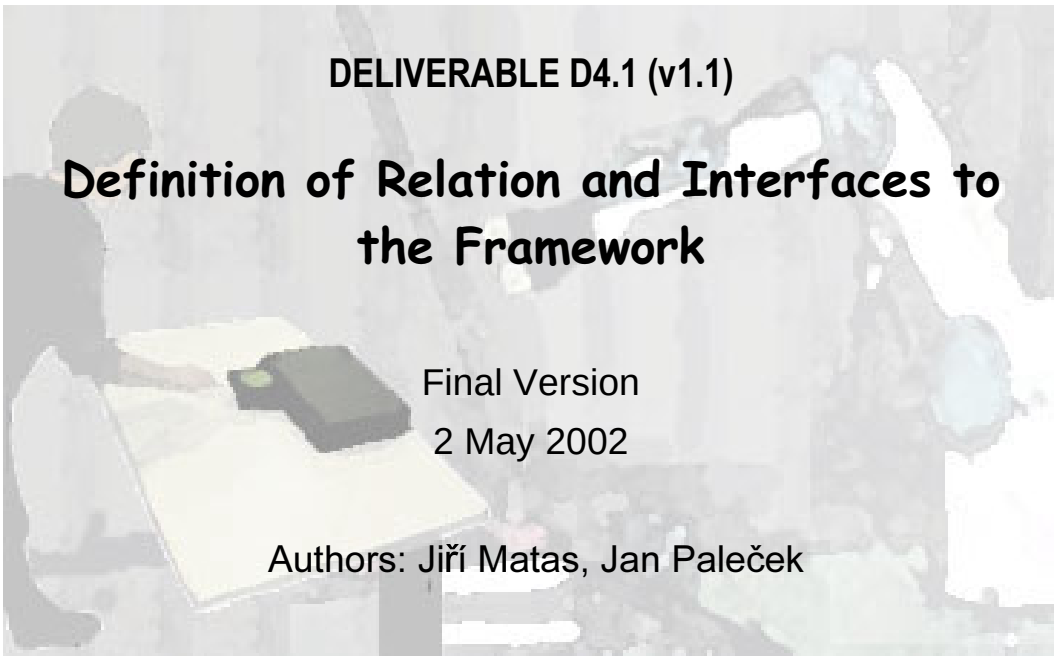

DELIVERABLE D4.1 (v1.1)

Definition of Relation and Interfaces to the Framework

Final Version

2 May 2002

Authors: Jiří Matas, Jan Paleček

Project acronym: ACTIPRET
Project full title: Interpreting and Understanding Activities of Expert Operators for Teaching and Education

Action Line IV.2.1: Real Time Distributed Systems (Cognitive Vision)
Contract Number: IST-2001-32184

Contents

1 Introduction

This deliverable document provides initial specification of the object recognition module and its relationship to other components of the *ActIPret Demonstrator* (AD). It summarises the work carried out thus far under Task 4.1 (Conception including interface definitions).

Object recognition, initially identification and later in the project categorisation and detection are of major importance for the cognitive vision process. Recognised object and object categories form the basic 'words', whose spatial and temporal relationships define actions.

The recognition method we propose to use in the *ActIPret system* is built on top of the novel concept of distinguished regions [1]. Initial experiments in wide-baseline stereo [2], [3], in object recognition on standard datasets [4] and in image retrieval [5] show performance superior to the state of the art. Moreover, the method has some attractive features making it suitable for a cognitive vision system.

We do not intend to provide an in-depth description of the approach here; the interested reader is referred to the references listed below. A brief outline of the method using local affine frames on distinguished regions is given in Section 3. However, it is worth mentioning two properties of the method important in the cognitive vision system:

1. all the representation are *learned* from (possibly multiple) examples. This very important in a system where flexibility and reuse in different, a priori unknown applications and environment is expected. The issue of learning is interest in a cognitive system in its own right.
2. the method is *fast enough to support* processing of selected frames from a *video* sequence. It is not real-time. However, response time in tens of seconds can be expected without heavy optimisation. In this deliverable, we do not provide details of the method and the interested reader is referred to the cited documents.

The main objective is:

1. to specify interfaces to other ActIPret modules
2. to provide basic insight about our method
3. to show very preliminary results on data related to the intended ActIPret scenario

The rest of the deliverable is structured as follows. First, in Section 2, relationship to other modules is discussed. So far unresolved issues are listed in Section 3. This part of the deliverable is intended as our contribution to the iterative improvement of the specification and design of the ActIPret framework. Preliminary results on data from an ActIPret CD player scenario are presented in Section 4. The deliverable is concluded in Section 5. The IDL-description is given together with all other IDL description of components in Deliverable D1.1, Appendix B.

2 Relationship to framework

The object recogniser is a module with well-defined interface and functionality. As a first approximation, its interface to other modules in the frameworks can be described in very simple terms. On the one hand, the module processes image provided by the Image Server. On the output, information about presence of objects known to the system in the scene is passed to the Reasoning Engine and the Object Relation Generator.

2.1 Relationship to other work packages

Two work packages are related to this task, WP 3 and WP 6:

- **WP 3:** spatial relationship generator is considered related to this WP, because relationships can control processing of object recognition. For example, the aggregate relationship defines one or more SOIs (Spaces Of Interest) where object recognition is supposed to operate. A similar argument holds for the detection components.
- **WP 6:** attentive control, makes use of the information about detected objects. However, due to the component-based set-up of the framework, the individual vision processing component (e.g. object detection and tracking, object recognition, hand tracker) will directly impose these view requests.

The rationale behind this set-up is that it is only the 2-D vision components that know about which views are required to solve their tasks. To transfer this knowledge to other components would be dangerous, since the same functionality would be located in two components. On the other hand, to inform a 2-D vision processing component that there is another object that needs to be attended when calculating the view request seems a natural extension to the capability of a vision processing component. Every 2-D vision processing component needs to request views in order to take into account potentially occluding objects.

All relationships of the WP4 module are represented in Figure 1 by arrows, with the exception of the view controller link, marked by the hashed square. The interface service provided by the recognition module have been specified in detail as IDL in Deliverable 1.1.

The basic functionalities are:

Phase	Functionalities
Learning	processed outside ActIPret system
Expert	Camera/View/Pose Services and Reasoning Engine
Tutor	Camera/View/Pose Services and Reasoning Engine

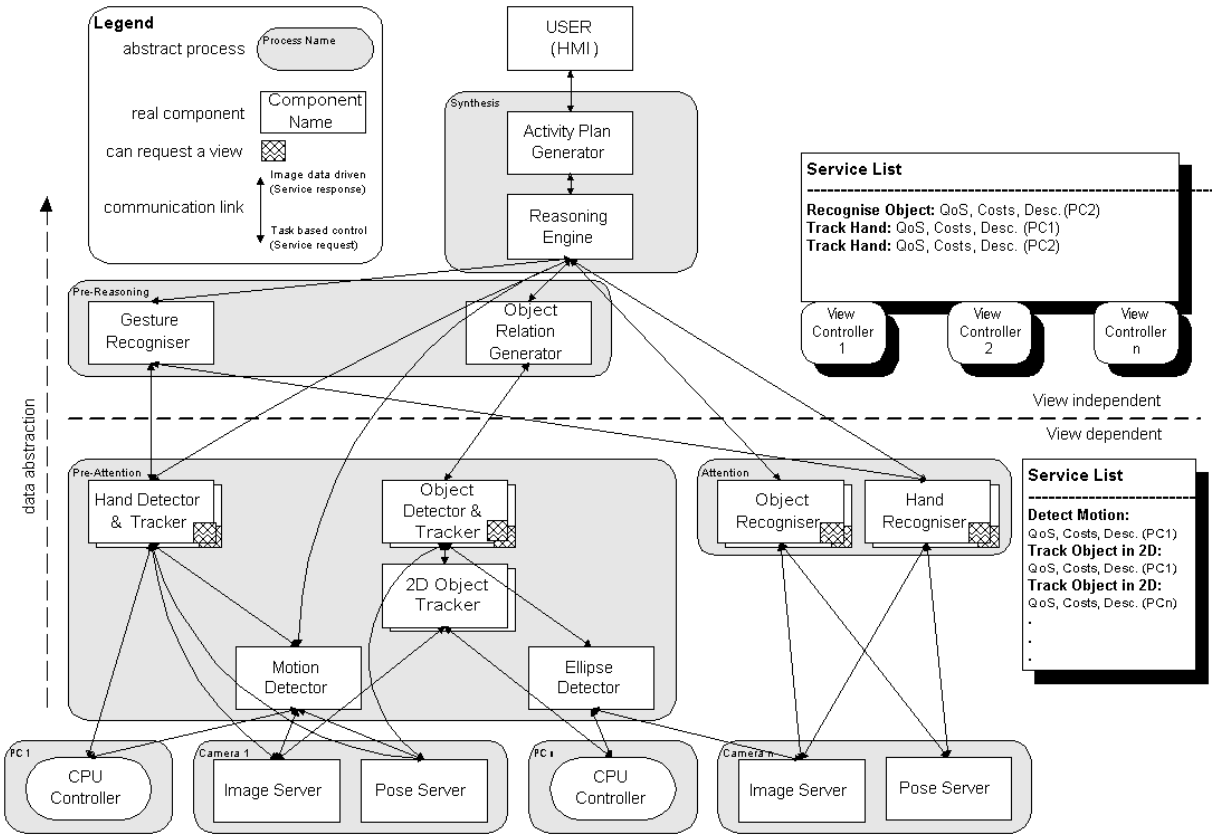

Figure 1: Example set-up.

3 Internal functionality

In the recognition method proposed by CMP, objects are represented as collection of so-called distinguished regions. On each distinguished region, a local affine-invariant reference frame is constructed. Objects are detected by correlation of normalised colour functions inside the affine frames.

The concept of a distinguished region (DR) has been introduced and formally defined in [1]. In the paper we show that distinguished regions are very good candidates for matching. Two new types of distinguished regions, the *Separated Elementary Cycles* of the Edge Graph (SECs) and the *Maximally Stable Extremal Regions* (MSERs) have been proposed. For both types, an efficient (near linear complexity) and practically fast detection algorithm is available.

Experimentally the usefulness of the proposed *DRs* has been shown on complex wide-baseline stereo problems [2][1], image retrieval and most importantly, object recognition [4].

An example of detection and localisation using the local affine frames on distinguished regions is shown on two image of a scene that was chosen as a first approximation to the ActIPret Scenario. The images depict 6 different CDs. In the first image (Figure 2), the CD are shown at a smaller scale and without occlusions. The second image (Figure 3) contains occlusion. Using default settings only one type of distinguished regions, five out of six CD are successfully recognised and localised. Only the featureless and specular CD, lying at the top

left in the first image is not recognised. Note that neither colour information nor segmentation was used

In a second experiment (Figure 4 and Figure 5), the same approach was used to detect skin of a human performing the 'CD insertion' task. The distinguished regions were the Maximally Stable Extremal Regions (MSERs), but this time not operating on raw intensities. A probability map of skin colour presence was used instead.

Open issues:

1. So far, we have been developing a system enabling rough localisation in the 2D image of the objects of interest. In order to establish spatial relationships, 3D information is needed. Interface to these services has been defined in Deliverable 1.1, but so far the implementation remains an open problem. Possible options include calibrated stereo on regions corresponding to detected objects or structure from motion.
2. A second group of issues relates to the view control. Although clearly the object recognition may (and will) request a different view of the scene to facilitate recognition, the maintenance of the world model and the services of the view controller have not been specified in full yet.
3. A third group of topics requiring detailed investigation is the cooperation between the recognition module and the Object Detector and Tracker. The tracker will be initialised often by location and representation provided by the recognition module. Interface to these services are yet to be defined.

Figure 2: Detection and localisation of CDs on the table.

Figure 3: Detection and localisation CDs with occlusion.

Figure 4: Detection skin.

Figure 5: Detail on DVD player.

4 Conclusion

In this first stage of the project we specified interfaces to other ActIPret modules and provided basic insight for partners in the consortium about our method for object recognition (SEC, MSER).

We mentioned difficulty of transition between 2D and 3D recognition. We defined basic approaches of cooperation with other services.

We showed preliminary results on real experimental images related to the intended ActIPret CD player scenario (detection and localisation CDs; detection human skin).

5 References

[1] Matas, J.; Chum, O.; Urban, M.; Pajdla, T. - *Distinguished Regions for Wide-baseline Stereo* (CMP-TR-2001-33)

[2] Matas, J.; Chum, O.; Urban, M.; Pajdla, T. - *Distinguished Regions for Wide-baseline Stereo,* Submitted to British conference on machine vision, Cardiff, 2002

[3] Matas, J; Odrzalek, S. – *Local affine frames for wide-baseline stereo*, Accepted to International Conference on Pattern Recognition, Quebec, Canada

[4] Obdrzalek, S, Matas, J. – *A novel framework object recognition*, Submitted to British conference on machine vision, Cardiff, 2002

[5] Obdrzalek, S Matas, J. – *Local affine frames for image retrieval*, Accepted to the Challenge of Image Retrieval, London, UK, 2002