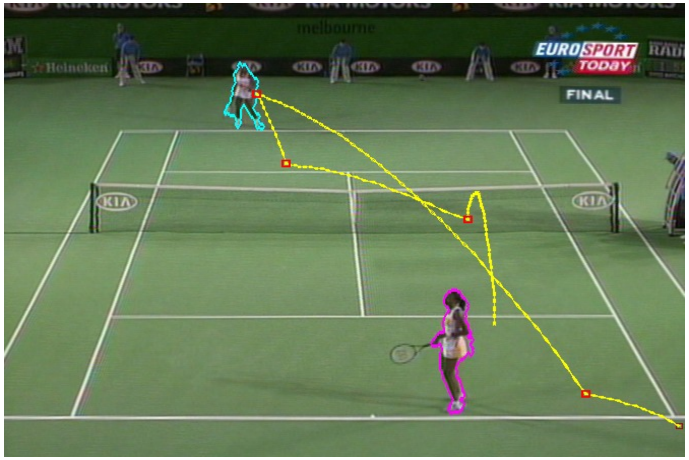
# ACASVA: Adaptive Cognition for Automated Sports Video Annotation

PROJECT OVERVIEW

David Windridge

**CVSSP,**
**University of Surrey,**
**Guildford, UK**



# Project Overview

1. ***Objectives & Motivation***

2. ***Key Themes***

3. ***Cognitive bootstrapping***

4. ***Inspirations***

5. ***Work Plan: axes of collaboration***

# 1. Objectives & Motivation

**Practical goal:** generalise existing tennis video annotation system => *transfer learning where relevant.*

- Accomplished via the ***mono-modal*** and ***cross-modal*** bootstrapping of high-level visual/linguistic structures

*-parallels innate human capabilities?*

=> ***<u>psychological component</u> of ACASVA***

*<u>WHY SPORT VIDEO?</u>*

- Rich in terms of visual structures and rule-induction possibilities, all of which have strong linguistic correlates

# 1.1 Scientific & Engineering Motivations

**Engineering problem:**

How do we connect low-level representations to the abstract structures governing environment - can autonomously specify a formal '*grammar*' of entities/agents common to audio and vision?

**Scientific problem:**

a) How are visual grammars organised and employed in the learning problem?

b) How grammars modified by prior linguistic knowledge of the domain.

c) At what stage does formal abstraction of visual features takes place?

d) How do visual grammars map onto linguistic grammars?.

e) How inferred high-level linguistic concepts (e.g. rules of an unfamiliar game) influence lower-level visual learning (e.g. gaze-specification)?

# 2. Key Themes

**1. Cognitive Bootstrapping**

-induction of the representational hierarchy

-passing of hypotheses ***up*** and ***down*** the representational hierarchy

**2. Cross-Modality**

-passing hypotheses ***between*** differing modes of representational hierarchy

-discovering common grammar underlying Visual/Aural grammars?

(agent/object/frame semantics are baseline *a priori* assumption in ACSAVA)

-new event detection as DISTINCT from outlier detection (cf DIRAC)

Event is new if disparity between (weakly-constrained) unimodal classifier output and fused (highly constrained) classifier output, provided unimodal confidence high

## 2. Cross-Modality -cont.

## Model of Cross-modal Hypothesis Passing (Sound-> Vision):

The semantic parser (first-order logic-based) constitutes phrases as a relation between:

***agents*** (e.g. players, officials),

***objects*** (e.g. the ball),

**actions** (e.g. serving),

***reference-entities*** (e.g. the court),

***abstract quantities*** (e.g. match-score)

E.G. The commentators phrase **``*Henman hits the ball into the net*''** refers to:

**Agent:**: ``*Henman*'', **object**:``*ball*'', **action**: ``*hits*'', **reference objec**t: ``*the net*''.

Can extract semantic relations between the vocabulary entities (a *'player'* always *'hits*' a '*ball*', but not vice-versa) to render above phrase as a predicate sequence:

***agent****(player); **object**(ball); **action**(hits); **reference-entity**(net)*

=> can then be related to the high-level visual predicates *(eg hypothesise that high visual saliency object at termination of ball trajectory correlates with the spoken word 'net)'*

## 3. Transferable Learning Between Game Domains

- Is it easier to port learning from the top or bottom of the hierarchy?

- Also true for humans?

## 4. Interdisciplinarity - Speech/Vision/Psychology

”Cognitive bootstrapping lies at the interface of the cognitive science and machine learning disciplines, being both a mechanism for learning performance optimisation as well as a fact of human cognition. The concept is one that is susceptible to empirical definition via the methods of experimental psychology, and to direct stochastic and structural modelling by machine learning techniques.”

**- Two-way Scientific/Engineering collaboration:**

a) Engineering in assistance of scientific endeavour

*(eg behavioural mining of eye-tracking data)*

b) Science to inspire & benchmark engineering solutions

*(eg gaze base weak supervision, determining the order in which tennis rules induced, determining what is a hard problem)*

# 3. Cognitive Bootstrapping

## 3.1 Cognitive Bootstrapping in machine learning:

### Main Features

**1. Simultaneous Learning of Domain Model *AND* aptest *Mode* of Representation**

- Issues of Philosophical foundation/Underdetermination of solutions

=> need objective constraints on representation

- use **efficiency** (eg complexity of player model is dictated by induced rules), **action-relevance**

**2. Representational Sumbsumption**

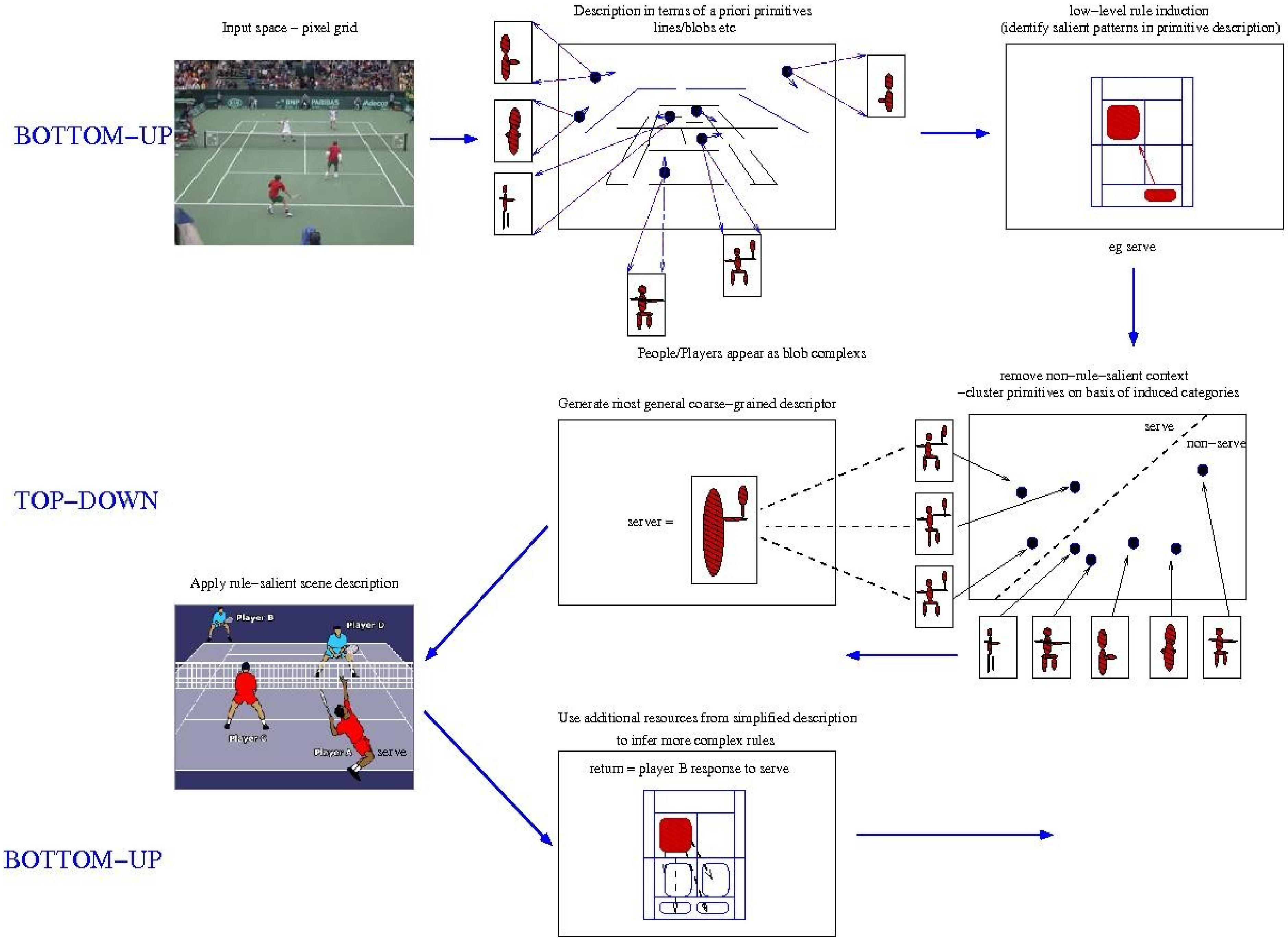
- progressive abstraction/decontextualisation of scene-description parameters

**3. Inter-level Feedback of Representational Hypotheses**

- top-down feedback is most novel aspect of ACASVA

**4. Potential to Relax Low-level Global Coherence/Consistency Requirements**

Only require top-down consistency -cf Rensink?



BOTTOM–UP

Input space – pixel grid

Description in terms of a priori primitives
lines/blobs etc

low–level rule induction
(identify salient patterns in primitive description)

eg serve

People/Players appear as blob complexs

TOP–DOWN

remove non–rule–salient context
–cluster primitives on basis of induced categories

Generate most general coarse–grained descriptor

server =

Apply rule–salient scene description

BOTTOM–UP

Use additional resources from simplified description
to infer more complex rules

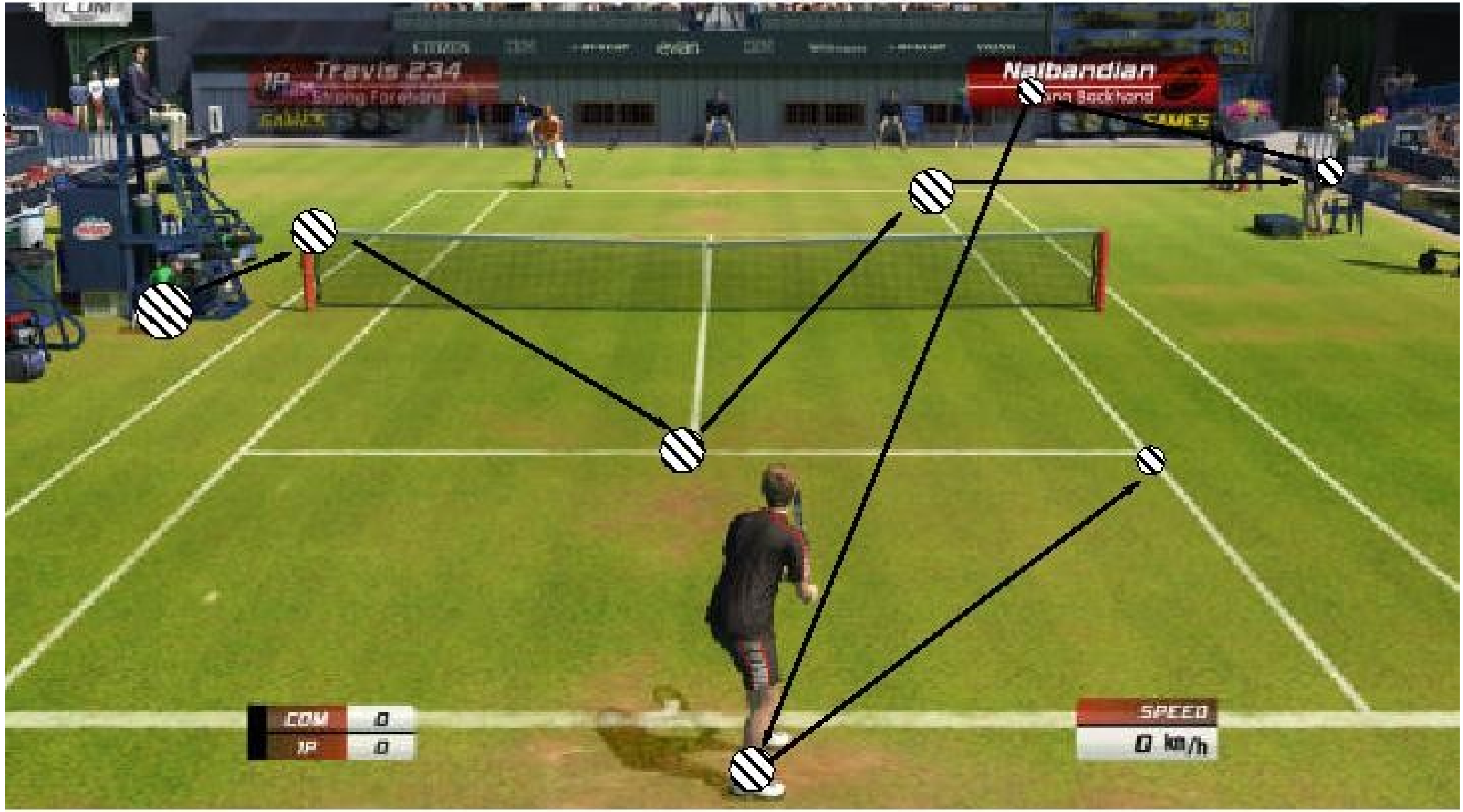
return = player B response to serve

# 3.2 How Might Cognitive Bootstrapping Manifest Itself in the Human Domain?

1. Initially observer attends to multitude of low-level features driven by visual/aural saliency
2. Observer induces partial game rules
3. Attention becomes focussed on rule-relevant entities
4. Focussed-attention permits deeper rule quantification (can assess independently)
5. Goto 3

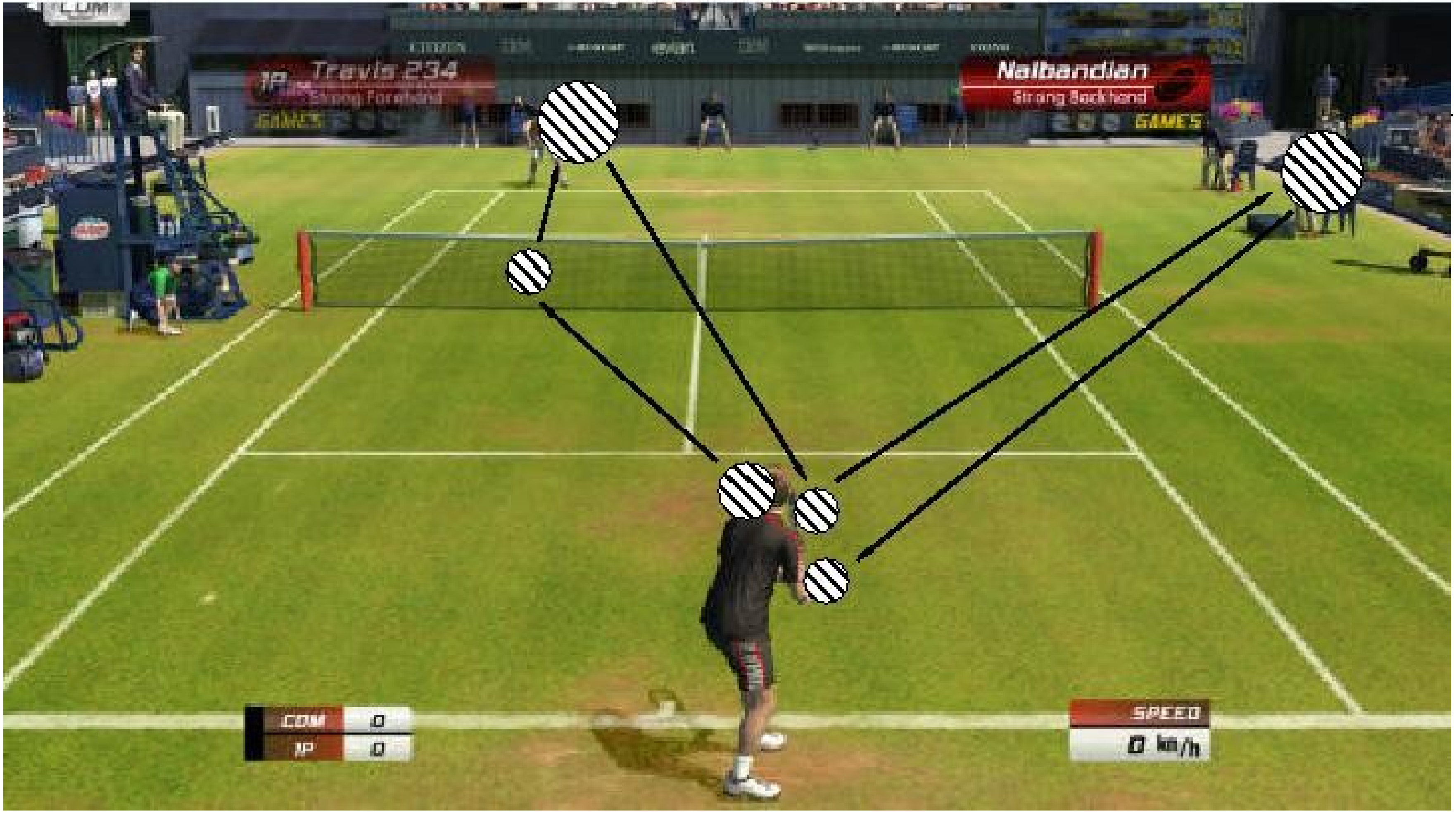
**=> should be apparent in gaze behaviour:**

# Visual (edge, corner, motion)–Saliency–based Fixation



(Note lower dwell times on average)

# Rule–Saliency–based Fixation



# 3.2 Theoretical Basis for Cognitive Bootstrapping

**Have an *a priori* scene description** ***eg edge/line detections, blob tracks etc***

- Not appropriate for task of game annotation

*(want 'Player A hits the ball into the net'-type description).*

- Is there an ***objective measure*** of the 'best' scene-description that can gives this type of output *without it being specified in advance?*

=> ***YES***: Minimum Description Length (MDL) :

***Best Hypothesis H = min$_H$ Length(H) + Length(Data|H)***

In **parametric** terms want to minimise:

**cost-of-scene-description=cost-of-parametrization+cost-of-parametric-description**

***cost-of-parametrization*** *is min. bit length of algorithm required to specify parametrized entities:* **cost-of-parametric-description** *is number bits required to describe scene using these parameters*

Argue most compact MDL description/coordinatisation of a sport match at a particular time ***IS IN TERMS OF GAME RULES - EG:***

***match****{1|2|3}*, ***se****t{1|2|3}*, ***game****{1|2|3}*, ***serve_end****{top|bottom}*, ***serve_end****{left|right}*,
***player_A_location****{1111|1112|..}*, ***player_B_location****{1111|1112|..}*, ***rally_nº****{1|2|..}*
+ ***player/ball offset terms*** *(ie deviations from expected location}*

=> parameters function as ***indexicals w***ithin the subsumption hierarchy

**IF** we have inferred game rules correctly, offset terms are confined with small area (a square, roughly), so that the above description is close to description length minimum.

**IE MDL description= Most Rule-Salient Scene Description**

Can approximate MDL (Kolmog.) complexity-gives ***Bayesian Information Criterion:***

***BIC cost= nº parameters*ln(nº samples$^{0.5}$) - ln max liklihood(Data|parameters)***

Hierarchical parameters ***co***-index lower-level parameters (eg through progressive sub-factoring), so that nº of parameters at level ***i*** appropriate to describing data of intrinsic dimensionality ***D*** is ***~ (n samples)^(1-i/D)***

**=>** ***the number of parameters decreases with increasing i***

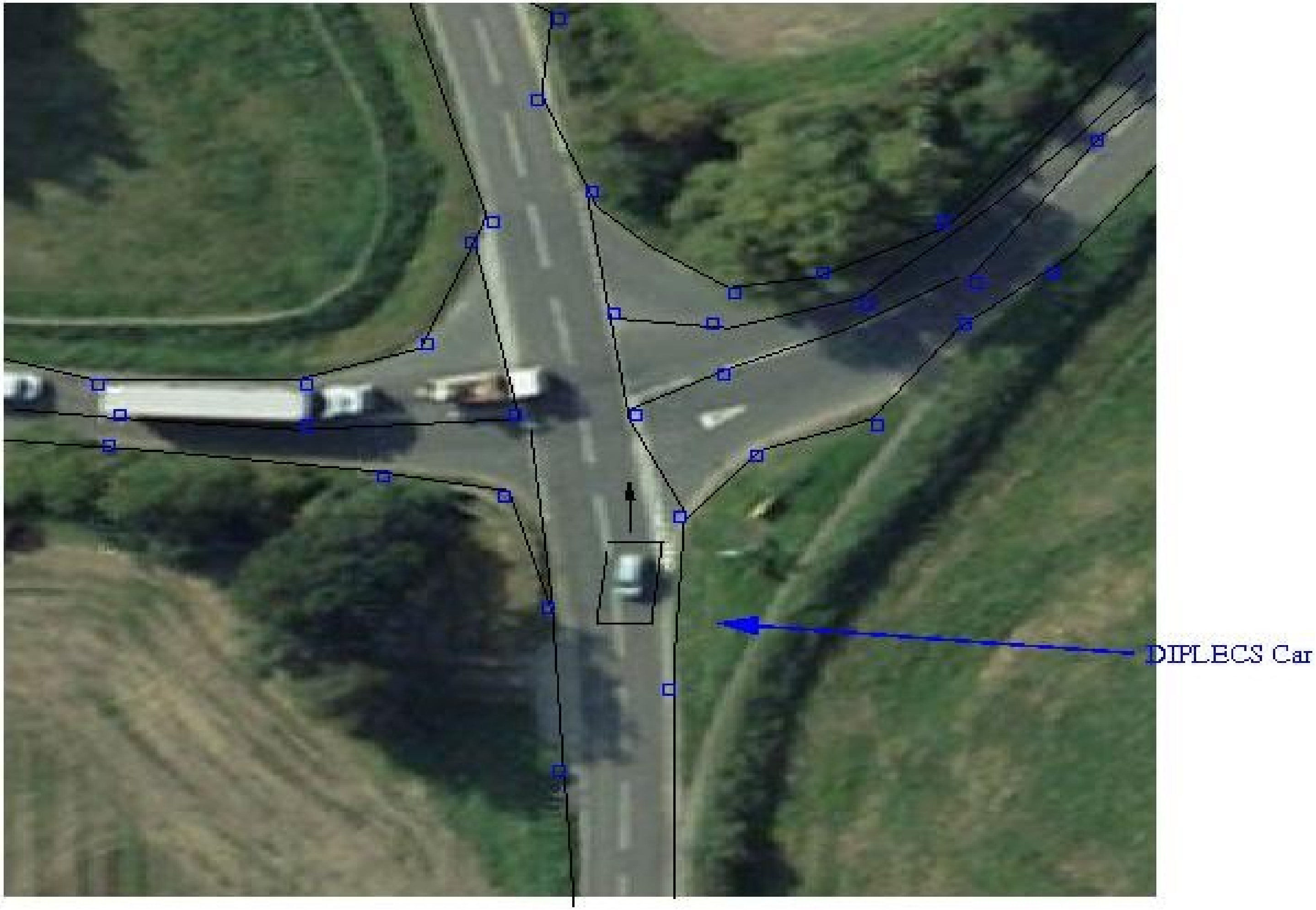
Also, key criterion for generating additional parameters is that they *reduce uncertainty* => ***max liklihood(Data|parameters)*** *increases with i.*

***Hence min BIC cost is dominated by max liklihood considerations***

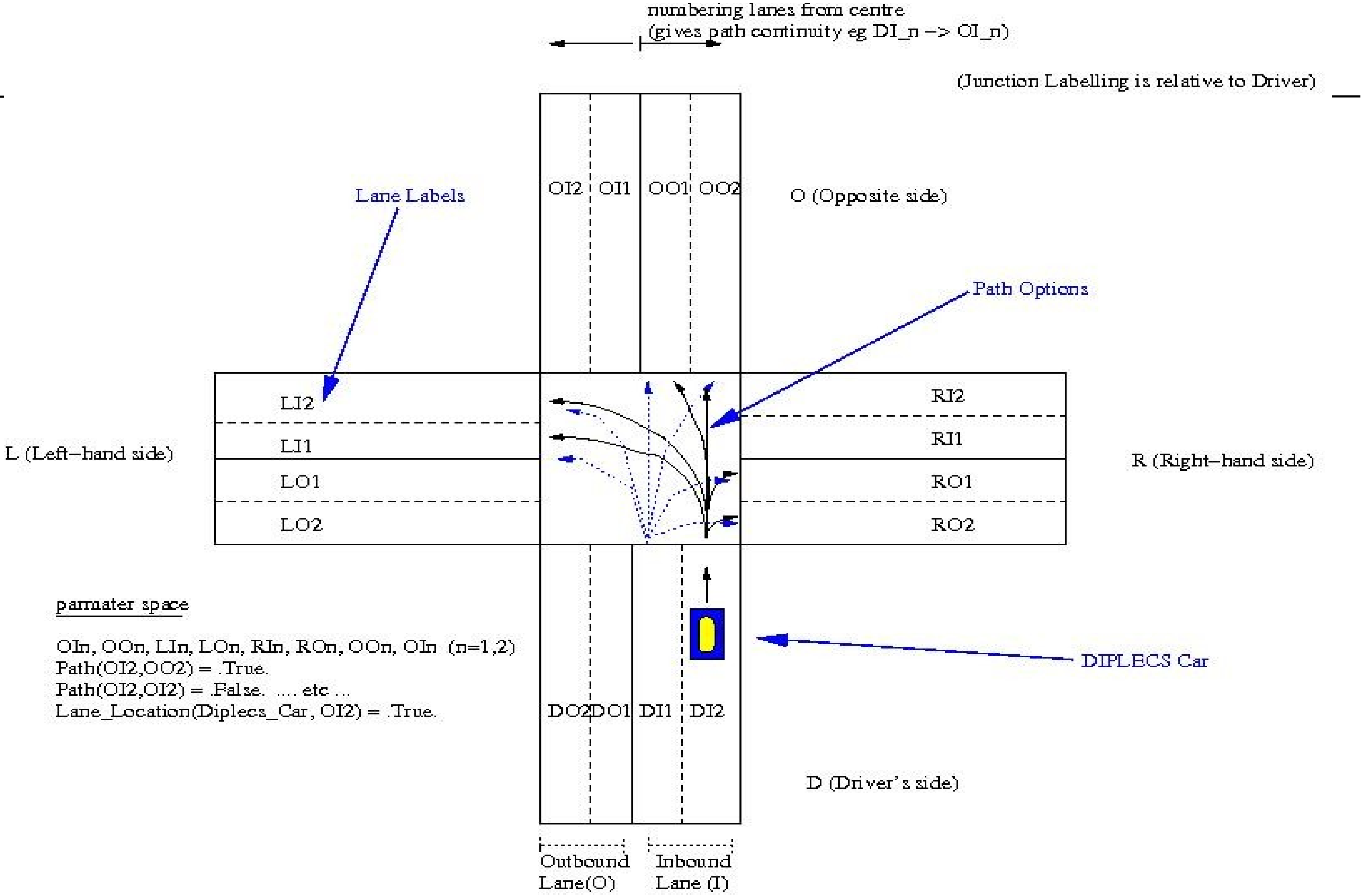
Illustration of Parametric Subsumption:

# Cross–Road Descriptors: level 1

parametric description: (px1,py1,px2,py2,px3, py3,px4,py4 ... , carx, cary, velx, vely)



# Cross–Road Descriptors: level 2



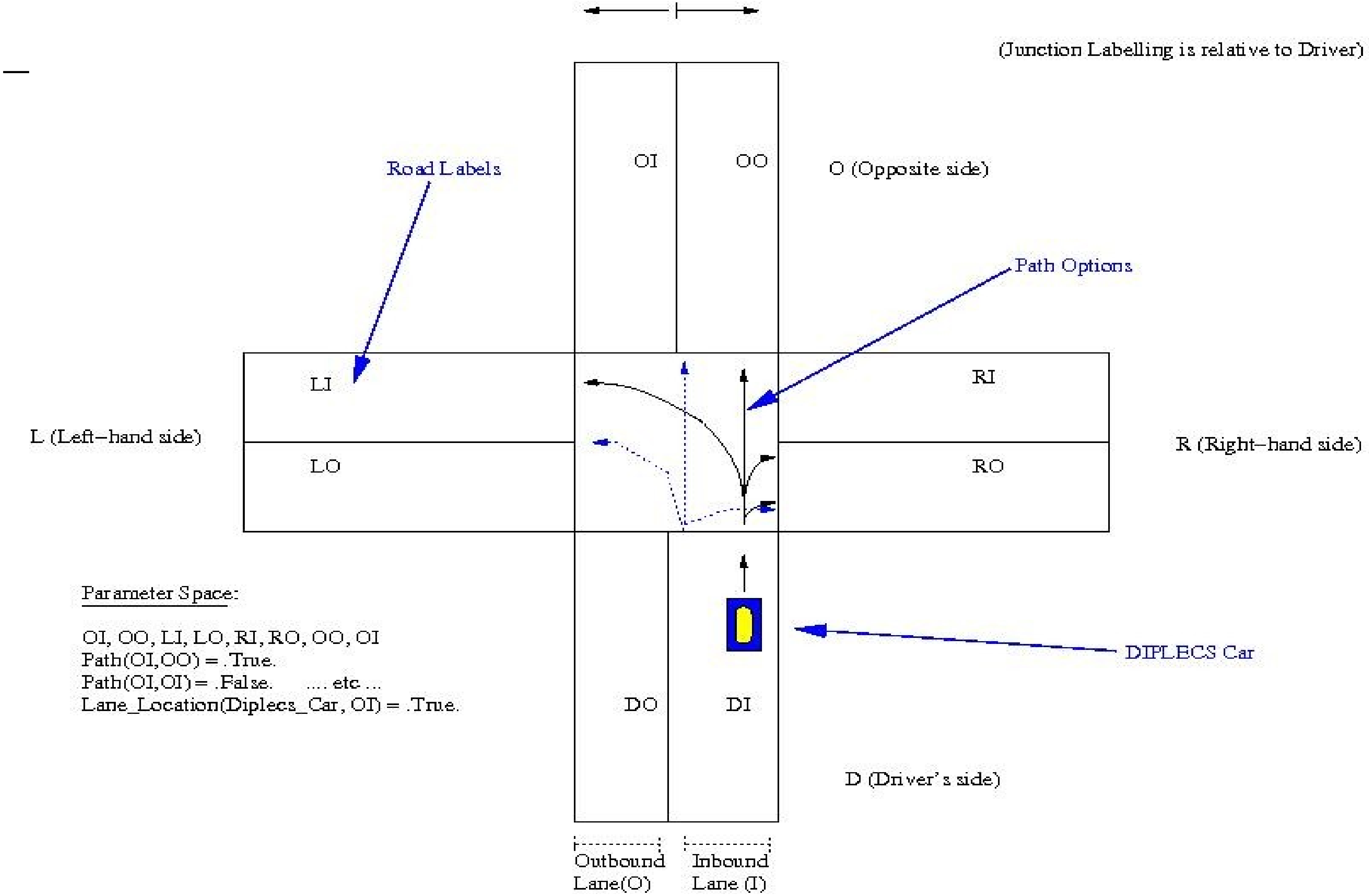
numbering lanes from centre
(gives path continuity eg DI_n –> OI_n)

(Junction Labelling is relative to Driver)

parmater space

OIn, OOn, LIn, LOn, RIn, ROn, OOn, OIn (n=1,2)
Path(OI2,OO2) = .True.
Path(OI2,OI2) = .False. .... etc ...
Lane_Location(Diplecs_Car, OI2) = .True.

# Cross–Road Descriptors: level 3

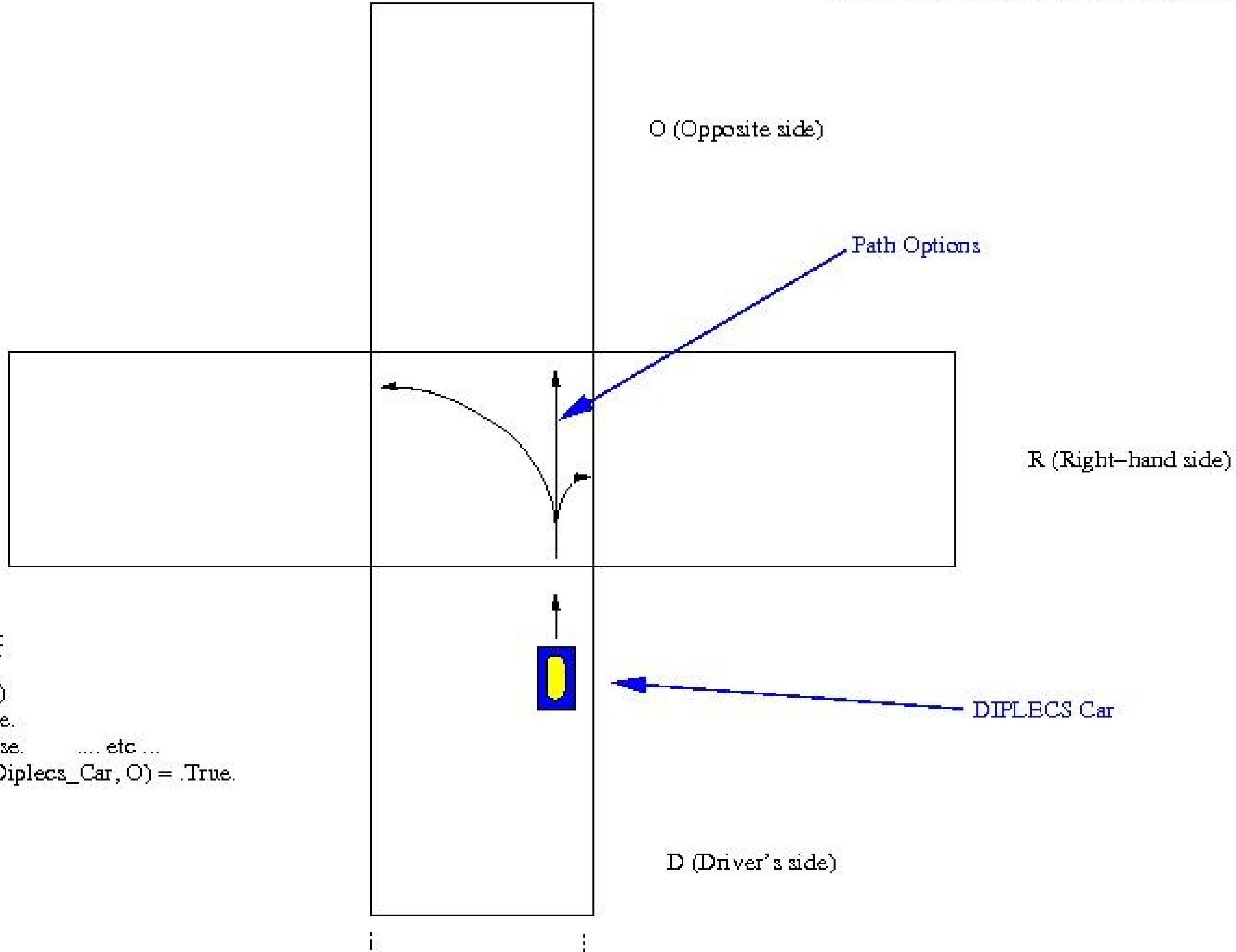


Parameter Space:

OI, OO, LI, LO, RI, RO, OO, OI
Path(OI,OO) = .True.
Path(OI,OI) = .False. .... etc ...
Lane_Location(Diplecs_Car, OI) = .True.

# Cross–Road Descriptors: level 4

(Junction Labelling is relative to Driver)



Parameter Space:

Road(O, L, R, D)
Path(O,L) = .True.
Path(O,O) = .False.       .... etc ...
Lane_Location(Diplecs_Car, O) = .True.

**So, utilising parametric subsumption, we have that:**

**Most Rule-Salient Set of Scene Description=MDL=BIC=Max L**

**Usually, Expectation Maximisation (EM) renders max liklihood tractable**:

(iterate between allocating data to individual parameters on basis of parametruc expectation and then maximising likelihood of parameters given this data allocation)

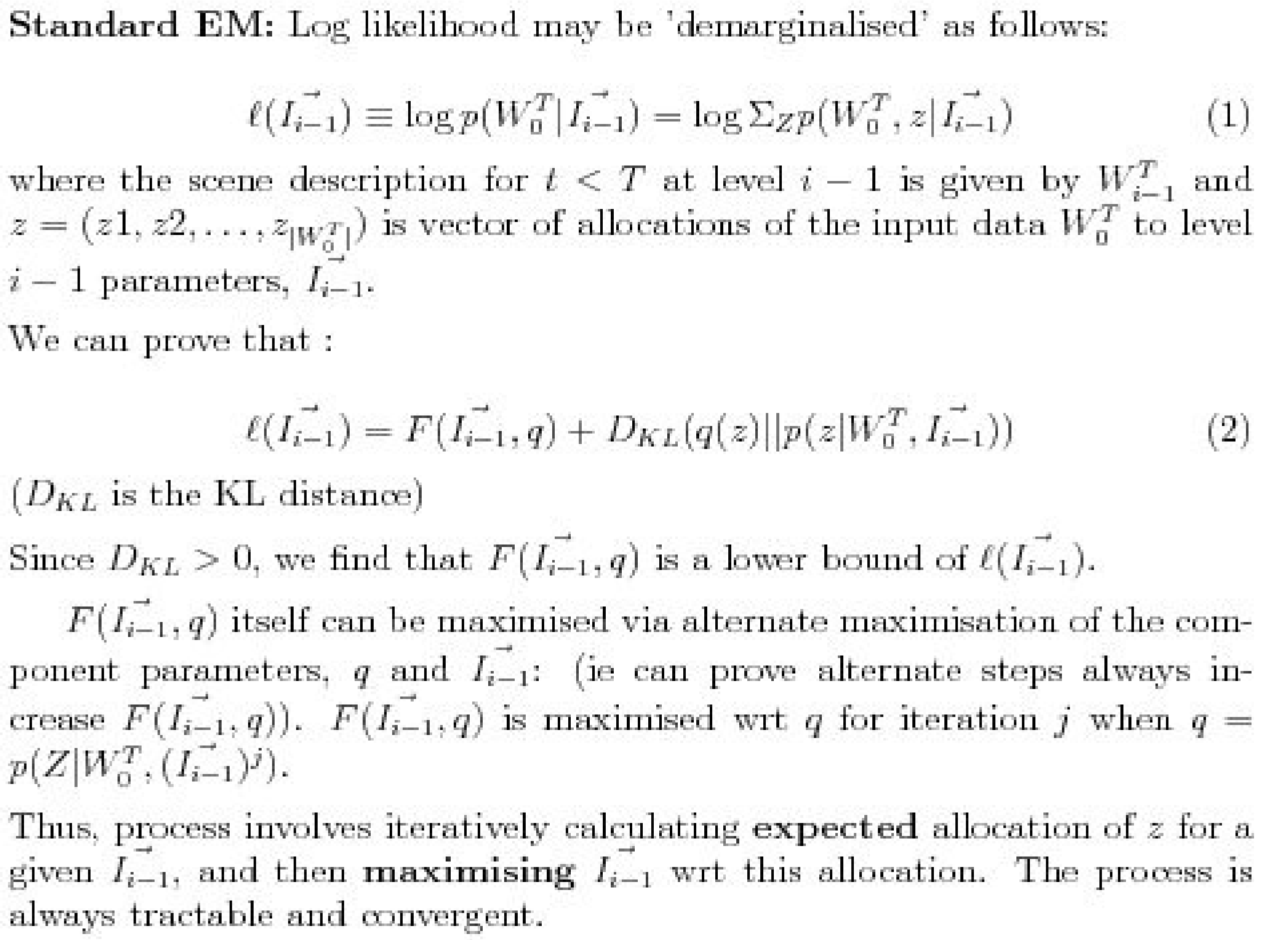
**However parametric-subsumption means cannot use *straightforward* EM**

- need to iterate between (standard) 'horizontal' EM and (new) 'vertical' EM (wherein data is allocated to ***subsumptive*** parametrisation on basis of expectation, and likelihood of ***novel*** parameters is maximised on the basis of this allocation)

=>Equivalent to top-down, bottom-up ***cognitive bootstrap iteration***

*-allows us to prove that Cognitive Bootstrapping is* ***convergent:***

## So, how do we make EM '2-dimensional', to take into account parameter subsumption?



**Standard EM:** Log likelihood may be 'demarginalised' as follows:

$$\ell(\vec{I}_{i-1}) \equiv \log p(W_0^T|\vec{I}_{i-1}) = \log \Sigma_Z p(W_0^T, z|\vec{I}_{i-1}) \tag{1}$$

where the scene description for $t < T$ at level $i-1$ is given by $W_{i-1}^T$ and $z = (z1, z2, \ldots, z_{|W_0^T|})$ is vector of allocations of the input data $W_0^T$ to level $i-1$ parameters, $\vec{I}_{i-1}$.

We can prove that :

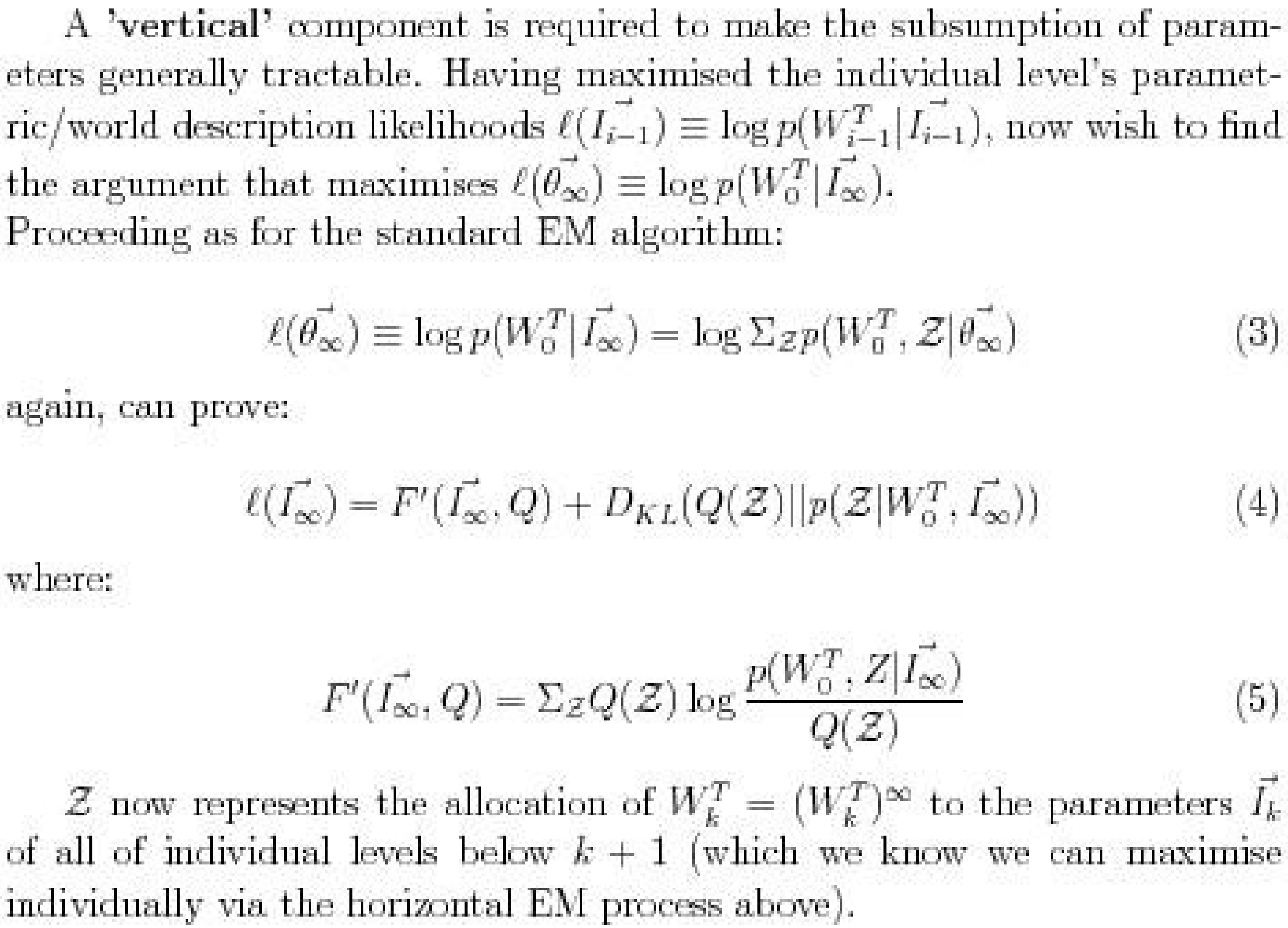
$$\ell(\vec{I}_{i-1}) = F(\vec{I}_{i-1}, q) + D_{KL}(q(z)||p(z|W_0^T, \vec{I}_{i-1})) \tag{2}$$

($D_{KL}$ is the KL distance)

Since $D_{KL} > 0$, we find that $F(\vec{I}_{i-1}, q)$ is a lower bound of $\ell(\vec{I}_{i-1})$.

$F(\vec{I}_{i-1}, q)$ itself can be maximised via alternate maximisation of the component parameters, $q$ and $\vec{I}_{i-1}$: (ie can prove alternate steps always increase $F(\vec{I}_{i-1}, q)$). $F(\vec{I}_{i-1}, q)$ is maximised wrt $q$ for iteration $j$ when $q = p(Z|W_0^T, (\vec{I}_{i-1})^j)$.

Thus, process involves iteratively calculating **expected** allocation of $z$ for a given $\vec{I}_{i-1}$, and then **maximising** $\vec{I}_{i-1}$ wrt this allocation. The process is always tractable and convergent.



A **'vertical'** component is required to make the subsumption of parameters generally tractable. Having maximised the individual level's parametric/world description likelihoods $\ell(\vec{I_{i-1}}) \equiv \log p(W_{i-1}^T|\vec{I_{i-1}})$, now wish to find the argument that maximises $\ell(\vec{\theta_\infty}) \equiv \log p(W_0^T|\vec{I_\infty})$.
Proceeding as for the standard EM algorithm:

$$\ell(\vec{\theta_\infty}) \equiv \log p(W_0^T|\vec{I_\infty}) = \log \Sigma_{\mathcal{Z}} p(W_0^T, \mathcal{Z}|\vec{\theta_\infty}) \tag{3}$$

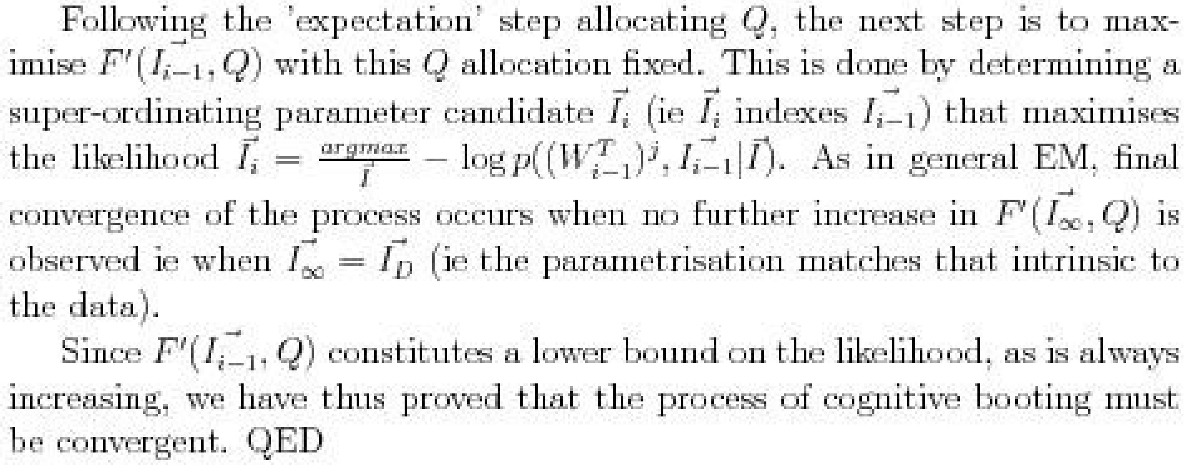
again, can prove:

$$\ell(\vec{I_\infty}) = F'(\vec{I_\infty}, Q) + D_{KL}(Q(\mathcal{Z})||p(\mathcal{Z}|W_0^T, \vec{I_\infty})) \tag{4}$$

where:

$$F'(\vec{I_\infty}, Q) = \Sigma_{\mathcal{Z}} Q(\mathcal{Z}) \log \frac{p(W_0^T, Z|\vec{I_\infty})}{Q(\mathcal{Z})} \tag{5}$$

$\mathcal{Z}$ now represents the allocation of $W_k^T = (W_k^T)^\infty$ to the parameters $\vec{I_k}$ of all of individual levels below $k+1$ (which we know we can maximise individually via the horizontal EM process above).



Following the 'expectation' step allocating $Q$, the next step is to maximise $F'(\vec{I_{i-1}}, Q)$ with this $Q$ allocation fixed. This is done by determining a super-ordinating parameter candidate $\vec{I_i}$ (ie $\vec{I_i}$ indexes $\vec{I_{i-1}}$) that maximises the likelihood $\vec{I_i} = \frac{argmax}{\vec{I}} - \log p((W_{i-1}^T)^j, \vec{I_{i-1}}|\vec{I})$. As in general EM, final convergence of the process occurs when no further increase in $F'(\vec{I_\infty}, Q)$ is observed ie when $\vec{I_\infty} = \vec{I_D}$ (ie the parametrisation matches that intrinsic to the data).

Since $F'(\vec{I_{i-1}}, Q)$ constitutes a lower bound on the likelihood, as is always increasing, we have thus proved that the process of cognitive booting must be convergent. QED

***IE 2D-EM is process of iteratively projecting progressively higher-level representational hypotheses onto sensor data, and testing for optimality***

=>Since EM is convergent, Cognitive Bootstrapping is **convergent**

***- iterative top-down, bottom-up process always reaches BIC minimum (possibly local)***

***NOTE - Two ways in which new high-level description can modify low-level description:***

1. Via the downward action of prior probabilities: $p(W^T_{i-1}|W^T_i) \neq p(W^T_{i-1})$

   If new parameters are informative, can give rise to an improved separation between existing modes characteristics – ***ie existing indexicals now refer to better distinguished entities.***

2. Where resources are limited & low-level parametrisation is computationally expensive, can coarse graining over, or omit parametrisation of low-level entities not distinguished by higher level parameters.

*eg A subset of parametrisation instances can be classed as* ***'background'*** *due to not being rule-salient, such that no further parametric allocation (eg tracking) is carried out on them.*

**One Other Cognitive Bootstrapping Issue:**

Choice of Base-level descriptors (input space)

Helps to choose *a priori* domain carefully:

- ***want features that universalise across all court games as far as possible***

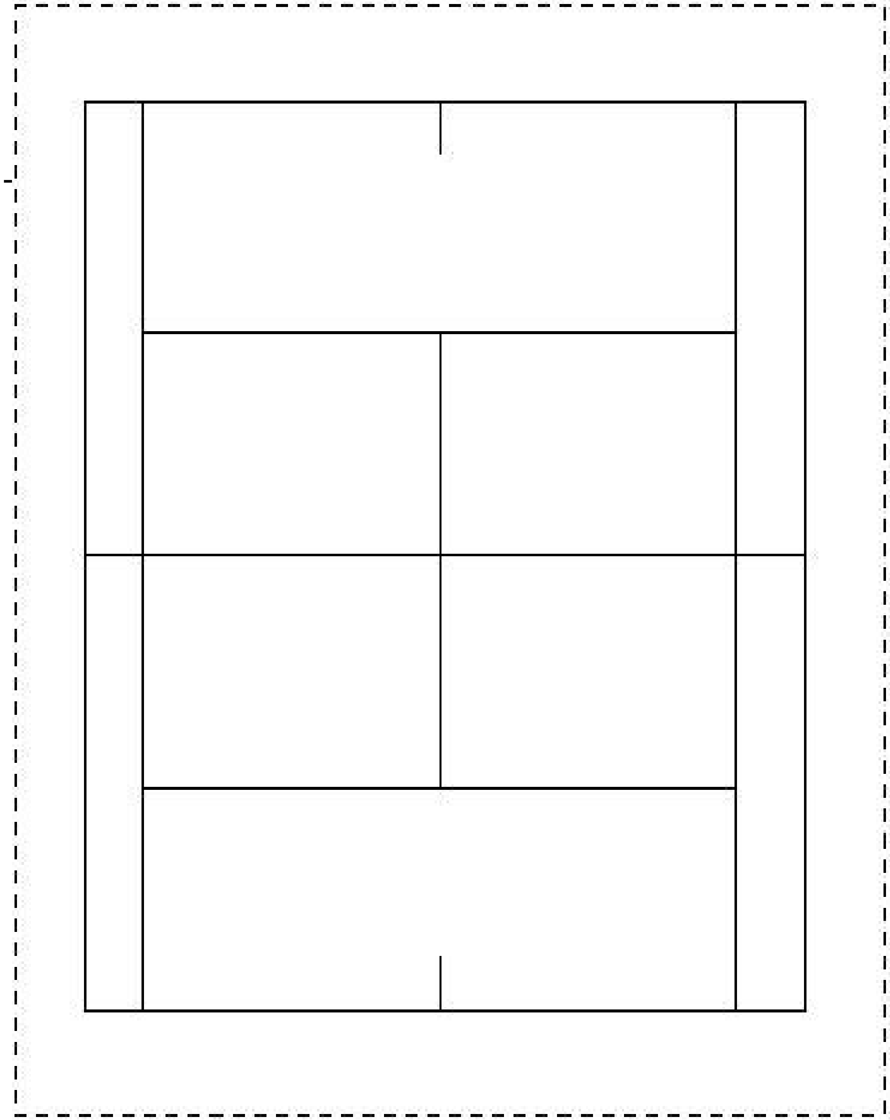
  ***IE SUCH THAT UNIVERSALISATION IS IMPLICIT IN PARAMETRIC SUBSUMPTION***

Example: universal court area descriptors

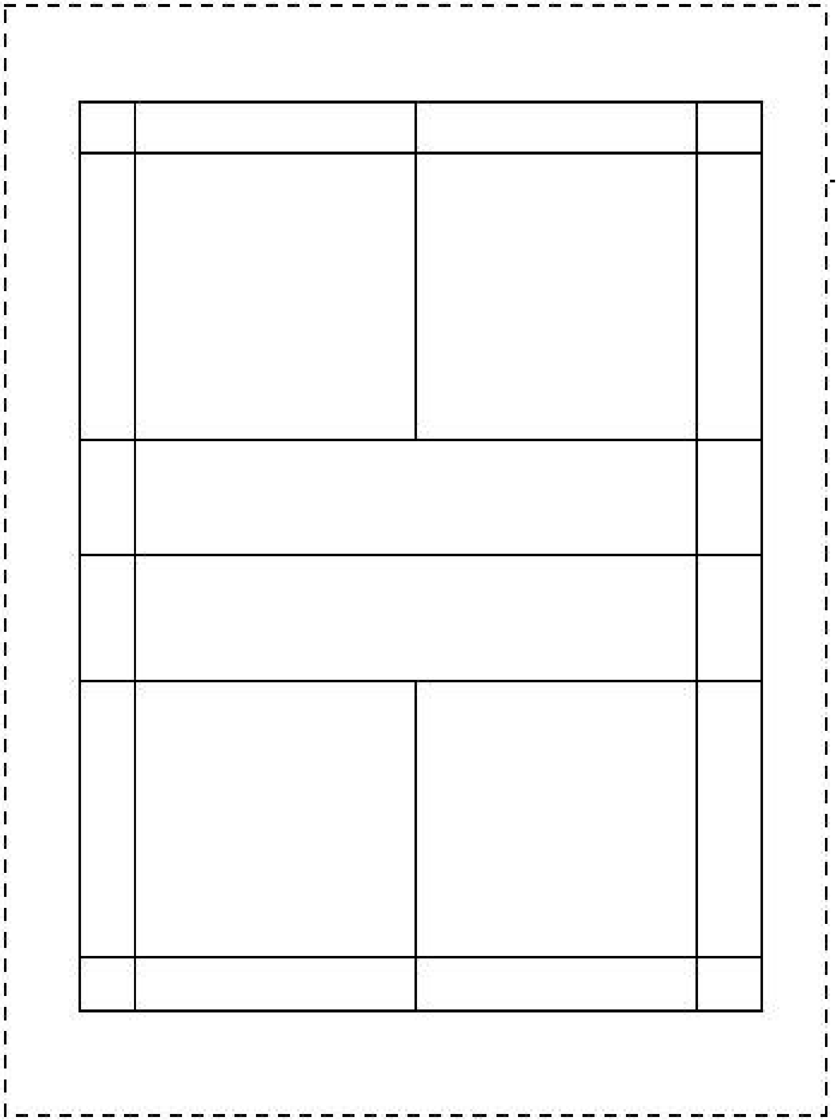
Tennis

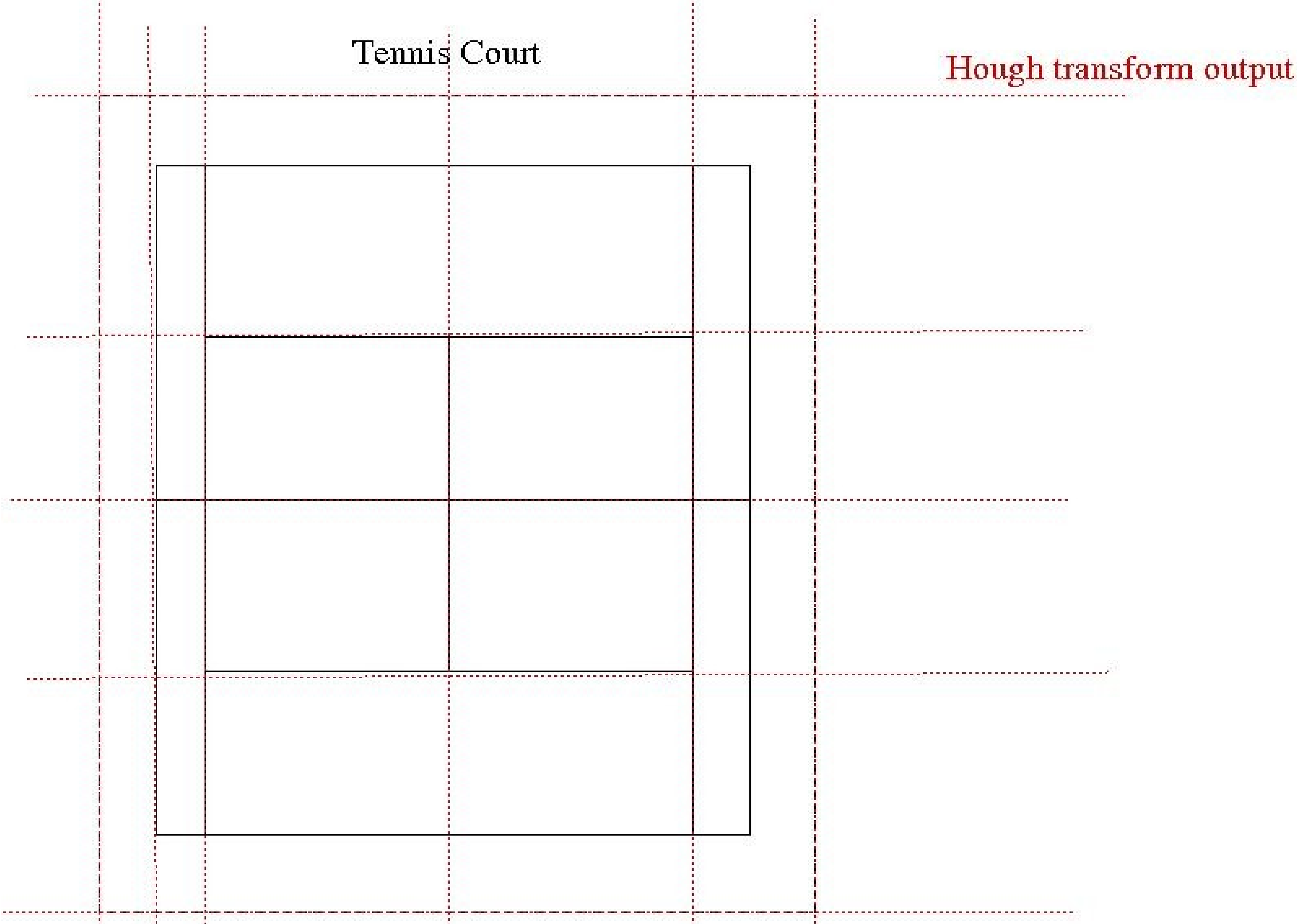
Badminton

Tennis Court

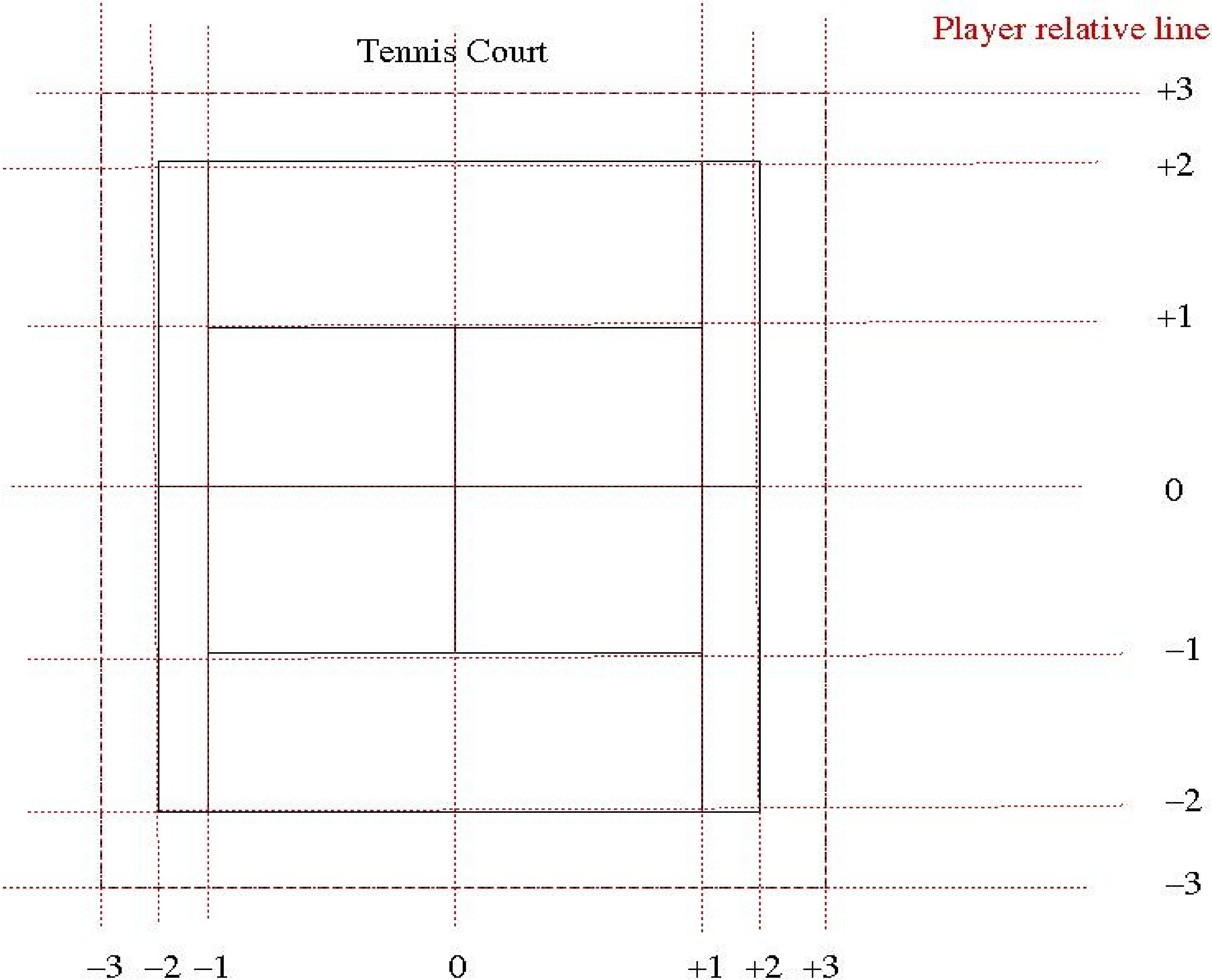


Badminton Court

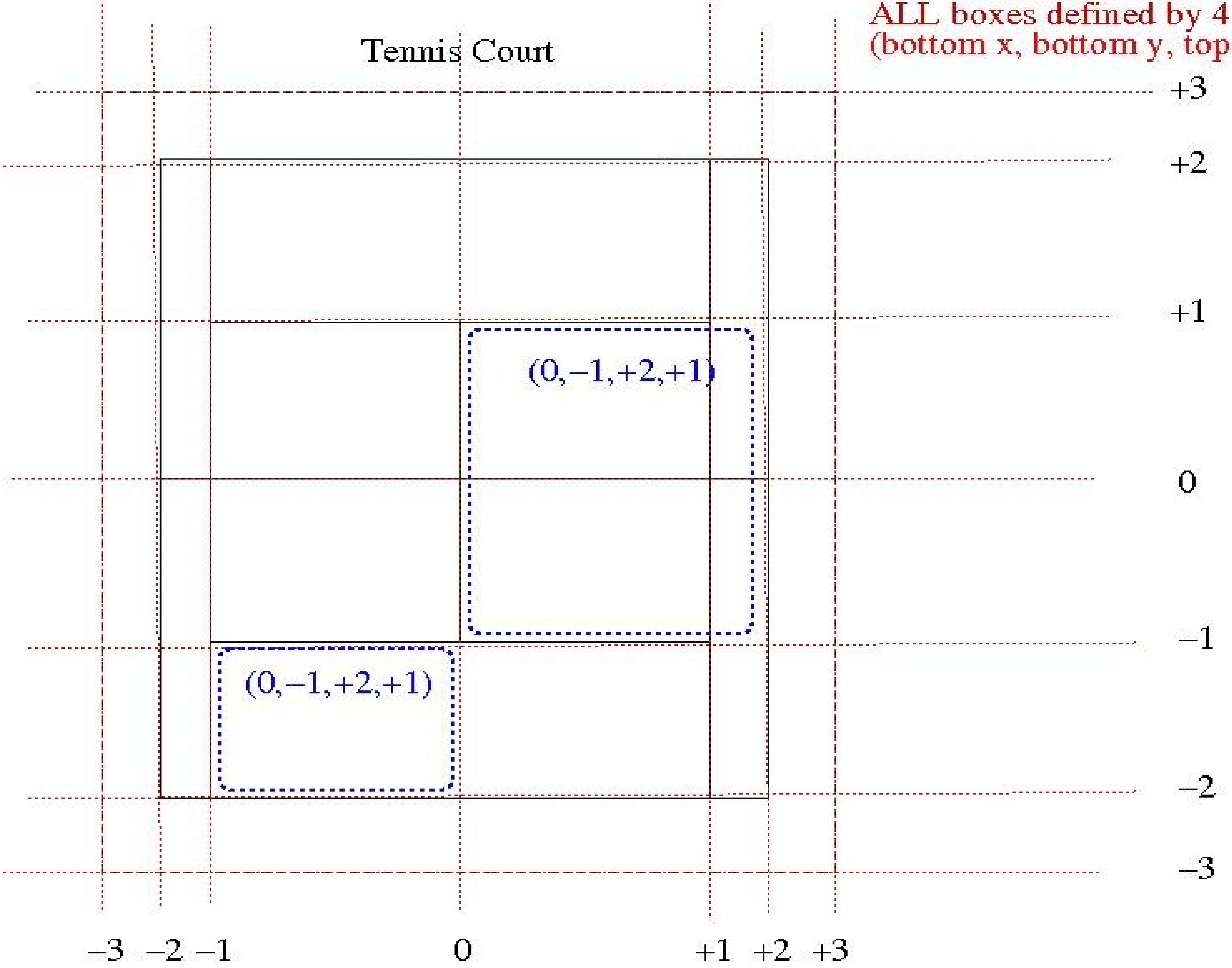


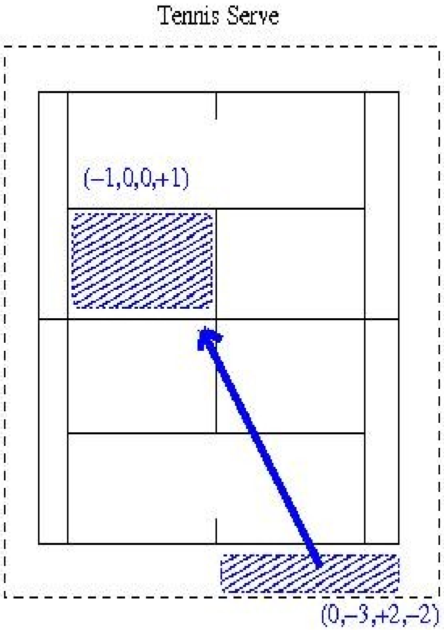

Tennis Court
Hough transform output

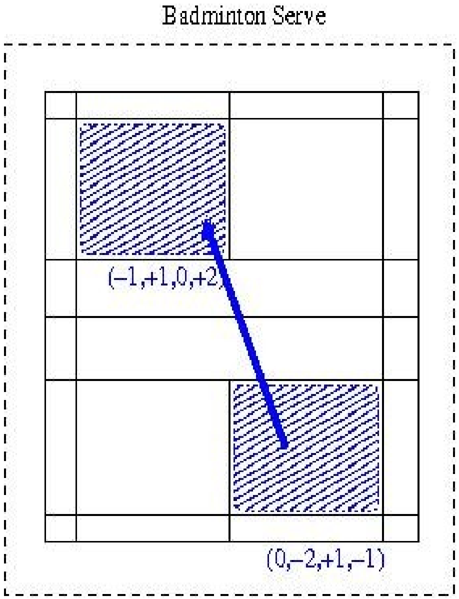
Player relative line–labelling



ALL boxes defined by 4 ordinates:
(bottom x, bottom y, top x, top y)







Serve defined as:
(−1,0,0,+1) <–(0,−3,+2,−2)

Serve defined as:
(−1,+1,0,+2) <–(0,−2,+1,−1)

Heirarchy of visual descriptors:

| | | |
|---|---|---|
| pixel level: | (25,33,36,43) <– (32,5,73,13) | (24,43,32,53) <– (34,14,44,23) |
| Hough box: | (−1,0,0,+1) <–(0,−3,+2,−2) | (−1,+1,0,+2) <–(0,−2,+1,−1) |
| Hough sign: | (−ve,+ve,−ve,+ve) <– (+ve,−ve,+ve,−ve) | (−ve,+ve,−ve,+ve) <– (+ve,−ve,+ve,−ve) |
| H. sign contract: | (−ve,+ve) <– (+ve,−ve) | (−ve,+ve) <– (+ve,−ve) |

Parametric Subsumption

Indentical Serve Description from This point of Heirarchy

# 4. Inspirations – Related projects

## 1. COSPAL

-Parametric subsumption of representation, P-A learning

## 2. Modelling information seeking by integrating visual and semantic and memory maps

(University of Nice, University Joseph Fourier, Centre national de la recherche scientifique)

## 3. ViSiCAST

- First-order logic-based semantic DRS to mediate between visual & oral grammar

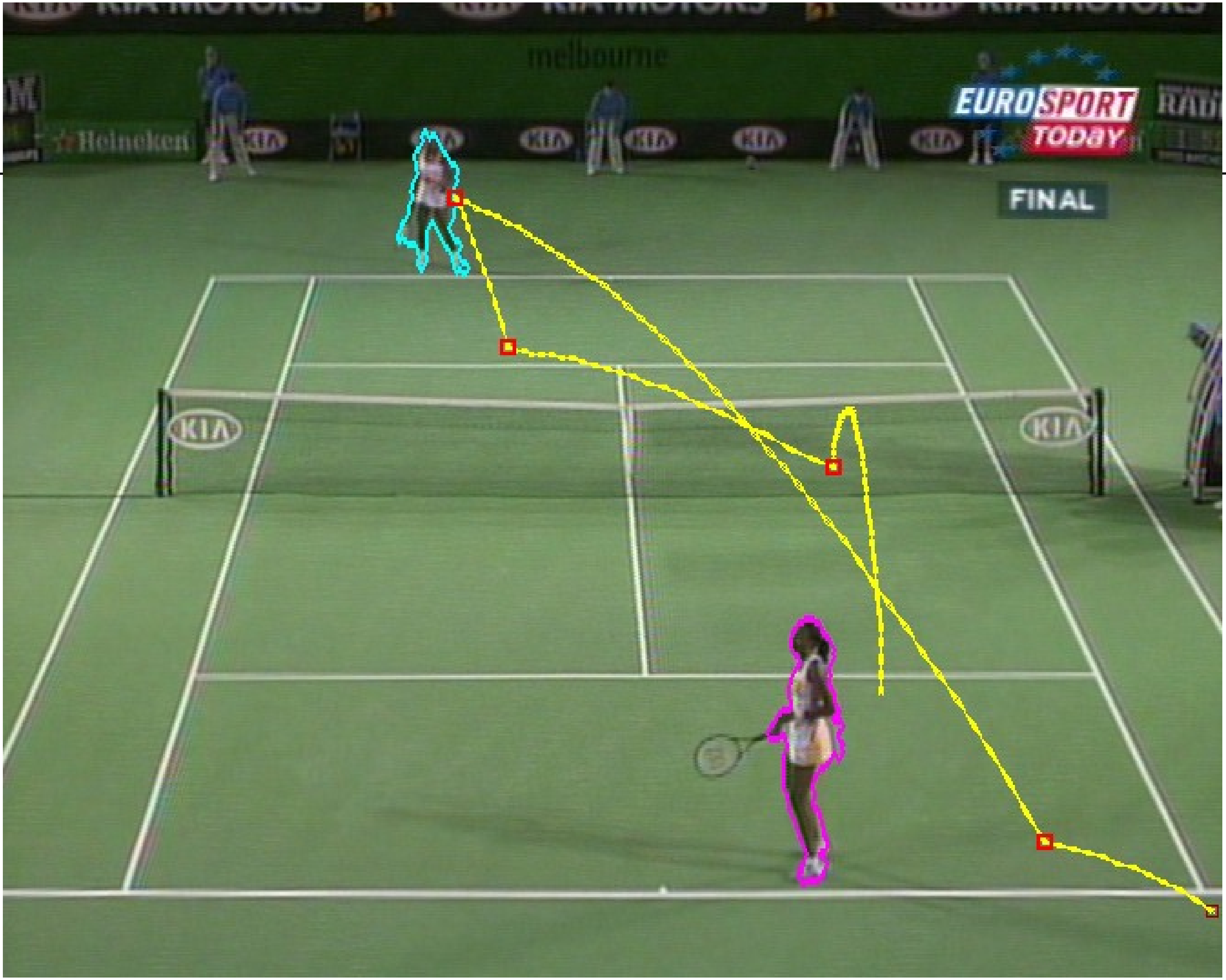
## 4. VAMPIRE-Visual Active Memory Processes & Interactive REtrieval,

- A***ctive memory system***, operating at several different semantic levels, leading to a degree of machine understanding when applied to specific scenarios.

- Demonstrate the Active Memory concept by applying it to the analysis and browsing of a tennis video. Annotation provided at all levels, from shot detection to a complete breakdown of the scoring during the match.

## 5. DIRAC -Detection & Identification of Rare Audio-visual Cues

- Developing environment-adaptive autonomous artificial cognitive systems that will detect, identify and classify rare events from multiple active information-seeking audio-visual sensors.


KIA MOTORS
melbourne
Heineken
KIA
EUROSPORT
TODAY
FINAL

# 4.1 **COSPAL**

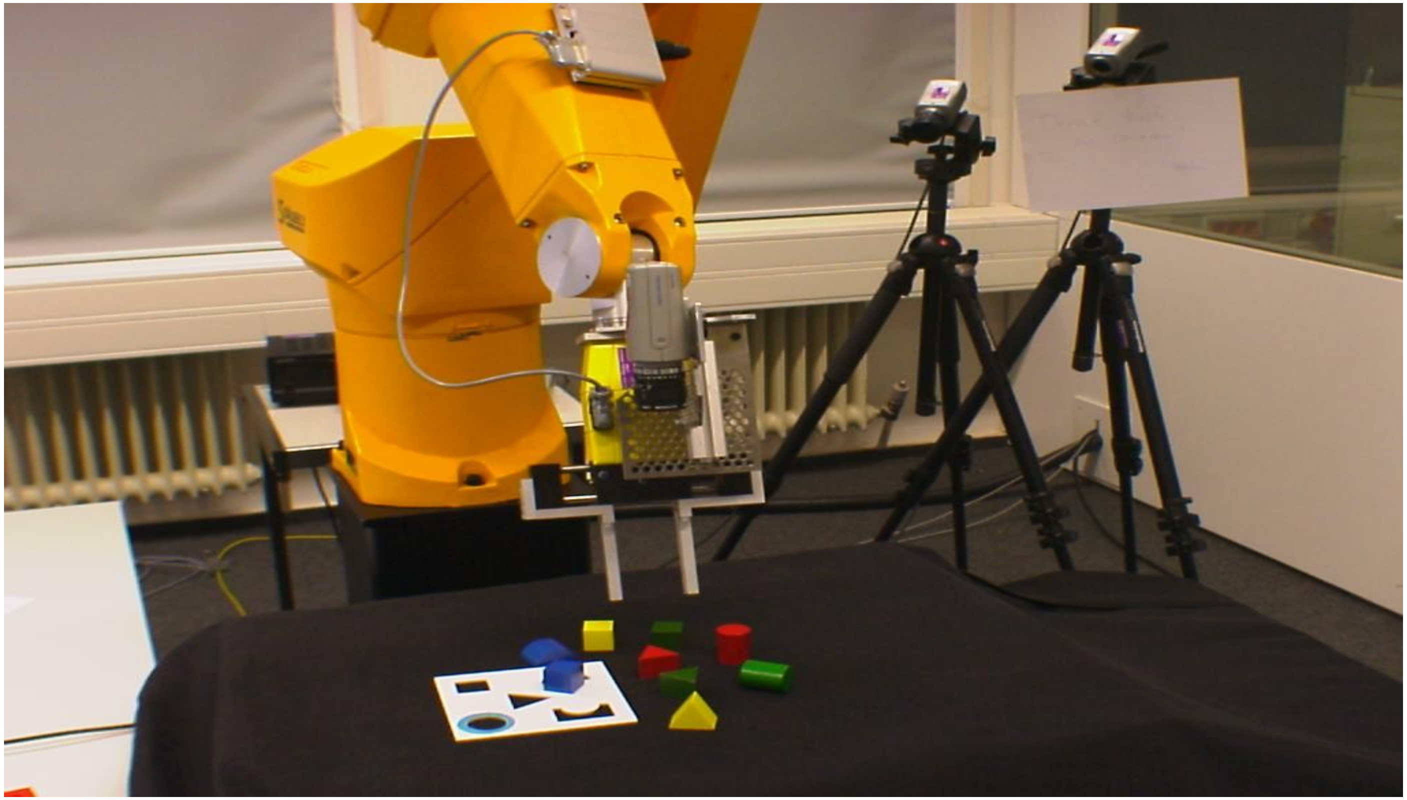
In **perception-action** (P-A) learning, action *precedes* perception

- to reduce complexity, only consider perceptual states that **distinguish** results of agent actions

=> The creation of novel symbolic representation of the world can only occur in relation to the heirarchical acquisition of *novel abstract action capabilities*.

- *Perceptual framework can be **falsified** by exploratory actions in same way as the agent's external world model*

**COSPAL test domain:** The 'shape-sorter' puzzle



# 1) Bootstrapping Representations in the Logical Domain

Three principle stages:

**<u>1) Randomised exploratory activity</u>**

carried out in terms of ***all levels*** of current world-model/representational-framework

**<u>2) Induction of gen. rules governing action legitimacy</u>**

(legitimacy: action does what agent intends)

**= *<u>affordance-based model of world</u>***

**<u>3) Remapping of Perceptual Variables</u>** to represent action model in most efficient manner **(=P-A condition)**

**{Action Possibilities} <=> {P_revised} X {P_revised}.**

**<u>REPEAT</u>**

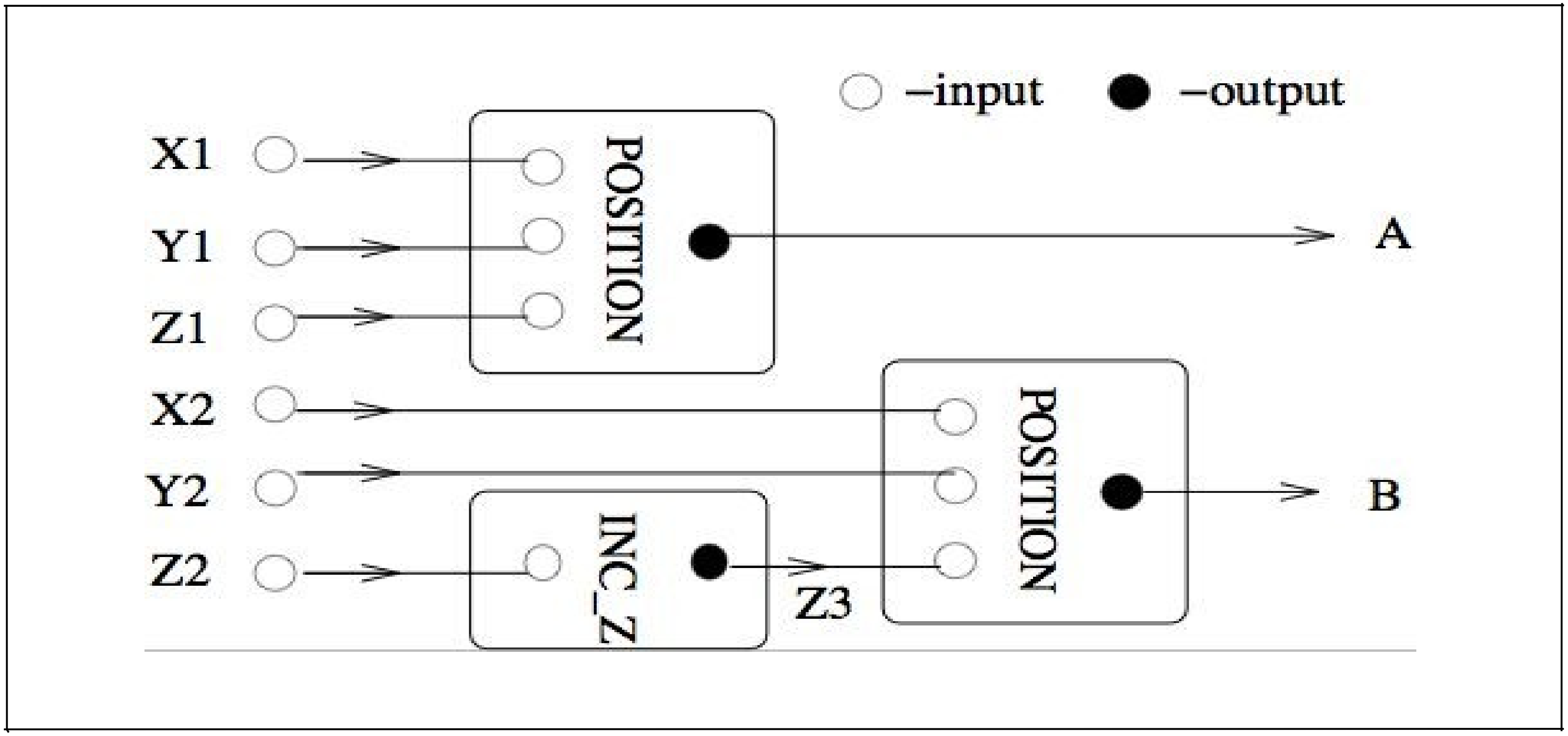
# Remapping the Percept Space After ILP (Example).

- Suppose that after limited exploration we have inferred a partially accurate representation of the shape-sorter rules in terms of the *a priori* predicates '**position**' (positional occupancy) and '**inc_z**' (vertical adjacency):

```
move(X1,Y1,Z1,X2,Y2,Z2) :- position(A,X1,Y1,Z1),
 inc_z(Z3,Z2), position(B,X2,Y2,Z2)
```

- (ie '*entity labelled A can be moved onto entity labelled B*')

- Can represent clause I/O structure as the following schematic:



- By defining predicates as functionally ***reversible***, possible to instantiate ***all*** of the legitimate moves via the variables (A,B) rather than the original (X1,Y1,Z1,X2,Y2,Z2).

- ie **move(X1,Y1,Z1,X2,Y2,Z2) => move(A,B)**

=> Have implicitly remapped the original three-dimensional percept space (X, Y, Z) into two new one-dimensional 'spaces' (A & B).

*{Action Possibilities} <=> {P_revised }X {P_revised}.*

=> Have reconceived the percept space in terms of the high-level concepts ***objects*** and ***surfaces***, rather than the lower-level concept ***position***.

# Cognitive Bootstrapping as Active Learning

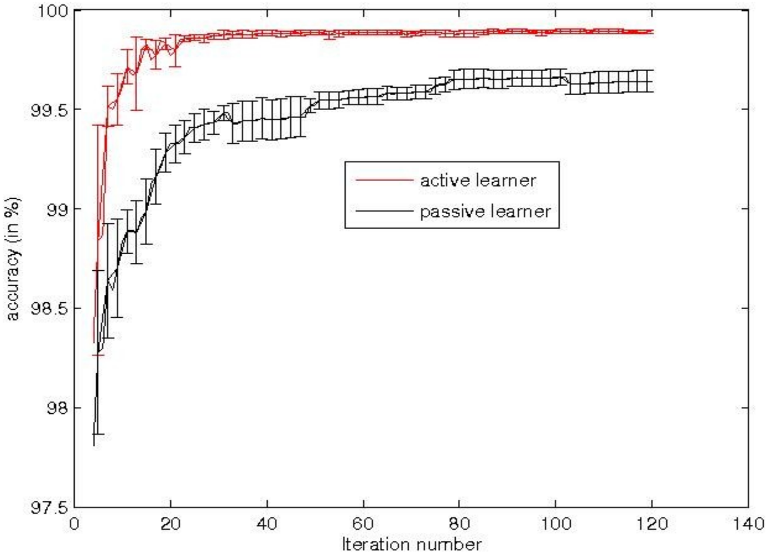
- Employ **remapped** percept space to (randomly) propose percept states to which exploratory movement can take place.

  (ie drive exploration from newly-inferred higher-level perceptual categories)

- *should cause faster convergence on objective model because actions capable of falsifying existing models are found far sooner.*

# Experimental Results

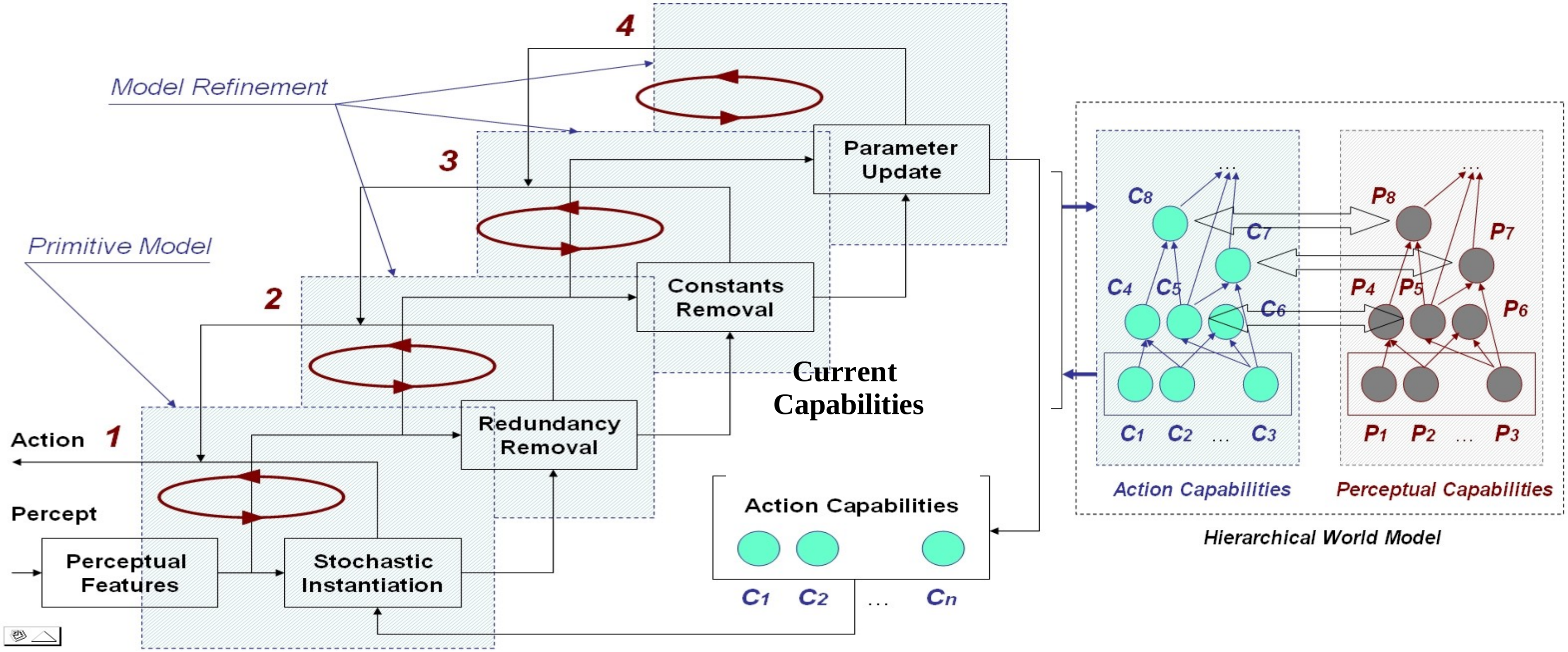
- Active Learning Performance verses Passive Learning Performance:



- Convergence ~7 times faster on average (defined as time taken to attain <0.5 percent of maximum accuracy).

  (also higher maximum achieved)

# The Sub-Logical Bootstrapping Architecture

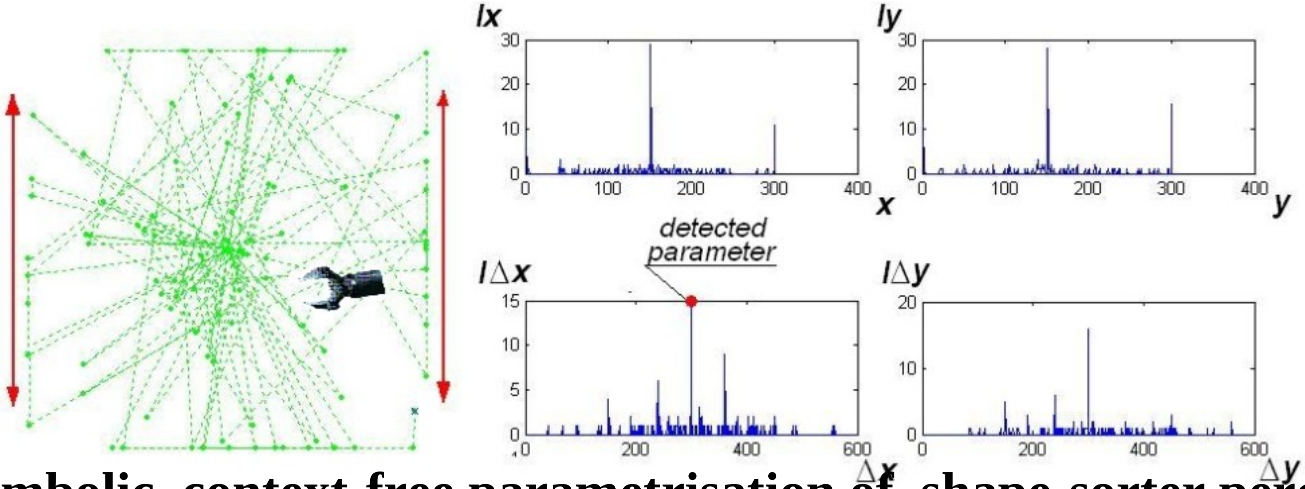


- Involves two stages: 1) aquiring a primitive motor capability and 2) model refinement.
- Four stage cascade process: each step is a separate perception-action cycle.
- Again, the world model is a hierarchical structure of behavioural models and perceptual concepts.

## STAGE 1) primitive goal detection and determination of motor constraints via perceptual variables:

**Observe supervised shape-sorter solving:** determine behaviour histogram in ***a priori* perceptual parameter space**

- use **saliency thresholds** to determine *goal parameters*



=> **symbolic, context-free parametrisation of shape-sorter perceptual goals**

- can then be mapped onto parameters of existing motor capabilities by stochastic exploration

**STAGE 2)** Perform stochastic exploration within the domain of **generalisations** of existing capabilities to find an action model instantiation satisfying the current goal.

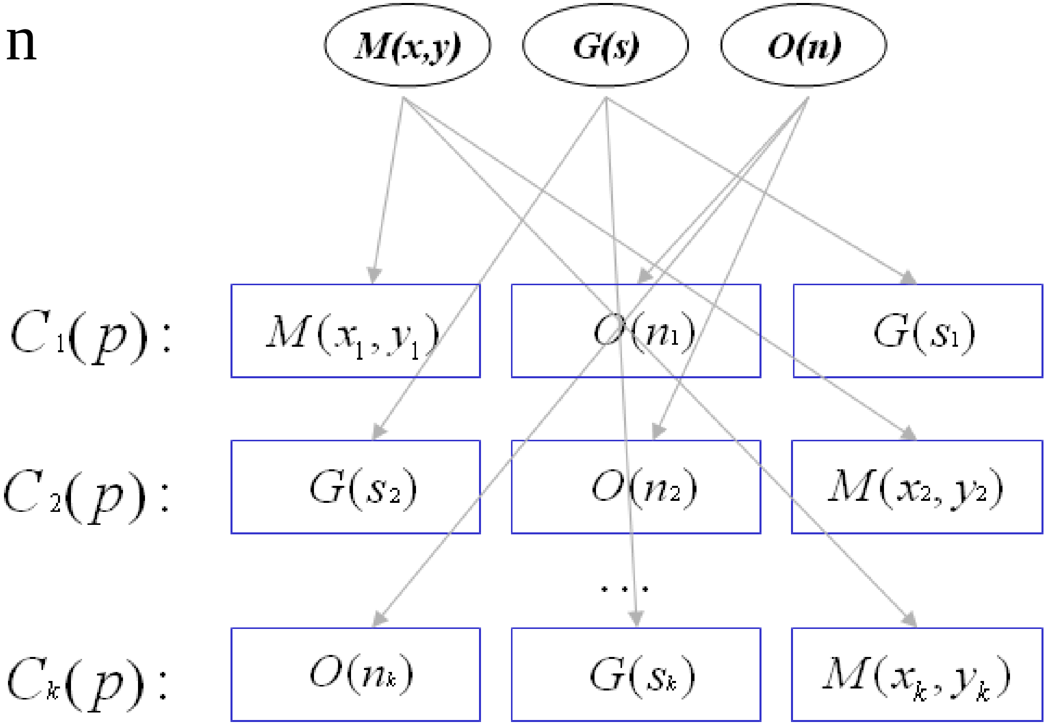
Progress towards goal measured by introducing a perceptual **distance function**.

= >do this by randomly concatenating **sequences of existing motor capabilities** indexed by primitive symbolic goals

O(n) =Move to Object n

G(s) =Grip/Ungrip

M(x,y)= Move to (X,Y)



# We then generalise the behavioural capabilities by mapping into a **more compact parameter domain:**

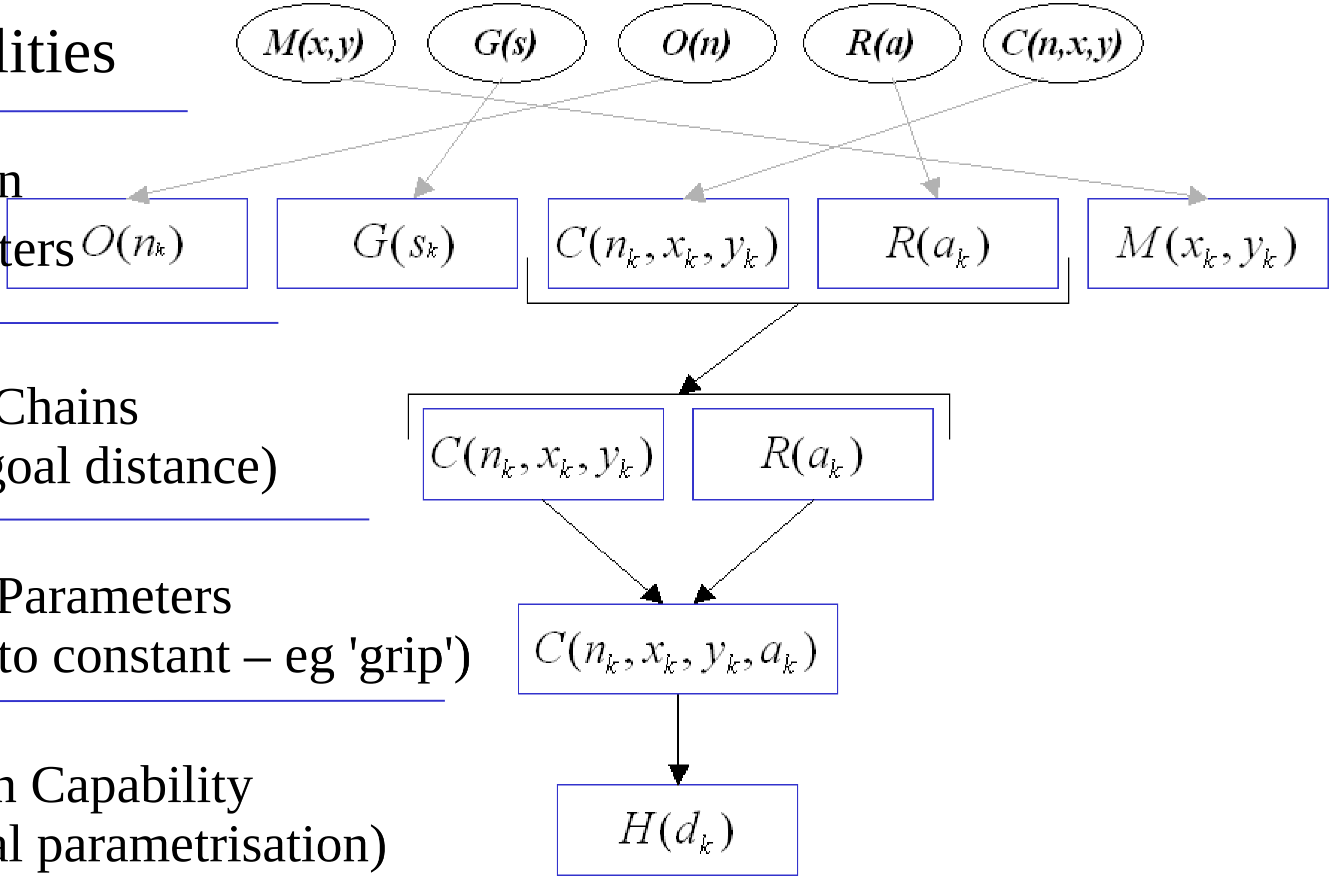
Established Capabilities

Random Instantiation
of Sequence Parameters

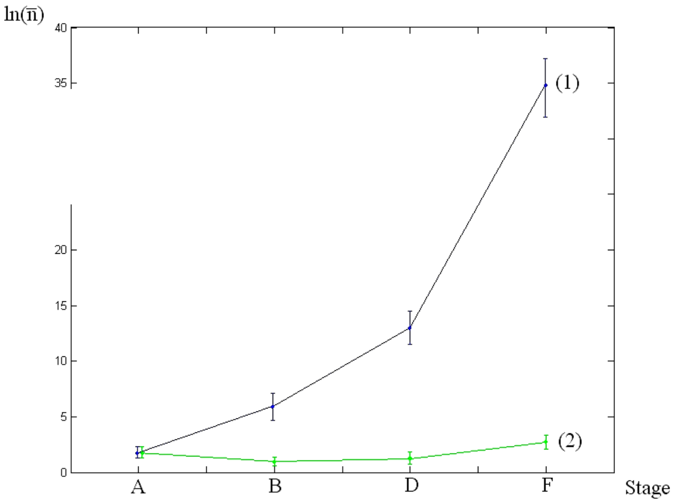
Remove Redundant Chains
(no or -ve effect on goal distance)

Remove Redundant Parameters
(always instantiated to constant – eg 'grip')

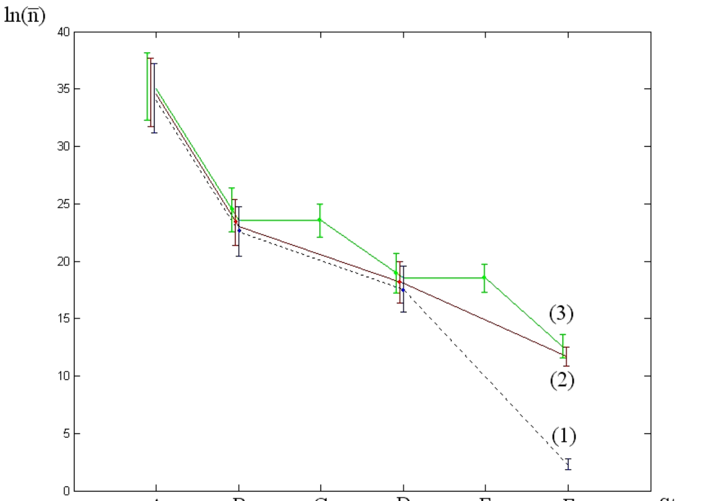
Define Novel Action Capability
(with own perceptual parametrisation)



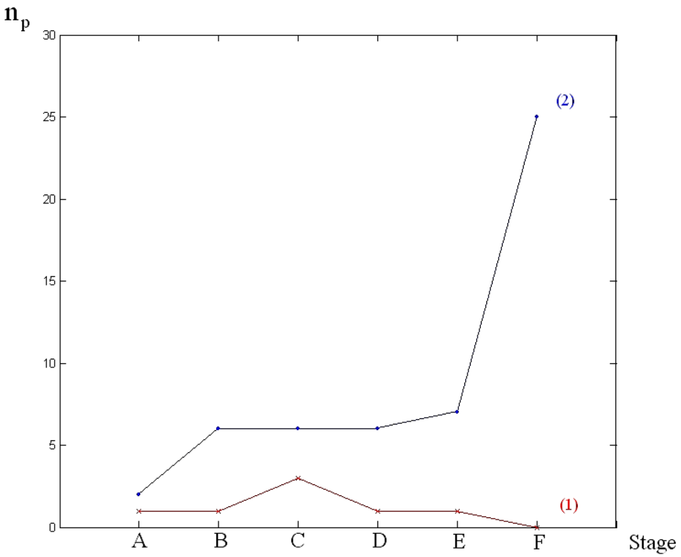
# Results



**1) Computational requirements of learning progressive puzzle-competences for Bootstrapping/Non-Bootstrapping algorithms**

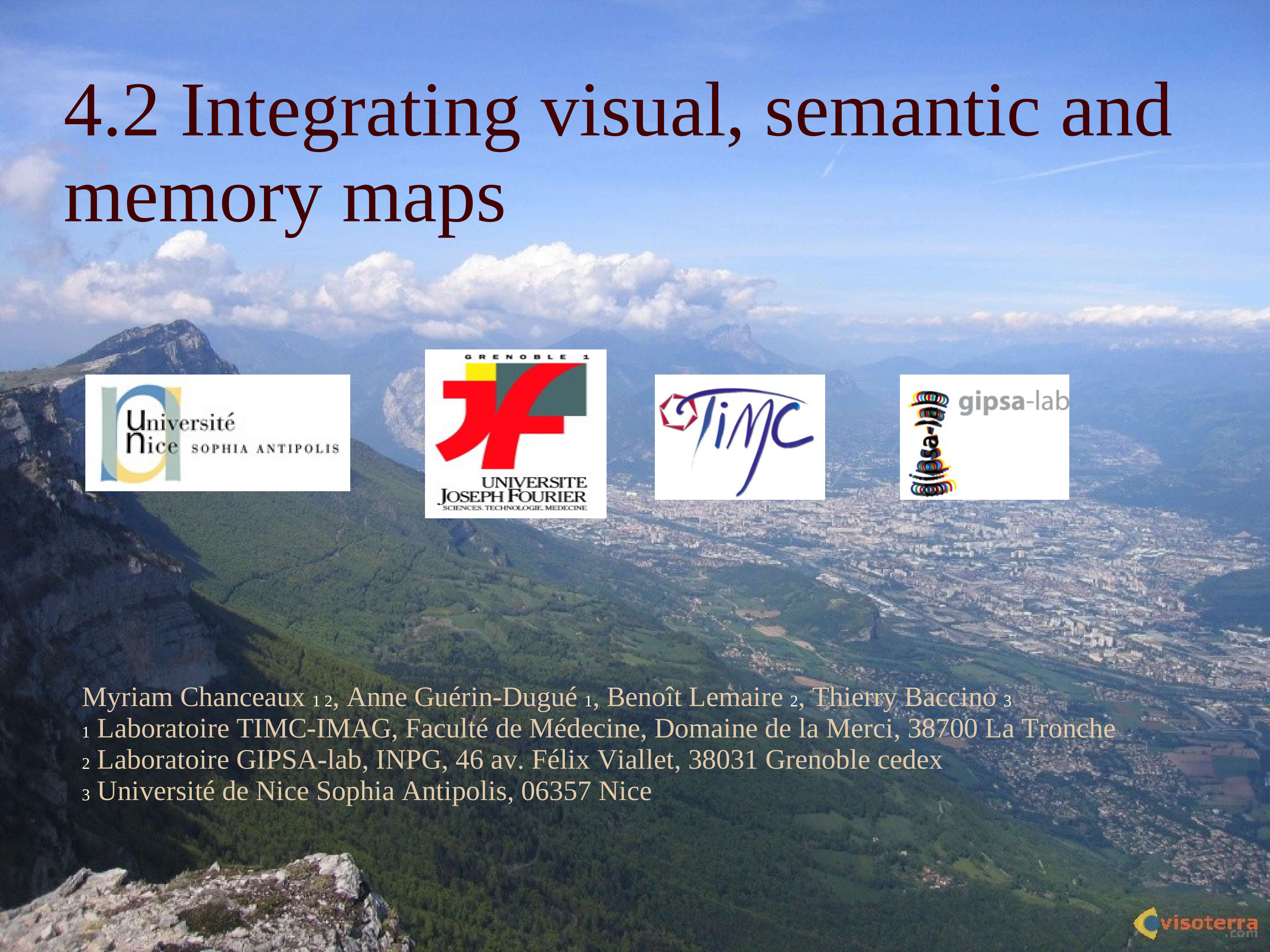


**2) Puzzle Solving Rate per Competence Level**:
(1) Simple hierarchically-guided environment
(2) Simple autonomous environment
(3) Autonomous environment with distractors



$n$=number of P-A cycles
$n_p$=number of parameters

**3) Evolution of motor parameters in the Bootstrapping/Non Bootstrapping regimes**



# 4.2 Integrating visual, semantic and memory maps

Myriam Chanceaux [1,2], Anne Guérin-Dugué [1], Benoît Lemaire [2], Thierry Baccino [3]
[1] Laboratoire TIMC-IMAG, Faculté de Médecine, Domaine de la Merci, 38700 La Tronche
[2] Laboratoire GIPSA-lab, INPG, 46 av. Félix Viallet, 38031 Grenoble cedex
[3] Université de Nice Sophia Antipolis, 06357 Nice

# Information seeking: 3 processes

- Visual:
  - Model based on the feature integration theory (Treisman & Gelade, 1980)
  - Itti's model (Itti & Koch, 2000)
- Semantic:
  - Latent Semantic Analysis, Colides (Kitajima, Blackmon, & Polson, 2000)
- Guidance memory:
  - Inhibition of return (IOR) (Klein, 1988)

# Our model: Integration of 3 maps

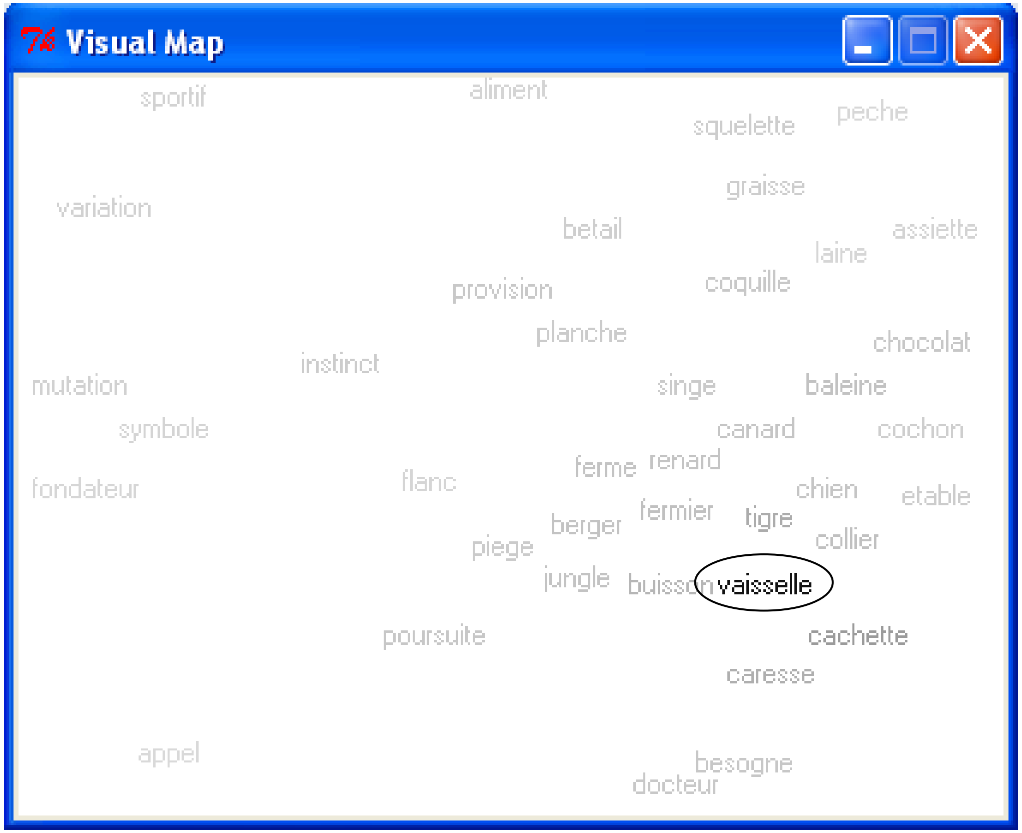
For a given fixation : compute the weights of each words and choose the next fixated word
The map are conditionned on the current fixation and the memory map depends on all the previous fixations

v Visual SaliencyMap + s Semantic Map + m Memory Map = Main Map

v + s + m = 1

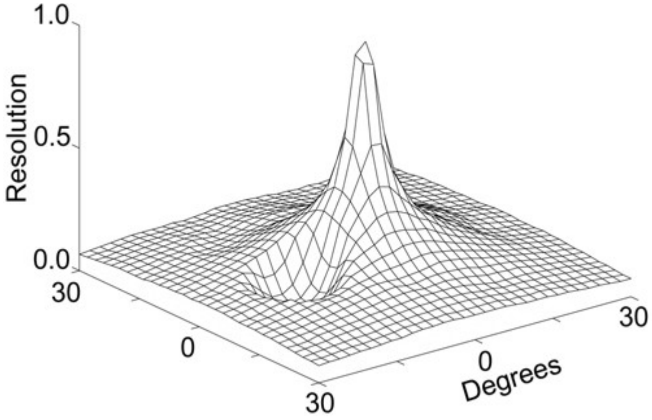
# Visual map

a classical visual saliency map multiplied by a filter corresponding to the visual acuity per degree of eccentricity.

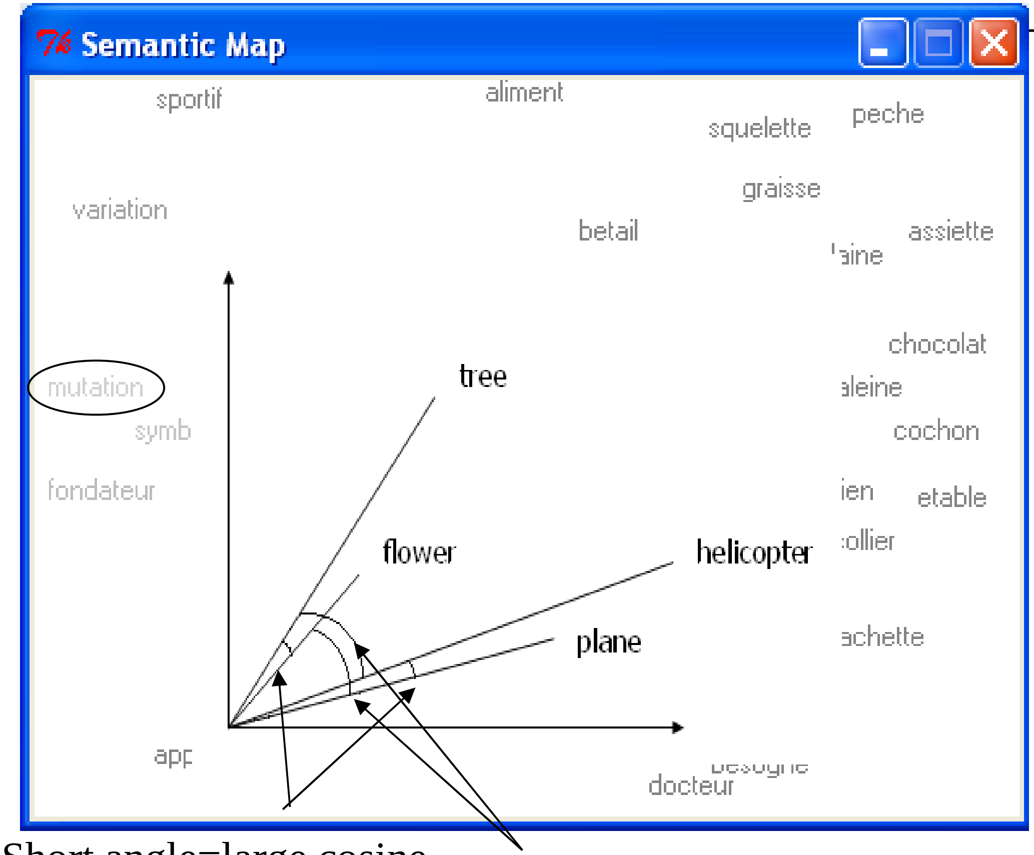


Weight(w) = Saliency(w)*VAcuity(w, currentfix)

Saliency(w) = nbChar(w)/9*(fontSize(w)/19)^2



# Semantic map

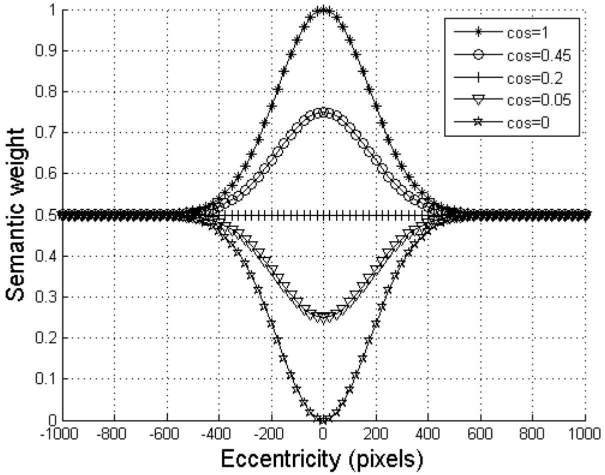


Short angle=large cosine

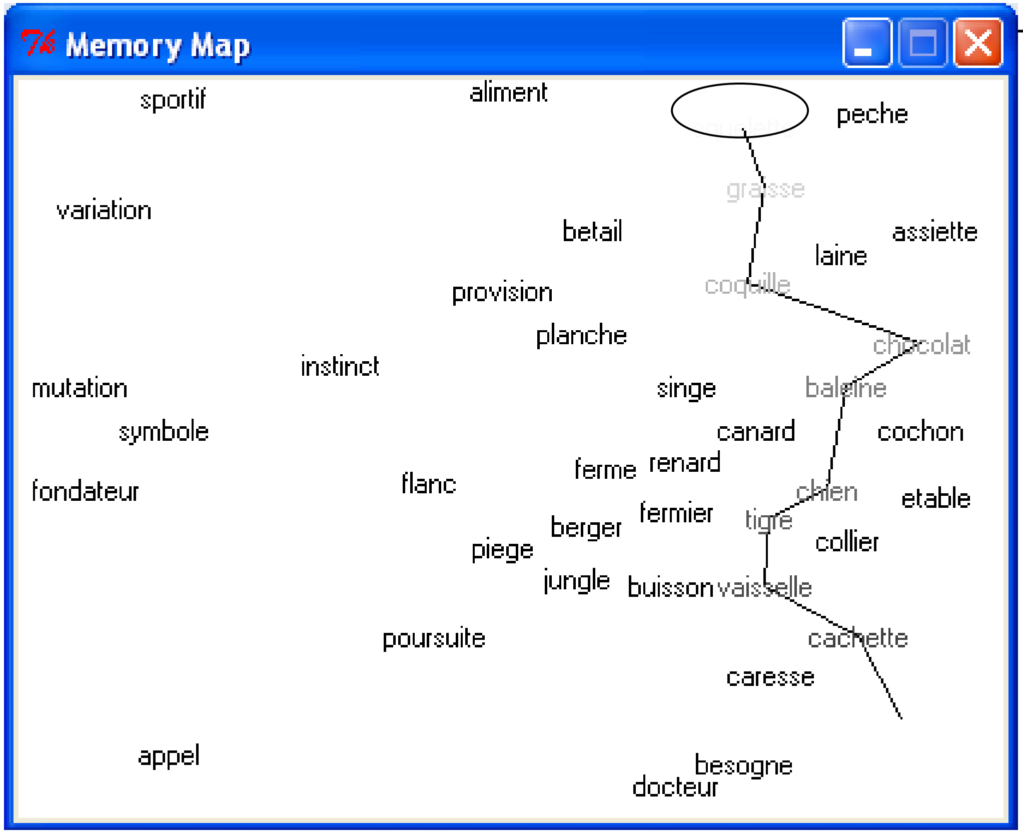
value=semantically close

Large angle=short cosine value=semantically far away

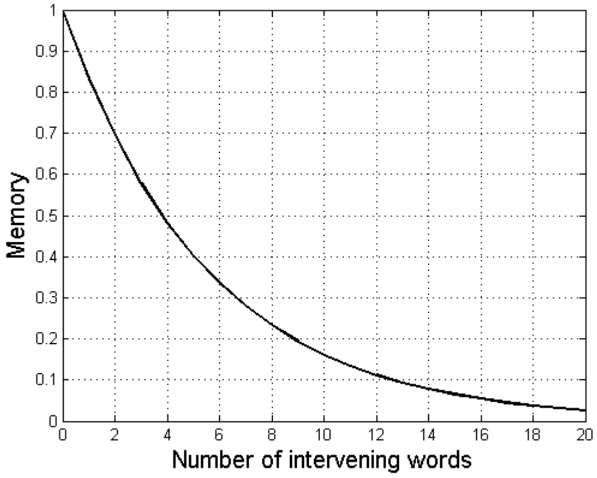
- If the fixated word is semantically close to the definition: neighbourhood
- If the fixated word is semantically away from the definition: neighbourhood
- If the fixated word is neutral with regard to the definition (cos = 0.2): it does nothing



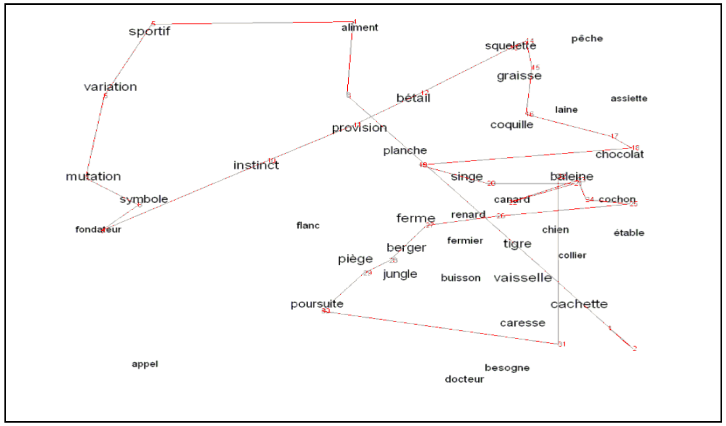
# Memory map

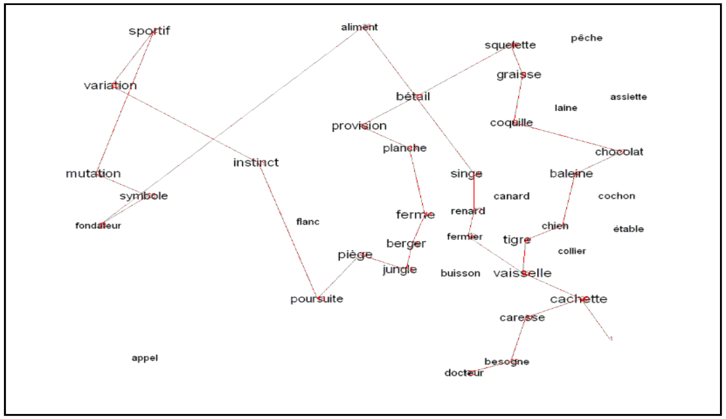


- Inhibition Of Return
- Forgetting mechanism
- Memory of the path



# Validation





- Based on experimental data:
  - 43 subjects, scanpaths registered with eyetracker: EyeLink II
  - 18 trials/subject
  - 3 visual conditions (within-subject)
  - 2 semantic conditions (between-subject)
- Need of high level variables to compare data and model
  - number of fixations until target is found
  - average angle between saccades
  - rate of spatial progressive saccades

# What are the "good" parameters for the model?

## All combinations of parameters V, S and M were tested

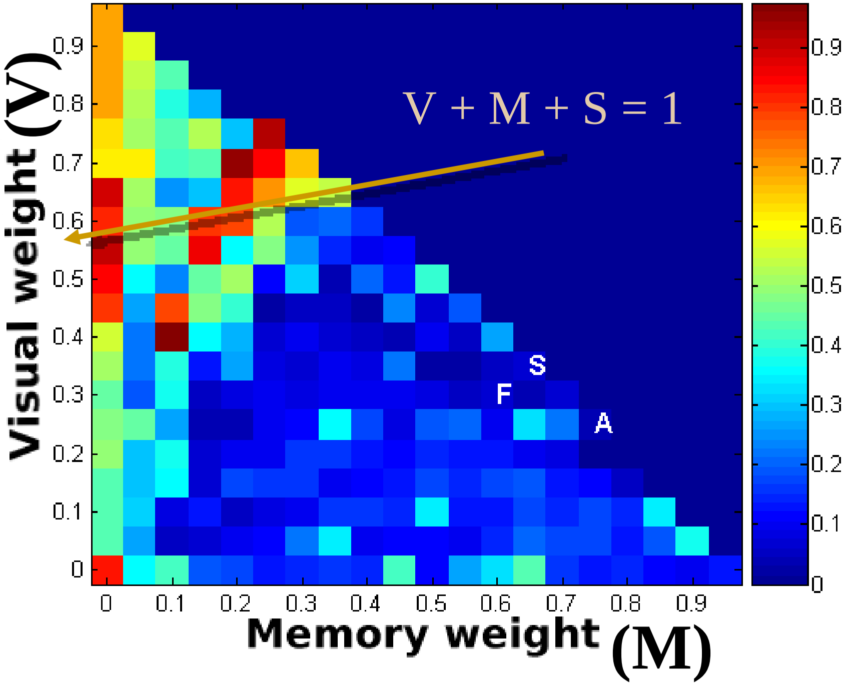
Weights with minimum error are the following:

– number of fixations until target is found (F): V = 0.3, S = 0.1, M = 0.6;

– average angle between saccades (A): V = 0.25, S = 0, M = 0.75;

– rate of progression saccades (S): V = 0.35, S = 0, M = 0.65;

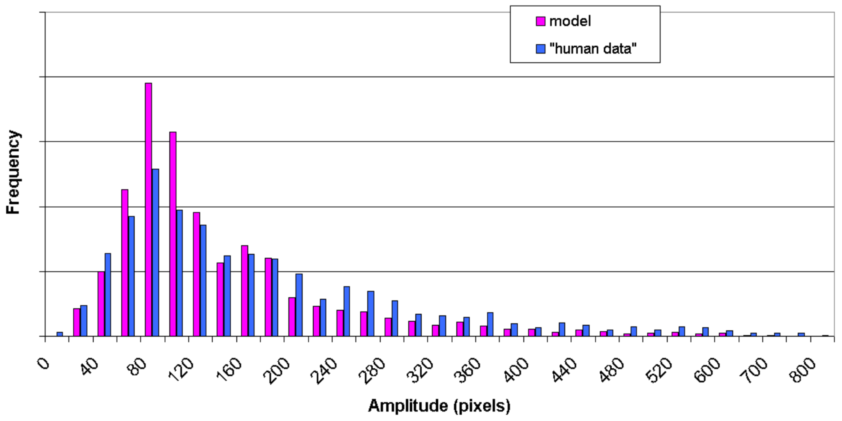
Average relative errors for all values of V and M



Comparison of human distribution of saccades with the best model.

V = .30, M = .60, S = .10

Distribution of saccade amplitude



# 5. Work Programme

**WP1: Data Collection (CVSSP, DoP & CMP)**

T1.1 Obtain machine-learning footage;

T1.2 Obtain psychology footage;

T1.3 Obtain annotation data-base.

**WP2: Psychological study of human learning in the sports video domain (DoP)**

T2.1 Determine the number and nature of the differing gaze-attention behavioural classes w.r.t familiar sport footage;

T2.2 Determine the evolution of these classes w.r.t unfamiliar footage;
T2.3 Compile a functional model of cognitive bootstrapping in human cognitive domain.

**WP3: Development and maintenance of cognitive video annotation system (CVSSP)**

T3.1 Generalise reference feature recogniser, object and agent trackers;

T3.2 Generalise interface to rule set module;

T3.3 Build any additional structures pertinent to automated cognition as suggested by WP2;

T3.4 Add Rule induction module;

T3.5 Add high-to-low-level cognitive feedback module;

T3.6 Add perceptual remapping module;

T3.7 Add match-commentary segmentation module;

T3.8 Carry out code maintenance and optimisation.

## WP4: Evaluation of transferable learning ability in the visual domain (CVSSP)

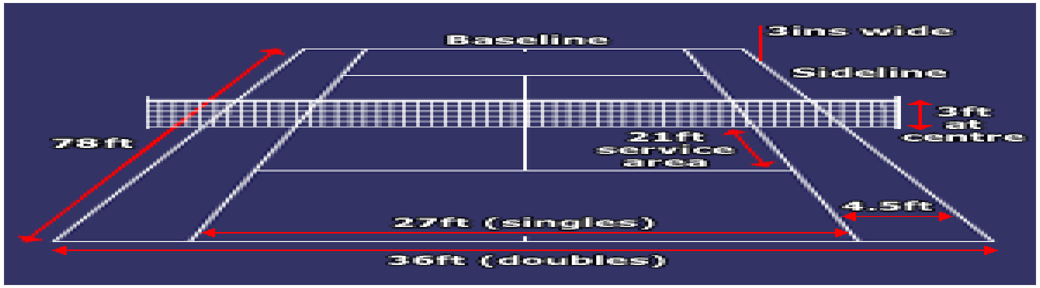
1. Transferable learning from tennis singles to tennis doubles}. These have identical low-level visual features, but differing rule protocols, serving to test logical induction given a fixed visual feature set.

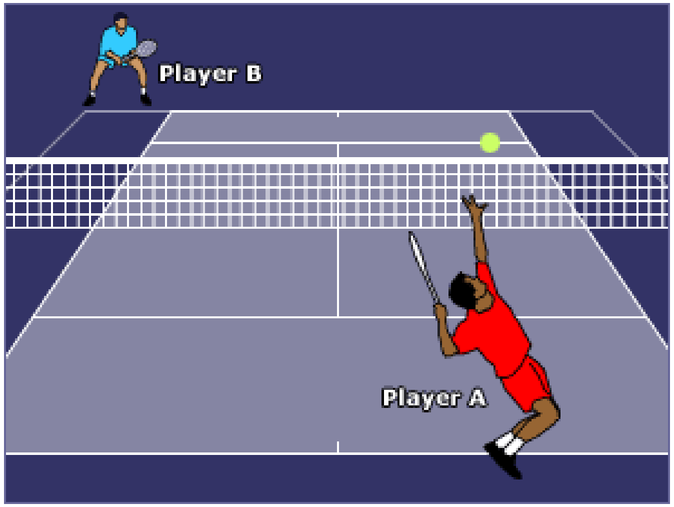
2. Transferable learning from singles tennis (Wimbledon) to singles (French Open)}. These have identical high-level rule protocols, but differing low-level features.

3. Transferable learning from singles tennis to doubles badminton}. These games significantly differ at both the high {\it and} low-levels of representation: hypotheses must propagate both up and down the learning hierarchy.

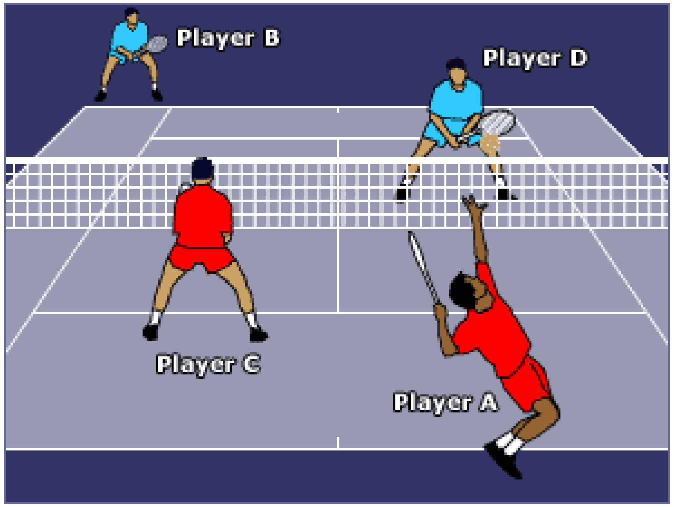
T4.1 Evaluation of transferable learning of low-level vision primitives;
T4.2 Evaluation of transferable learning of high-level game-representations;

T4.3 Evaluation of bidirectional learning transfer for full cognitive bootstrapping module.


Baseline
3ins wide
Sideline
3ft
at
centre
78ft
21ft
service
area
4.5ft
27ft (singles)
36ft (doubles)


Player B
Player A

====>


Player B
Player D
Player C
Player A

**WP5. Investigation of the effect of linguistic moderation on human learning (DoP/CMP)**

1. How prior verbal descriptions of game protocol affect the induction of low-level visual representation.
2. How  exposure to match commentary affects the induction of high and low level representations}.

T5.1 Determine how prior descriptions of game protocol affect the induction of low level visual features;

T5.2 Determine how exposure to match commentary affects induction of high and low-level visual features;

T5.3 Establish how visual grammar schematics correlate with the abstract match-commentary representation of WP6.

**WP6: Integration of Speech and Language Technology (CMP/CVSSP)**

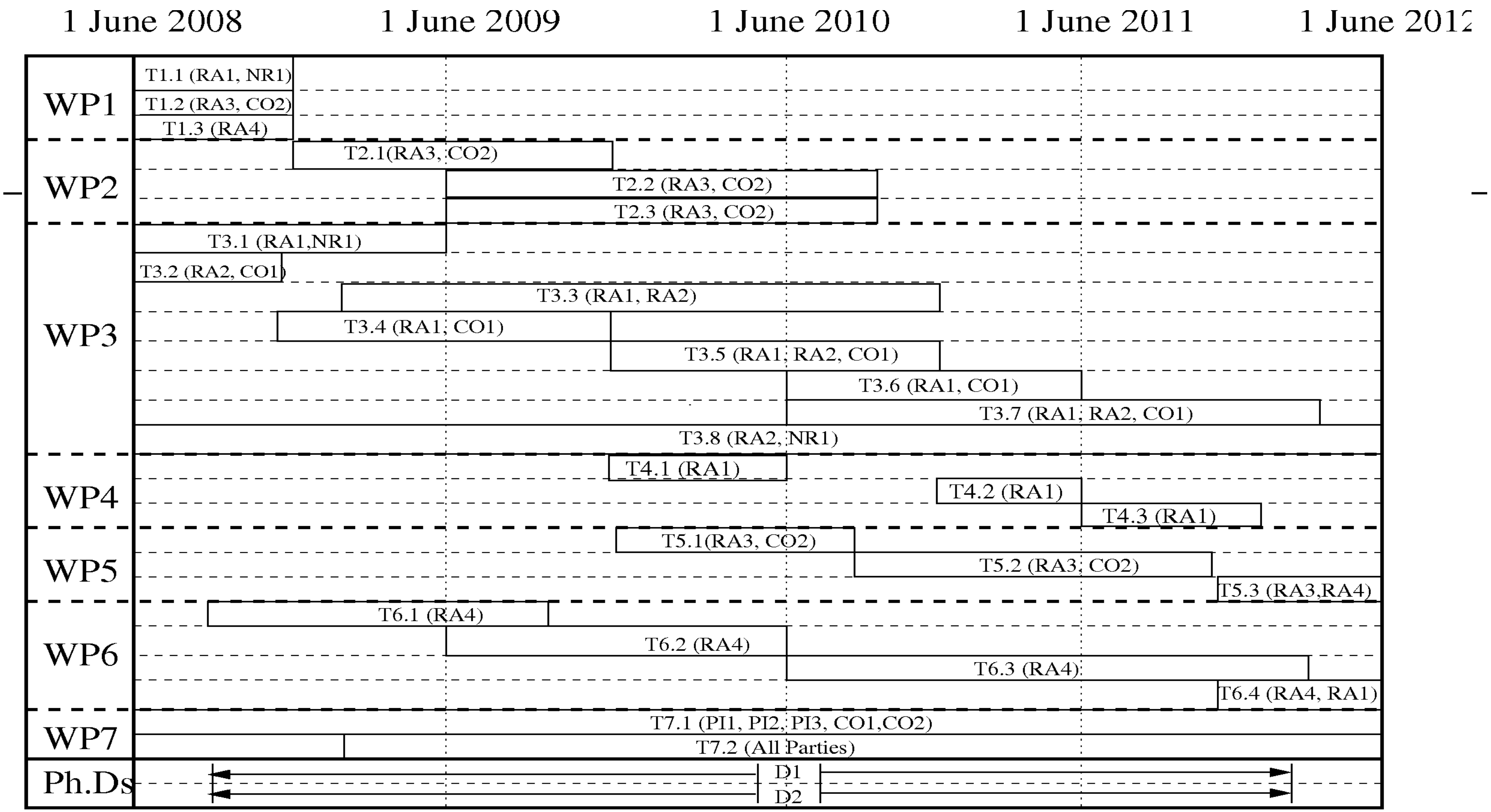
T6.1 Construction of initial speech-recognition system for transcribing audio;

T6.2 Construction of initial semantic parser for transposing commentary in action/ agent/referent/abstraction predicate-terms;

T6.3 Modification of parser & recognition systems to incorporate cross-modal bootstrapping;

T6.4 Construction of demonstrator.

**WP7: Management & dissemination (CVSSP)**



1 June 2008 | 1 June 2009 | 1 June 2010 | 1 June 2011 | 1 June 2012

WP1
- T1.1 (RA1, NR1)
- T1.2 (RA3, CO2)
- T1.3 (RA4)

WP2
- T2.1(RA3, CO2)
- T2.2 (RA3, CO2)
- T2.3 (RA3, CO2)

WP3
- T3.1 (RA1,NR1)
- T3.2 (RA2, CO1)
- T3.3 (RA1, RA2)
- T3.4 (RA1, CO1)
- T3.5 (RA1, RA2, CO1)
- T3.6 (RA1, CO1)
- T3.7 (RA1, RA2, CO1)
- T3.8 (RA2, NR1)

WP4
- T4.1 (RA1)
- T4.2 (RA1)
- T4.3 (RA1)

WP5
- T5.1(RA3, CO2)
- T5.2 (RA3, CO2)
- T5.3 (RA3,RA4)

WP6
- T6.1 (RA4)
- T6.2 (RA4)
- T6.3 (RA4)
- T6.4 (RA4, RA1)

WP7
- T7.1 (PI1, PI2, PI3, CO1,CO2)
- T7.2 (All Parties)

Ph.Ds
- D1
- D2

KEY

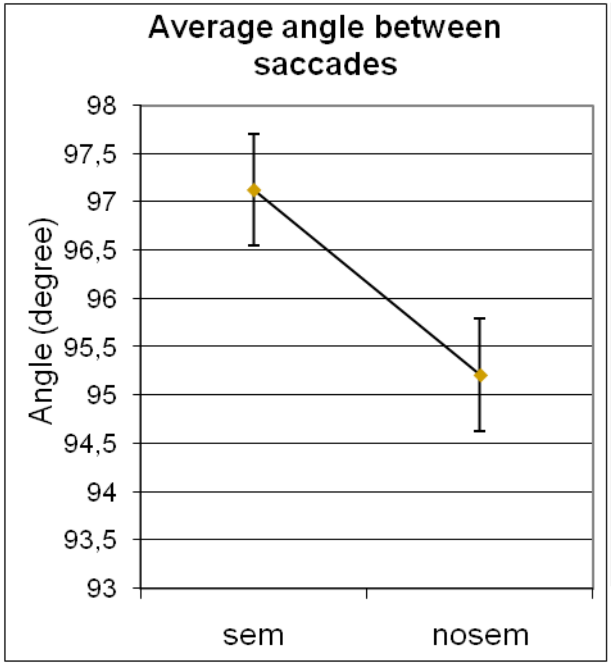
RA1 – CVSSP Research Associate
RA2 – " " "
RA3 – DoP Research Associate
RA4 – CMP Research Associate
D1 – CVSSP Ph.D. (Research Topic: Cognitive Bootsrapping)
D2 – CVSSP Ph.D. (Research Topic: Cross–Modal Grammars)

NR1 – CVSSP Named Researcher (20% time commitment)
CO1 – CVSSP Co–Investigator (60% time commitment)
CO2 – DoP Co–Investigator (10% time commitment)
PI1 – CVSSP Principal Investigator (10% time commitment)
PI2 – DoP Principal Investigator (5% time commitment)
PI3 – CMP Principal Investigator (10% time commitment)

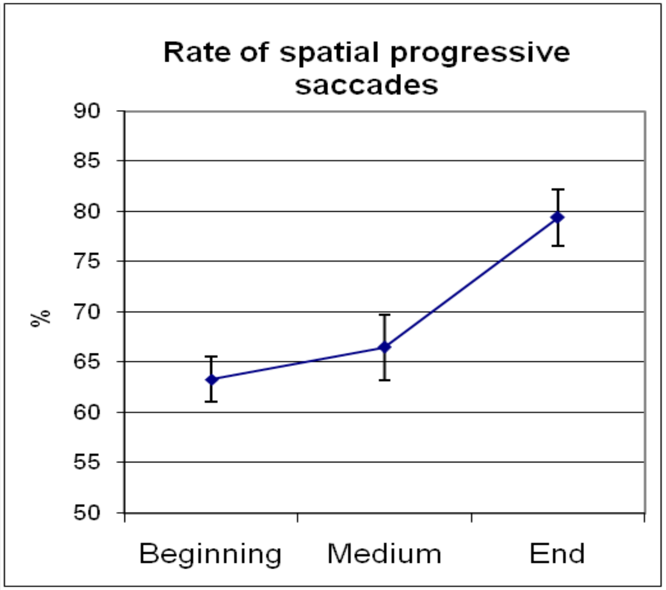
**Nominal job allocations at CVSSP**
**(with many horizontal/vertical overlaps)**

**Josef:** Principle Investigator
**Me:** Project Coordinator
**Teo:** action classification + eye-tracking behavioural mining
**Aftab:** rule induction/top-down feedback
**Fei:** ball/shuttlecock tracking
**Ibrahim:** HMM universalisation, cross-modality
**Bill:** code-base overseer

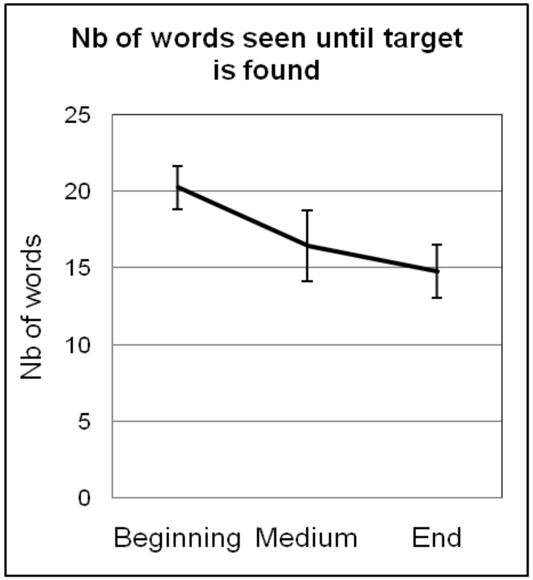
# Results on data subjects



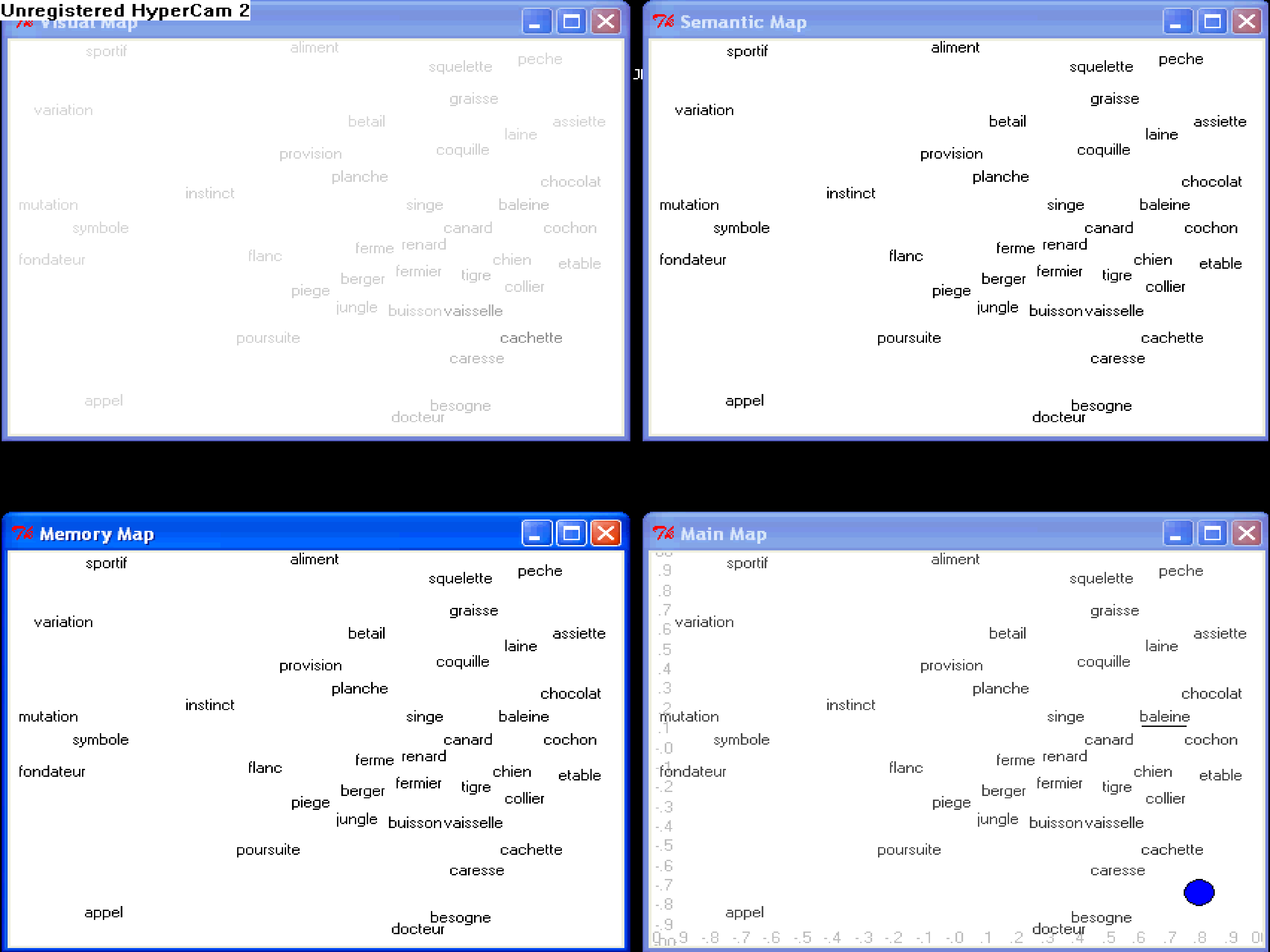
Significant difference (p=0.036)



Significant difference (p<0.001)

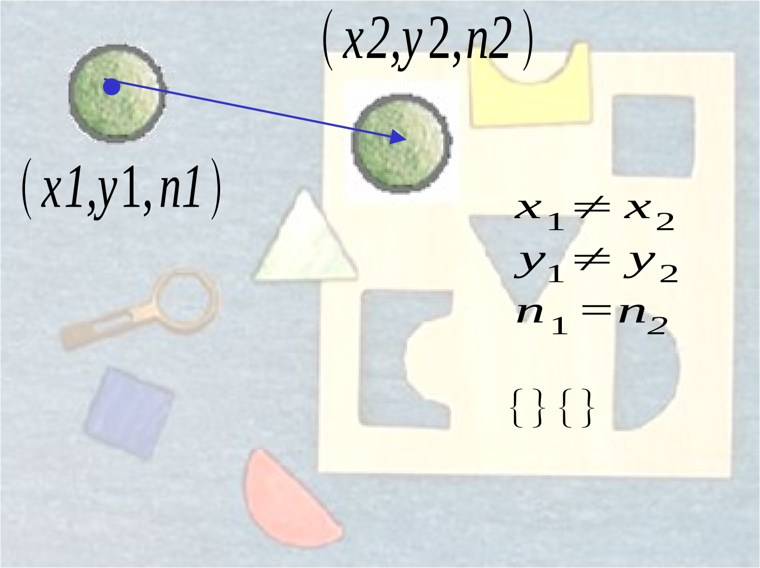


Significant difference (p<0.02)


Visual Map
sportif
aliment
squelette
peche
variation
graisse
betail
assiette
laine
provision
coquille
planche
chocolat
instinct
mutation
singe
baleine
symbole
canard
cochon
ferme
renard
fondateur
flanc
chien
etable
berger
fermier
tigre
collier
piege
jungle
buisson
vaisselle
poursuite
cachette
caresse
appel
besogne
docteur
Semantic Map
sportif
aliment
squelette
peche
variation
graisse
betail
assiette
laine
provision
coquille
planche
chocolat
instinct
mutation
singe
baleine
symbole
canard
cochon
ferme
renard
fondateur
flanc
chien
etable
berger
fermier
tigre
collier
piege
jungle
buisson
vaisselle
poursuite
cachette
caresse
appel
besogne
docteur
Memory Map
sportif
aliment
squelette
peche
variation
graisse
betail
assiette
laine
provision
coquille
planche
chocolat
instinct
mutation
singe
baleine
symbole
canard
cochon
ferme
renard
fondateur
flanc
chien
etable
berger
fermier
tigre
collier
piege
jungle
buisson
vaisselle
poursuite
cachette
caresse
appel
besogne
docteur
Main Map
.9
.8
.7
.6
.5
.4
.3
.2
.1
-.0
-.1
-.2
-.3
-.4
-.5
-.6
-.7
-.8
-.9
sportif
aliment
squelette
peche
variation
graisse
betail
assiette
laine
provision
coquille
planche
chocolat
instinct
mutation
singe
baleine
symbole
canard
cochon
ferme
renard
fondateur
flanc
chien
etable
berger
fermier
tigre
collier
piege
jungle
buisson
vaisselle
poursuite
cachette
caresse
appel
besogne
docteur
-.9
-.8
-.7
-.6
-.5
-.4
-.3
-.2
-.1
-.0
.1
.2
.3
.4
.5
.6
.7
.8
.9

# Supervised Learning Framework

Symbolic goal: an element of the symbolic perceptual domain (ie a subset of the original feature space with the **redundant context eliminated**).
=> Goals are therefore parametrised by the **features salient to the current learning scenario.**



eg Movement of object n1 onto n2 involves 6 perceptual parameters (x1,y1,n1,x2,x2,n2)

- however the goal can be specified by **just 2 parameters:** (n1,n2).