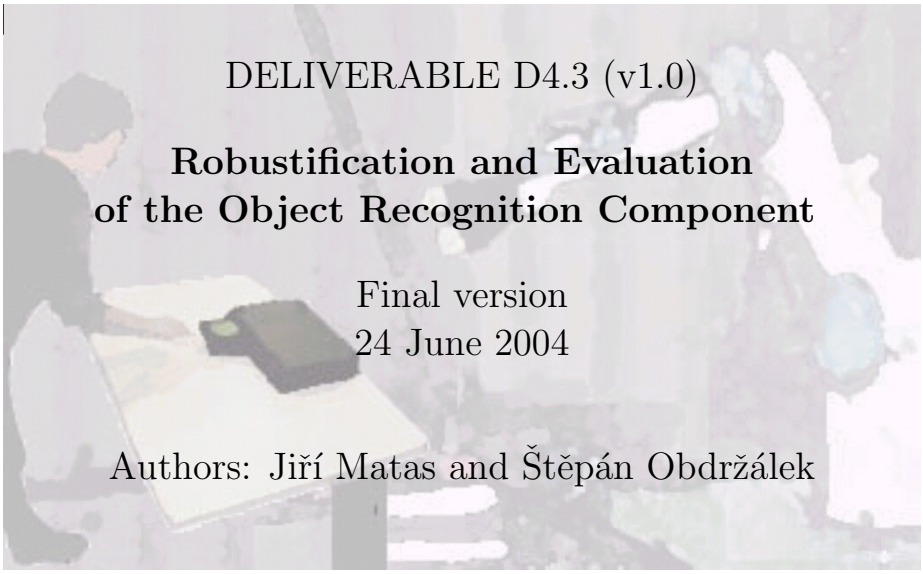


DELIVERABLE D4.3 (v1.0)

Robustification and Evaluation of the Object Recognition Component

Final version
24 June 2004

Authors: Jiří Matas and Štěpán Obdržálek

Project acronym:	ACTIPRET
Project full title:	Interpreting and Understanding Activities of Expert Operators for Teaching and Education
Action Line IV.2.1:	Real Time Distributes Systems (Cognitive Vision)
Contract Number:	IST-2001-32184

1 Introduction

In the Actipret framework, the Object Recognition Component reports presence and location of object of interest in the scene. Besides standard performance measures of a recognition system – the recognition rate and the false positive rates – the process of activity interpretation requires the object recogniser to operate in "near-real" time. Precision of localisation is also important, since gestures and activity interpretation depends on co-locations of objects.

The requirements for object recognition are contradictory. To achieve maximum recognition rate and highly precise localisation, complex recognition strategies have to be employed. The flexibility of the Actipret framework prevents us to rely on fast, object-specific, hand tailored approaches. On the other hand, the near-real time performance requirement limits the generality and complexity of the recognition method. A compromise must be sought. We propose to use an *adaptive* recognition strategy. We first run the recognition system in a mode that maximises the recognition rate. After analysing the processes that lead to correct recognition, only the smallest subset of recognition processes guaranteeing an acceptable recognition rate is selected for further operation.

Two improvements towards robustification of the recognition approach are presented. First, the distinguished region detection is generalised. Besides intensity MSER, extremal regions with other ordering of RGB values are used, yielding lower false negative rate. The second improvement is in decision making: the background model reflecting spatial dependencies (a development towards a fully Markov model) lowered the false positive rate.

Through adaptation, the fastest set up of the recognition system achieving desired performance v. speed trade-off is found. The adaptation process is posed as constrained optimization.

The text is structured as follows: we first briefly review the object recognition component. Next, we describe the newly developed low level detector of variants of Maximally Stable Extremal Regions (MSERs) that enhance recognition performance (at some non-negligible computational expense). The focus then switches to the the adaptation scheme for selecting subsets of the variants of the MSERs. The robustness of the components is further enhanced by careful modelling of object boundaries (Section 5). In section 6, the object recognition component performance is evaluated on sequence acquired at two different locations (Profactor, CMP). The reported is concluded in Section 7.

2 Structure of the Object Recognition Component

.

The structure of the recognition algorithm is visualised in to Figure 1. The MSER-LAF (Locally Affine Frames on Maximally Stable Extremal Regions) method proceeds as follows:

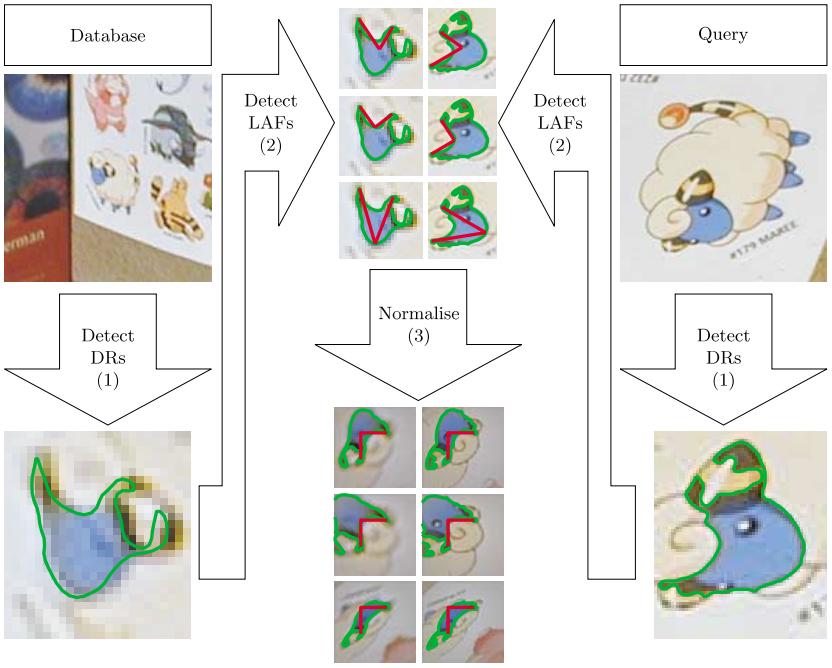


Figure 1: The structure of the recognition algorithm

1. Detect **Distinguished Regions** (DRs). MSERs used here, but any process producing image regions stable under affine transformations can be exploited.

2. Build **Local Coordinate Systems** (LAFs), applying various affine covariant constructions.

3. Define a **Measurement Region** (MR) in terms of the local coordinate systems. A square $\langle -1, 2\rangle \times \langle -1, 2\rangle$ is used.

4. **Geometric Normalisation:** Transform MRs of individual LAFs into a canonical form.

5. **Photometrically Normalise** RGB values in MRs.

6. Represent the normalised MRs by low frequency DCT coefficients.

7. **Local Correspondences** are established by correlation the DCT coefficients.

8. **Verification** of the object detections hypothesised by local correspondence.

(a) Original Image

(b) Intensity projection

(c) Ordering RGB values by saturation

(d) Ordering by projecting onto R-B axis $v = (r, g, b)^T(1, 0, -1)$

Figure 2: Examples of scalar projections of a RGB image

3 Generalised Extremal Region detection by Choice of Ordering of Image Pixels

The Maximally Stable Extremal Regions (MSERs) are highly repeatable regions that can support invariant recognition. MSERs are common on objects used in the Actipret scenario, nevertheless, we failed to detect MSERs on some objects of interest, especially in the case of occlusion.

The MSER-LAF method can be robustified by generalising the concept of extremal regions. In MSER detection, pixels are ordered by intensity. But intensity can by viewed as a specific case of mapping of RGB values onto some totally ordered set, here the set of positive real numbers. Different orderings of RGB values yield different distinguished regions on objects. The novelty had the following beneficial consequences:

- The richer representation of objects results in better recognition rates.
- Many regions and objects as a whole in Actipret scenes (and possibly in general) are distinguished w.r.t saturation, but not intensity.
- To maintain the same robustness, less Local Affine Frames (LAFs) per distinguished region are needed.
- Matching time depends quadratically on the number LAFs. LAF reduction leads to overall speed-up of the recognition process (approximately

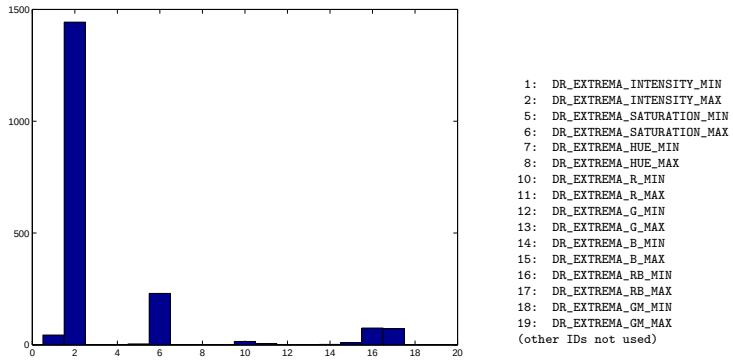


Figure 3: Statistics gathered on the the P0020 sequence. Distribution of regions contributing to object detection according to region type.

50% reduction of time).

- Interestingly, the overall speed of the recognition is improved, even though multiple DR detections are executed.

Qualitative analysis of the performance improvement is given in 6.

4 Adaptation of Object Recognition to Environmental Conditions

The adaptation algorithm is motivated by the following observations:

- Robustness of the recognition method is increased by employing multiple constructions of distinguished regions and local affine frames.
- But the computation speed is adversely affected.
- Not all of the constructions are always necessary.
- Only a subset of constructions that facilitate recognition should be computed. This subset depends on the data and therefore cannot be preset.
- *The subset must be adaptively selected during the recognition process*

The observations listed above are consistent with the statistics collected on two Actipret test sequences, as shown in Figures 3 and 4. We see that that on the Profactor sequences, as few as one or two types of regions provide recognition without a significant loss in performance (see Fig 3). For the CMP sequence, more constructions are required to maintain the robustness (see Fig. 4). The need to dynamically adapt the set of constructions is a consequence of the difference of utilities of constructions utility in CMP and Profactor sequences.

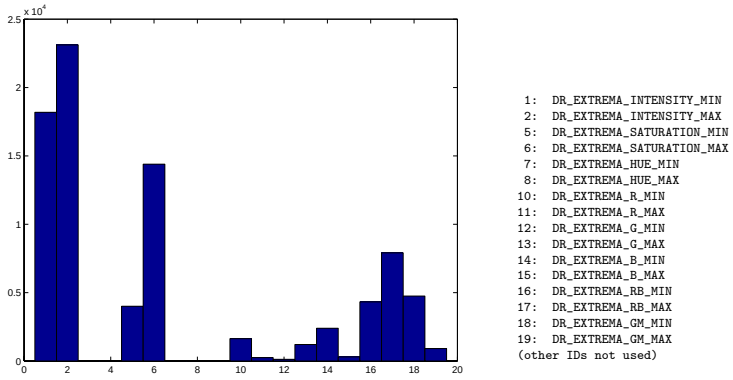
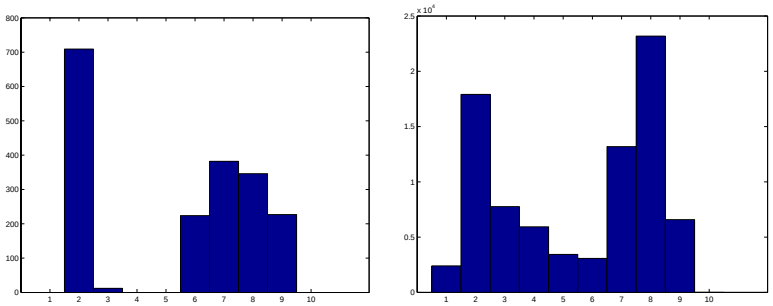


Figure 4: C0000 sequence: Distribution of distinguished regions contributing to object detection according to region type.



(a) P0020 sequence (b) C0000 sequence

Figure 5: C0000 sequence: Distribution of LAF construction contributing to object detection according to LAF type.

The observations made about utility of different distinguished region types are also valid for Local Affine Frame constructions. The utility of a LAF construction depends on the objects involved and on the environment (scene background, illumination condition, camera characteristics, e.g. noise). The statistics are visualized in Figure 5.

4.1 The adaptation algorithm

The adaptation algorithm is an optimization scheme consisting of the following steps:

1. Recognize objects using *all* available LAF constructions and DR types, i.e. in maximally robust (but slow) configuration of the method.
2. Consider recognized objects as true positives, i.e. ground truth. Time consuming verifications of the object hypotheses can be exploited.
3. For every object i identify the set S_i of all correspondences that would lead to correct recognition.
4. Search for a configuration of the method (combination of constructions and settings, e.g. thresholds) that minimizes expected computation time subject to the condition of retaining the recognition rate.

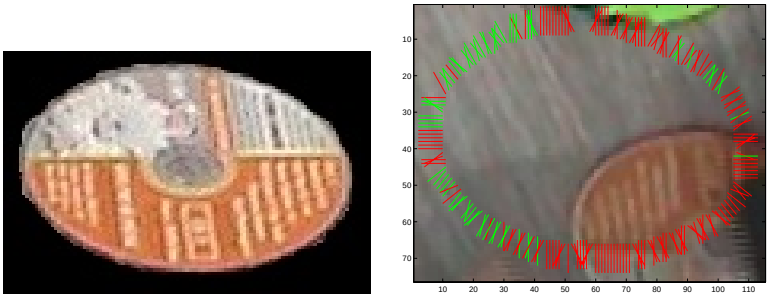
The final step is implemented as local optimization with an objective function that seeks a minimum of an expected computation time, which is estimated from the following quantities:

- The number of DR processes applied.
- The selectivity parameters affecting the number of regions created
- The number of LAF constructions.

The configuration is defined as acceptable if For at least $\alpha\%$ of the objects there are at least β correspondences retained. Would we denote $C_i \subset S_i$ the subset of correspondences that would be computed on the i-th object given the current configuration, $|C_i| >= \beta$ for at least $\alpha\%$ of the objects. In the current implementation, the settings are $\alpha = 95\%$, $\beta = 3$.

5 Robustifying Verification of a Hypothesised Model Occurrence - Modelling Spatial Dependencies on Object Boundaries

In the Actipret recognition component, two types of errors can occur. False positives - hallucinated objects not present in the scene, and false negatives - missed objects of interest. The methodology described in the previous section



(a) Model image (b) Rejected mismatch

Figure 6: An example of a situation where properties along the object boundary helps rejection of false matches

sequence	images	total (per image)	user (per image)	system
P0020_L0	223	1m18.620s (0.35s)	1m03.80s (0.28s)	0m04.98s
P0020_R0	223	1m05.548s (0.29s)	0m51.92s (0.23s)	0m03.67s
P0023_L0	210	1m31.012s (0.43s)	1m10.21s (0.33s)	0m04.90s
P0023_R0	210	1m16.637s (0.36s)	0m58.36s (0.28s)	0m02.41s
P0024_L0	202	1m26.046s (0.43s)	1m07.26s (0.33s)	0m03.19s
P0024_R0	202	1m14.063s (0.37s)	0m55.27s (0.27s)	0m02.99s

Table 1: 17.6.04 (without CD player detection)

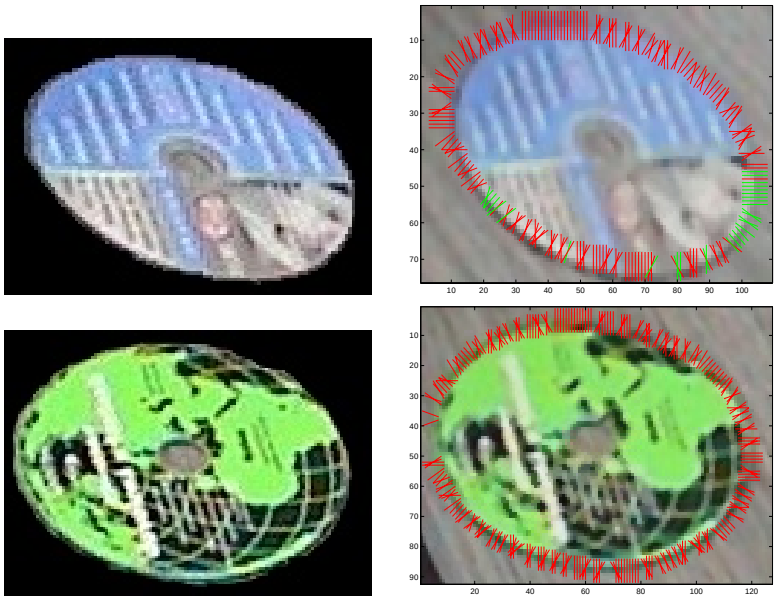
is aimed at minimizing false negatives. In this section, a highly selective strategy for object hypothesis verification is presented. Precise verification has the potential to significantly reduce false positives.

Our objective is to use the knowledge, obtained during the training phase of the recognition system, of the boundary of the object of interest. Probability of observing an RGB value *outside* the object is different for the 'background' situation (statistically continuous) and in the presence of 'object' of interest. From a decision-theoretic point of view, we are not only modelling the $P(observation|Object)$ probability, but a model of $P(observation|Background)$ is included. We are currently working on formalisation of the boundary continuity within a Markovian framework.

The impact of the improved verification step is shown in Figure 6. In the original method, the gray table was misinterpreted as the gray part of the CD. RGB values match in a large part part of the image area on which the CD was projected.

6 Speed and Performance Evaluation

Information about the speed of the object recognition, the June 2004 version (excluding the detection of static objects - the CD player), is presented in Table



(a) Model images (b) Examples of correct matches

Figure 7: Examples of model hypotheses confirmed by the boundary properties

sequence	images	total (per image)	user (per image)	system
P0020_L0	223	2m49.020s (0,76s)	2m06.05s (0.57s)	0m11.57s
P0020_R0	223	2m44.175s (0.74s)	2m05.37s (0.55s)	0m10.37s
P0023_L0	210	2m48.037s (0.80s)	2m07.59s (0.60s)	0m11.61s
P0023_R0	210	2m46.800s (0.79s)	2m06.86s (0.60s)	0m10.00s
P0024_L0	202	2m39.073s (0.79s)	1m58.14s (0.58s)	0m11.78s
P0024_R0	202	2m38.557s (0.78s)	2m03.60s (0.61s)	0m08.70s

Table 2: **17.2.04 (without CD player detection)**

1. Compared with the speed of the recognition component in February, a more than two-fold speed-up is achieved. This is a consequence of three factors: the application of the adaptation algorithm, a faster re-implementation of the detector of distinguished regions and a slightly higher speed of the processor.

Importantly, the speed up was achieved *without* compromising robustness of the system. This is documented in the standardised tables summarizing object component performance on prototypical Actipret test sequences from Profactor, ACIN and CMP:

- The recognition results compared against manually obtained ground-truth.

Sequence: P0020 Frames: 222
Poor Location: $\delta^2 = 65$

Camera L0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	120	135	15	11.1	13	10	204
cdplayer	222	222	0	0.0	0	0	222
ejectButton	202	222	20	9.0	20	20	202
tray	75	82	7	8.5	0	0	215

Camera R0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	120	137	17	12.4	13	12	204
cdplayer	222	222	0	0.0	0	0	222
ejectButton	171	222	51	23.0	51	51	171
tray	76	85	9	10.6	0	0	213

Sequence: P0023 Frames: 209
Poor Location: $\delta^2 = 65$

Camera L0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	71	120	49	40.8	11	11	160
cdplayer	209	209	0	0.0	0	0	209
ejectButton	191	209	18	8.6	18	18	191
tray	57	65	8	12.3	0	0	201

Camera R0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	108	120	12	10.0	12	10	195
cdplayer	209	209	0	0.0	0	0	209
ejectButton	138	209	71	34.0	71	71	138
tray	59	67	8	11.9	0	0	201

Sequence: P0024 Frames: 201
Poor Location: $\delta^2 = 65$

Camera L0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	110	120	10	8.3	11	7	187
cdplayer	197	201	4	2.0	4	4	197
ejectButton	181	201	20	10.0	20	20	181
tray	26	54	28	51.9	0	0	173

Camera R0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-intel	113	122	9	7.4	12	8	188
cdplayer	194	201	7	3.5	3	3	194
ejectButton	146	201	55	27.4	51	51	146
tray	13	50	37	74.0	0	0	164

Sequence: C0000

Frames: 237
Poor Location: $\delta^2 = 65$

Camera L0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-blue	236	237	1	0.4	0	0	236
cd-cyan	237	237	0	0.0	0	0	237
cd-orange	57	76	19	25.0	20	13	211
cd-phone	188	198	10	5.1	0	0	227
cd-world	237	237	0	0.0	0	0	237
cdplayer	86	237	151	63.7	30	30	86
ejectButton	105	237	132	55.7	11	11	105

Camera R0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-blue	211	237	26	11.0	26	26	211
cd-cyan	237	237	0	0.0	0	0	237
cd-orange	61	85	24	28.2	35	17	195
cd-phone	182	217	35	16.1	0	0	202
cd-world	237	237	0	0.0	0	0	237
cdplayer	161	237	76	32.1	69	69	161
ejectButton	138	237	99	41.8	92	92	138

Sequence: C0005

Frames: 356
Poor Location: $\delta^2 = 65$

Camera L0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-orange	290	297	7	2.4	5	1	345
cd-world	184	211	27	12.8	0	0	329
cdplayer	65	356	291	81.7	41	41	65
ejectButton	62	356	294	82.6	44	44	62

Camera R0

Object	Detected	Appear	FN	FNr (%)	FP	Loc-	Correct
cd-orange	189	293	104	35.5	106	97	243
cd-world	180	223	43	19.3	0	0	313
cdplayer	263	356	93	26.1	19	19	263
ejectButton	167	356	189	53.1	115	115	167

Appear	# of frames where object appears	A
Detected	# of frames where object detected	D
FN	# of false negatives (missed object)	FN=A-D
FNr	false negative rate	FNr = 100*FN/A
FP	# of false positives (hallucinated detection)	
Loc-	wrong location, # of objects farther than squared Euclidean distance	$\delta^2 = 65$

7 Conclusions

Two improvements towards robustification of the recognition approach are presented. First, the distinguished region detection is generalised. Besides intensity MSER, extremal regions with other ordering of RGB values are used, yielding lower false negative rate. The second improvement is in decision making: the background model reflecting spatial dependencies lowered the false positive rate.

To maintain efficiency of recognition after the introduction of a number of extremal region detectors, an adaptation scheme is proposed. The recognition system adaptively modifies its behaviour to match the environment conditions and exploits only recognition processes necessary for maintaining a low recognition rate.

We have shown experiemntally that the adaptation scheme, together with improved implementation, is capable of robust operation on Actipret sequences acquired in three different labs. The increase in robustness is not achieved at the expense of speed of the recognition component.