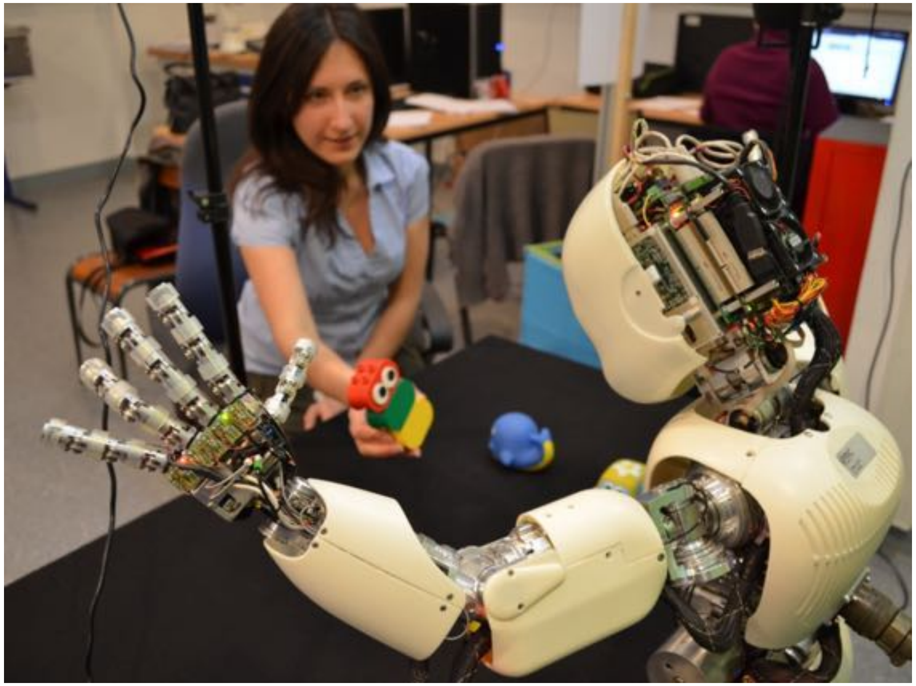
Modéliser les objets

- Pour les reconnaître
- Sans supervision
- Sans base de donnée

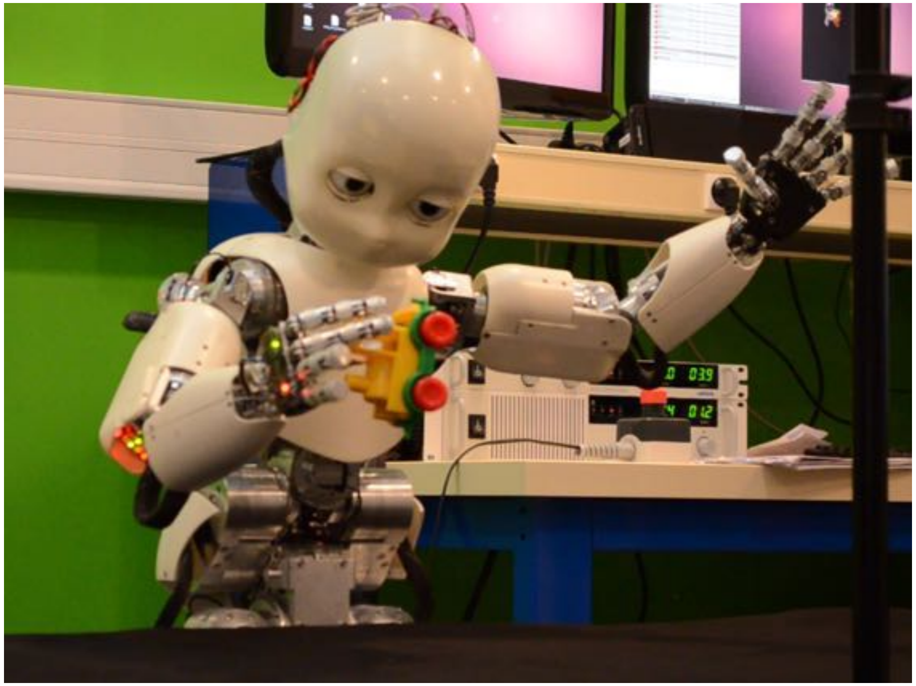
S'inspirer des enfants

- Observation
- Interaction sociale
- Action

Approche développementale



Apprentissage par observation

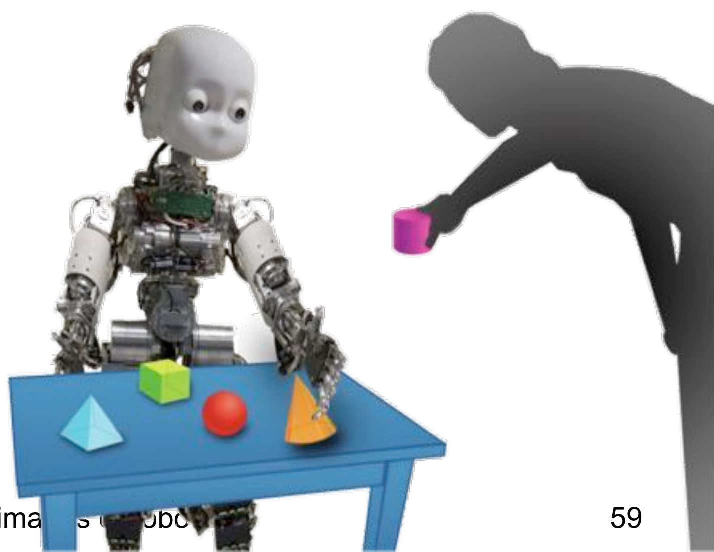


Apprentissage par manipulation

Approche

1. Segment the visual space,
2. learn the appearance of physical entities,
3. categorize
   - robot parts,
   - human parts,
   - objects,
4. improve object learning

   through manipulation



System overview

Arm joints

4. Categorization

- robot,
- human,
- objects

Entity model

3. Learning multi-view entities

RGB-D sensor

1.Detection of proto-objects

2. Learning appearances of views

System overview

Arm joints

4. Categorization

- robot,
- human,
- objects

Entity model

3. Learning multi-view entities

RGB-D sensor

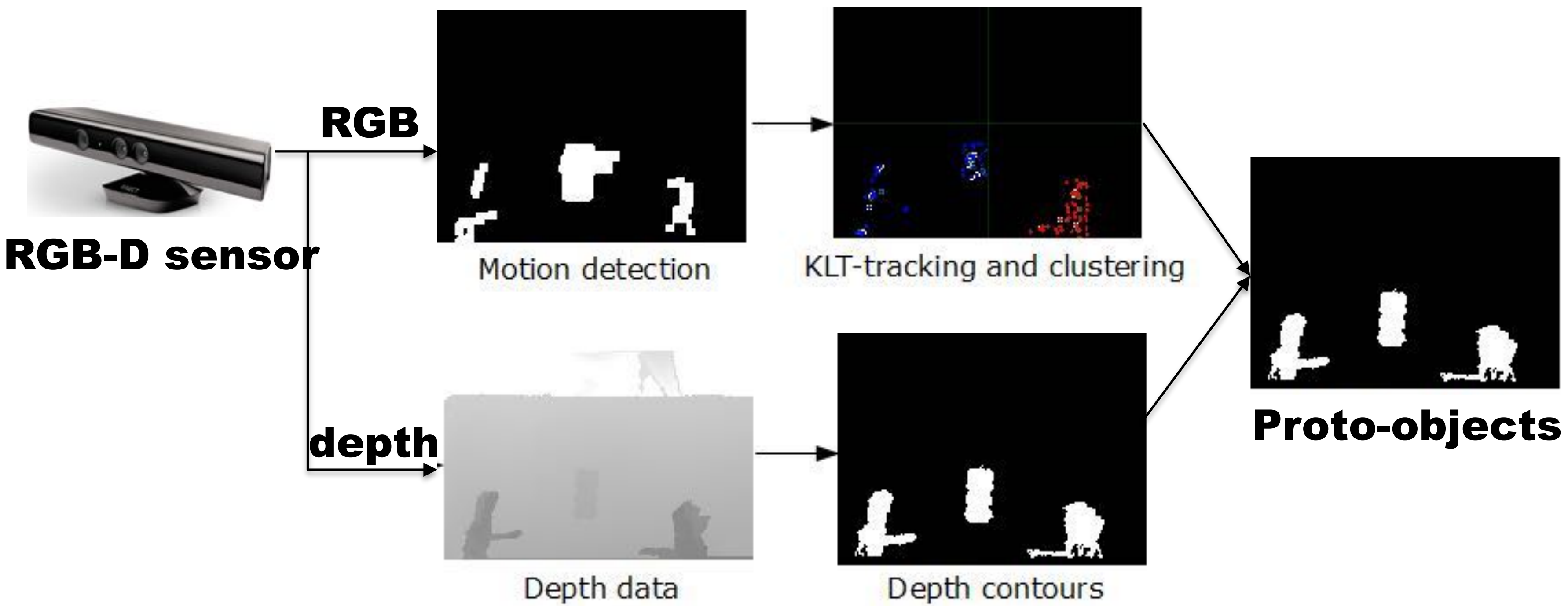
1.Detection of proto-objects

2. Learning appearances of views

1. Segmentation

Proto-objects

- units of visual attention with 'objecthood' characteristics
- detected from coherent motion and appearance
- segmented using depth contours



System overview

Arm joints

4. Categorization

- robot,
- human,
- objects

3. Learning multi-view entities

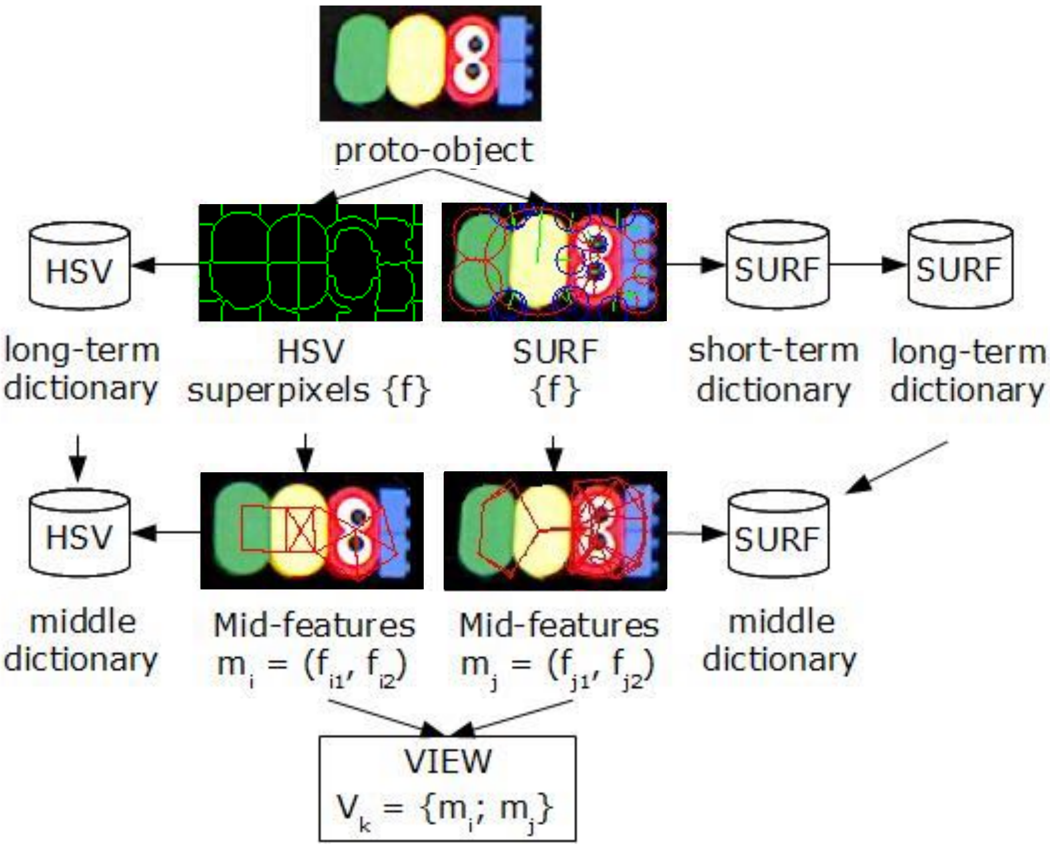
Entity model

RGB-D sensor

1.Detection of proto-objects

2. Learning appearances of views

2.1. Learning appearance of views



Low-level

complementary features

Mid-features

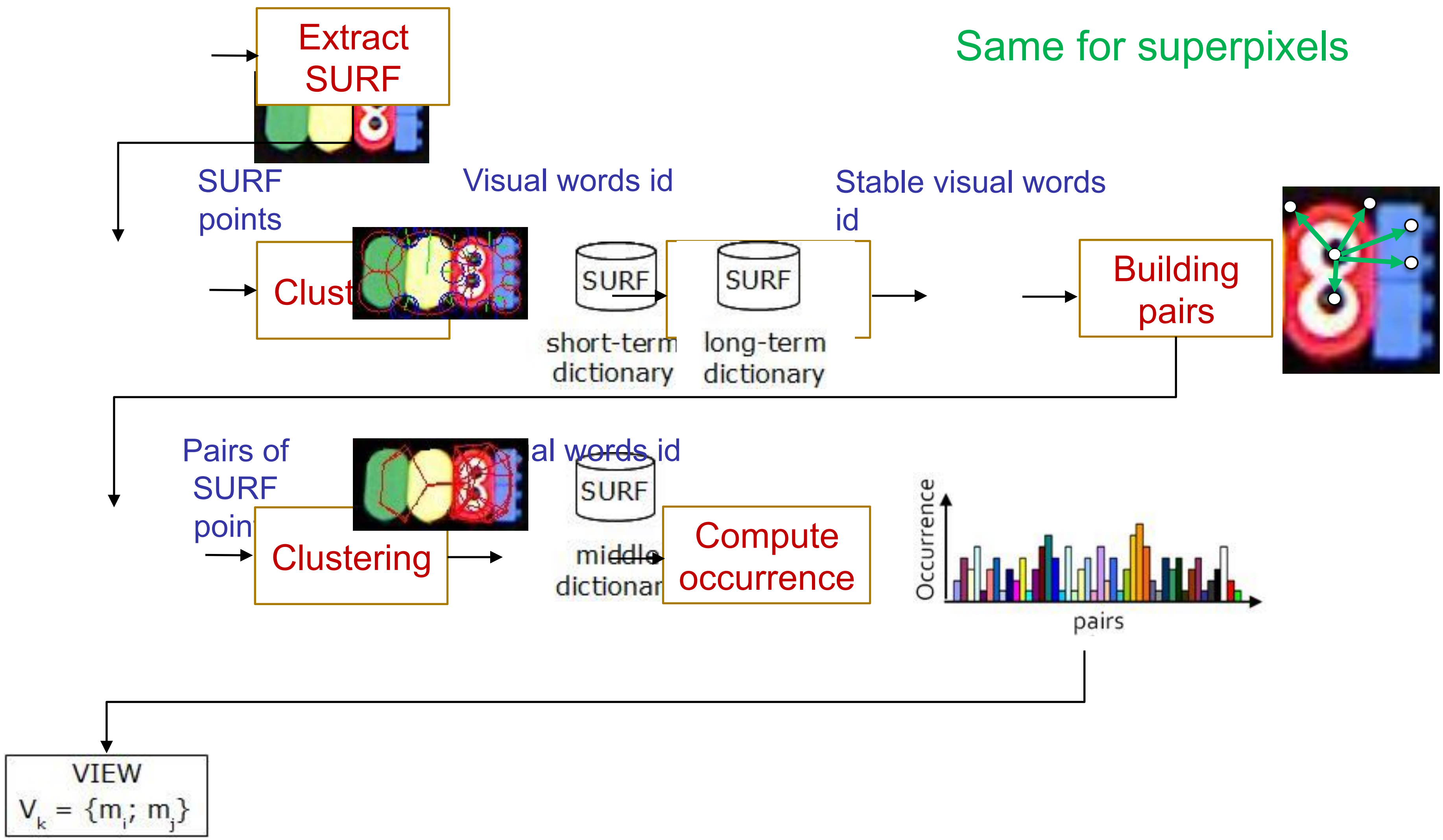
introducing local geometry

Bag of mid-features

based on BoW

2.1. Learning appearance of views

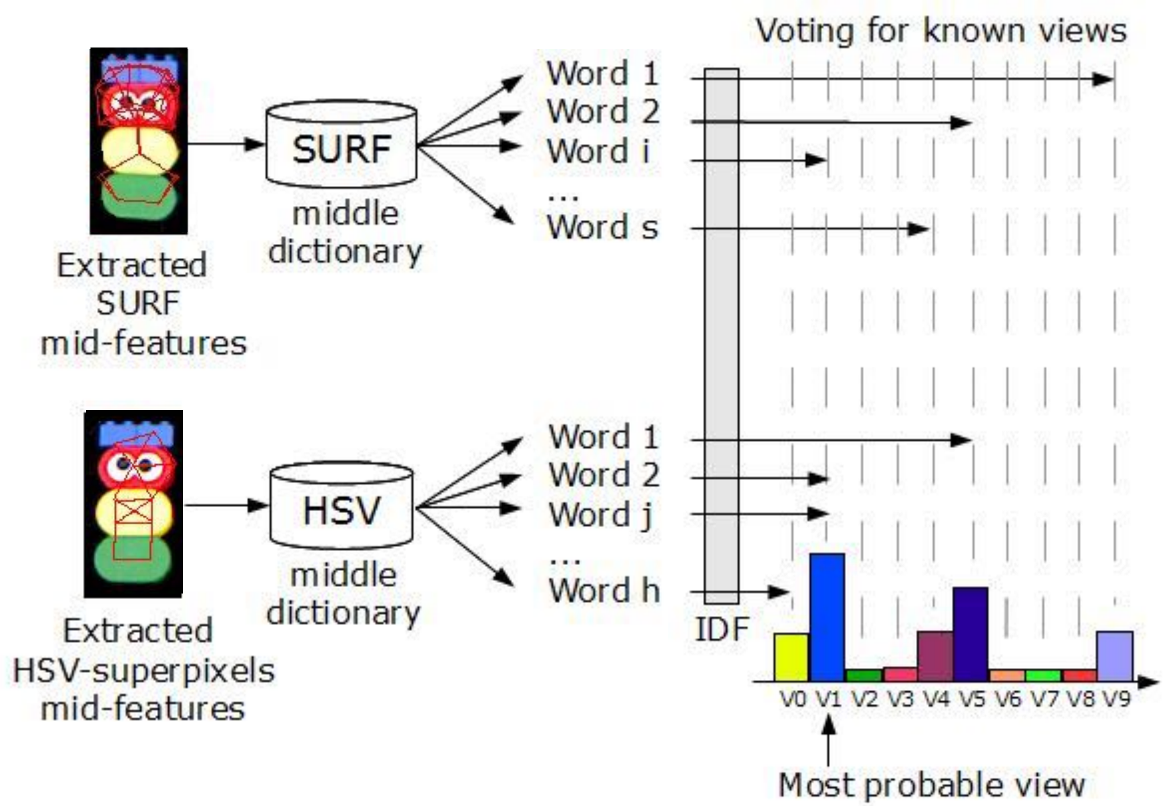
Same for superpixels



2.2. Recognizing views

TF-IDF learning based on

- Mid-Feature-Frequency – Inverse-View Frequency Recognition
- maximum likelihood computed through a voting method



System overview

Arm joints

4. Categorization

- robot,
- human,
- objects

Entity model

3. Learning multi-view entities

RGB-D sensor

1.Detection of proto-objects

2. Learning appearances of views

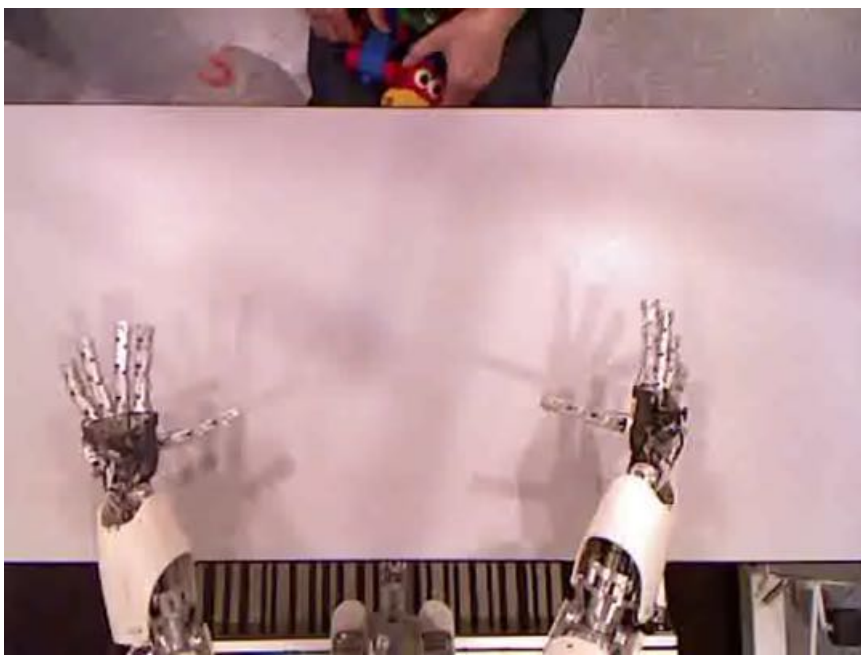
3. Learning multi-view entities

Based on tracking

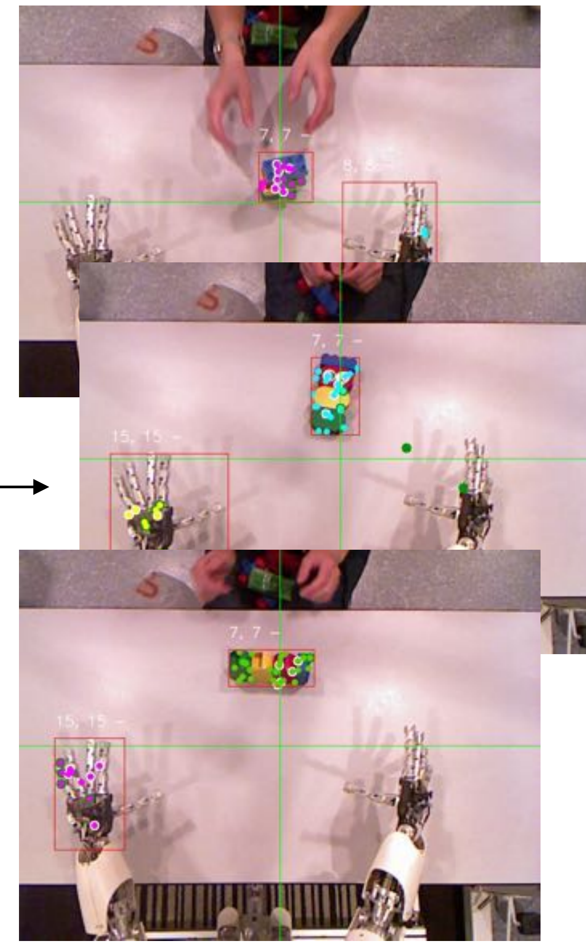
- detected views are added to the entity model

Based on view-recognition

- from the occurrence frequency
  of a current view among learned entities

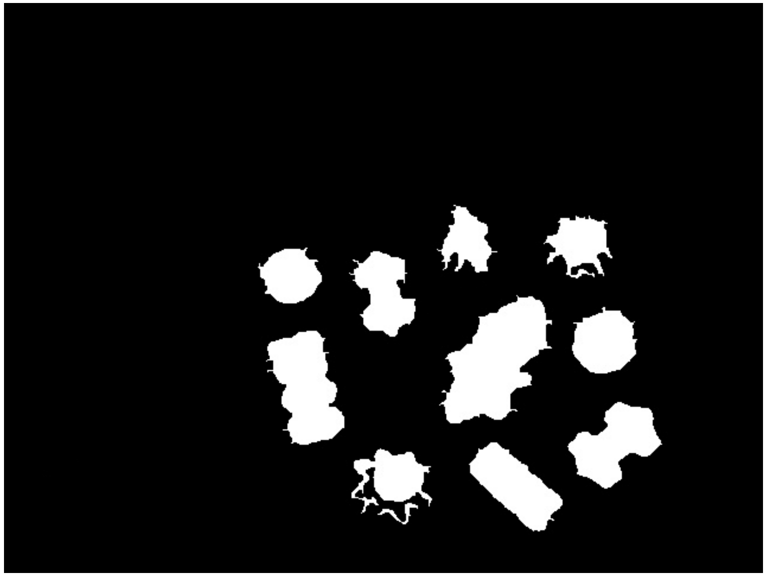


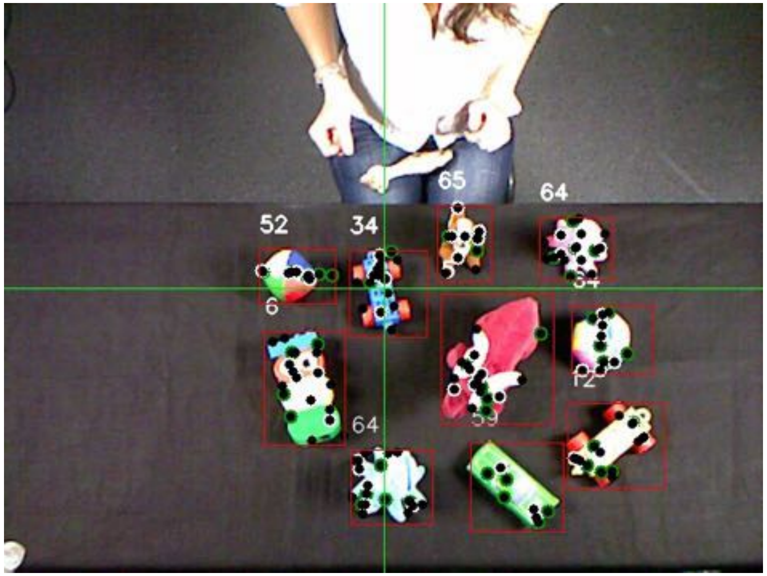
OBJECT
On = {Vk}



Simultaneous tracking

10 objects detected, tracked, recognized as different entities

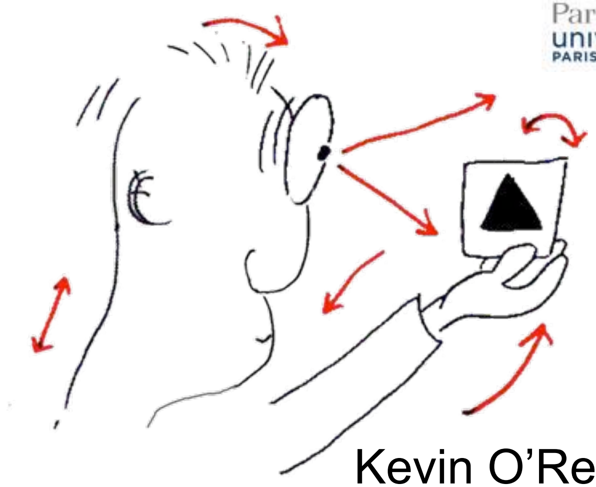




Agir pour reconnaitre

Théories sensori-motrices

- Couplage de la perception et l'action
- La perception dépend de l'action



Kevin O'Regan

Utilité de l'action pour apprendre des objets

- Découvrir des propriétés physiques
- Améliorer la segmentation
- Découvrir des points de vues
- Besoin de reconnaître soi/autre/objets

System overview

Arm joints

4. Categorization

- robot,
- human,
- objects

Entity model

3. Learning multi-view entities

RGB-D sensor

1.Detection of proto-objects

2. Learning appearances of views

4. Categorization

Robot category

– has high mutual information between the sensory data and proprioception

$$H(Lc) = -\sum_{l} p(l) log(p(l)),$$

$$MI(Lc; Ac) = Hc(Lc|Ac) - H(Lc)$$

$$Hc(Lc|Ac) = -\sum_{a} p(a) \sum_{l} p(l|a) log(p(l|a)),$$

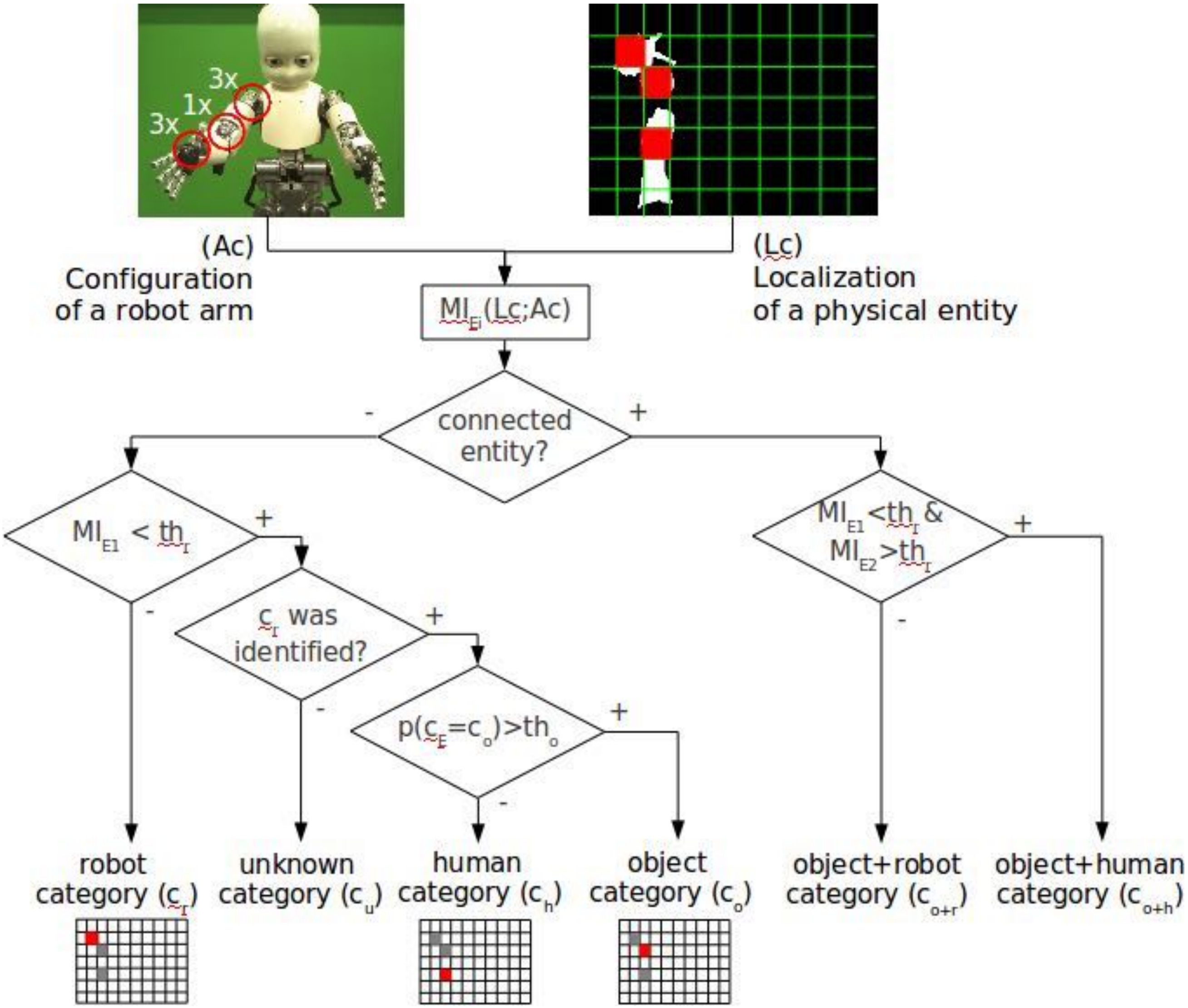
Object category

– 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



Robot model

Object model

Human model

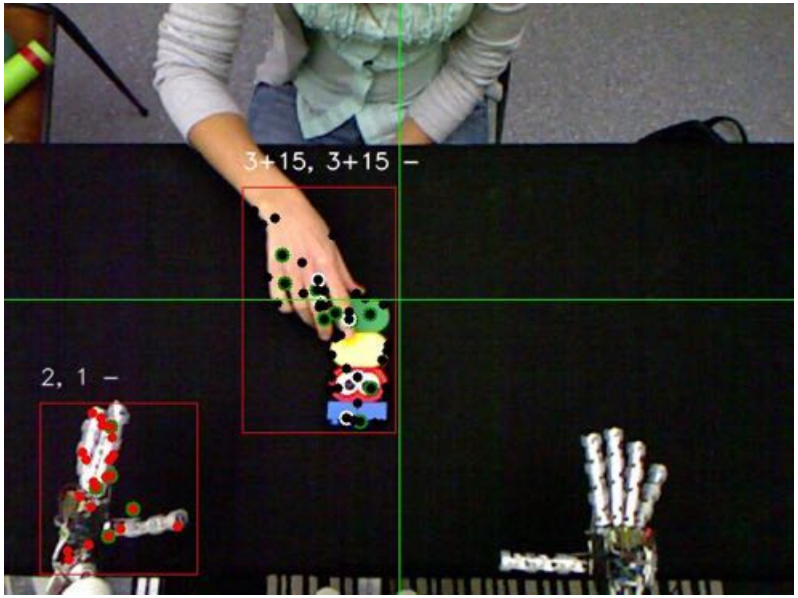
Object learning with manipulation

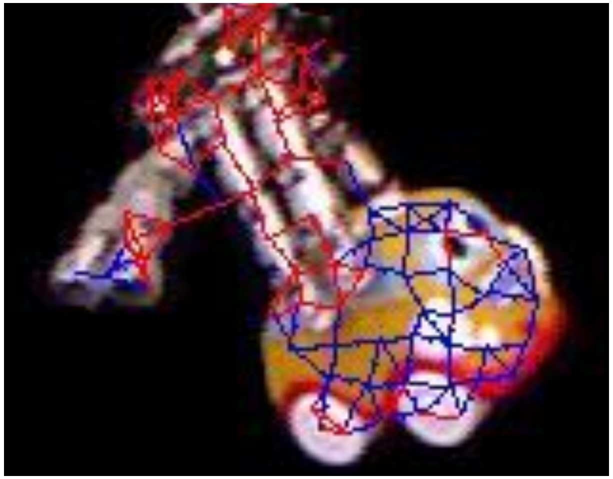
Identifying objects hold by robot/human

- Hand holding an object compose a single moving blob.

Segmentation

- identify the most probable object
- eliminate features that belong to it
- check a presence of another object, using remaining features





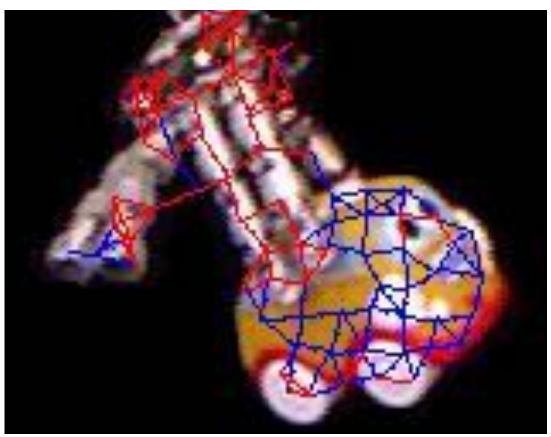
Object learning with manipulation

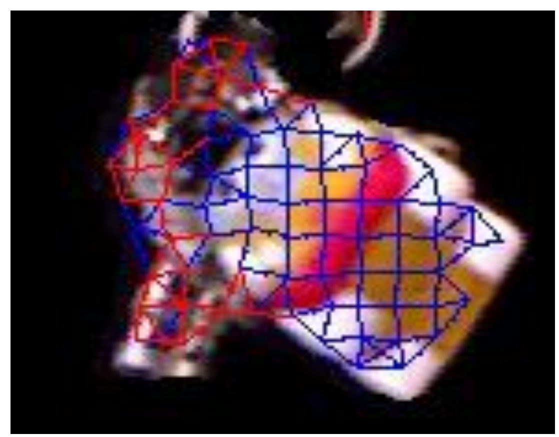
Updating manipulated object

- If connected entities are identified as (r+o) category, the model of the grasped object is updated
  - with a recognized non-robot view,
  - with a new view created from features that do not belong to a robot view.

Robot view (red lines)

Object view (blue lines)





Robot view (red lines)

Object view (blue lines)

Object model


ENSTA ParisTech
ISIR
UPMC
CNRS
DEVELOPMENTAL OBJECT LEARNING
THROUGH MANIPULATION
AND HUMAN DEMONSTRATION
MACSi
Natalia Lyubova, David Filliat, Serena Ivaldi

Effets de la manipulation

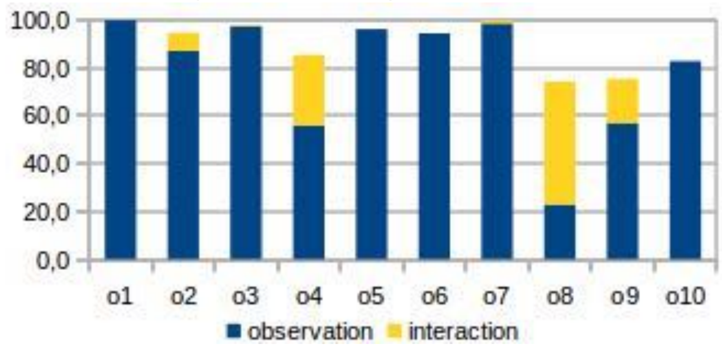
Learning through observation

- objects can be associated with several entities
  - major - the most frequent
  - pure - given to this object but never to others
  - noisy - associated with several objects

Learning through manipulation

- entities are merged,
- noisy entities are deleted

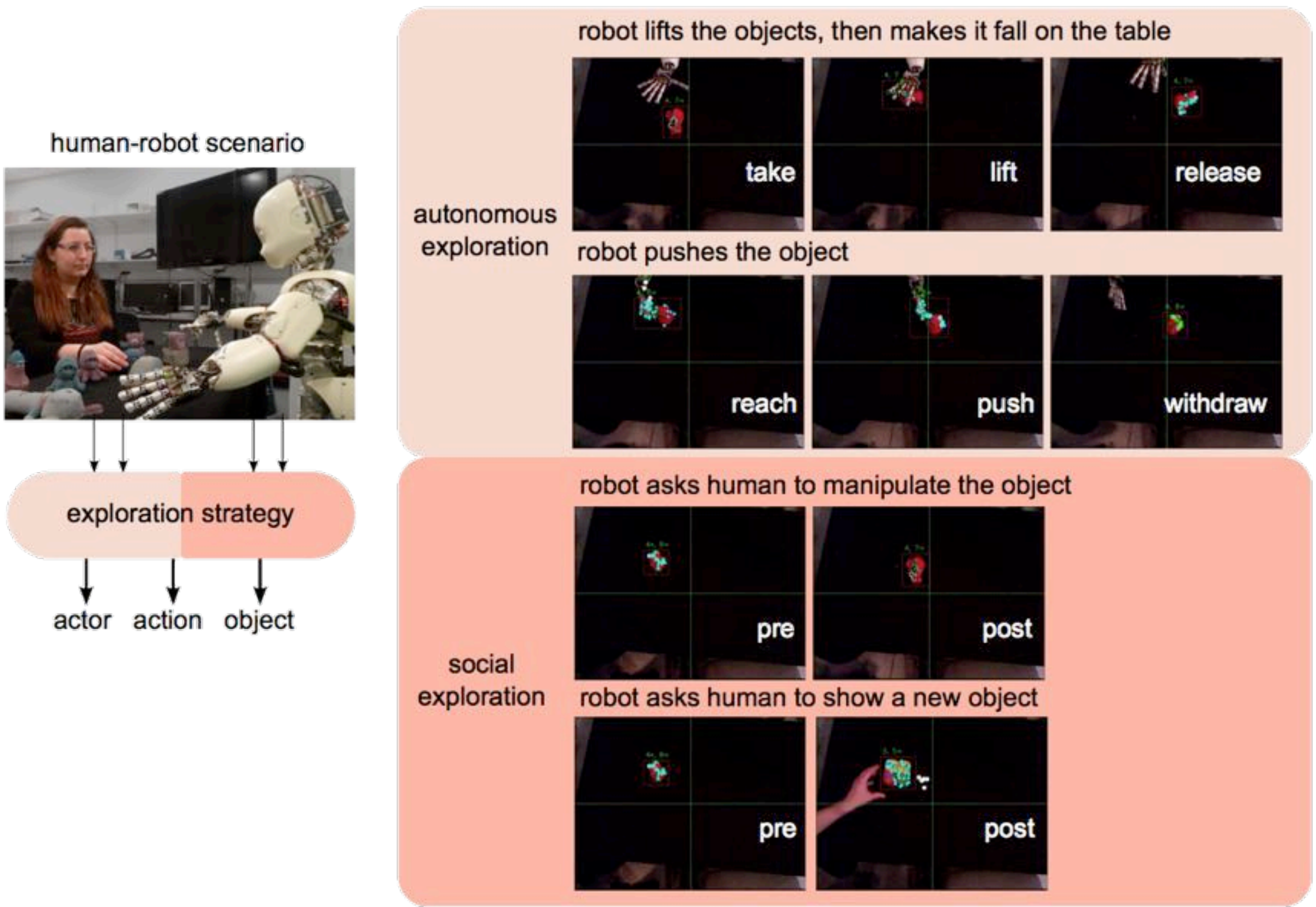
Object recognition rate



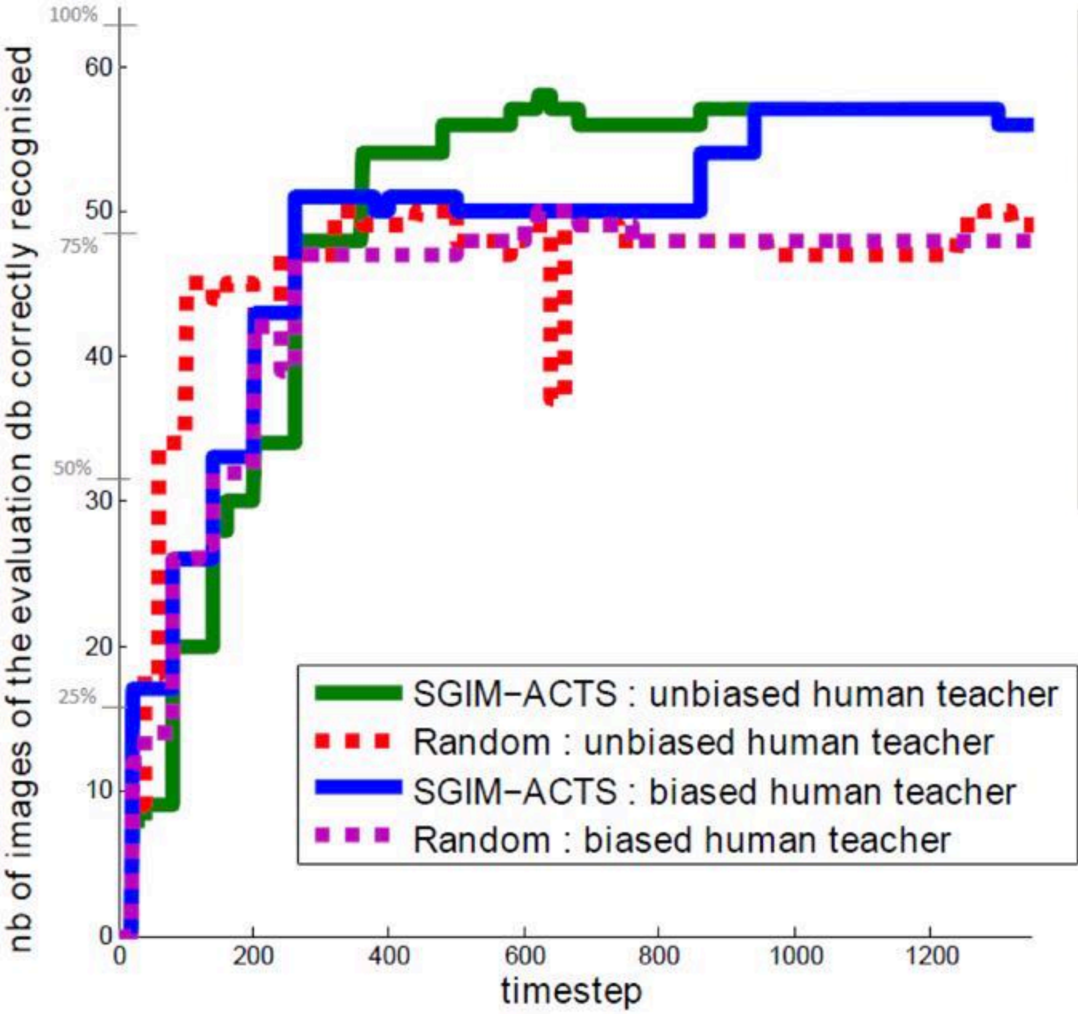
Learning through observation (blue color), TakeLiftFall manipulation (yellow color).

Interaction physique et sociale

Fusion motivations intrinsèques et apprentissage social



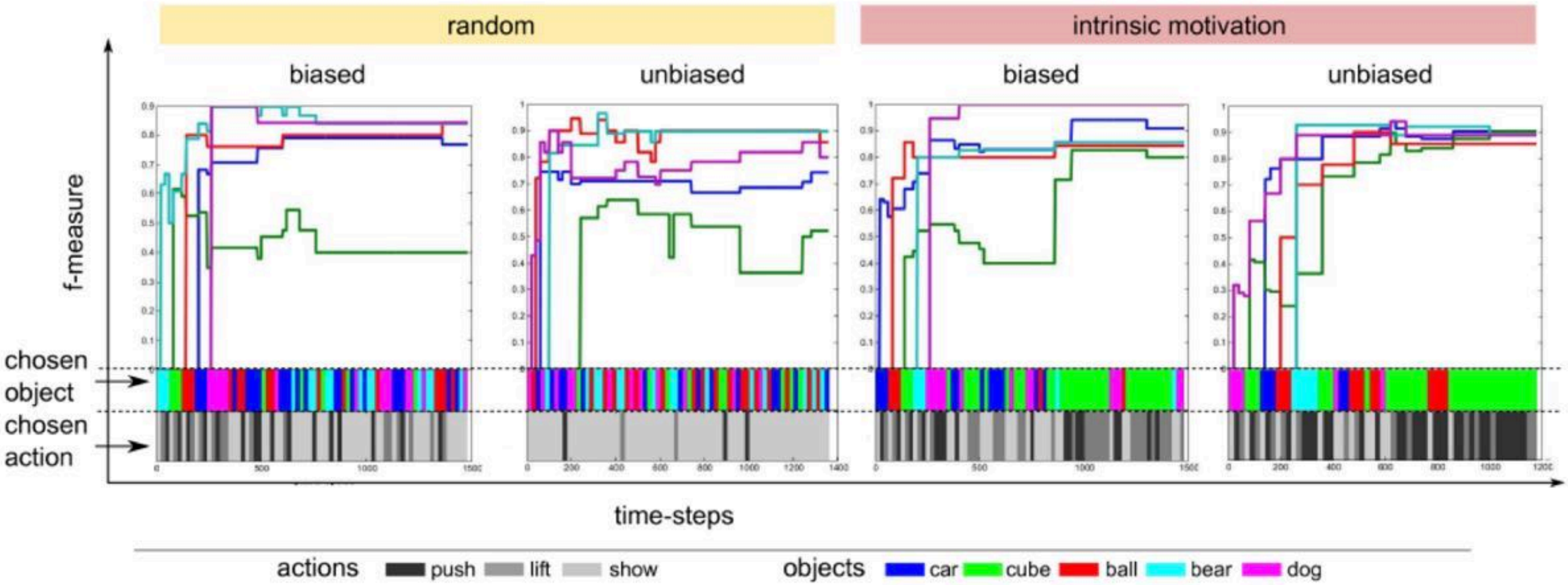
Interaction physique et sociale



Amélioration des performances par le choix actif des actions et des objets

Interaction physique et sociale

Biais vers les objets les plus difficiles et les actions les plus efficaces



Plan du cours

Introduction
- Robotique développementale
- Modèles de sacs de mots visuels

Cartographie - Localisation
- Application des sacs de mots visuels

Apprendre à interpréter des images
- Distinguer soi / non soi
- Modéliser soi/objets/humain

Apprendre à chercher des objets
- **Apprentissage de saillance visuelle**

Apprendre à éviter des obstacles
- Prédiction de profondeur en video monoculaire

Motivations

Human vision

- Foveal/peripheral field of view
- Active perception

Computer vision

- Each pixel has the same resolution
- Image is processed as a whole

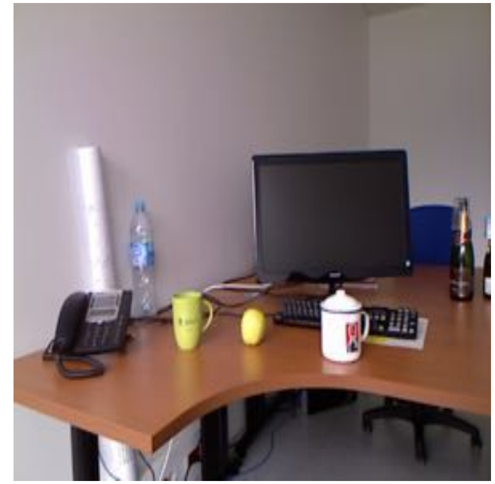
Robotics vision

- Exploit Computer vision algorithms
- Action can be used 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

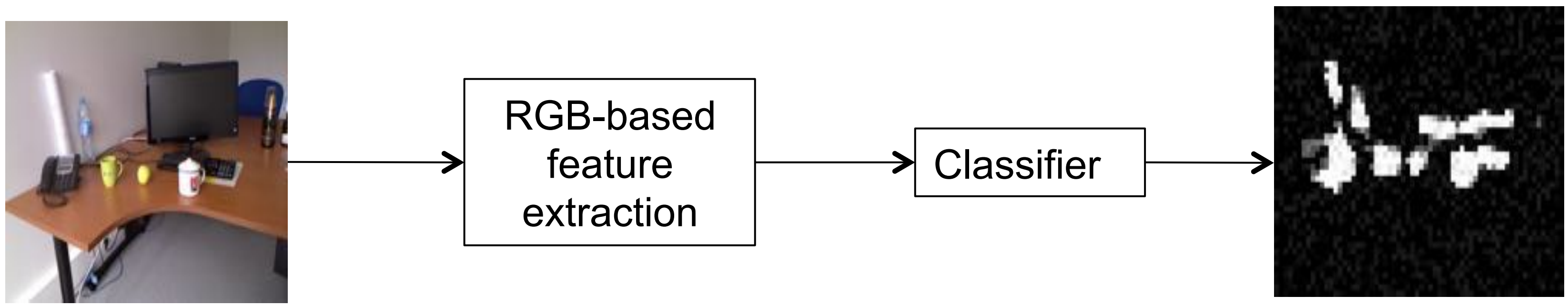
- Incremental learning
- Guiding learning through action selection

Using foveal and peripheral vision alternance

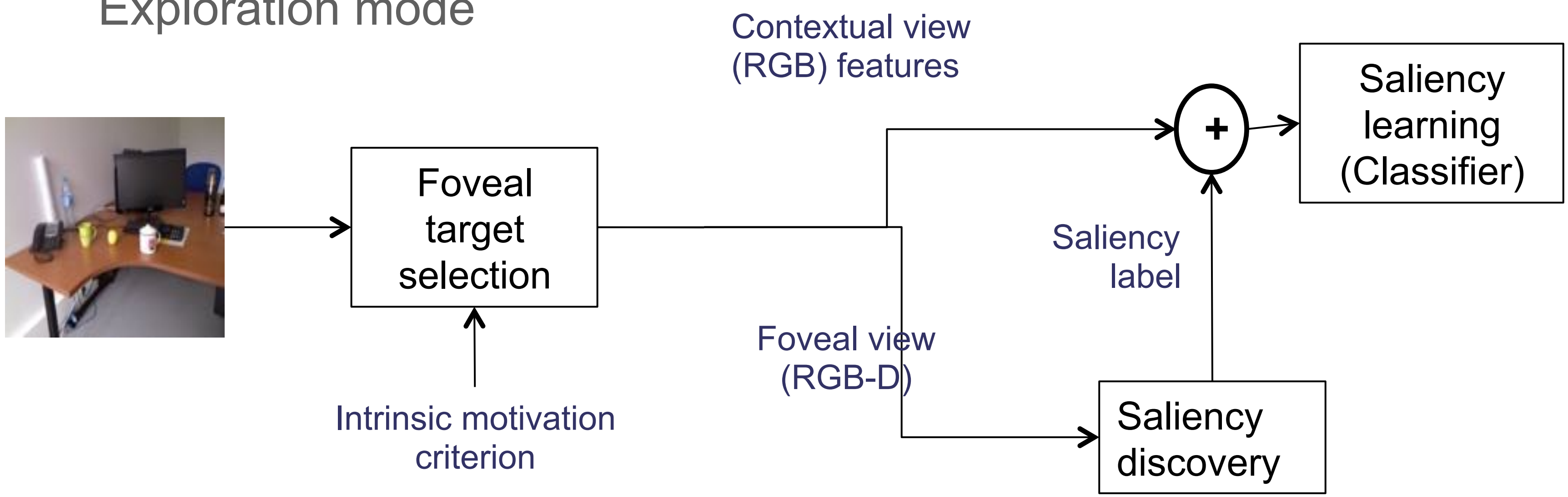
- The fovea provides accurante and reliable data
- This data is used to learn saliency in the periphery

Proposed approach: General architecture

Exploitation mode

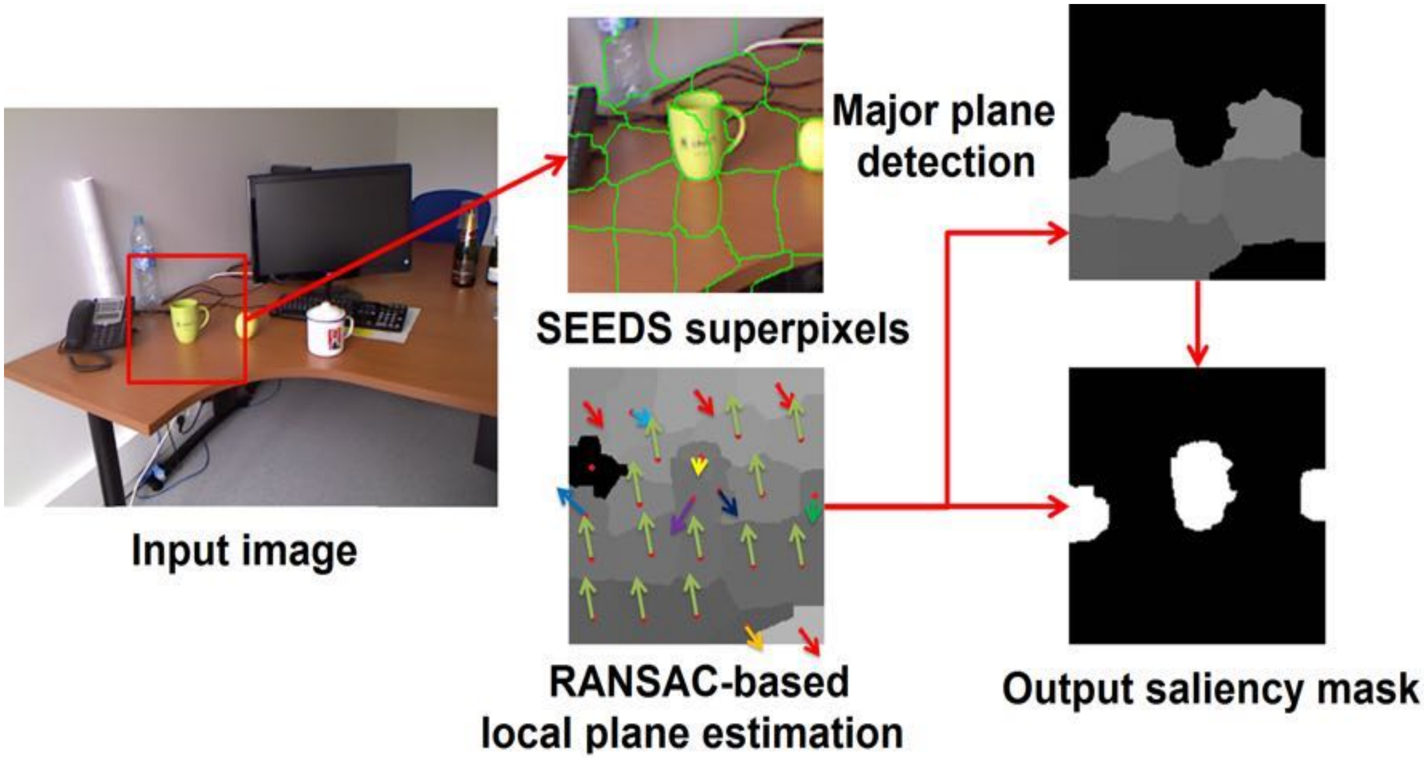


Exploration mode

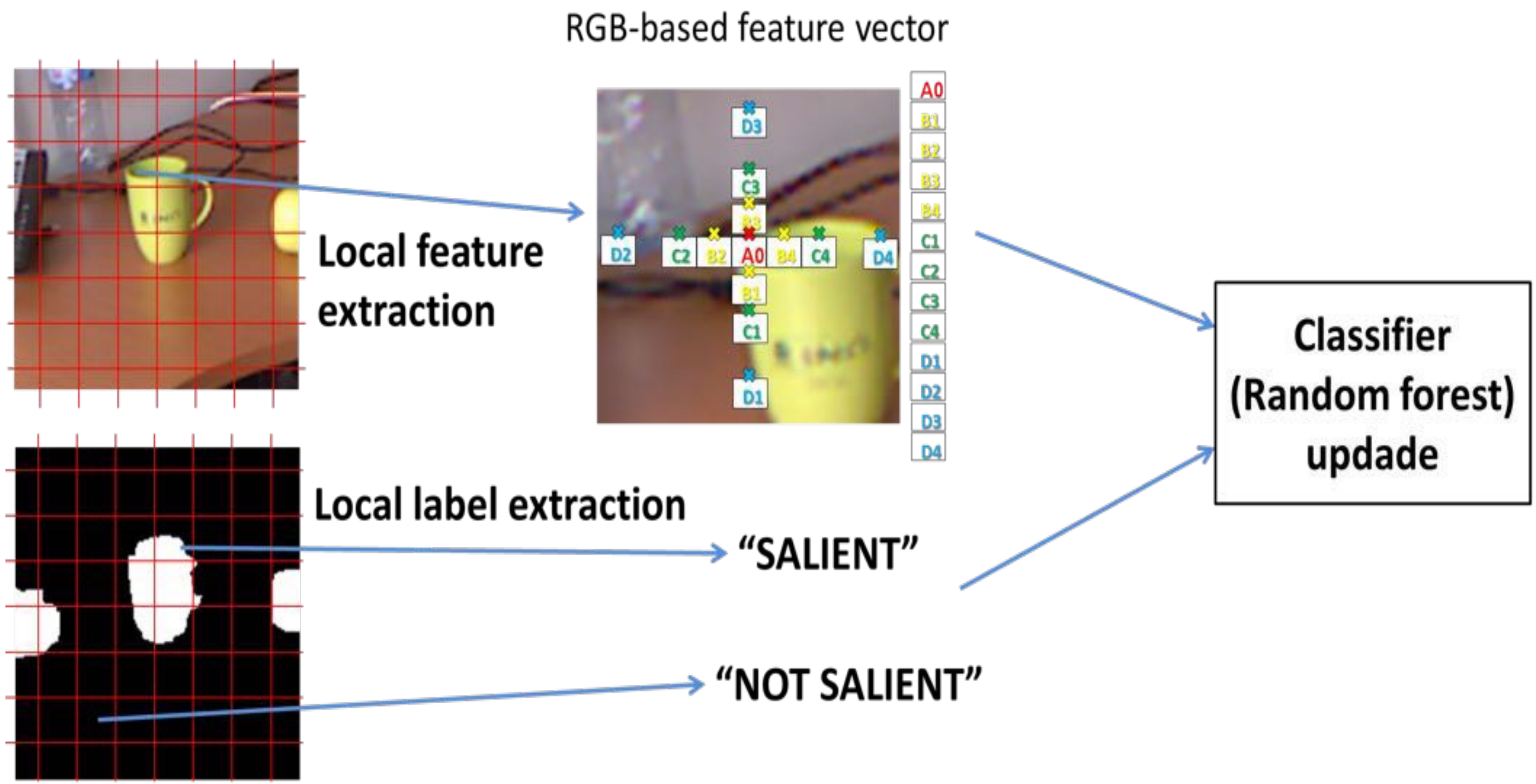


Proposed approach: Salient elements discovery

- Salient elements are objects lying on plane surfaces
- Processing is too expensive to be processed on the whole frame



Proposed approach: Saliency learning

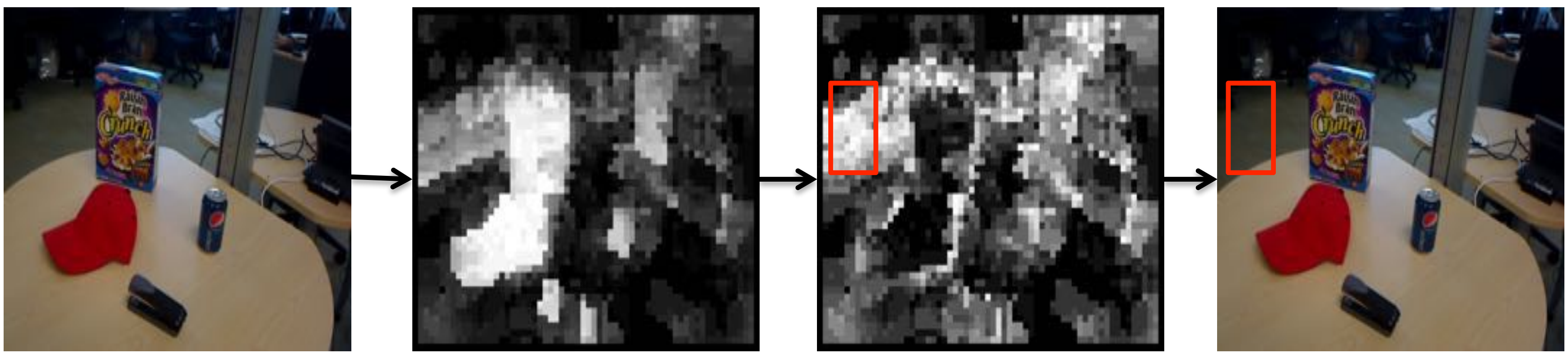


Proposed approach: Intrinsically motivated exploration

Select the fovea with an intrinsic motivation criterion

Uncertainty

- Select the foveal area based on saliency map fuziness

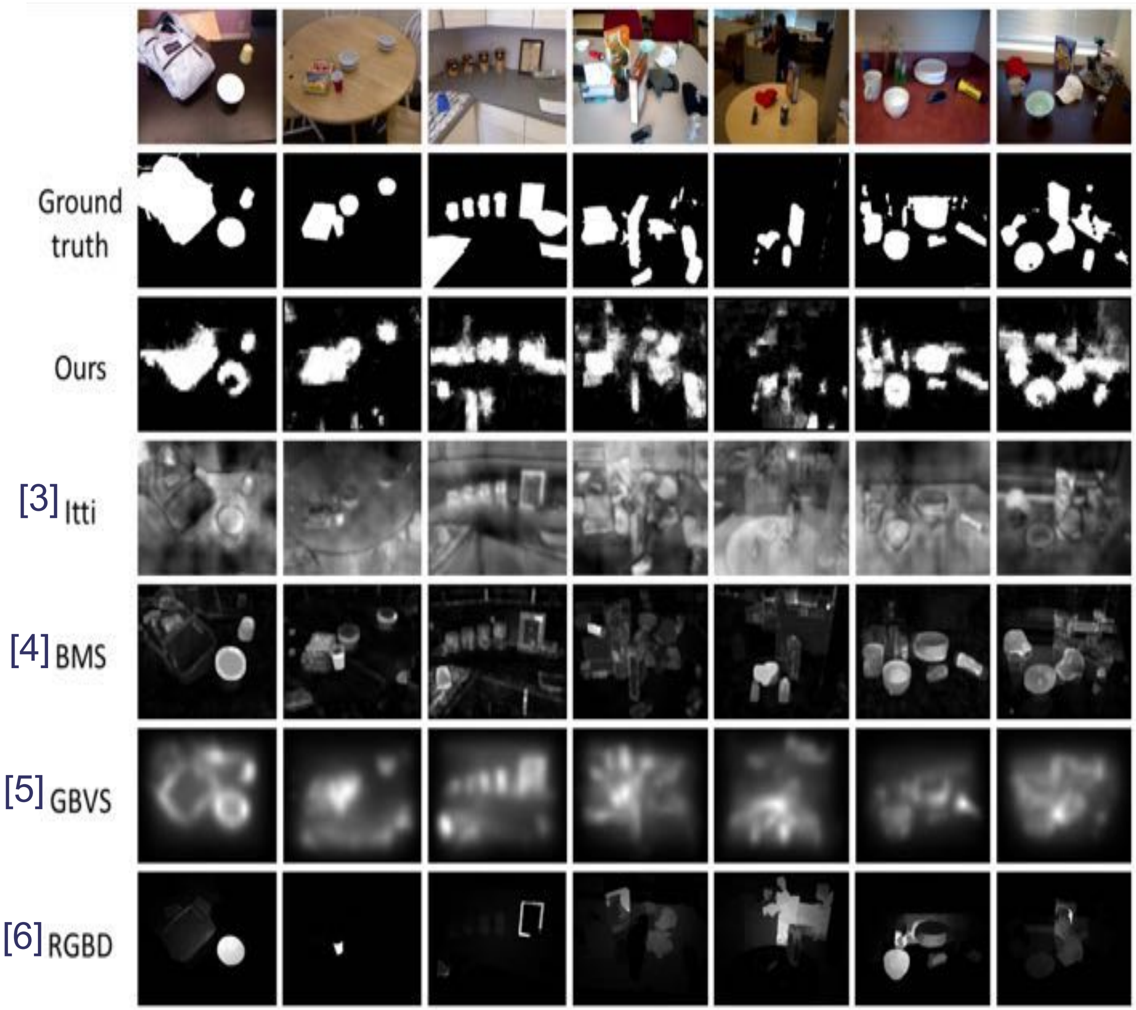


Novelty

- 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

Using RGB-D scenes[7] and GIT [8] public datasets

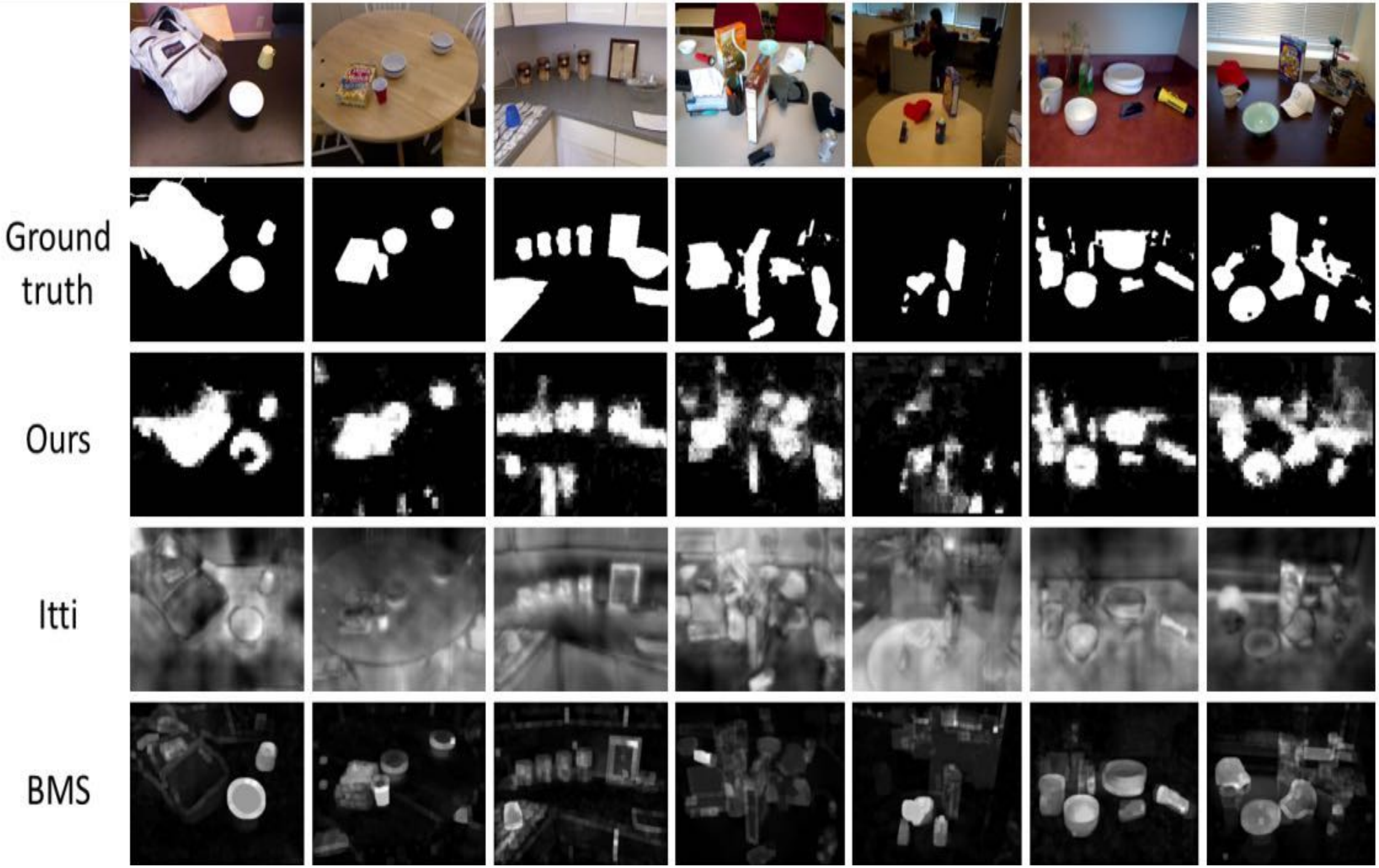


- Ground truth manually determined
- Compared with 4 methods

Visually

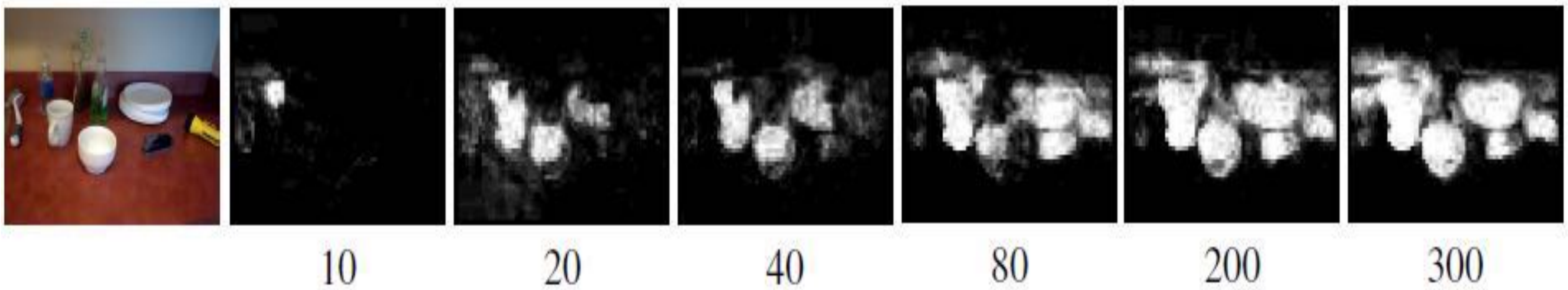
- Our method looks most like ground truth
- Provides rough idea of object shape

Experimental results: Comparison to state of the art

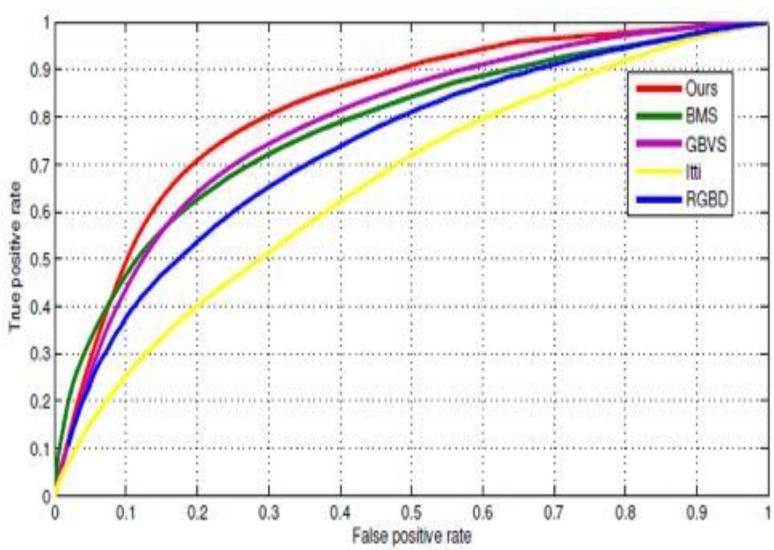


Experimental results: Comparison to state of the art

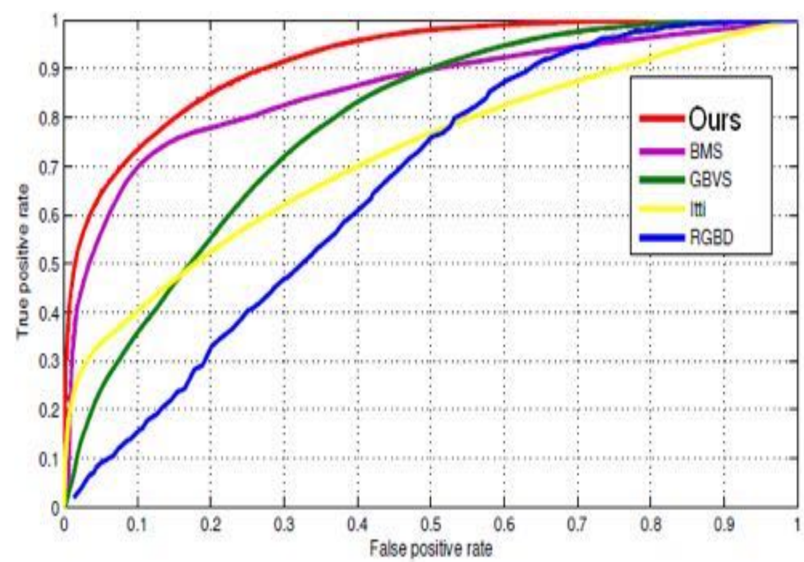
Saliency learning evolution



ROC curves on still images

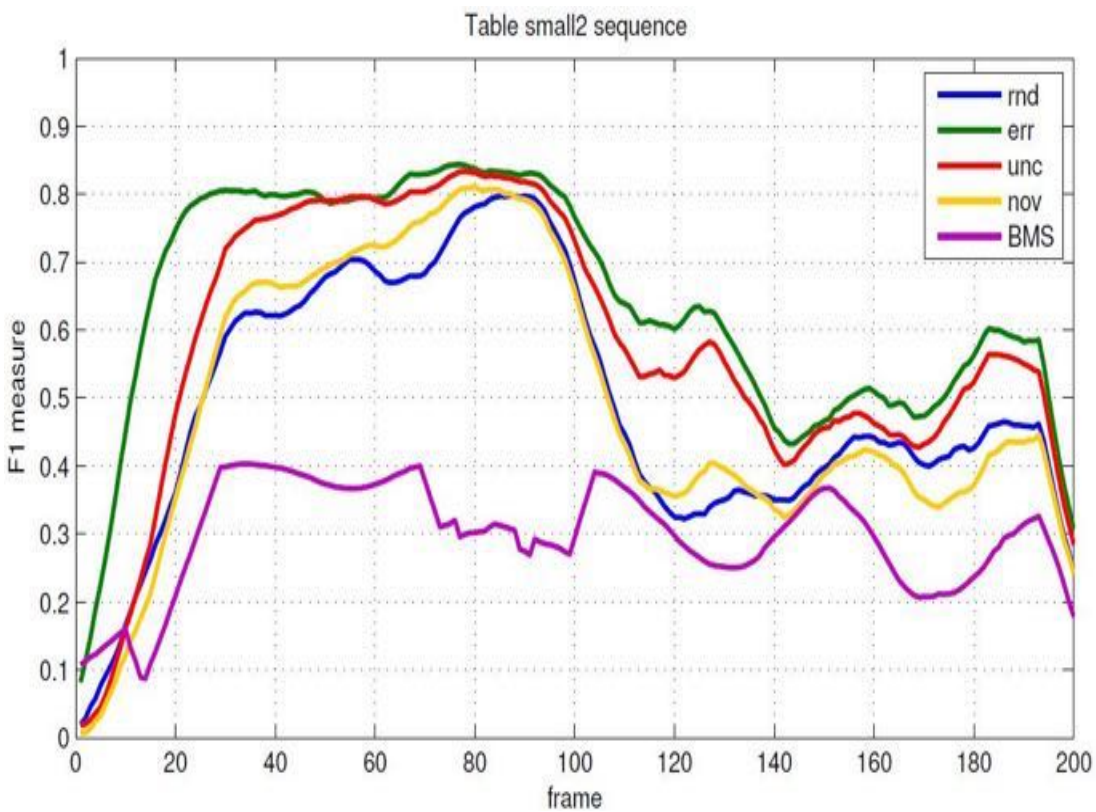


ROC curves on video



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

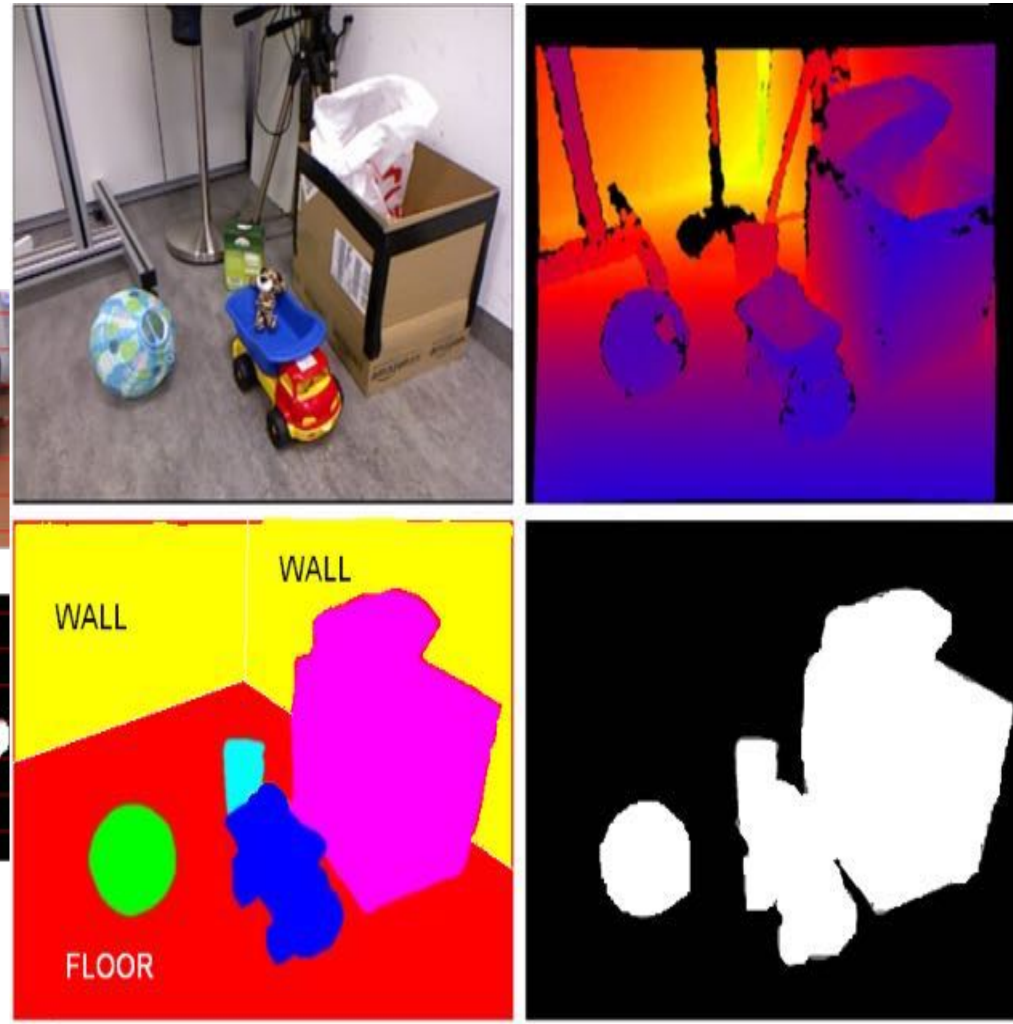
Application on a mobile robot

Segmentation

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

Learning

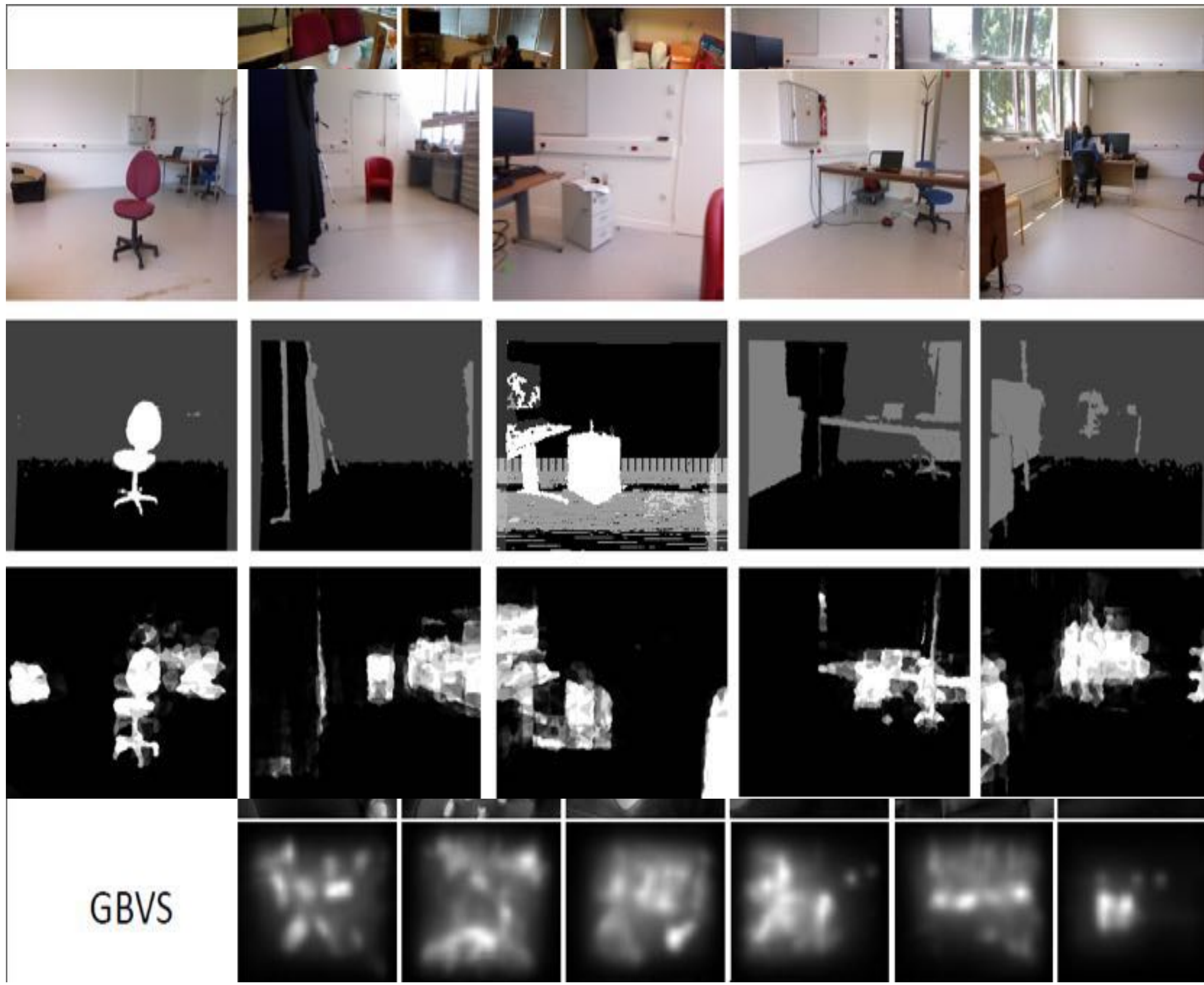
- Features : RGB-based
- Labels : based on the segmentation result



Experimental results

State-of-the-art

Segmentation vs saliency



Experimental results

Exploration strategies

The robot moves in its environment

– how to explore better than randomly ?

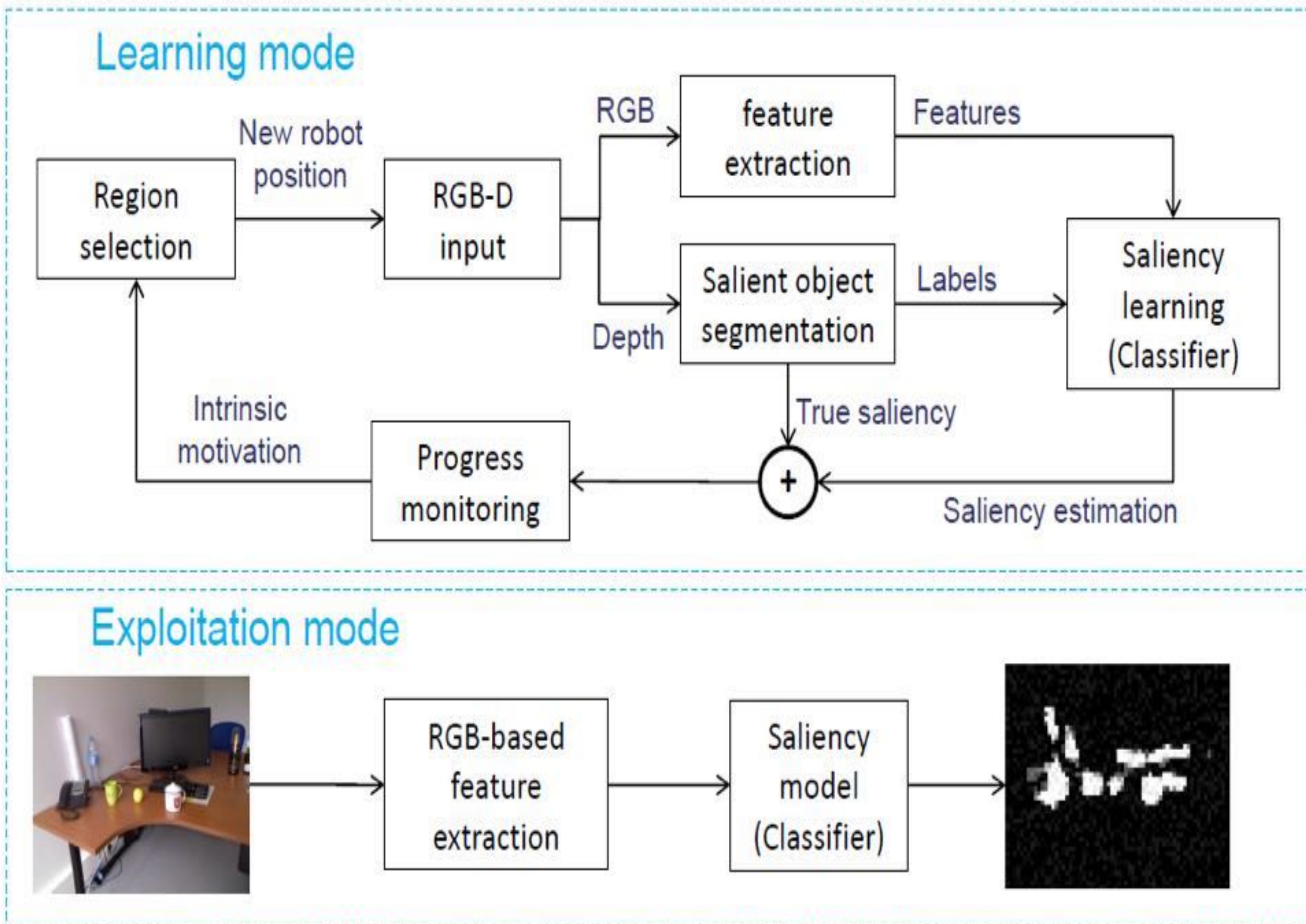
Early stategies : Select samples in the input image

– Based on novelty or uncertainty scores
– 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



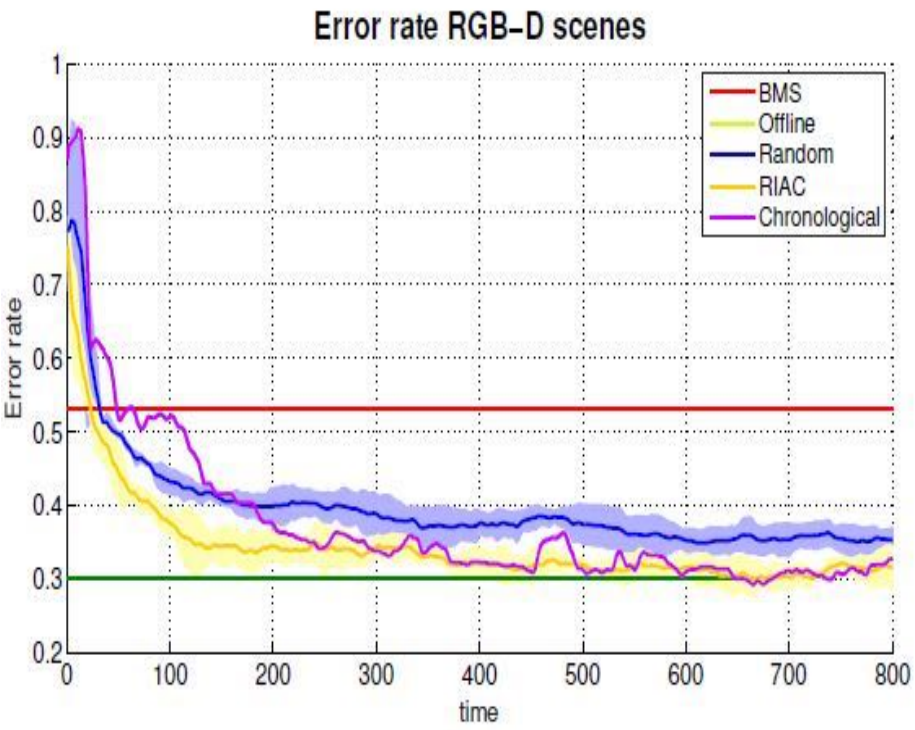
Results

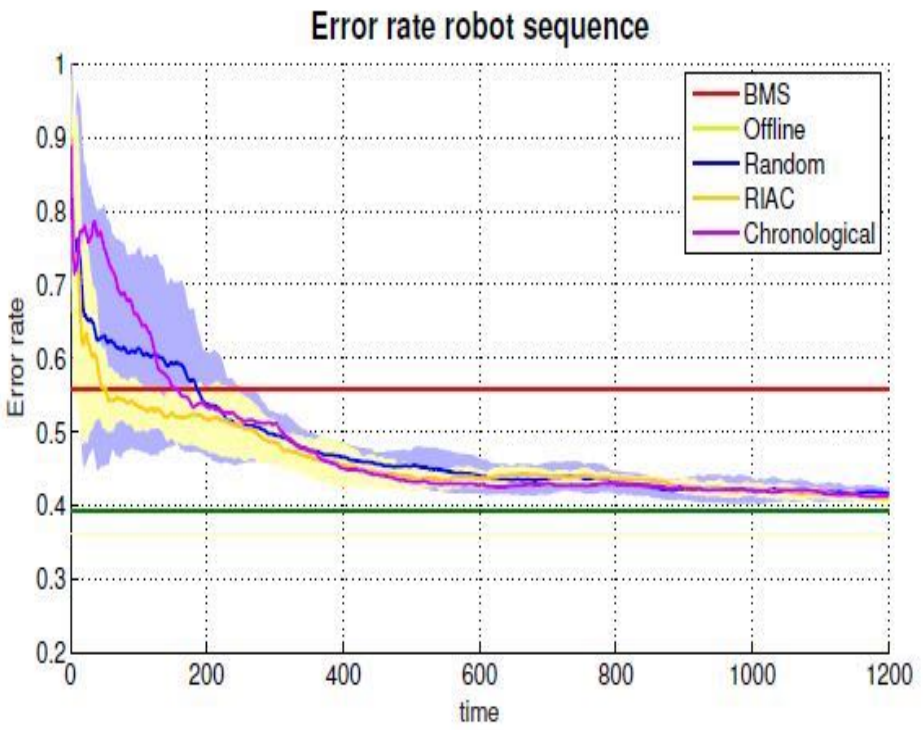
Behavior illustration in simulation

Results

Evolution of the error rate

- IAC improves learning speed
- Similar final performances





Plan du cours

Introduction

- Robotique développementale
- Modèles de sacs de mots visuels

Cartographie - Localisation

- Application des sacs de mots visuels

Apprendre à interpréter des images

- Distinguer soi / non soi
- Modéliser soi/objets/humain

Apprendre à chercher des objets

- Apprentissage de saillance visuelle

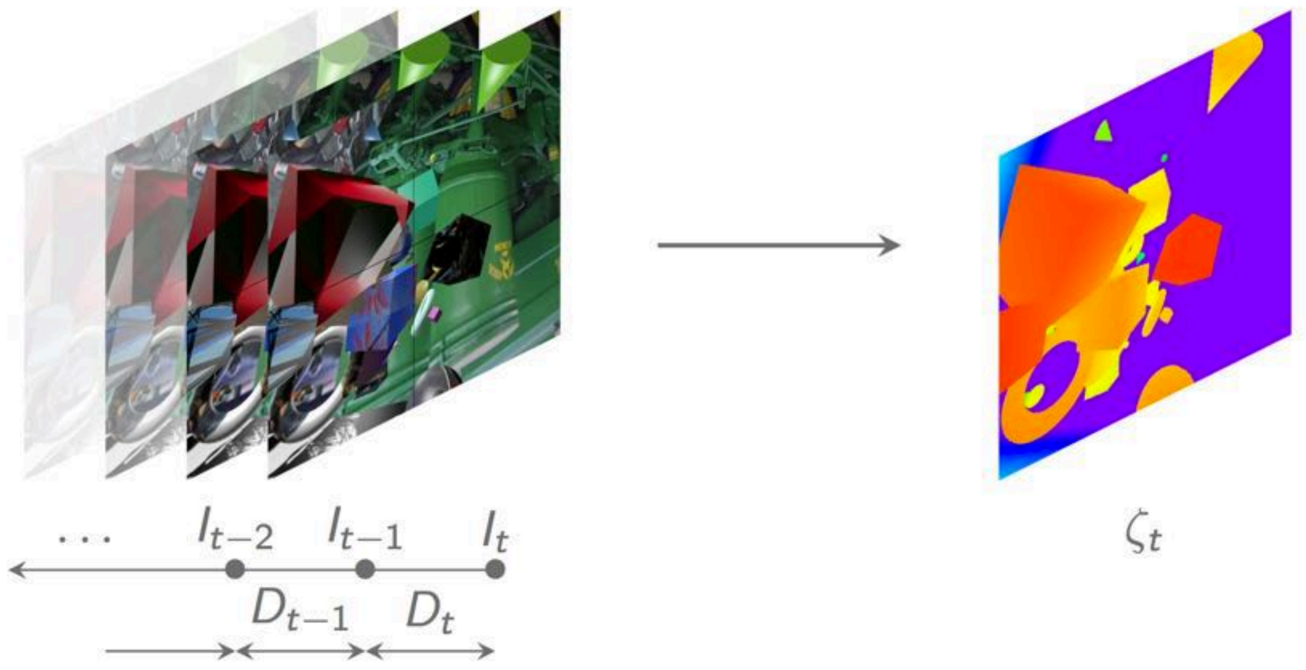
Apprendre à éviter des obstacles

- **Prédiction de profondeur en video monoculaire**

Prédiction de profondeur

Objectifs

- Evitement d'obstacles sur un drone
- Exploitation vision monoculaire
- Prédiction de la profondeur depuis la vidéo



Prédiction de profondeur

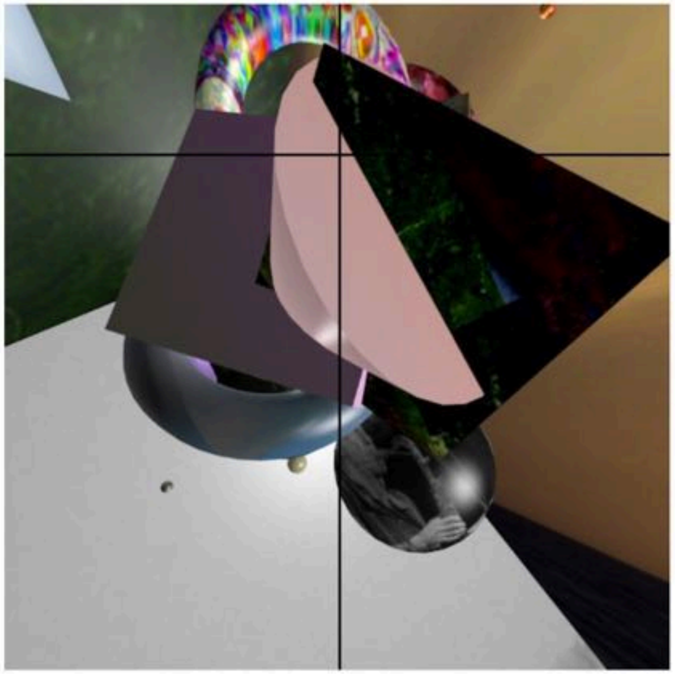
Exploitation du contexte

- Séparation rotations/translations difficile
- Exploitation de vidéos stabilisées en rotation
- Systèmes optiques/numériques

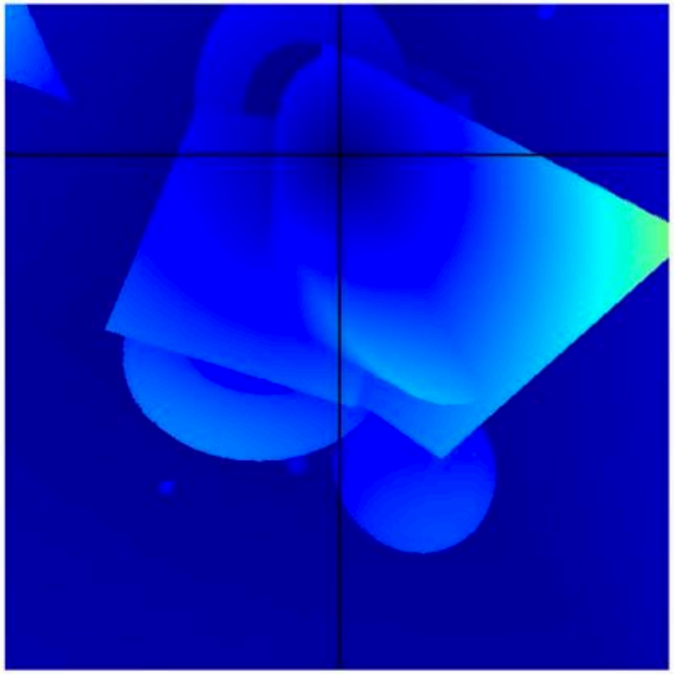
Prédiction de profondeur

Perte d'information au point de fuite

- Disparité = 0
- Besoin de régulariser
- -> Apprentissage direct de la profondeur



Input Images



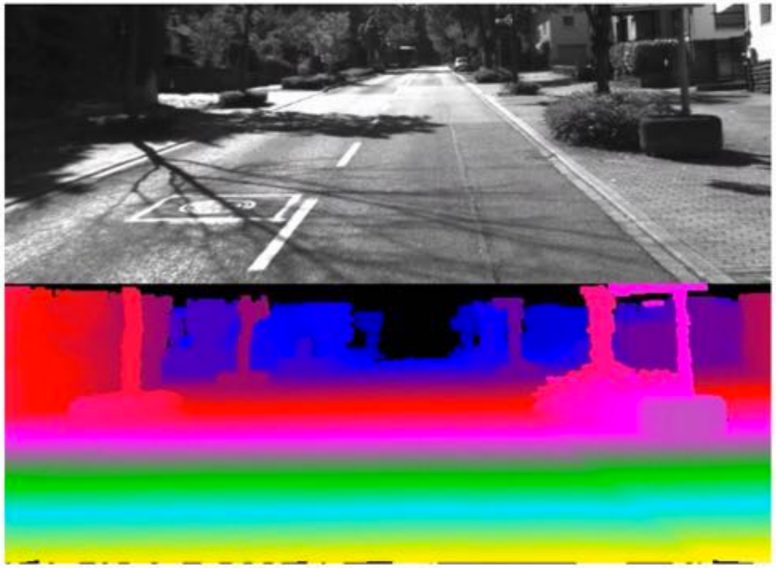
Disparity Map δ

Prédiction de profondeur

Données d’apprentissage

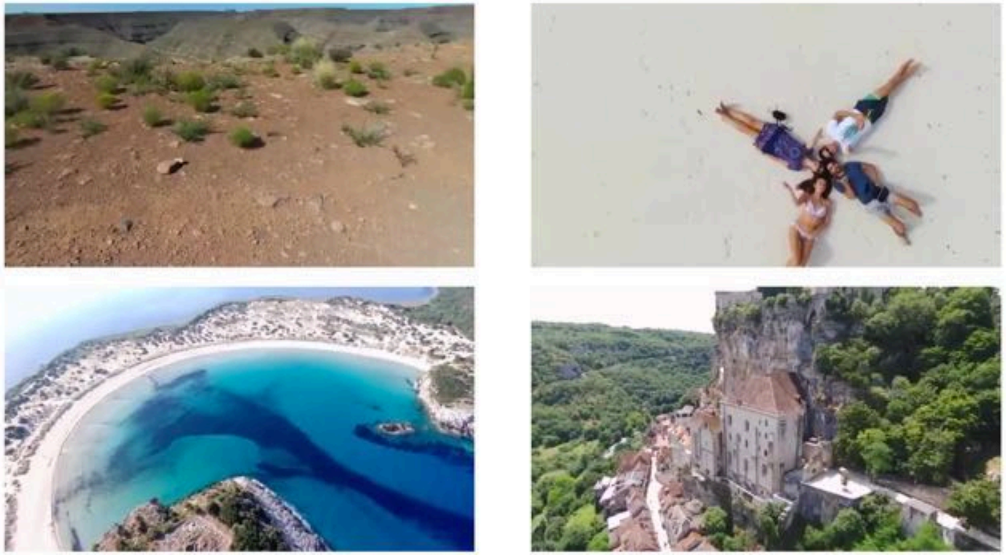
- Besoin de points de vue variés
- Besoin d’images annotée

Pdv contraint / verité terrain



KITTI Dataset

Pdv varié / pas de verité terrain



Drone

Prédiction de profondeur

Exemple d'apprentissage en stéréo

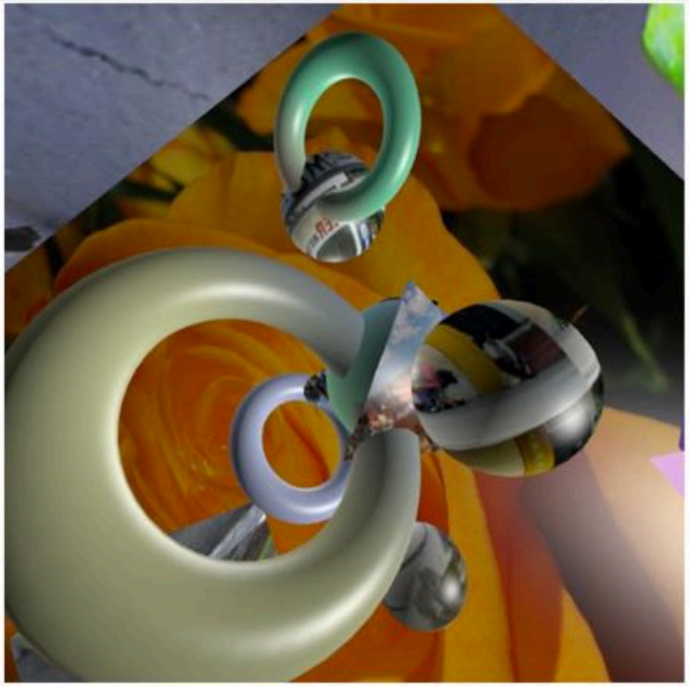
- Auto supervision : profondeur prédite des deux images doit correspondre
- Apprentissage spécifique à l'application

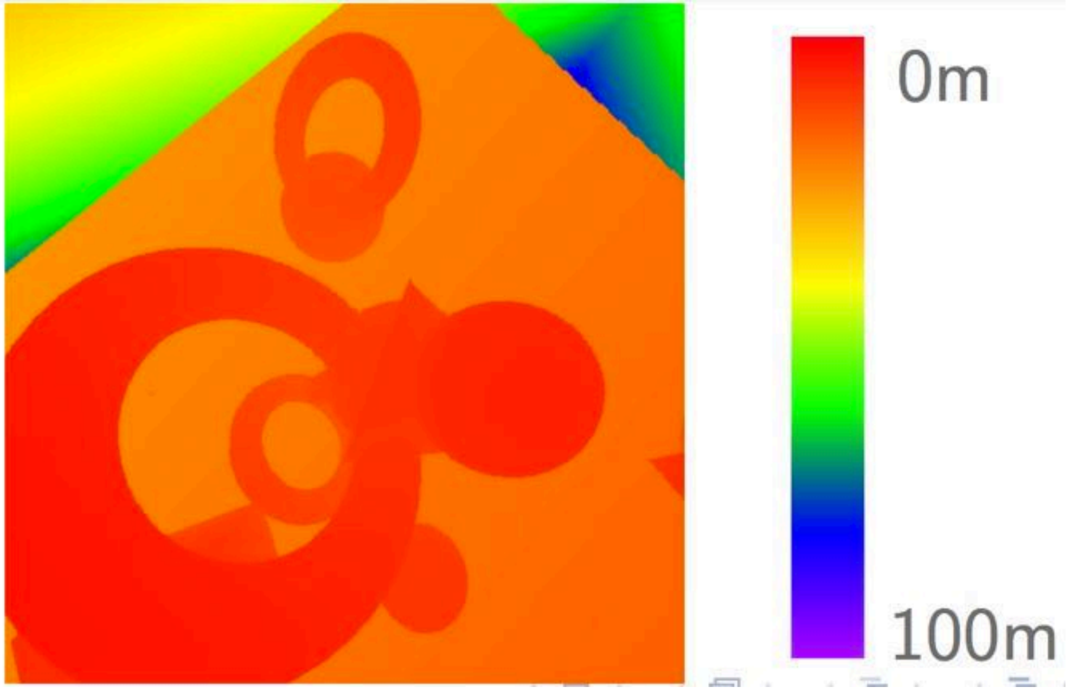
Unsupervised Monocular Depth Estimation with Left-Right Consistency
Godard, Mac Aodha and Brostow

Prédiction de profondeur

Apprentissage en simulation: Still box dataset

- Mouvement du drone aléatoire, pas de rotation
- Scènes rigides, texture / formes aléatoires
- Profondeur depuis image seule impossible

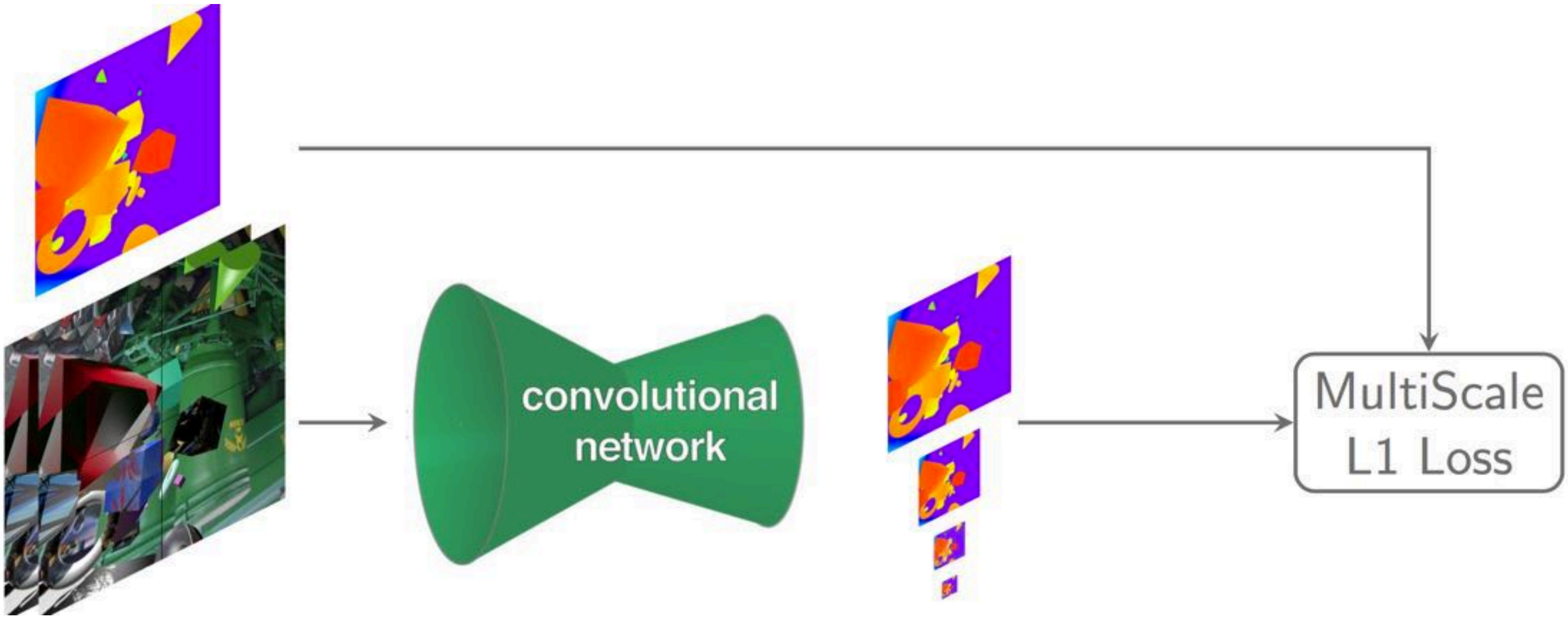




Prédiction de profondeur

Apprentissage via réseau convolutionnel

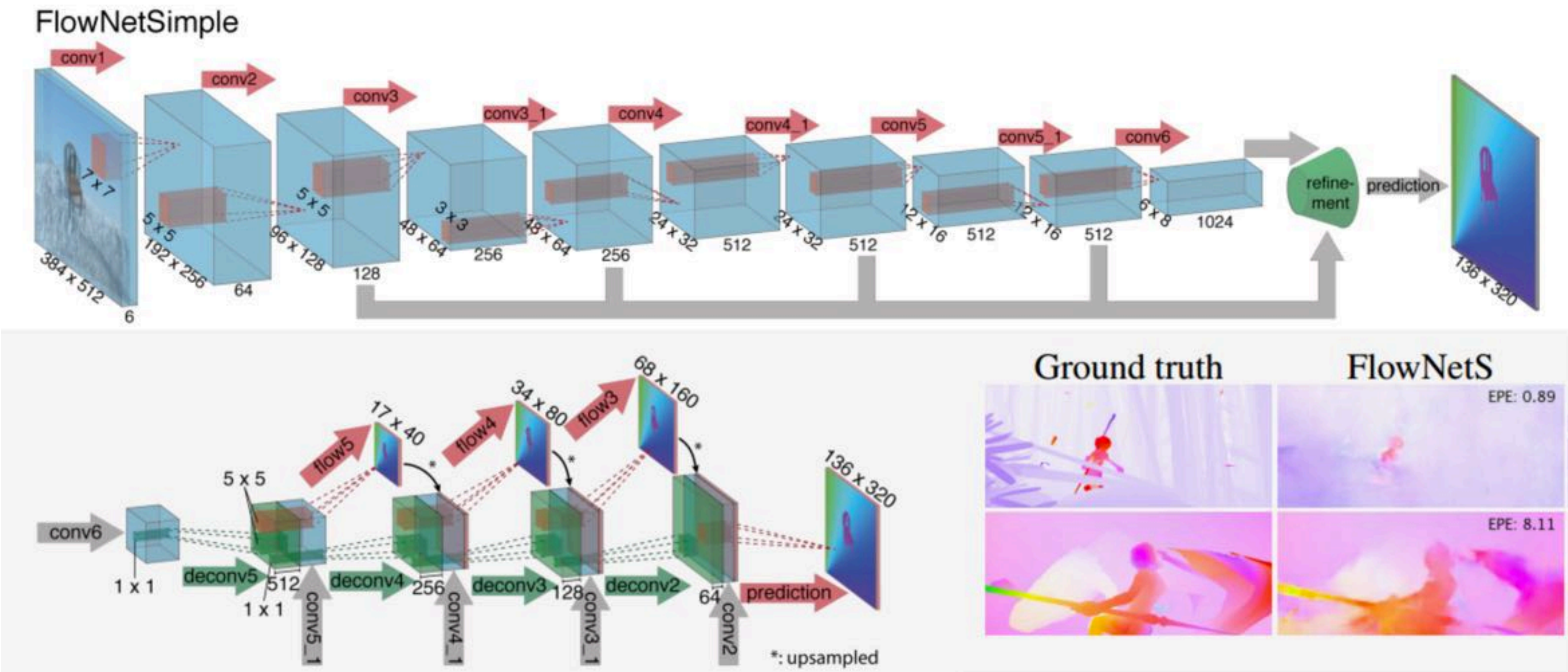
- Disparité pas suffisante (pb au point d'expansion)
- Apprentissage direct de la profondeur
- Généralisation possible à des images réelles (exploite le mvt)



Prédiction de profondeur

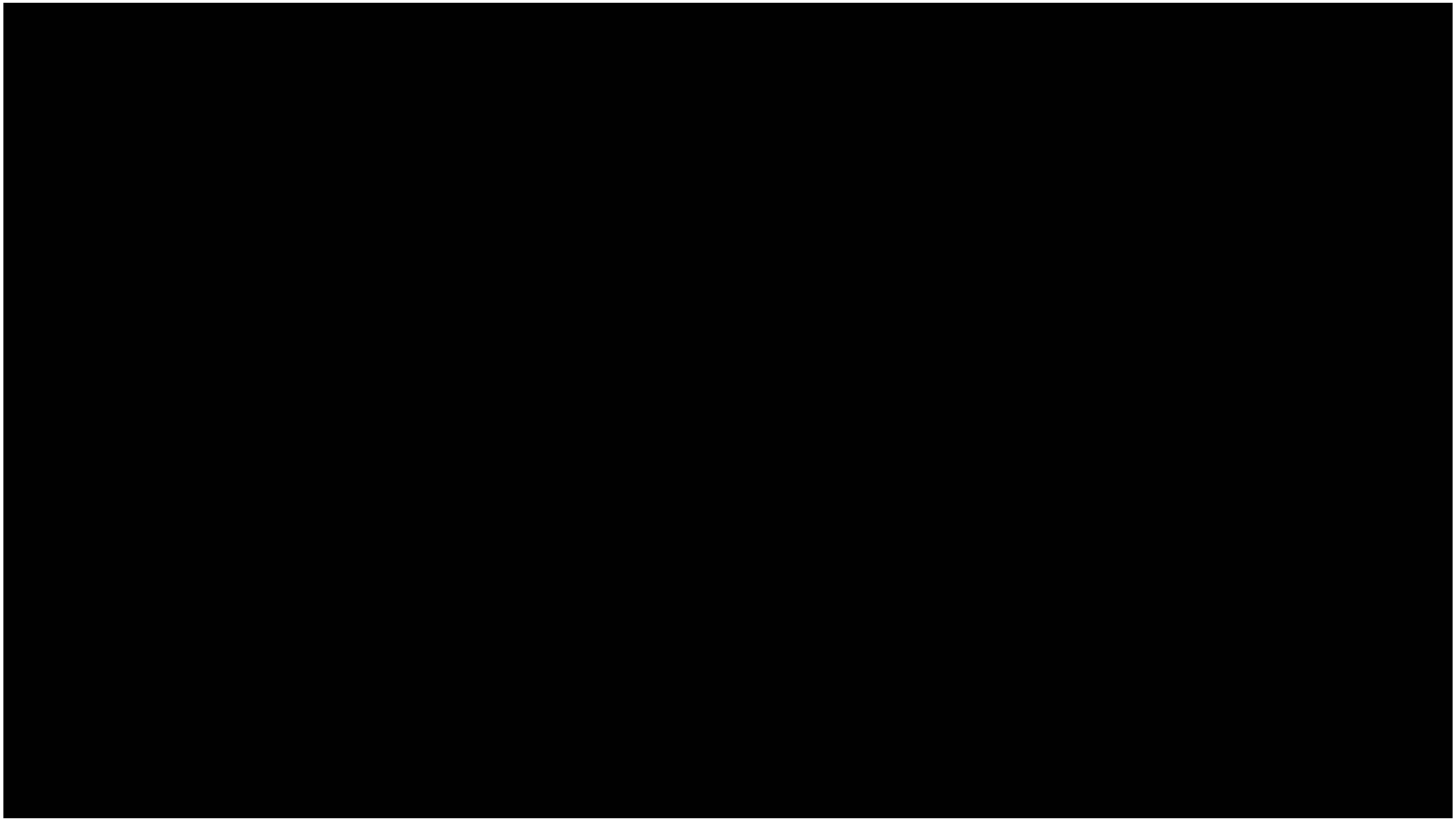
Réseau dérivé de FlowNetSimple

- Structure encodeur/décodeur



Prédiction de profondeur

Apprentissage via réseau convolutionnel



Conclusion

Interprétation d'image en robotique

- Localisation,
- Guidage,
- Reconnaissance d'objets,
- Recherche d'objets
- Evitement d'obstacles
- …

Capacité à agir

- Explorer un environnement
- Choisir un point de vue
- Obtenir une information de supervision

pour mieux interpéter

MERCI