

DIPLECS Deliverable D7.1
(Prototype system for bootstrapping formalised strategies and associated logical representations)

Grant Agreement number: 215078

Project acronym: DIPLECS

Project title: Dynamic Interactive Perception-action LEarning in Cognitive Systems

Funding Scheme: Small or medium-scale focused research project

Date of latest version of Annex I against which the assessment will be made: 5/10/2007

Period covered: from 1/12/2007 to 30/05/2009

Responsible Beneficiary: David Windridge, Surrey University UK

Tel: +44 (0) 1483 68 6048

Fax: +44 (0) 1483 68 6031

E-mail: d.windridge@surrey.ac.uk

Address: CVSSP, University of Surrey, Guildford, GU2 7XH

Contents

1	Introduction	3
1.1	Objectives	3
1.2	Approach to System Construction	5
1.3	Context of Deliverable D7.1 within WP7	6
2	Work Details: Part 1 - Annotation Specification and Implementation	7
2.1	Parallels between the Highway Code & ECOM models	7
2.2	Division of Video Annotation Protocols	10
2.3	Creation of Tool for Projective Ground-Plane Tracking	11
2.4	Expert Annotation of Intentional States	14
3	Work Details: Part 2 - Context-Free Machine Learning	16
4	Work Details: Part 3 Logic Infrastructure Construction	22
4.1	Details of Deductive Logical Highway Code Implementation	26
4.1.1	Details of Logical ECOM Implementation	30
4.2	Combining Logical Consistency Considerations with Decision Tree Outputs	31
5	Work Details: Part 4. The Top-Down Feature Respecification Module	36
6	Summary	38
6.1	Summary of Activities	38
6.2	Future Activities	39

1 Introduction

1.1 Objectives

WP7, the System Supervision and Strategies work package, is concerned with the determination, and indication, of the optimal overall approach to driving at the most structurally abstract levels of the perception-action hierarchy. While it is possible to specify certain of these strategies in advance (for instance, via a formal representation of the official Highway Code in first-order logic), it is necessary to establish a complete *cognitive* model of intentionality with respect to which the Highway Code can be considered to act as a constraint.

Both intentions and driving rules must be thus expressed in terms of the corresponding perceptual entities, albeit described in structural terms (for example, traffic light states). The strategy induction system must consequently be capable of correlating the representative hierarchy with the induction of driving rule protocols. In this respect, the derivation of human driving strategies is a direct continuation of the DIPLECS cognitive bootstrapping architecture, albeit in structural terms. It must thus be capable of reconciling abstract deduction with pre-symbolic input presented by the various detectors generated in other work packages; in other words, it must address the issues of symbol grounding [3] and symbol tethering [7].

In the context of its practical application, the system arising from WP7 must thus, as its central mode of operation, be able to infer the current driving strategy adopted by the driver based on the external visual scene and the driver's intentions with respect to it. The system must hence be able to correlate motor and signal inputs with the visual features employed by the driver determined via eye-tracking in relation to the detected features of the external driving scene ¹.

Deliverable D7.1 thus aims to collate human driving strategies in formalised, protocol-expressible-terms, via perceptual entities of the appropriate level. To do this we employ the ECOM model developed by ARMINES. The delivered system must further be capable of ascending and descending the formalised representational hierarchy as required by the driving context. Hence, any visual features present to the forward camera must be capable of being fed through the full perceptual hierarchy until a high-level scene representation is reached that is encompassed by a recognised driving protocol. This is, to an extent, an interfacing and quality-control endeavour, ensuring that any redundancies, incomplete-

¹Findings from experimental psychology [4, 5, 6] suggest that attention within dynamic complex scenes relates primarily to scene transitions. Coherence Theory [6] indicates that if perception forms coherent scene representations, they must be *virtual* representations, such that it only appears to higher cognitive levels as if all objects simultaneously have coherent representations [5]. The idea of static, coherent world models is hence replaced by that of a dynamic representation sensitive to task-demands and observer-expectations

nesses or contradictions in the protocol-relevant scene-descriptors are resolved in order to produce an intentional representation that is consistent in any circumstances, given the potential safety cost associated with inducing no overall strategy.

It is hence a bottom-up process, in so far as it concerns logically establishing spatio-temporal consistency in the detected features. However, it is also potentially a top-down process in that the global consistency checks on the basis of well-determined prior knowledge also have the possibility of reweighting and respecifying detector inputs. It ought thus be possible to establish that a particular detector (for instance, a 'stop' sign detector) is inaccurate or faulty on the basis of global scene context, or that (in full bootstrap mode) two individually detected components of a sign (for instance a circle and a number from a speed limit sign) are in fact part of the same entity on the basis of their logical co-dependency. For the current system we address the former capability under simulated detector failure conditions.

D7.1 is thus a proof-of concept prototype system for the deductive implementation of formalised strategies, and for the top-down bootstrapping of associated logical representations.

D7.1 thus consists of four distinct modules (reflected in the delineation of the remaining sections):

1. An intentional annotation system (for ground-truthing of training data in terms of the ECOM model).
2. An atemporal, context-free intentional learning system (based on decision trees for determining *instantaneous* clause-like decision rules).
3. A Highway Code & ECOM based logical deduction system (for determining long-range *a priori* logical consistency classes of decision tree outputs).
4. A top-down feature respecification module (for re-weighting feature detectors on the basis of global logical consistency).

For this proof of concept system, all data is ground truthed with only simulated experimental noise. The later deliverable D7.2 will address *in situ* functionality.

Ultimately, the system will be able to parse the difference between the computer vision representation of the world and the inferred cognitive representation of world of the DIPLECS-car driver with respect to inferred driving intention in order to generate appropriate warnings.

eg 'you're attempting to turn the wrong way down one-way street'

We thus work on the assumption that inappropriate Regulating/Monitoring-level ECOM intentions are all validly critiquable (being high-level, implicitly conceptual intentions),

while inappropriate Tracking-level ECOM intentions (eg getting too close to the driver ahead) are not. This is consistent with road safety findings since lower-level errors generally require instantaneous responses, and any verbalised warning would be unhelpful (eg 'you're getting too close to the car in front'), and could even be dangerously distracting. It is only at the higher ECOM levels that such warnings would be useful, when there is time to process the input and formulate a response.

1.2 Approach to System Construction

The distinct components of Deliverable D7.1 are constructed in the following order:

1. We firstly formalise and render mutually-consistent the Highway Code and ECOM models as a list of rules intermediate between first order logic and the English-based terms in which they were first formalised. We then compile a comprehensive list of low-level predicates relating to the ECOM model of Regulating/Monitoring intentions and the corresponding highway-code-derived world-model. These will correlate with the various possible detector inputs (ie they will relate to visual features, signal inputs, motor inputs and gaze positions).

2. We then build all relevant *a priori* hierarchical predication and constraints in terms of PROLOG-based first order logical clauses describing the ECOM and Highway Code models. For example, the hierarchical relation between lane and road occupancy can be formulated:

$$\forall n \text{ In_Lane}(n, \text{car}_i) \Rightarrow \text{In_Road}(\text{car}_i)$$

3. We next obtain ground truth per-frame data for all of the low-level predicates. For this purpose, we select junction scenarios as constituting the most complete exemplars of ECOM intentionality with their conditional logic dependencies (for example, on traffic-light states). Cross roads, in particular, act as a superset of all other road junction situations, with their exhaustive pathing options. World annotation is conducted via machine-learning techniques, while intentional annotation is carried out by human experts via the interface with ARMINES, and manually correlated with the machine-learning output as a quality control measure.

4. We then extend the annotation to all of the intermediate and high-level predication by application of the rules determined in stage 2, giving rise to a per-frame binary feature vector of scene and intentional descriptors at progressively abstracted levels of the representational hierarchy.

5. Through the application of decision-trees to the above training-data, we obtain unordered/context-free association rules for determining intentional classes and sub-classes given the signalling/motor/feature inputs.

6. We then employ an online PROLOG-based system to perform deductive logical implication with respect to the ECOM & Highway code models in order to consistency-check spatio-temporal compositions of intentional classes with respect to *a priori* environmental/intentional rules. We thus optimise the per-frame decision-tree in terms of global context.

7. We finally assess the top-down potential of the system to perform autonomous predicate respecification by reweighting feature-detectors if indicated on the basis of global logical consistency. We thereby complete the cognitive bootstrapping loop.

1.3 Context of Deliverable D7.1 within WP7

The current deliverable arises from work of WP 7.1 (Determination of human strategies and corresponding symbolic representations) and WP7.2 (Deduction of Appropriate strategy for any given situation).

Completion of WP 7.2 essentially requires the implementation of the ECOM control hierarchy in logical terms (with tracking and regulating intentions appropriately modelled) in a manner consistent with the relevant highway code subset.

Crucial to this approach is the interface with ARMINES, providing both the psychological annotation categories within the intentional hierarchy, as well as the annotation itself. Through this collaboration within DIPLECS, we are thus able to transfer knowledge from psychology to the technical system specification in a concrete fashion. This is very much an interactive process, and one that we anticipate will be ongoing throughout the project.

We can thus envisage WP 7.1 as a specifying a set of classification problems relating to driver intentionality, which then becomes an output specification (with an appropriate degree of logical closure and consistency checking) in WP 7.2. This is reflected in our two-tiered approach to machine learning in the current deliverable: in essence, the decision tree learning captures the contingent correlation between driver inputs/world states and the ECOM intentions, while the logic system defines the *a priori* correlation. The combination of the two effectively allows us to ensure that abstract symbolic entities are properly grounded in the noisy detector inputs.

Deliverable D7.1 (in which intention is deduced from the current situation and the gaze, signalling & control behaviour) hence constitutes a proof-of-concept demonstration operating in certain constrained scenarios, which will have the capacity to be built-upon

as the project progresses (ie a 'prototype' bootstrap strategy/logical representation system will be presented at month 18 in accordance with the Technical Annex).

The deliverable D7.2 will be a logically-complete and consistent *in situ* induction/deduction system.

2 Work Details: Part 1 - Annotation Specification and Implementation

The ECOM model consist of four layers of control (or concurrent loops), only three of which are appropriate to DIPLECS: Monitoring, Regulating & Tracking. For the current purposes, we adopt the principle that Tracking level behaviour manages the continuous activity undertaken to keep the vehicle within a specific, discrete conceptual configuration (eg car-order within a lane). From a driver's perspective it refers to minor modifications of car speed, direction of car, intended distance from the car in front or back, or the lateral position on the road. In the case of an experienced driver these actions are predominantly a matter of physical reflex without high-level conscious attention. (However, in the case of an inexperienced driver these Tracking behaviours may conceivably be enacted at the Regulating level). Regulating intentions hence provide an input to the Tracking control-loop to perform a specific, high-way code relevant action, e.g changing lane. Other regulating intentions include intentionally stopping and turning right/left at a junction, and as such can, where necessary, be linked hierarchically.

2.1 Parallels between the Highway Code & ECOM models

The UK highway code formalises the interface between driver intentionality and the physical world. As all legal drivers must be aware of it, it can be considered as constituting an *a priori* link between the two. As such, it delineates certain concrete entities in both domains.

In particular, physical entities deemed relevant by the high way code include: all road using entities (pedestrians, cars, bicycles etc); *relative* orientations, velocities, and lane-positions of these entities, absolute velocities and orientations of these entities (ie relative to the road); all signals made by these entities; all signs; all traffic signal entities (traffic lights etc, independently of state); all traffic signal states (red light, amber light etc); all road markings (including lane/road-dividers/painted signs); all junctions (cross-roads, t-junctions, roundabouts). There are also two more general physical entities: Country driving areas and Residential driving areas.

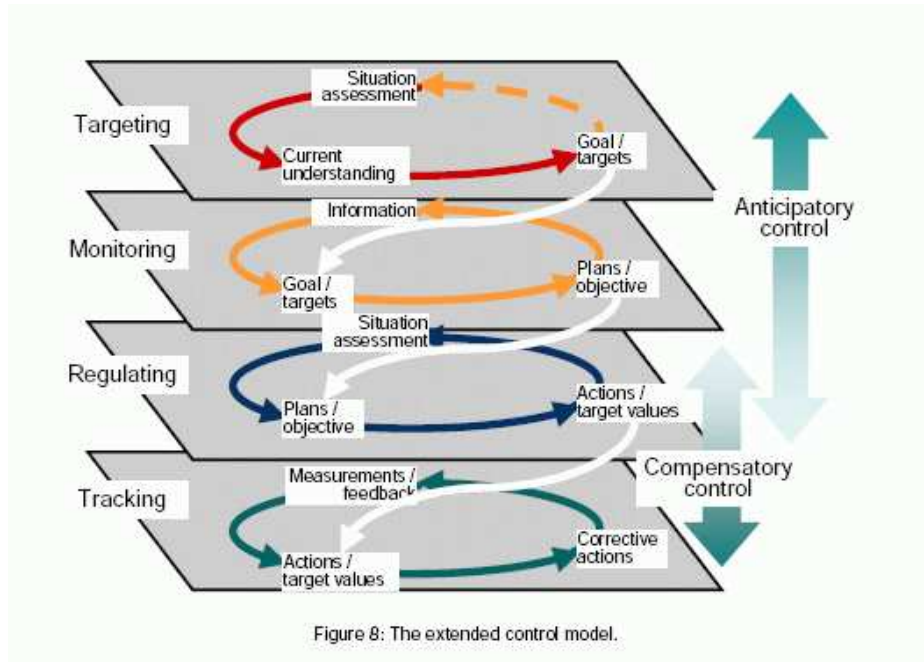


Figure 1: ECOM model layers



Figure 2: Example Highway Code rule

Cognitive entities/states deemed relevant by the high-way code include, at the lower-level, the proximal object of attention and the direction of attention (eg whether the driver looks behind before changing lane). High-level cognitive entities include immediate intentions (Stopping at Junction, Turning Left at Junction etc) and situational anticipations

(assumptions about what other road users intentions are, or how the road situation will change).

As such, the Highway code strongly correlates with certain levels of the ECOM model developed at ARMINES - in particular: Monitoring, Tracking and Regulating. Regulating intentions (Stopping at Junction, Turning Left at Junction etc) are relatively few in number and relate to those concretely-formed intentions of which we are aware and can communicate, but which are free of navigational assumptions (being subsumed by them). These are therefore typically the intentions that we formulate in relation to road signs and signals.

Tracking-level cognitive entities on the other hand (eg the proximity/direction of object of attention) are required in low-level perception-action for maintaining the current configuration (eg maintaining separation and lane-position). In general, once we are proficient at these tasks, we do not have a strong awareness of the intention to carry them out. However, being subsumptively related to the higher intentions, they are implicit within these more conscious acts.

Intentionality is thus evidenced in the current work through driver motor inputs and its correlation with eye-gaze (thus defining a perception-action cycle). Beyond our employment of the ECOM model, we do not, however, make any strong cognitive claims about the cognitive nature of conscious intentionality.

Approximately paralleling this ECOM-based subsumption hierarchy [1], there is also to be found a hierarchical structuring of the driver's representation of the world implicit in the Highway code. For instance, associated with the Monitoring/Regulating ECOM levels, the Highway code assumes the existence of ordered sets defining lane structure (defining the cross-section of the road orthogonal to direction of travel), topological connectives defining the junction types (cross-road, t-junction, roundabout), topological label spaces defining the location of the key road-signs and traffic signals, and temporal topological label attribute sets (defining eg the traffic light states [*red* → *amber* → *green*]). At the Tracking level, there are the car-relative and road-relative Cartesian vector-spaces of road users, along with their velocity attributes. In fact, the presence of Cartesian-like spaces within the Highway code hierarchy can be treated as a key cut off within DIPLECS, determining whether structural or stochastic methods are the most appropriate.

The objective for the logical deduction system embodying all *a priori* knowledge is therefore to combine the ECOM model of Tracking, Monitoring and Regulating with the high-level representational structures implicit in the highway code in order to build a complete, syntactically-closed logical model of road activity. At the same time, we aim to allow maximum scope for the active bootstrapping of novel bottom-up representations.

Note that an intention-based logical implementation of the road situation potentially implies parallel treatment between the DIPLECS car and other vehicles. While we have more direct access to the DIPLECS car driver's intentionality via gaze measurements that

relate to external behaviours, we must infer other driver’s intentional states purely from their external behaviour. The relationship between the various representations (and representations of representations) is given in figure 3. Most of this complexity will be managed internally via the logic system in its final iteration.

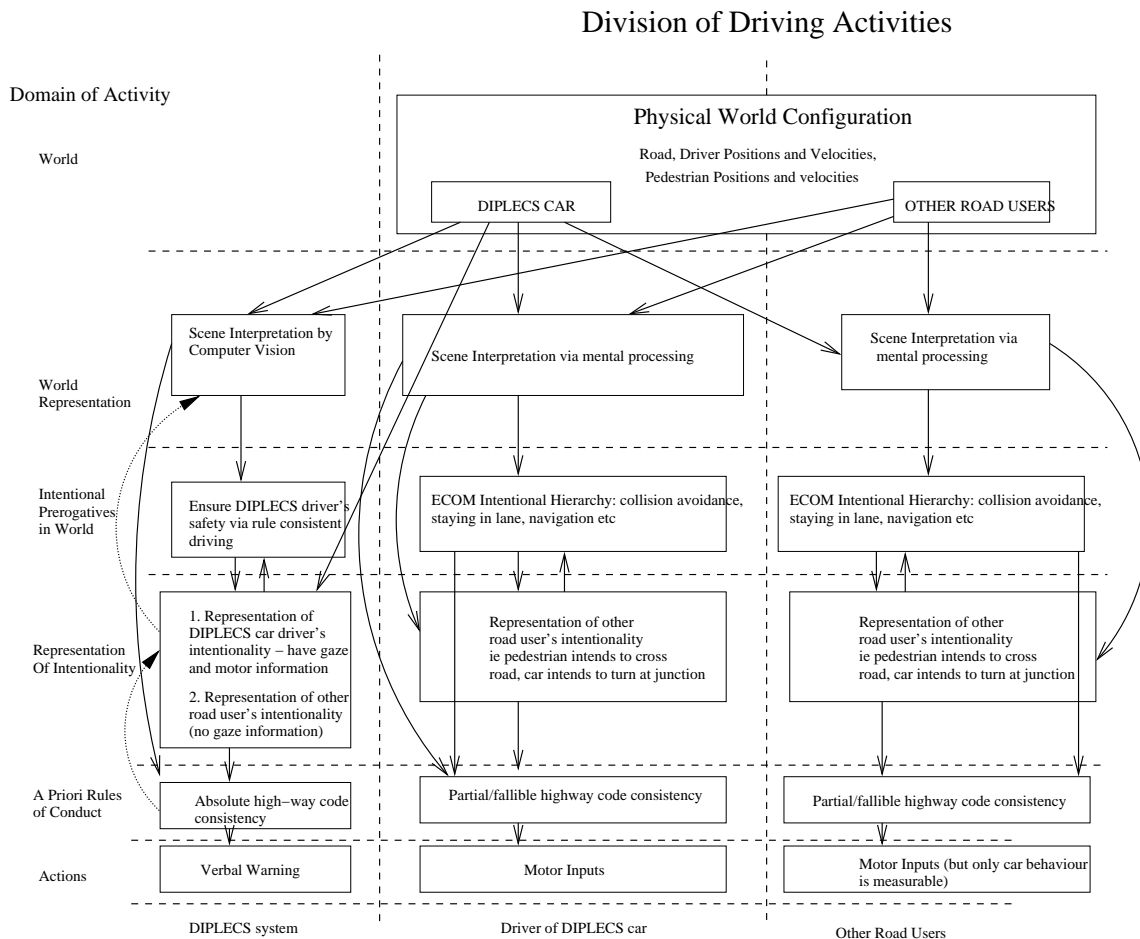


Figure 3: Relation between the various representations

2.2 Division of Video Annotation Protocols

In approaching the mapping of the ECOM model onto the high-way code-relevant entities, we focus exclusively on providing per-frame Regulating/Monitoring Level ground-truth information (this being most appropriate to the logical induction undertaken by WP7), with

Tracking-level behaviour (slowing, accelerating, car-following, maintaining lane position) treated as a potential classifier input at a later stage.

As well as the Tracking verses Regulating/Monitoring category distinction, another essential category distinction implicit in the ground-truth annotation of the junction scenarios chosen for annotation will be 'ground-plane' verses 'view-plane' high-level scene description. This distinction encompasses the notion that we are not concerned with the absolute position and orientation of certain indicator features, such as sign and lights, but only with the fact that they can be seen and relate to certain road lanes. They are therefore bounded in the view plane. Other indicator entities, such as road-arrows and lane-markers are intrinsically positional in their relation the world, and consequently their outlines are considered as bounded with respect to the ground plane.

This is illustrated in fig. 4.

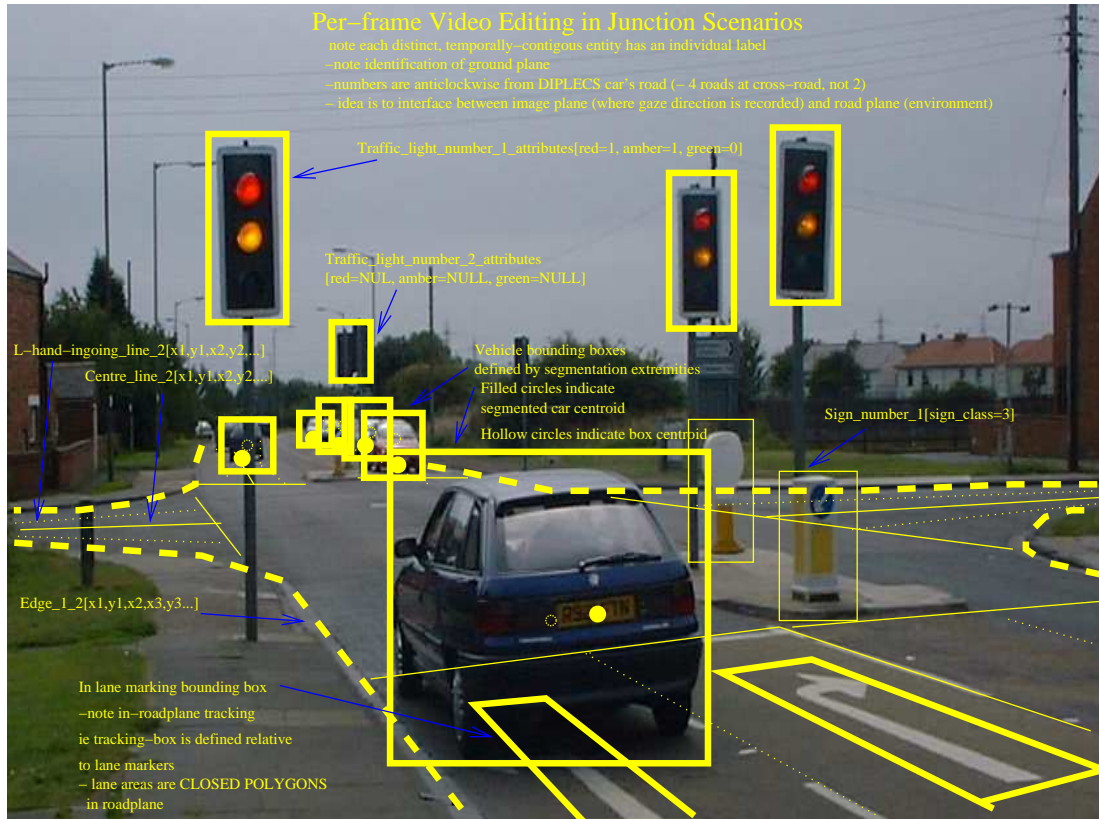
Making this category distinction enables the direct coupling of computer-vision entities with cognitive intentionality (deduced via control and signal inputs, as well as gaze-behaviour) in a manner directly congruent with the ECOM intention/control model as interpreted above. Gaze behaviour is hence characterised, on a per-frame basis, via logical-entity (eg sign, traffic-light) bounding-box transitions within both the ground-plane and view-planes. (More complex behaviour depending on dwell-time characteristics within the ground-plane and view-plane bounding boxes can thus be modelled later if required).

Per-frame annotation of the mobile ground plane entities (cars, pedestrians) and static indicators (signs, lights) is undertaken via manual placement of bounding boxes. However, The propagation of junction topologies and zones throughout the video footage in order to establish correlations with the gaze direction cannot be treated in this way, and a special tool is required for reliable ground-plane bounding box-propagation.

2.3 Creation of Tool for Projective Ground-Plane Tracking

The ground-plane tracking tool for propagating key ground-plane entities was found to be most effectively accomplished via the following five-stage methodology:

1. Temporally aggregate LIDAR Data to give an approximate delineation of junction-outlines within inner-urban areas (cf fig 5).
2. Histogram and drift Correct aggregate LIDAR data to further distinguish road outline and differentiate it from traffic-trajectory noise (cf fig 6).



Car bounding box has corner [dx,dy] rescaling attributes as well as [x,y] positional attributes ie Car_number_1_box[x1,y1,x2,y2,y3,y4,dx1,dy1,dx2,dy2,dx3,dy3,dx4,dy4]
 - similarly for sign/road-marking/traffic-light bounding boxes
 (note that a null is returned for the relevant corner attribute when occlusion occurs ie Car_number_1_box(x1,y1, Null, Null, x3,y3 ...) -segmented centroid is as observed
 Car itself has signalling attributes ie Car_number_1_attribute[left-indicator=0, right-indicator=1, brake-indicator=1] and a (road-relative) orientation estimation
 DIPLECS car also has motor-control attributes -ie Car_number_0_motor_input[steering-wheel-position=-3, brake-intensity=5, break-foot-hovering=1], also gaze, etc
 (DIPLECS car is always number 0 and has NULL bounding box attributes)

Figure 4: Ground plane and View plane annotation

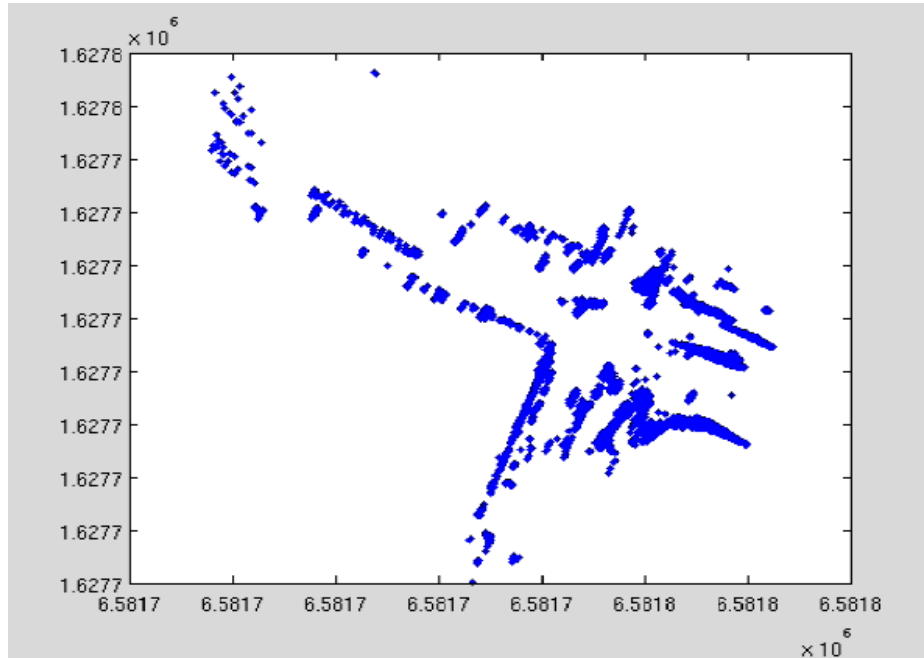


Figure 5: Temporal aggregation of LIDAR Data

3. Edge-detect and Hough Transform histogram with high angular suppression to obtain predominating road vectors. That is, we obtain a Canny edge detected image such that for intensity values with coordinates (x_0, y_0) in the image plane, and Hough intensity $H(r, \theta)$ defined via the mapping:

$$r(\theta) = x_0 \cdot \cos \theta + y_0 \cdot \sin \theta \quad (1)$$

we apply a selection criterion such that the top two line candidates (ie highest density bins) are subject to the constraint that they are $> 30^\circ$ apart in the θ ordinal ie:

$$\{(r_1, \theta_1), (r_2, \theta_2)\} : \operatorname{argmax}_{r_1, \theta_1, r_2, \theta_2} H(r_1, \theta_1) + H(r_2, \theta_2) \text{ s.t. } |\theta_1 - \theta_2| > 30^\circ \quad (2)$$

This is illustrated in fig 7.

4. Fit junction topology and pedestrian-crossing/lane structure to $\{(r_1, \theta_1), (r_2, \theta_2)\}$ on the basis on *a priori* knowledge of their absolute number (but not scale or orientation). (cf fig 8 and 9).

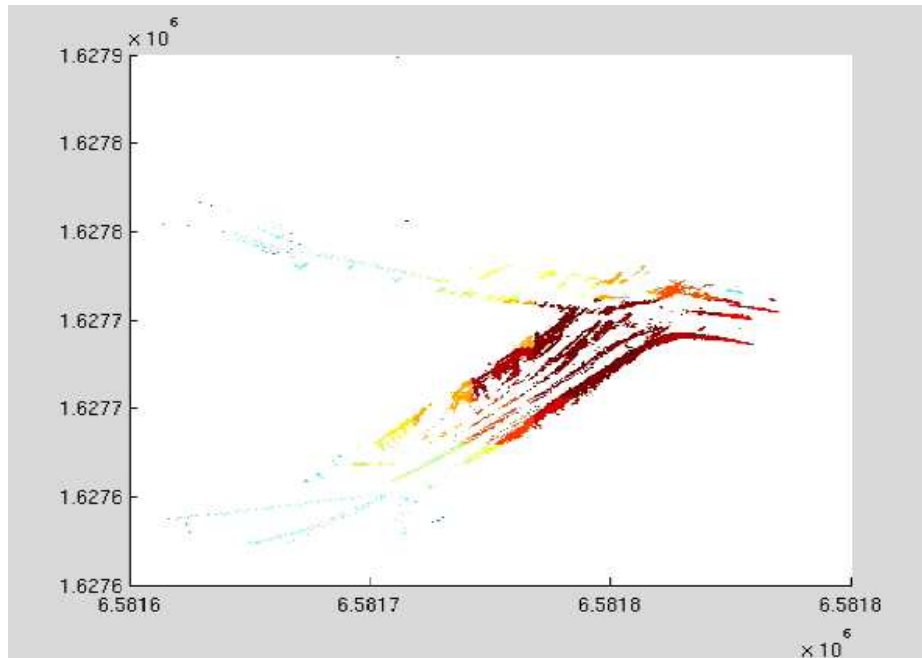


Figure 6: Histogram/drift Correction of aggregate LIDAR data

5. Application of approximate view-plane transformation matrix in order to project the junction topology into screen frame for further small-scale adjustments of car-height/camera orientation etc. This idea as illustrated in the frame sequence of figures 10 to 12.

The outputs of the above process are then the projected junction-lane bounding boxes on the driver's view plane, along with the per-frame gaze occupancies of these entities. These are then added to the view-plane bounding-box annotation.

This data can then be utilised to formulate the appropriate model of attention via the research-interface with ARMINES, and to establish intentional correlations such that gaze can be mapped onto the symbolic-logical visual categories (traffic-light states, signs etc).

2.4 Expert Annotation of Intentional States

The intentional states are expert annotated along the lines of the ANVIL annotation model used in WP6. This one-dimensional, temporal annotation is based on the observed context and the pre-existing partial rule formulations, based on the empirically-motivated psychological model (cf fig 13).

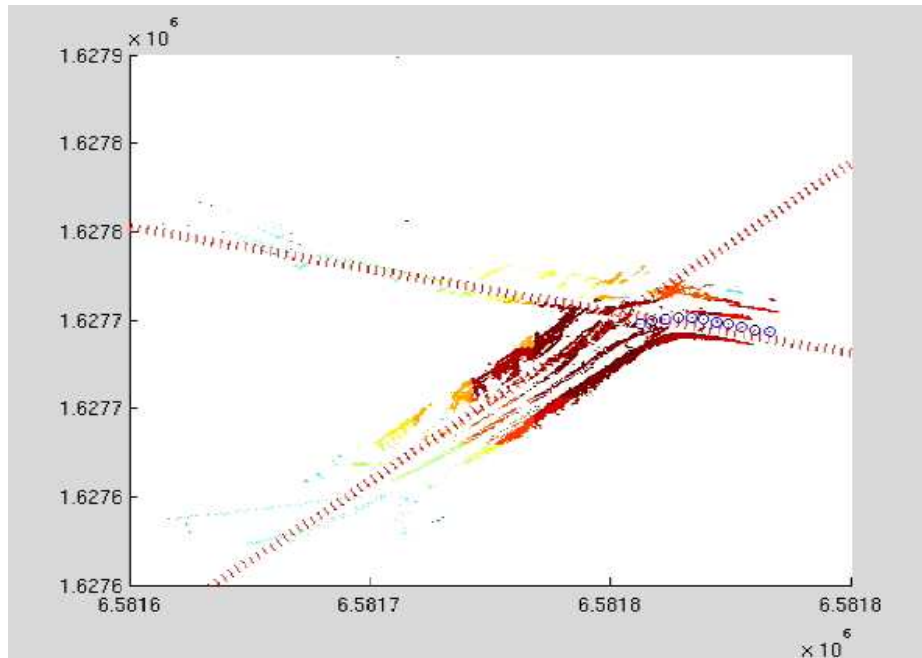


Figure 7: Edge-detect and Hough Transform histogram

There is hence the potential for disparities to exist between the human annotation and the later explicit formalisation of ECOM in first-order logical terms. That is, we allow the possibility for expert intuition to be more valid than theory-based formulation.

Once this is done per-frame for all ECOM levels, and gaze has been mapped onto the symbolic-logical visual categories (traffic light states etc) via bounding box occupancy tests, the composite data serves as a training set for for the next stage, in which intention is deduced from gaze, signalling & control behaviour with respect to the changing road configuration.

Outputs are thus the lane bounding box labels with per-frame gaze occupancies: these are supplemented with car/pedestrian lane occupancies and the intentional states arising from the ARMINES annotations, giving all per-frame Monitoring/Regulating-Level information.

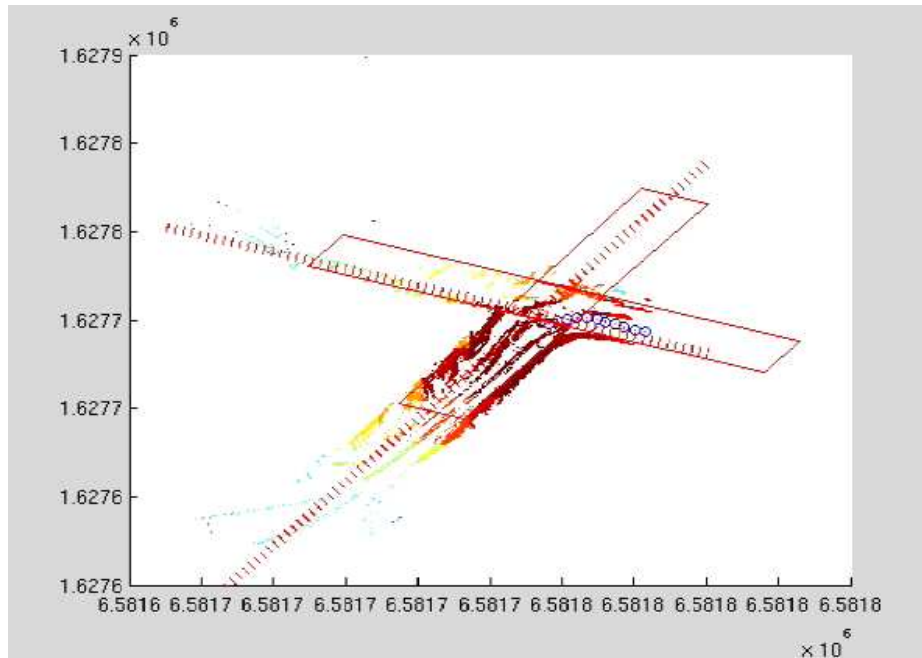


Figure 8: Fit junction topology

3 Work Details: Part 2 - Context-Free Machine Learning

A number of inner-urban junction scenarios within Dataset 8 were selected for ground-truth annotation on the basis of their provision of good-quality LIDAR data, and exemplification of the Regulating and Monitoring intentions of the ECOM control model, along with their conditional logic dependencies on environmental entities such as traffic lights and signs.

In the current experiment, 6 cross-road traversing scenarios are considered, consisting of 2 cases each of left-turning, right-turning, and 'straight-over' junction traverses. This set is a nearly maximally complex representation of the driving situation, in the sense that it is selected to effectively exhaust to all possible 'configuration-changing' driving scenarios: that is, all other driving situations (lane changes, roundabout-traverses, etc) can be considered degenerate instances of these cases.

Individual levels of the ECOM hierarchy, l , are considered to be mutually temporally exclusive; individual levels, however, may be simultaneously operative. (More generalized levels at the top of the hierarchy will tend to be operative over a longer period of time than sub-level intentions). We thus see that the intentional classification problem is one of

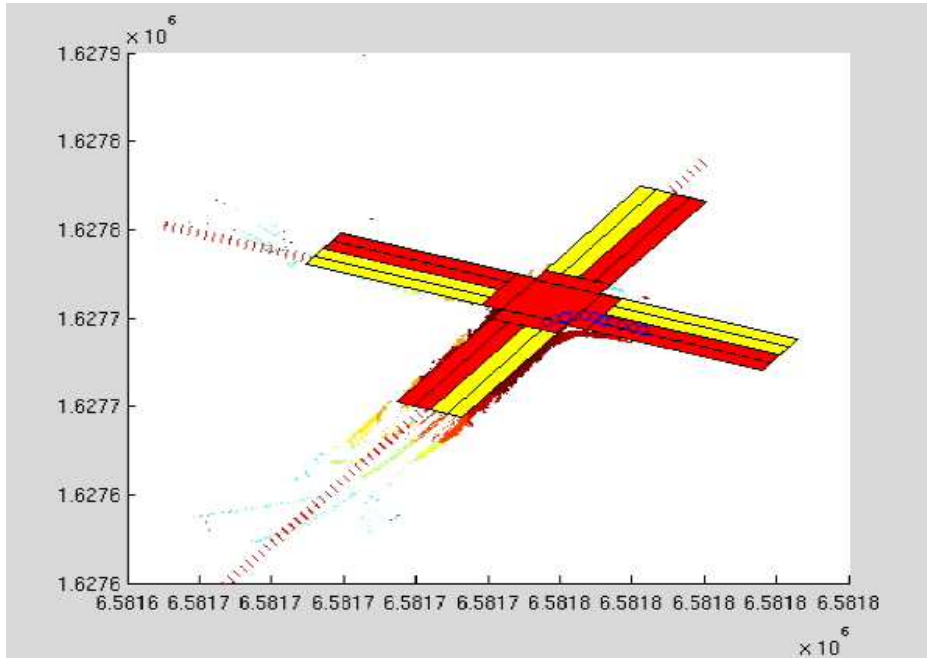


Figure 9: Fit pedestrian-crossing/lane structure

simultaneous categorization of the unique item i^l applicable within each level (with the inclusion of a null item if necessary). I.e., for the totality of individual frames we need to solve the mapping problem:

$$\forall l, X \rightarrow i^l : l = \underset{n}{\operatorname{argmax}} \{p(i^n|X)\} \quad (3)$$

(X being the feature vector).

Note that there is, in general, a strong downward dependency amongst levels ie:

$$P(i_{l_n}|i_{l_m}, X) \neq P(i_{l_n}|X)P(i_{l_m}|X) \text{ if } m > n \quad (4)$$

Building a robust and locally self-consistent inference system means being able to accommodate potentially inconsistent information at any arbitrary level of the hierarchy. Resolving inconsistency ideally means determining the intentional class attributions all the way down to the lowest level of the hierarchy (where, in general, consistency is likely to be weakest, given the abundance of perceptual attributions).

Ultimately this relationality will be determined via the first-order deductive process applied to the feature vectors; the initial acontextual approach treats individual levels as

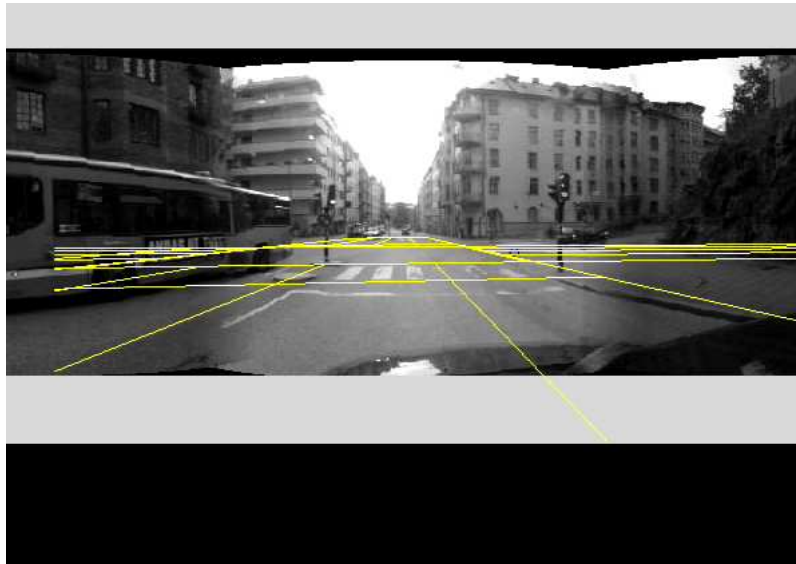


Figure 10: Project the junction topology into screen frame (1)

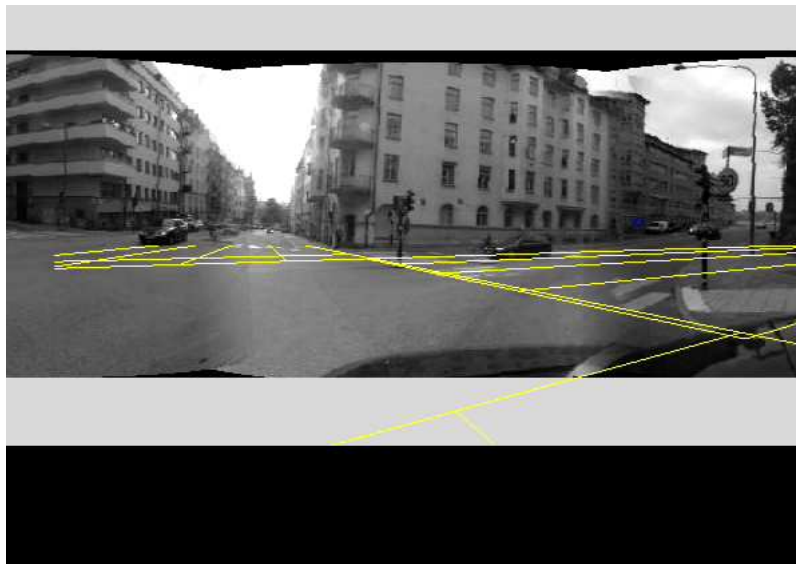


Figure 11: Project the junction topology into screen frame (2)

distinct classification problems in their own right.

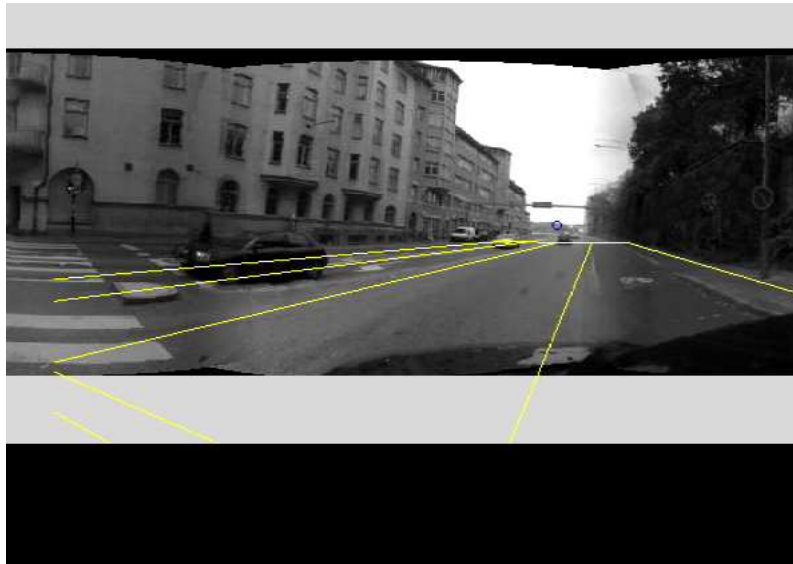


Figure 12: Project the junction topology into screen frame (3)

INITIAL IMPLEMENTATION OF HIERARCHICAL ECOM INTENTIONAL MODEL					
Task	Environmental Condition	Driver Perceptual Condition	Ordered Sub Tasks	Environmental Condition	Driver Perceptual Condition
Turn Left	At T junction OR at Xroad	Identified Junction	<ol style="list-style-type: none"> 1. Stop 2. Set Indicator 3a. Attain Higher Speed 3b. Turn 5. Stop 6. Attain higher speed 7. Get in lane 	Red light ahead If no in-lane direction signs Sub task 1. has taken place Car has not yet reached speed of any vehicle in front (Tracking sub-sub task) If traversed lane will not be clear during maneuver Sub task 5. has taken place AND traversed lane will be clear during maneuver If over threshold of left hand road	Spotted red light No In-lane signs spotted Intentionally stopped Has looked left Has seen car in traversed lane Intentionally stopped
Turn Right etc	At T junction OR at Xroad etc	Identified Junction	etc	etc	

Figure 13: Initial ECOM formulation

However, we seek to migrate relational descriptors down to the feature level as far as possible so as to avoid implementing a recursive deductive process at the logic level (which can lead to overlong computation times for an *in situ* system. We thus seek to provide a *maximally populated* hierarchical feature domain in which ECOM-like behaviors can

be more rapidly determined by standard pattern-recognition approaches. To this end, as a proxy for metric-temporal logic, we include an additional binary feature X'_t for each standard feature X_{t_c} activated at time t_c such that:

$$\forall t > t_c - 1 : X'_t = X_{t_c} \text{ if } ([\forall t_c < t : X_{t_c} = 0] \ \& \ [X_{t_c} > 0]) \text{ else } X'_t = 0 \quad (5)$$

We also generate a full hierarchy of binary features so that, for instance gaze attention directed at a particular lane also includes the road that contains it, and the junction that contains the road, etc (cf fig 14).

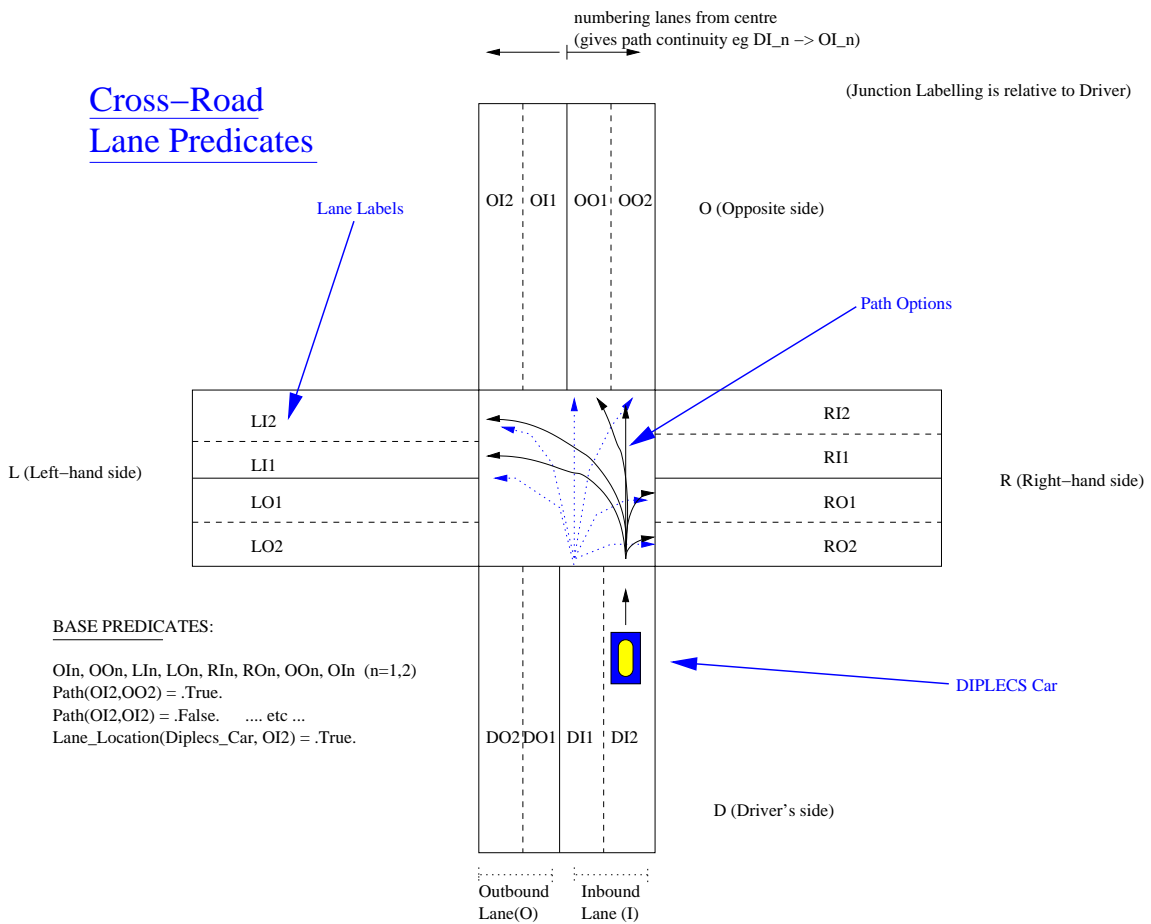


Figure 14: Junction composition

There are also more generalized hierarchical features such as 'driver-ward', 'leftward',

and so on, to allow for the possibility of more coarse-grained relations to be captured by the classification process.

Thus the learned ECOM classifiers can use the representation-action hierarchy as required when presented with attentional behavior located at arbitrary points within it (ie useful deduction need not proceed on the basis of having obtained a consistent cluster of maximally-specific predications about the environment); in this way, para-consistencies can be more readily accommodated. (Other work undertaken in support of the development of this notion of cognitive bootstrapping is detailed in [8] and [9]).

We hence obtain a logical feature vector of 772 descriptors for each frame of data (cf fig 15).

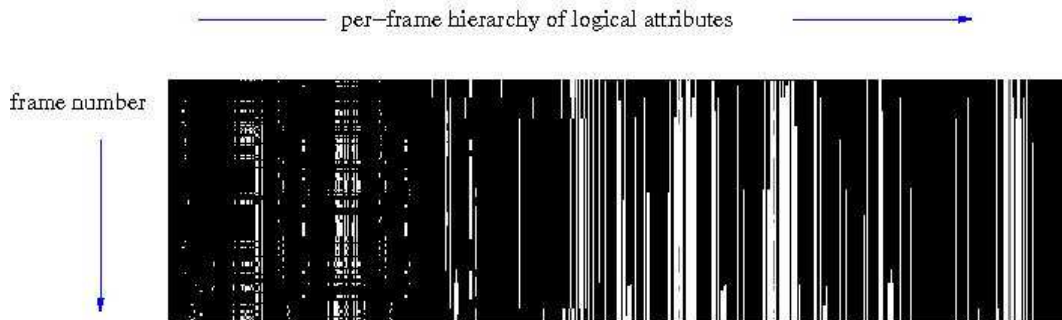


Figure 15: logical feature vector

We use a decision-tree algorithm for classification on the basis of its readily-interpretable results. In particular, it has the characteristic of defining a decision rule that is directly translatable into first-order logic.

We use leave-one-out cross-validation to test the acontextual classifiers. Trees are pruned by via iterative removal of non-leaf nodes and performance testing. This results in an average per-level accuracy of of between 9.1 and 30.16 percent on the ground truthed data (we omit the first level 'Traverse junction' intention as there are no negative examples

in the data). The decision tree algorithm shows good performance at the higher levels of the intentional hierarchy (cf Table 1). Full description of the intentional levels is given in the next section.

Table 1: Percentage misclassification rates for each scenario

Level	Straight on 1	Straight on 2	Left turn 1	Left turn 2	Right turn 1	Right turn 2
2	0	0	0	0	0	0
3	10.27 ± 0.072	17.07 ± 0.15	9.10 ± 0.064	7.16 ± 0.043	29.88 ± 0.166	7.48 ± 0.046
4	15.38 ± 0.063	17.07 ± 0.15	14.13 ± 0.067	11.29 ± 0.028	30.16 ± 0.168	10.56 ± 0.035
5	16.30 ± 0.066	19.89 ± 0.079	16.81 ± 0.067	11.29 ± 0.024	25.71 ± 0.077	12.40 ± 0.041

Decision trees for ECOM levels 2 to 5 are depicted in figs 16 to s 19.

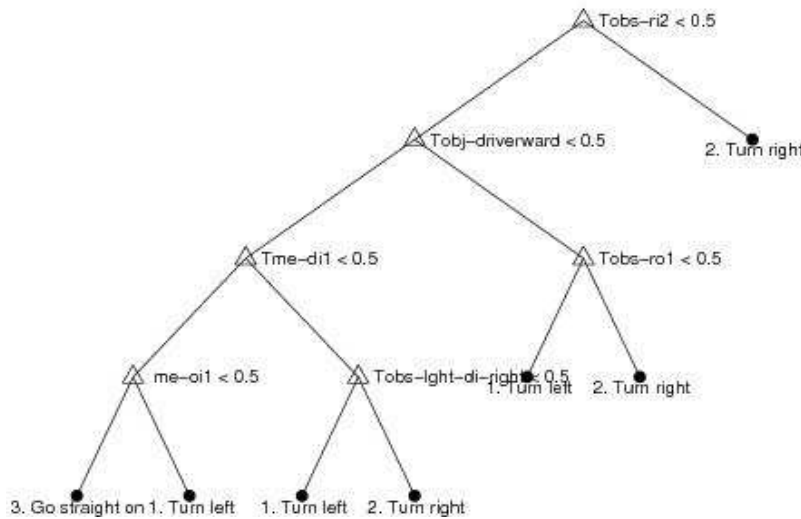


Figure 16: Level 2 decision tree

4 Work Details: Part 3 Logic Infrastructure Construction

The deductive logic system extends the above intentional detection system to accommodate rule-like behaviours relating to Highway-code based intentional configuration changes that cannot be fully captured by stochastic correlation.

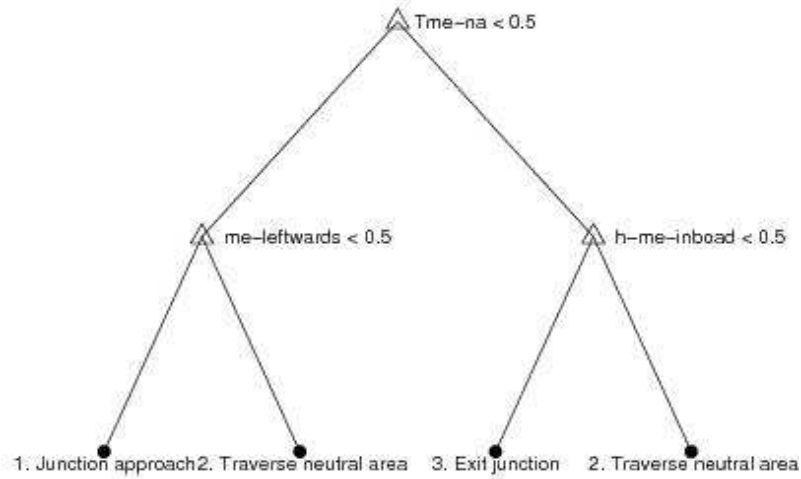


Figure 17: Level 3 decision tree

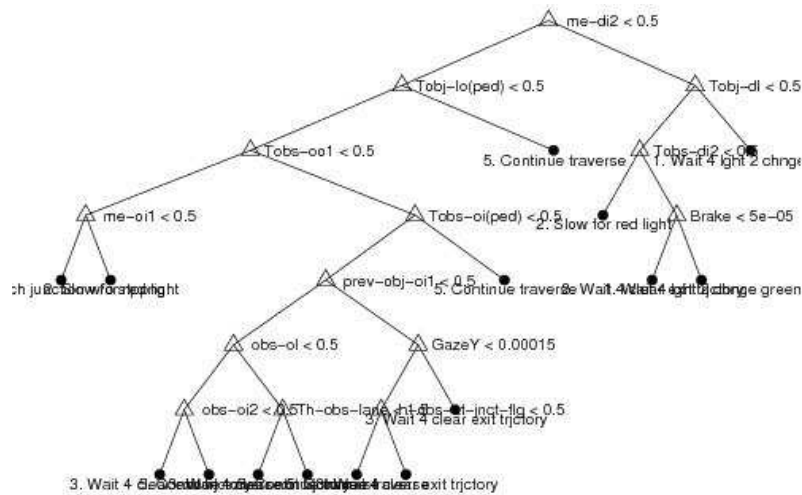


Figure 18: Level 4 decision tree

When placed in situ (the requirement of the next deliverable), the activity of the logical infrastructure within the strategy/supervision system will threefold; in the most typical mode of operation, the system will construct a logically consistent world-model from the computer-vision system’s input and use the conditional dependencies from the driver’s gaze, signal and control inputs to determine the operating intention and sub-intention of

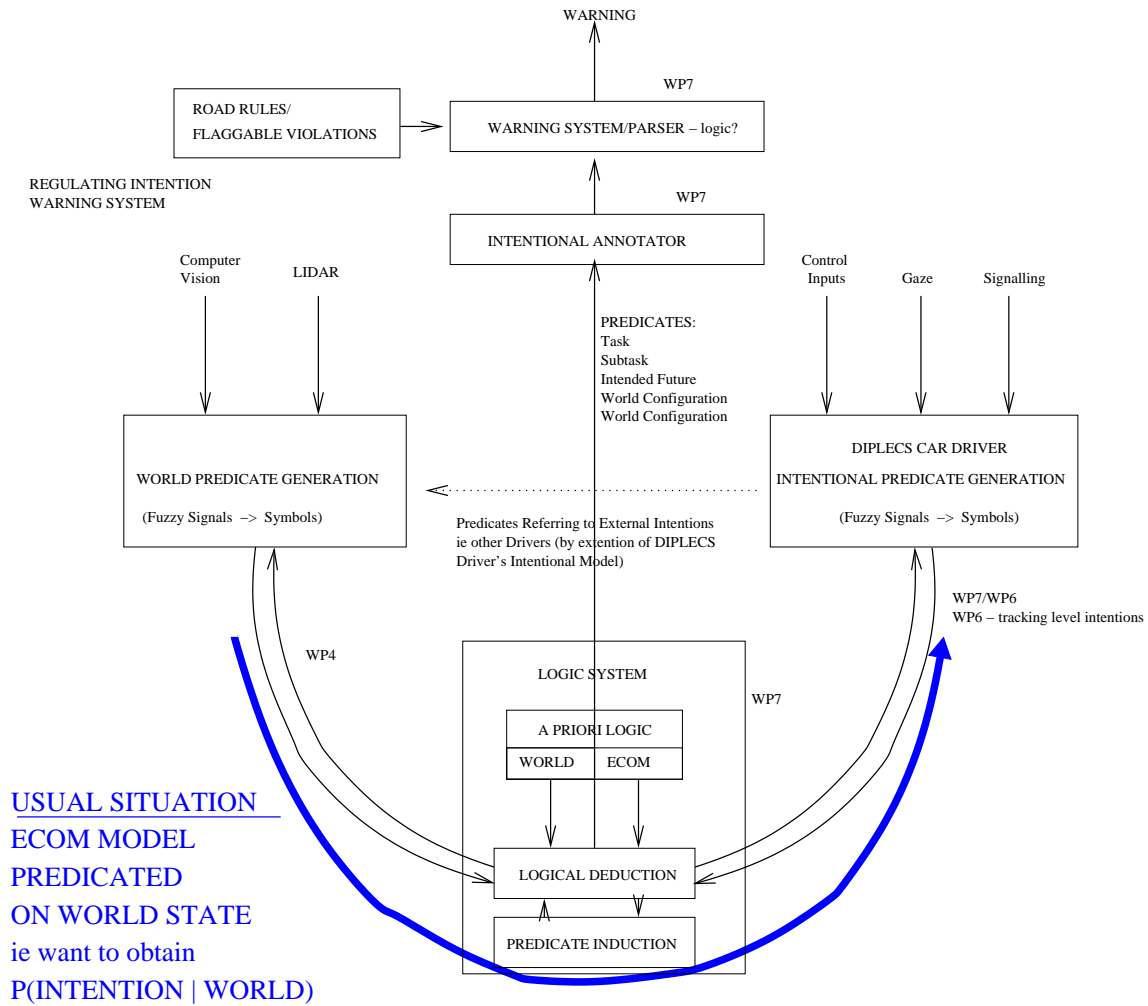


Figure 20: Usual situation: Intentional model predicated on World Model

Predicate/rule implementation may be subject to modification on the outcome of attempts to ascend the annotation/classification hierarchy (e.g. if it transpires that the visual detectors cannot differentiate individual cars from individual pedestrians, but can differentiate the logically-coarse category of 'road-users' from the logically-coarse category of 'non-road users', then the predicate-specification would need to reflect this. Ideally, such predicate respecification would be carried out automatically; however, this constitutes a very ambitious goal for WP7, and the final demonstrator will not be dependant on this).

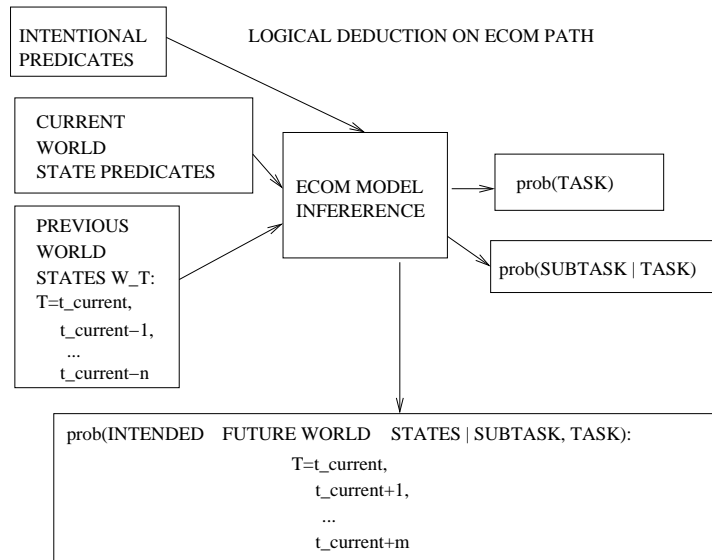


Figure 21: Dependencies in usual situation

For the present deliverable we focus on the issue of global logical consistency of predicates, and proof-of-concept demonstration of the potential for cognitive bootstrapping in terms of the re-weighting of feature detectors on the basis of this consistency.

We build the system initially in first-order predicate logic with *a priori* logical predication applied via a top-down approach, starting with the most general world predicates and clauses. Hierarchical relations at this world-description level are thus defined via the application of simple *a priori* rules wherever possible: for instance, a pathing clause linking all in-bound roads to all out-bound roads. The deductive system is hence coded in SWI PROLOG with a recursive predicate structure designed to maximally assist debugging and code-base expandability. This is briefly explained below for the purely physical (ie topological) aspects of junction logical description.

4.1 Details of Deductive Logical Highway Code Implementation

The junction predicates are defined hierarchically as follows:

1. We start with the most general predicate "JUNCTION", having four distinct sub-predicates deriving from their constituent road topologies defined relative to the driver for maximum generalisability (i.e. Cross-Road, Left-stemmed-T-junction, Right-stemmed-T-junction, Symmetric-T-junction). Straight roads are treated as junction connectives (irre-

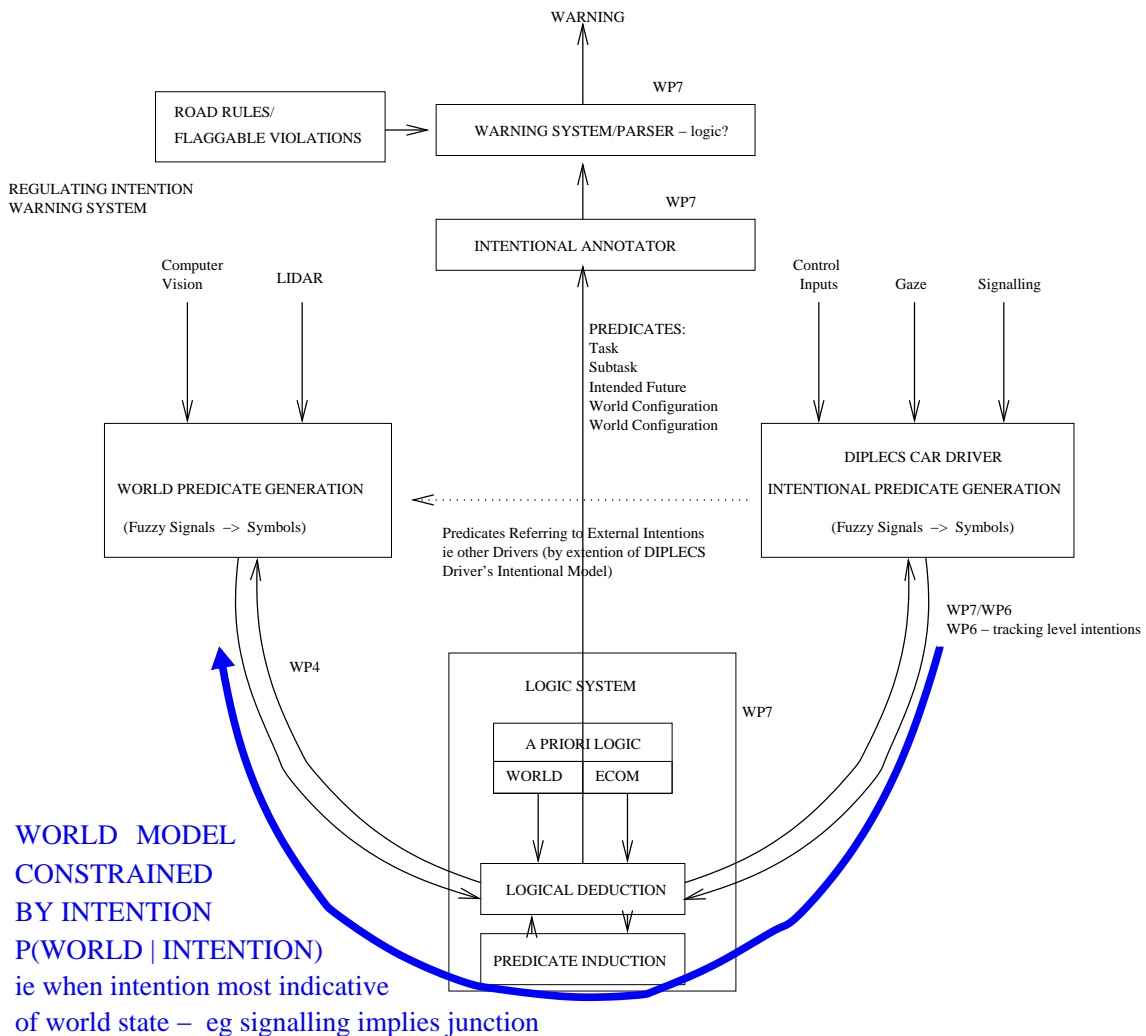


Figure 22: Reverse situation: World model predicated on Intentional Model

spective of length).

2. The predicate "ROAD" is of four different sub-types according to coarse-grained angular-orientation in relation to the driver (of the DIPLECS car, though with extensibility to other drivers).
3. Each road has one or two LANE-DIRECTION predicates drawn from the set: $\{In_bound, Out_bound\}$.

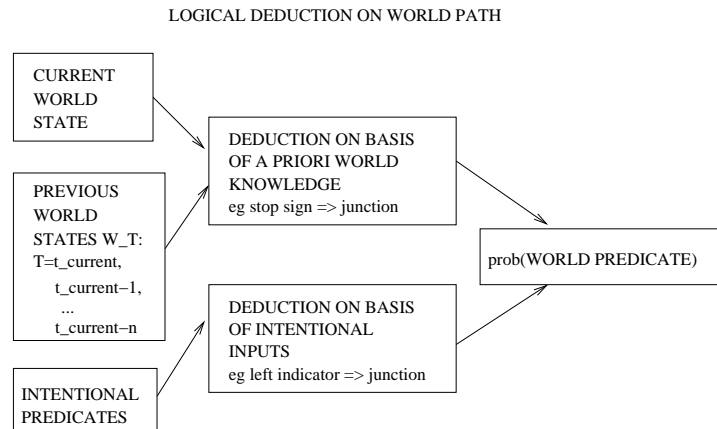


Figure 23: Dependencies in reverse situation

4. Each lane-direction has associated a particular number of in-bound/out-bound lanes, n . *Specific* lanes within the total set of inbound and outbound lanes are specified by a distinct indicator argument, m .
5. The three-argument predicate PATH (determining lane connectives) has two different subtypes, namely legal path and physical path. The latter covers the range of physical possibility, while the former defines the legal routes that may be followed by each car while crossing a junction or moving from one point to another. Lane numbering is defined so that pathing obeys the simplest possible general rules (eg $\forall m \text{ In_bound}(m) \rightarrow \text{Out_bound}(m)$)

Other modules implement the "SIGN and TRAFFIC-LIGHT" predicate-structures, as well as the *a priori* rules (clauses) covering their interrelation with the topology (e.g., the fact that a T-junction sign implies the confluence of two roads), to be followed by the *a priori* legal restriction clauses. *A priori* rules and predication relating to driver knowledge are implemented in the deductive system under a separate set of predicate headings dealing with ECOM dependencies (eg so that knowledge of signs both persists and relates to road conditions for the duration of the junction traverse, while allowing that knowledge of traffic-light states must be treated temporally).

A function 'MAIN' activates the logic system in which the different road types, numbers of lanes and lane positions of any detected cars at any given time are specified, along with any of these objects that are visually attended to, as well as all of the control inputs applied to the DIPLECS car. Also included is any primary classification of animating in-

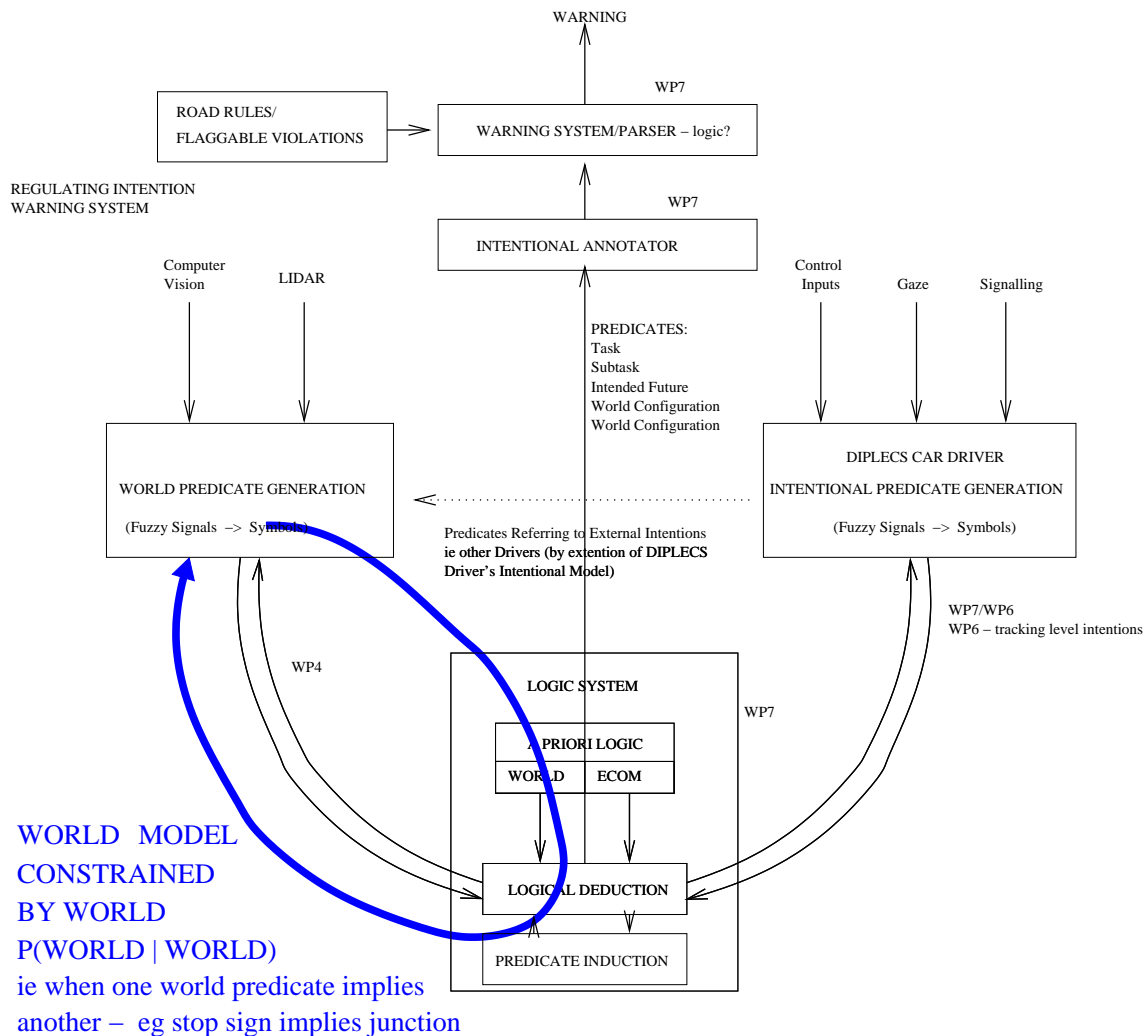


Figure 24: 3rd situation: part of World model predicated on rest of World model

tention. That is, the full extent of instantaneous symbolic detections are presented to the system, irrespective of their completeness or consistency. The PROLOG system then runs through a deductive process in order to generate the logically-legitimate configuration possibilities, as well as all possible legal or physical paths for overall temporal consistency; it also outputs the active ECOM intention and sub-intention (at the Monitoring/Regulating level).

Although the system in being constructed a top-down, forward deductive manner, it

remains possible to implement the reverse case; eg the system can deduce junction topology from car paths and vice versa. The animating idea is thus that, in the predominant operation of the deductive process, predicate definition is made as generic as possible in order to be able to easily ascend/descend through the hierarchy as required when presented with input predicate specifications located at arbitrary points within the representation-action hierarchy. Using flexible top-down/bottom-up hierarchical deduction thus enables the intentional model to inform the vision model and vice-versa.

4.1.1 Details of Logical ECOM Implementation

Corresponding to the above is the logically-formalised implementation of ECOM level structures. This is indicated as follows (with only major conditionality shown for simplicity):

level 1.1 navigate junction

level 2.1 turn left (*if lane(diplecscar) = dl(n) & prev(lane(diplecscar) = di(n))*)

level 2.2 turn right (*if lane(diplecscar) = dr(n) & prev(lane(diplecscar) = di(n))*)

level 2.3 go straight on (*if lane(diplecscar) = do(n) & prev(lane(diplecscar) = di(n))*)

level 3.1 junction approach (*if lane(diplecscar) = di(n)*)

level 3.2 traverse neutral area (*if lane(diplecscar) = na*)

level 3.3 exit junction (*if lane(diplecscar) = di & prev(lane(diplecscar)) = na*)

level 4.1 wait for light to change to green (*if 3.1 & stopped & red or red + amber light seen*)

level 4.2 slow for red light (*if red light seen*)

level 4.3 wait for clear exit trajectory (*if (3.1 & green & (stopped|slowing)) or (3.2)*)

level 4.4 approach junction without stopping (*3.1 & green light seen*)

level 4.5 continue traverse (*if was(4.3) & 3.2*)

level 5.1 monitor traffic while stopped at light (*if look(mov_obj) and 4.1*)

level 5.2 monitor lights (*if look(light) and (4.1 or 4.2)*)

level 5.3 monitor exit while stopped (*if 4.1 and ([look(lo) and 2.1]*

or [look(ro) and 2.2]

or [look(oo) and ;2.3]

or (look(mov_obj) &

(lane(mov_obj) = di or look(mov_obj) &

lane(mov_obj) = di(ped)))

level 5.4 follow car in front (*if (lane(mov_obj) = lane(diplecscar)*

and often_look(mov_obj))

and path(mov_obj) = path(dipleccar) and (3.1 or 3.2 or 3.3)
and not(stopped))
 level 5.5 monitor traffic while traversing (*if look(mov_obj) and 3.2 and not(stopped)*)
 level 5.6 monitor traffic while exiting (*if look(mov_obj) and 3.3*)
 level 5.7 monitor traffic while approaching (*if look(mov_obj) and 3.1 and not(stopped)*)
 level 5.8 check exit while moving (*if not(stopped) and*
([look(lo) and 2.1]or[look(ro) and 2.2]
or [look(oo) and 2.3] or (look(mov_obj) &
(lane(mov_obj) = di or look(mov_obj) &
lane(mov_obj) = di(ped))
& [not(mov_obj) & not(lane(mov_obj) = di(ped))])))
 level 5.9 check lights (*if look(light) and not(5.2)*)

where:

di = driver's outbound road

mov_obj = moving object

na =neutral area

4.2 Combining Logical Consistency Considerations with Decision Tree Outputs

Having thus obtained the intentional/forward-camera-view-annotated binary vectors along with per-frame decision tree outputs, the task is to use this contextually with the PROLOG code implementing the prior logical structures of the Highway code and ECOM models.

We thus wish to use PROLOG to determine the consistency of current intentions/world-states with the bulk of preceding intentions/world-states by using Vitterbi-like searching according to which the longest/best path through intentional nodes is selected, weighted by the decision-tree outputs. This involves explicitly leveraging the layer dependencies within the logic (the logical clause structure dictates that intra-level intentions are mutually exclusive while inter-level intentions have complex AND/OR relations, with a strong dependency of lower-level intentions on higher level intentions by virtue of the explicitly subsumptive implementation of the ECOM model).

In principle, this should allow the higher levels of the logical hierarchy to be very robust to detection error - for instance, junction detection, being at the higher-end of both the visual and logical hierarchy is underpinned by the multiplicity of observations (lanes, roads etc) at the lower levels of the hierarchy. Consequently, while it is possible for a missed car signal to generate a false intentional reading at the lower levels dealing with

(for example) intra-junction traffic etiquette, a missed reading of the junction detector for one frame will not generate a corresponding error the higher levels of the ECOM model dealing with the broader subsumptive intention of junction traversal. In this way logical consistency can be seen as a level-based construct, with, in general, higher levels of the representational hierarchy exhibiting greater likelihood of mutual consistency.

To implement this temporal contextualisation, predicates from the current frame (ie the predicates associated with individual binary vector headings) are thus asserted as being only *currently* true. Hence, all active predication is considered true and all inactive predication is asserted as false, with previous frame data asserted as *previously* true (such that all negative predication is specified as previously false). This gives the four valued-temporal logic: (*Previously_True, Previously_False, True, False*).

In parallel with this, intentional predication for each level is determined acontextually for each frame by the decision trees. There are thus two distinct modes of pattern recognition employed in the composite system; the discriminative (ie the decision trees) and the generative (the *a priori* logical model). In effect, the logical model considers only spatial-temporal context and makes no allowance for stochastic variation, whereas the decision trees disregard context, and consider all predication irrelevant to classification as stochastic variation. Exploiting the difference between the two will allow us to later construct error-resilient contextual classifier of intention.

To take advantage of temporal context, a frame-buffer is triggered at the first detected instance of junction approach. The caching of previous states within this expanding temporal window means there is the potential of the recent past to modify predication concerning the less-recent past since we assert *all* prior consistent temporal states within the window, allowing preexisting ambivalent states to be resolved (ie the past is progressively determined by the future). Logical causation is thus not the same as temporal causation, being able to work in both directions.

However, there will, as indicated, always be certain error-containing frames (annotation error in the case of the ground-truth data) that are incapable of being resolved within any continuous driving narrative. These are cases in which no possible allocation of truth values to predicates with ambivalent scope allows the concretely asserted predication to be globally true. These can be straightforwardly rejected within PROLOG to leave a 'consistency set' of temporal-narrative exemplars against which future frame predication can be tested. Those frames that pass the test are iteratively added to the consistency set to extend the temporal narrative.

This means that the logic system is, in itself, capable of acting as a frame intention classifier: each of the intentions i^l on each of the levels, l , can be asserted, in turn, alongside the negation of all of the other intentions on the same level to test for consistency. ie we test the assertions:

$$\forall l, i \begin{cases} i^{l'} \rightarrow True : l = l' \\ i^{l'} \rightarrow False : l \neq l' \end{cases}$$

All of the consistent intention classes can then be given an equal weighting such that they sum to unity to create a pseudo-probability, and give a stochastic measure of the likelihood of the correctly predicted intentional output. Clearly accuracy would be expected to increase with the size of the consistency set.

This is indeed the case for the junction traverses, as indicated by the purely logic-based accuracy results for the six scenarios set out in figures 25 to 30. Here, the logical default 'straight on' assumption becomes falsified in the 'left turn' and 'right turn' scenarios as more temporal context is accrued. In particular, note how the accuracy figures increase with time for the latter four scenarios in comparison to the static accuracy values for the two former scenarios.

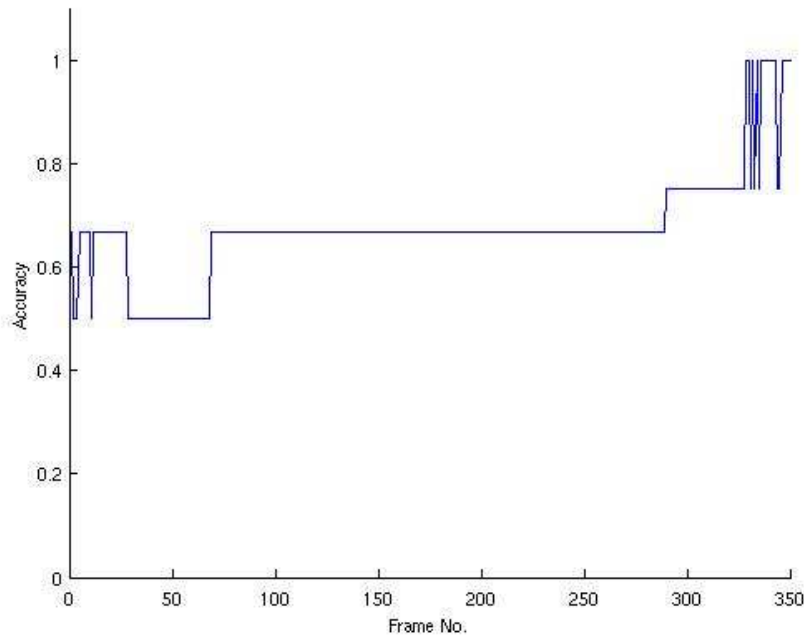


Figure 25: 1st Left-turn Scenario

Given that these results are purely logic based, it is thus possible to make relatively accurate prediction purely on the basis of prior knowledge of the conditionalities of the ECOM model and the constrained (highway code-based) nature of traffic scenarios.

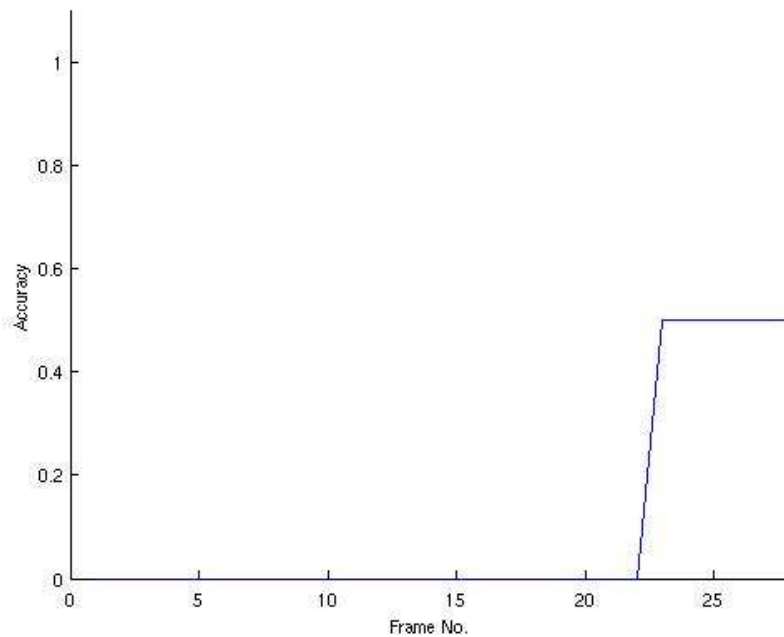


Figure 26: 2nd Left-turn Scenario

When per-frame decision tree accuracies² are superimposed directly on the above graphs, it is apparent the lack of temporal context affects per-frame accuracy (cf fig 31) for the first right-hand scenario). In particular, note the relative consistency of the decision tree accuracy over time.

However, it is also apparent that the decision tree output is as approximately accurate as the *a priori* logic model at the temporal outset. Since the errors made by the logic and decision-tree systems would appear to be uncorrelated, it ought thus to be possible to combine them to our advantage.

Decision-tree outputs are hence combined with logical deduction through their incorporation into the consistency testing and aggregation procedure. Under purely logic-based consistency testing it is, for instance, possible that error-containing sets are consistent with each other by chance, leading to a false aggregation of consistent frames, and irrecoverable system error. We hence use the decision tree outputs to set thresholds on acceptance for the consistency set by requiring the agreement of the decision-tree outputs with that

²Unlike the cross-validation tests, these decision trees are trained only on the initial three training sets for each scenario, so as to generate meaningful comparisons across the range of junction traverses.

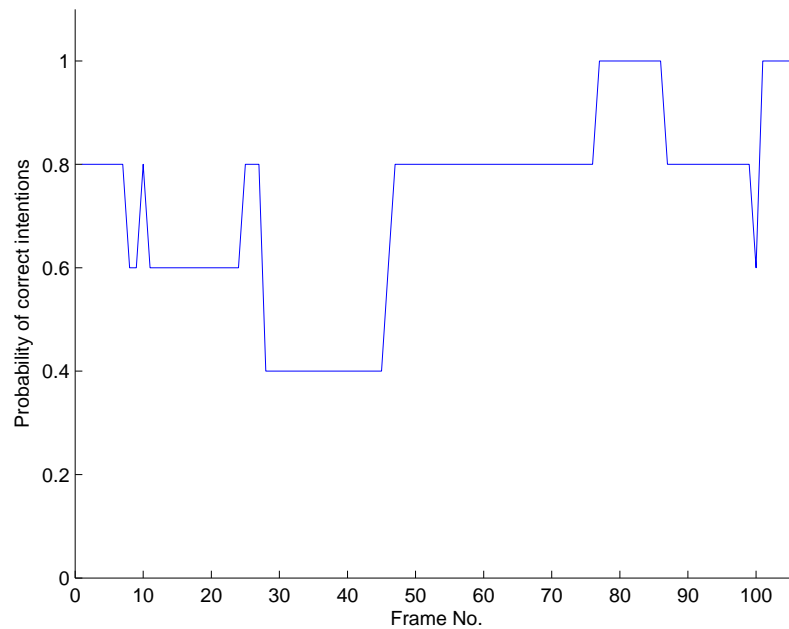


Figure 27: 1st Right-turn Scenario

of the logic in order for the frame to be added to the consistency set. Otherwise the logic output is given as output by default, but not added to the consistency set. When a fully-inconsistent frame is detected by the logic system, the output is switched instead to that of the decision tree.

With this decision-tree-based temporal-node-weighting, the accuracy of the composite system is very significantly greater than that of the individual systems, as depicted in fig 32 for the right-hand turn data set.

We have thus created an effective prototype system for composition of discriminative and generative classifiers of intentional behavior, utilising stochastic and structural techniques to combine global and local information.

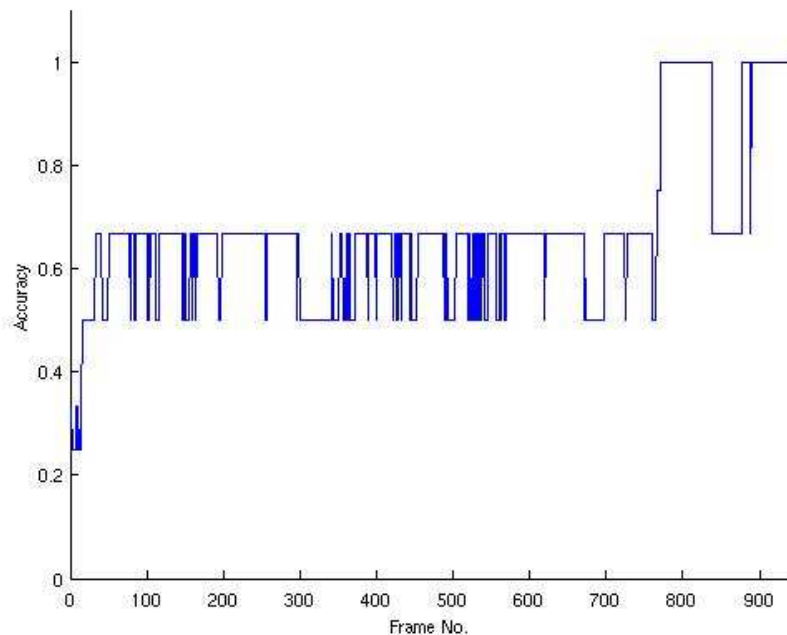


Figure 28: 2nd Right-turn Scenario

5 Work Details: Part 4. The Top-Down Feature Respecification Module

So far we have considered only the bottom-up aspects of the system (the major mode of operation), in which the world and intentional model is built-up from the detector inputs. However, the global consistency checks also have the potential to modify detector inputs in a top-down fashion, and thereby complete a full 'bootstrap' modelling of the world (which is to say, there might, in principle, exist the possibility of running the intentional induction system fully unsupervised). As an initial proof-of-concept of this idea, we wish to test the notion that the global consistency check is capable of providing useful top-down constraints on the feature-detectors under simulated failure conditions.

That is, we simulate the effects of noise on one of the features by randomly replacing binary values for this feature within the frame vector by arbitrary binary digits. We do this according to a uniform random distribution with an average of failure probability of one in every five frames.

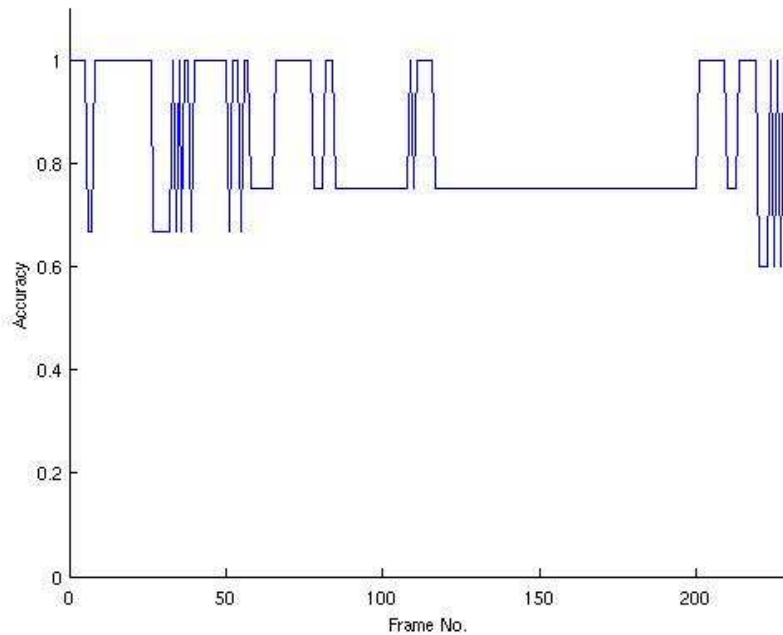


Figure 29: 1st Straight-over Scenario

Our aim is then to determine which of the 772 possible features is subject to this additive noise purely on the basis of global logical consistency.

We attempt this by sequentially removing feature predicates (ie detector outputs) and recalculating the associated predication of each frame for all of the 772 features and determining whether this improves overall global consistency as measured by the final size of the consistency set. We can then 'weight' feature detectors on the basis of this consistency (for the ground truth vector this a binary weighting; in practice, it is likely to require a fuzzy threshold). This will ultimately form the basis of the logic system's feedback to WP4.

The results of the application of this method are shown in fig 33 for a section of the features. We see that a clear peak in the magnitude of the consistency set exists for the error-compromised feature. It is also apparent that two other features exhibit this peak; this is because of the conditional dependencies that exist between features (such features tend to be in close proximity within the feature-vector by virtue of their similar characteristics; however note the feature vector itself is order-independent).

Critical to our argument is that this measure of global consistency correlates with the

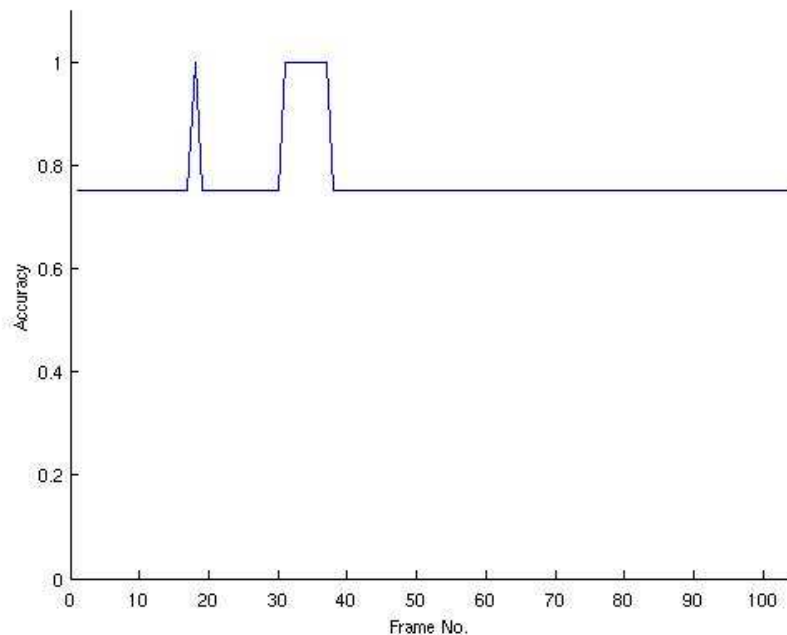


Figure 30: 2nd Straight-over Scenario

accurate prediction of the (unknown to the system) ground truth values. We see from fig 34 (giving the average accuracy over the whole time-sequence) that this correlation is remarkable.

Blanking the features associated with the peak in global consistency therefore acts to increase overall system accuracy in the manner intended; the system thus exhibits the capability of top-down and bottom-up bootstrap-induction of intentional and world models.

6 Summary

6.1 Summary of Activities

In summary, we have carried the following activities in the completion of deliverable D7.1:

1. Constructed a per-frame junction topology propagator to obtain ground-truth data for decision-tree-based machine learning, and carried out remaining annotation manually.

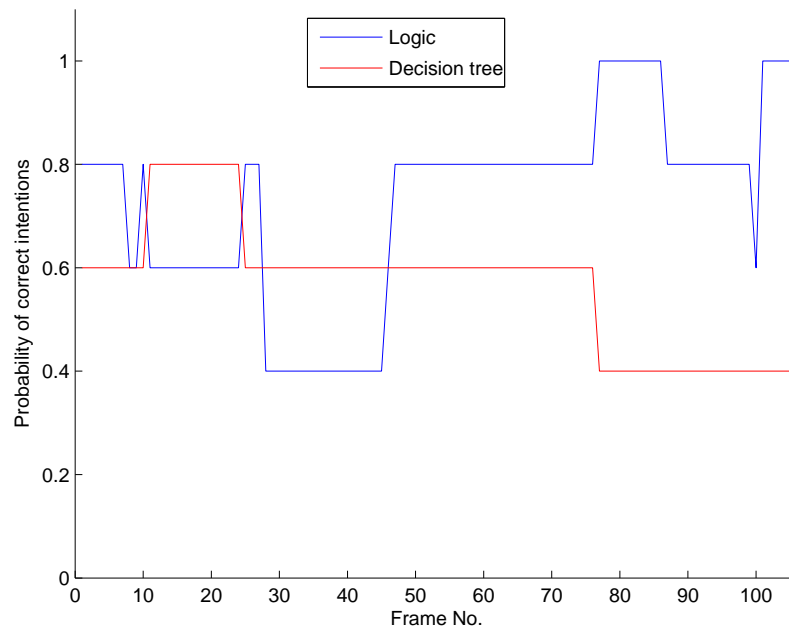


Figure 31: Comparison of *a priori* logic and decision-tree accuracy

2. Constructed an intentional deduction system based on *a priori* ECOM and Highway Code logic models.
3. Built a prototype system for determining intentions via global logical consistency, building on the above two activities.
4. Demonstrated the application of this system for top-down re-weighting of feature detectors on the basis of global logical consistency.

In doing so, we have demonstrated the possibility of applying bootstrap techniques within the strategy and supervision system, as well as providing a mechanism for classifying driver intention.

6.2 Future Activities

Deliverable D7.2 will require the prototype system to be applied in situ, in situations in which there are relatively few detected features (certainly less than the full ground-truth

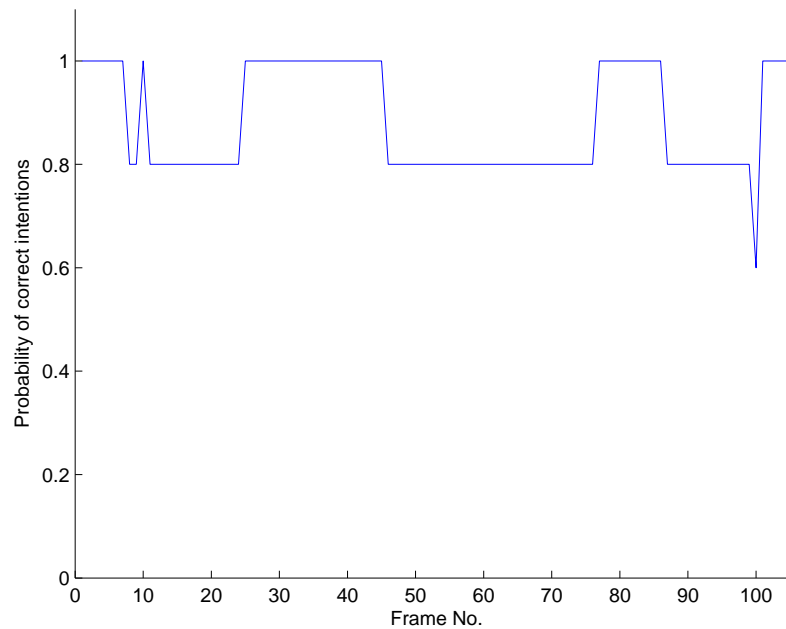


Figure 32: Combination of decision-trees with logical consistency constraints

set utilised here), and in which detector noise is likely to be apparent to differing degrees across the whole detector range. To do this, the system will have to relax the strong binary assumptions implicit in PROLOG and employ intermediate logic certainties. We plan to approach this initially via fuzzy logic applied to (at least) the detector inputs, such that multiple paraconsistent driving narratives can be simultaneously accommodated (perhaps employing an approach such as that of [2]).

Further refinements of the ECOM logical model will also be required; this will involve the ongoing collaboration axis between CVSSP and ARMINES.

References

- [1] R. A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 14(23), April 1986.
- [2] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):622–633, 2004.

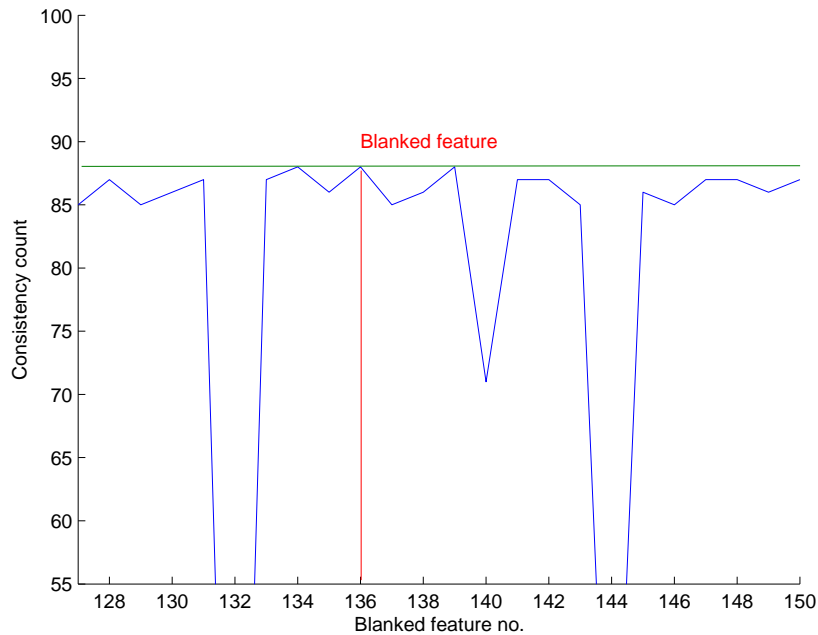


Figure 33: Size of consistency set verses feature number

- [3] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
- [4] A Hollingworth, C Williams, and J Henderson. To see and remember. *Psychon. Bull. & Rev.*, 8(4):761–768, 2001.
- [5] R. A Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.
- [6] R. A Rensink. Change detection. *Annual Review Psychology*, 53:245–77, 2002.
- [7] A. Sloman. Key ideas of semantic models, implicit definitions and symbol tethering, 2007. Retr. 1/12/7 from: www.cs.bham.ac.uk/research/projects/cogaff/talks/grounding.slides.ps.
- [8] D. Windridge and J. Kittler. A model for empirical validation in self-updating cognitive representation. In *Proceedings of Brain Inspired Systems 2008 (BICS 2008)*, Sao Luis, Brazil, 2008.

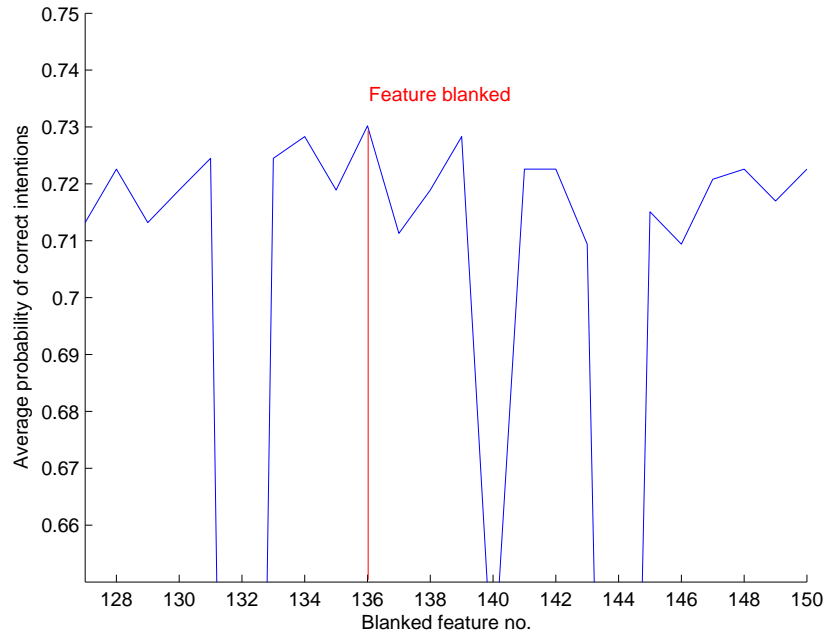


Figure 34: Accuracy verses feature number

- [9] D. Windridge, M. Shevchenko, and J. Kittler. An entropy-based approach to the hierarchical acquisition of perception- action capabilities. In *Proc. of 4th International Cognitive Vision Workshop ICVW 2008, (6th International Conference on Computer Vision Systems ICVS 2008)*, Santorini, Greece, 2008.