Predictive visual context in object detection*

Lucas Paletta

JOANNEUM RESEARCH
Institute of Digital Image Processing
Wastiangasse 6, 8010 Graz, Austria
lucas.paletta@joanneum.at

Abstract. This work discriminates external and internal visual context according to a recently determined terminology in computer vision. It is conceptually based on psychological findings in human perception that stress the utility of visual context in object detection processes. The paper outlines a machine vision detection system that analyzes external context and thereby gains prospective information from rapid scene analysis in order to focus attention on promising object locations. A probabilistic framework is defined to predict the occurrence of object detection events in video in order to significantly reduce the computational complexity involved in extensive object search. Internal context is processed using an innovative method to identify the object's topology from local object features. The rationale behind this methodology is the development of a generic cognitive detection system that aims at more robust, rapid and accurate event detection from streaming video. Performance implications are analyzed with reference to the application of logo detection in sport broadcasts and provide evidence for the crucial improvements achieved from the usage of visual context information.

1 Introduction

In computer vision, we face the highly challenging object detection task to perform recognition of relevant events in outdoor environments. Changing illumination, different weather conditions, and noise in the imaging process are the most important issues that require a truly robust detection system. This paper considers exploitation of visual context information for the prediction of object location and identity, respectively, that would significantly improve the service of quality in real-time interpretation of image sequences.

Research on video analysis has recently been focussing on object based interpretation, e.g., to refine semantic interpretation for the precise indexing and sparse representation of immense amounts of image data [13, 17]. Object detection in real-time, such as for video annotating and interactive television [1], imposes increased challenges on resource management to maintain sufficient quality of service, and requires careful design of the system architecture.

* This work is funded by the European Commission's IST project DETECT under grant number IST-2001-32157.

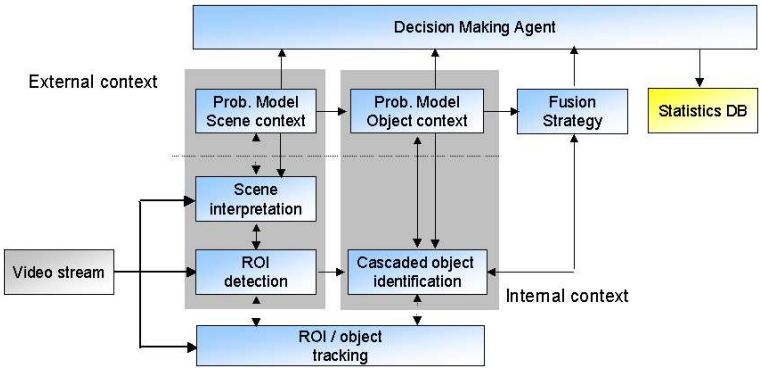


Fig. 1. External and internal visual context as a means to trigger discrimination processes for the purpose of object detection in video streams.

Fig. 1 illustrates how external and internal visual context are used to detect object information in a video stream. Rapid extraction of the context of a scene - the *global spatial context* with respect to object detection - might trigger early determination of regions of interest (ROI) and support careful usage of resources for more complex discrimination processes (Section 3.1). Within the ROI, object identification requires a grouping of local information [6]. In particular, the presented work describes how internal context from a configuration of object appearances - the *local spatial context* with respect to object detection - is exploited to distinguish collections of local measurements by means of their geometrical relations (Section 4). Finally, a federation of discriminatory processes is controlled by a supervising decision making agent [8, 22] to feed object information into a database for statistical performance evaluations.

Recent work on real-time interpretation applies attentional mechanisms to coarsely analyze the external context from the complete video frame information in a first step, reject irrelevant hypotheses, and iteratively apply increasingly complex classifiers with appropriate level of detail [32, 20]. In addition, context priming [31, 21] makes sense out of globally defined environmental features to set priors on observable variables relevant for object detection. Investigations on the binding between scene recognition and object localization made in experimental psychology have produced clear evidence that highly local features play an important role to facilitate detection from predictive schemes [4, 11, 5]. In particular, the visual system infers knowledge about stimuli occurring in certain locations leading to expectancies regarding the most probable target in the different locations (*location-specific target expectancies*, [10]).

Extraction of internal object context often optimizes single stage mapping from local features to object hypotheses [16, 25]. This requires either complex classifiers that suffer from the course of dimensionality and require prohibitive computing resources, or provides rapid simple classifiers with lack of specificity. Cascaded object detection [8, 32] has been proposed to decompose the mapping

into a set of classifiers that operate on a specific level of abstraction and focus on a restricted classification problem. We investigate the impact of local spatial context information on the performance of object detection processes, using a Bayesian method to extract *context from object geometry*.

The methodology to eploit visual context is embedded within a global framework on integrated evaluation of object and scene specific context (see [31], Section 2) with the rationale to develop a generic cognitive detection system that aims at more robust, rapid and accurate event detection from dynamic vision.

2 Visual context in object detection processes

In general, we understand *context* to be described in terms of *information that is necessary to be observed* and that can be *used to characterize situation* [7]. We refer to the ontology and the formalization that has been recently defined with reference to perceptual processes for the recognition of activity [6], and a Bayesian framework on context statistics [31], with particular reference to video based object detection processes.

In a probabilistic framework, object detection requires the evaluation of

$$p(\varphi, \sigma, \mathbf{x}, o_i|\mathbf{y}), \tag{1}$$

i.e., the probability density function of object o_i, at spatial location $\mathbf{x}$, with pose φ and size σ given image measurements $\mathbf{y}$. A common methodology is to search the complete video frame for object specific information. In cascaded object detection, search for simple features allows to give an initial partitioning into object relevant regions of interest (ROIs) and a background region.

The *visual context* is composed of a model of the *external context* of the embedding environment, plus a model of the object's *internal context*, i.e., *the object's topology characterized by geometric structure and associated local visual events* (e.g., local appearances) [31] so that local information becomes characterized with respect to the object's model (e.g., Fig. 5). Measurements $\mathbf{y}$ are separated into *local* object features representing object information $\mathbf{y}_L$ and the corresponding local visual *environment* represented by context features $\mathbf{y}_E$. Assuming that - given the presence of an object o_i at location $\mathbf{x}$ - features $\mathbf{y}_L$ and $\mathbf{y}_E$ are independent, we follow [31] to decompose Eq.1 into

$$p(\mathbf{y}|\varphi, \sigma, \mathbf{x}, o_i) = p(\mathbf{y}_L|\varphi, \sigma, \mathbf{x}, o_i) \cdot p(\mathbf{y}_E|\varphi, \sigma, \mathbf{x}, o_i). \tag{2}$$

Cascaded object detection leads to an architecture that processes from simple to complex visual information, and derives from global to local object hypotheses (e.g., [8, 32]). Reasoning processes and learning might be involved to select the most appropriate information according to an objective function and learn to integrate complex relationships into simple mappings. They are characterized by tasks, goals, states defined with respect to a model of the process, and actions that enable transitions between states [22], much in the sense of a decision making agent controlling discriminatory processes to improve quality of service in object detection (Fig. 1).

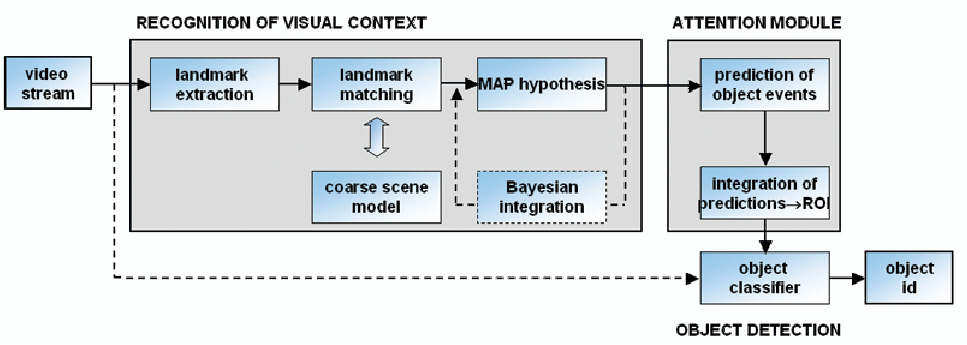


Fig. 2. Concept for recognition of external context for attention. Landmarks are extracted from the scene and matched towards a simlpe scene model. Bayesian recognition enables evidence integration over time and space. Attentive predictions on the location of embedded objects finally instantiate a complex object classifier ([21], Section 4) that verifies or rejects the object hypotheses.

3 External context for ROI detection

The concept is to propose attention from scene context using knowledge about forthcoming detection events that has been built up in repeated processing on the scene before. The knowledge which is derived from a simple scene model is activated from rapid feature extractions (e.g., using color regions) in order to operate only in those image regions where object detection events will most likely occur. The localization within an already modeled video scene is on the basis of a Bayesian prediction scheme. Recent investigations on human visual cognition give evidence for memory in visual search [12], underlining the assumption that already simple modeling mechanisms significantly support the quality of service in object detection.

3.1 Scene representations from landmarks

The basis for landmark based localization within a video scene is the extraction of discriminative and robustly re-locatable chunks of visual information in the scene. Landmarks have been efficiently defined on local greyvalue invariants [28], color and edge features [30], based on local appearance [29] and distinguished regions [18].

We apply an approach that rapidly extracts color and shape features but also considers the contrast of the extracted landmark region with respect to the corresponding features of its local neighborhood, being motivated by human perception, where, e.g., color is addressed by attentional mechanisms in terms of its diagnostic function [19]. Note that any other choice of local landmark representation would enable to pursue the methodology described in Sections 3.2, 3.3 as well.

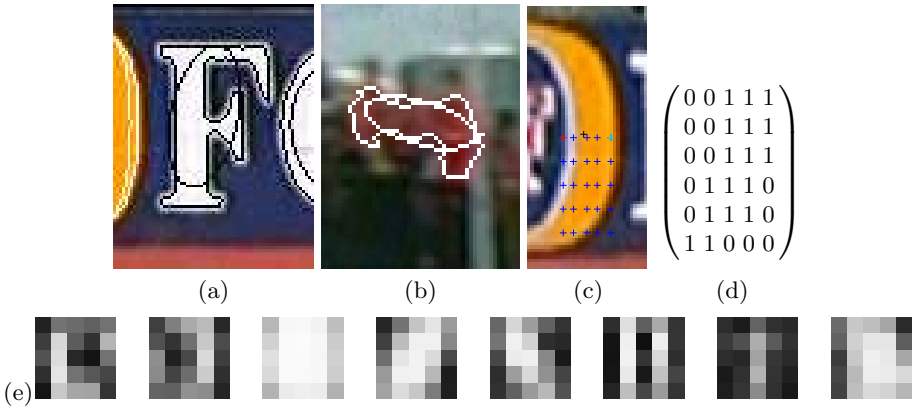


Fig. 3. Characteristic landmark features. (a,b,c) Color based ROIs for landmark definition, denoting the ROI border and the variance ellipsoid of the spatial distribution of ROI member pixels. (c,d) Class based extraction of shape: (c) Sampling (crosses) within the landmark region, (d) binary pattern received from color class based interpretation of the pixels sampled in (c), and attributed to class 4 in (e). (e) All prototypical patterns of shape to classify (d).

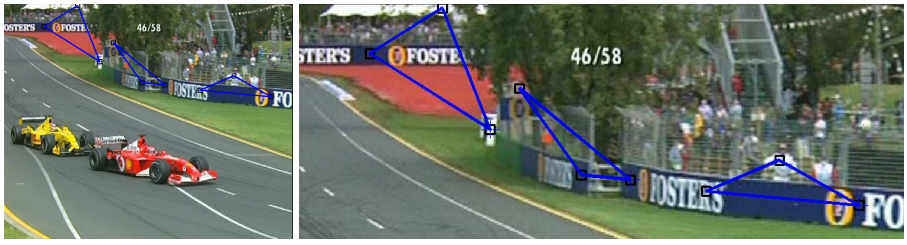


Fig. 4. Triple configuration of landmarks in a sample video frame using the landmark extraction described in Section 3.1.

In order to increase the discriminability of the locally extracted context, it is useful to combine landmarks into geometric configurations of 1-, 2-, and 3-tuples of landmarks (Fig. 4). Tuples of localized image properties own specific characteristics of scale invariance, ordering and topology [9] that make them attractive for landmark usage. Each single *landmark region* is encoded by a vector $\boldsymbol{\lambda}$ with landmark specific components $\boldsymbol{\nu}_i = (\mathbf{c}, \mathbf{n}, \mathbf{s}, \ldots)$, with features being vector-coded by color ($\mathbf{c}$), contrast ($\mathbf{n}$), and shape ($\mathbf{s}$). A 3-tuple *landmark configuration* denotes $\boldsymbol{\lambda} = [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \boldsymbol{\alpha}]^T$, where $\boldsymbol{\alpha}$ encodes the angles between landmarks $\boldsymbol{\nu}_i$.

3.2 Bayesian scene recognition

The goal of rapid scene modelling is to provide a simple and efficient encoding of the environment. The presented work is based on localization of a given landmark within a complete video sequence. The extracted landmark results in a hypothesis on representing a sample of a physical identity l_i of a landmark, i.e., the real landmark that generates a distribution of features from appearances to the observer. In our model, we pursue a framework of recognition and attribute each landmark sample $\boldsymbol{\lambda}$ to a physical landmark identity l_i and associated semantic blocks (frames) f_j in the reference (training) video sequence.

A simple scene model is rapidly generated from the frames of a video training sequence in terms of a list of landmark vectors $l_i \in \boldsymbol{\Lambda}$ that can be matched against a currently extracted landmark sample $\boldsymbol{\lambda_t}$. Scene recognition from interpretation of a landmark l^* is then computed via $l^* = \arg\min_{l_i} ||\boldsymbol{\lambda}_t - \boldsymbol{\lambda}(l_i)||$, which represents a nearest-neighbor matching to stored landmarks $\boldsymbol{\lambda}(l_i)$ in '$\boldsymbol{\lambda}$-space'.

In order to represent the uncertainty in landmark classification, the landmark l_i specific sample distribution is modelled using an unimodal Gaussian, $N_{l_i}(\boldsymbol{\mu_\lambda}, \boldsymbol{\Sigma_\lambda})$. The posterior interpretation of a landmark configuration $\boldsymbol{\lambda}$ is then outlined as follows,

$$P(l_i|\boldsymbol{\lambda}) = \frac{p(\boldsymbol{\lambda}|l)P(l_i)}{p(\boldsymbol{\lambda})} = \frac{p(\boldsymbol{\lambda}|l_i)\sum_{j=1}^{F} P(l_i|f_j)P(f_j)}{p(\boldsymbol{\lambda})}, \tag{3}$$

where $\boldsymbol{\lambda}$ denotes a sample landmark extraction from a test image, $P(l_i|\boldsymbol{\lambda})$ is the posterior with respect to a corresponding physical identity of a landmark, $P(l_i|f_j)$ is the probability for observing a physical landmark given a specific frame of the video sequence. To be precise, we require f_j to partition the space of landmarks l_i, which is the case in video block segmentation.

3.3 Contextual cueing to predict object detection events

Assuming that the scene has been repeatedly viewed and in a prevalent direction, each landmark configuration can be associated with a pointer to a succeeding object event that has been extracted before using any highly accurate, computationally expensive object identification method [32, 21]. In the scene model, a directional information in terms of an angle interval ($\beta \pm \sigma$), is provided in which the object event is completely embedded; β is in the direction of the center of the predicted detection event, and $\pm\sigma$ designates an angle interval so that the detection event is completely embedded within. This interval $\pm\sigma$ defines the standard deviation with respect to a one-dimensional normal distribution, i.e., $N(\mu_\beta, \sigma)$, that is defined geometrically normal to the straight line originating in landmark l_i with angle β. In total, these operations will define a probability density function (PDF) on the image, $p(\mathbf{x}|\Omega, l_i)$, with image locations $\mathbf{x}$ carrying confidence information about the support for a local object detection event, out of the set of objects, i.e., Ω, and in terms of a landmark specific *confidence map* (Fig. 6). However, in real-time implementations, Monte-Carlo sampling [15] would be appropriate to approximate the estimated PDFs.

To increase the robustness of the approach, we integrate the confidences from those landmarks $l_k \in K$ that have been consecutively visited in an observation sequence and been selected as estimators for the forthcoming object location, e.g., simply using a naive Bayes estimator,

$$p(\mathbf{x}(\beta)|\Omega, l_1, l_2, \ldots, l_K) = \prod_{k=1}^{K} p(\mathbf{x}(\alpha)|\Omega, l_k), \tag{4}$$

and thereby receive an incremental fusion of individual confidence maps. Fusion might use all those predictions $N()$ that correspond to the selected l_i giving $P(l_i|\boldsymbol{\lambda})$ (Section 3.2), weighting individual contributions according to the confidences given in Eq. 3 (Fig. 6).

4 Internal context from probabilistic structural matching

Recently, the requirement to formulate object representations on the basis of local information has been broadly recognized [26, 18]. Crucial benefits from decomposing the recognition of an object from global into local information are, increased tolerance to partial occlusion, improved accuracy of recognition (since only relevant - i.e., most discriminative - information is queried for classification) and genericity of local feature extraction that may index into high level object abstractions. In this paper we are using simple brightness information to define local appearances, but the proposed approach is general enough to allow any intermediate, locally generated information to be used as well, such as Gaussian filter banks [14], etc.

Context information can be interpreted from the relation between local object features [26] or within the temporal evolution of an object's appearance [2, 23]. Decomposing the complete object information into local features transforms $p(o_i|\mathbf{y}_L)$ into $p(o_i|\mathbf{y}_{L_1}, ..., \mathbf{y}_{L_N})$, N determines the size of the object specific environment. The *grouping* of conditionally observable variables to an *entity of semantic content*, i.e., a visual object, is an essential perceptual process [6].

The relevance of structural dependencies in object localization [26, 27] has been stressed before, though the existing methodologies merely reflect co-location in the existence of local features. The presented work outlines full evaluation of geometrical relations in a framework of probabilistic structural matching using Bayesian conditional analysis of local appearances as follows.

Geometrical information is derived from the relation between the stored object model - the trajectory in feature space - and the actions (shift of the focus of attention) that are mapped to changes in the model parametrization (e.g., change in viewpoint, i.e., $\Delta\varphi_j$). Figure 5 illustrates the described concept in the reference frame of the local appearance based object model. The geometry between local appearances is now explicitly represented by the shift actions a_i (deterministically causing $\Delta\varphi_j$) that feed directly into Bayesian fusion [22] by

$$P(o_i, \varphi_j|\mathbf{y}_1, a_1, \mathbf{y}_2) = \alpha P(o_i, \varphi_j|\mathbf{y}_1, a_1) p(\mathbf{y}_2|o_i, \varphi_j, \mathbf{y}_1, a_1). \tag{5}$$

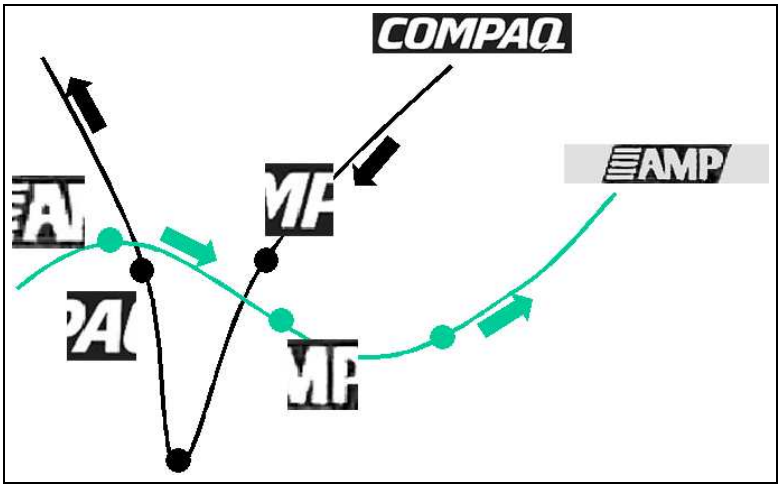


Fig. 5. Spatial context from the geometry of local information. A single appearance might give rise to evidence for multiple objects (at crossings of manifolds), even from a second measurement. The shift action to change a visual parameter of the manifold (e.g., a viewpoint change) and the associated appearances is then matched towards the manifold's trajectory in feature space to discriminate between object hypotheses.

Spatial context from probabilistic sturctural matching is now exploited using the conditional term $P(o_i, \varphi_j|\mathbf{y}_1, a_1)$: The probability for observing view (o_i, φ_j) as a consequence of deterministic action $a_1 = \Delta\varphi_1$ must be identical to the probability of having measured at the action's starting point before, i.e. at view $(o_i, \varphi_j - \Delta\varphi_1)$, thus $P(o_i, \varphi_j|\mathbf{y}_1, a_1) \equiv P(o_i, \varphi_j - \Delta\varphi_1|\mathbf{y}_1)$. Note that this obviously does not represent a naive Bayes classifier since it explicitly represents the dependency between the observable variables $\mathbf{y}_i, a_i$.

Furthermore, the probability density of $\mathbf{y}_2$, given the knowledge of view (o_i, φ_j), is conditionally independent on previous observations and actions, and therefore $p(\mathbf{y}_2|o_i, \varphi_j, \mathbf{y}_1, a_1) = p(\mathbf{y}_2|o_i, \varphi_j)$. The recursive update rule for *conditionally dependent* observations accordingly becomes,

$$P(o_i, \varphi_j|\mathbf{y}_1, a_1, \ldots, a_{N-1}, \mathbf{y}_N) = \alpha p(\mathbf{y}_N|o_i, \varphi_j) P(o_i, \varphi_j - \Delta\varphi_{N-1}|\mathbf{y}_1, a_1, \ldots, \mathbf{y}_{N-1}) \tag{6}$$

and the posterior, using $\mathbf{Y}_N^a \equiv \{\mathbf{y}_1, a_1, \ldots, a_{N-1}, \mathbf{y}_N\}$, is then given by

$$P(o_i|\mathbf{Y}_N^a) = \sum_j P(o_i, \varphi_j|\mathbf{Y}_N^a). \tag{7}$$

The experimental results in Figures 8 and 9 demonstrate that context is crucial for rapid discrimination from local object information. The presented methodology assumes knowledge about (i) the scale of actions and of (ii) the directions with reference to the orientation of the logo, which can be gained by ROI analysis beforehand.

5 Experimental Results

The object detection experiments were performed on 'Formula One' sport broadcast image sequences. The proposed object detection system first applies ROI detection based on contextual cueing from landmark configurations, supported by some color specific pattern classifiers [24, 21]. Within these detection regions, it extracts internal context from local features. The following paragraphs describe the recognition performance from (i) external context and (ii) using probabilistic structural matching (internal context).

(i) Context from landmark configurations The experiments were conducted on prediction of object detection in 'Formula One' broadcast videos. In particular, a video sequence of 71 frames (of 795×596 pixels) was used as training sequence and analysed to setup the scene model of the complete sequence, i.e., the interpretation of the landmark information, configurations, and the associated indexing and probabilistic interpretation for Bayesian scene recognition (Section 3.2).

The ROI color information was clustered into 12 Gaussian unimodal kernels via expectation maximization (EM) [3]. Shape patterns were clustered into 12 classes alike. The interpretation of this sequence resulted in 4351 n-tuple landmark registrations from 2123 physical landmark identities. The attribution to detection events was performed manually and under the assumption that this particular scene is captured by a specific camera motion (left to right) so that events are always encountered from one direction.

Via the localization of landmarks one can predict the successive detection event. The error in degree per single prediction is on average $2,6°$, $\pm 6,39°$ stdev). A direct hit rate of $93,7\%$ is achieved within the $2 \times \sigma_\beta$ interval (Fig. 7a). The resulting ROC curve (Fig. 7b) interprets the contextual cueing method in terms of a detection classifier, leading to excellent results with respect to its object detection performance. Finally, table 1 illustrates the gain in resources due to contextual cueing.

(ii) Context from geometry Spatial context from geometry can be easily extracted based on a predetermined estimate on scale and orientation of the object of interest. This is computed (i) from the topology of the ROI, and from (ii) estimates on $p_\varphi(\varphi|\sigma, \mathbf{x}, o_i, \mathbf{y}_E)$ and $p_s(\sigma|\mathbf{x}, o_i, \mathbf{y}_E)$ from global image features [31]. We present a recognition experiment from spatial context on 3 selected logos (Figure 8(a)) with local appearance representation as described above, and a 3-dimensional eigenspace representation to model highly ambiguous visual information. Figure 9 (right) demonstrates the dramatic decrease of uncertainty in the pose information for object o_2, i.e., $p(o_2, \varphi_j|\mathbf{y})$, from several steps of information fusion according to Eq. 6. Figure 8(b) illustrates the original and final distribution for all objects, $o_1 - o_3$.

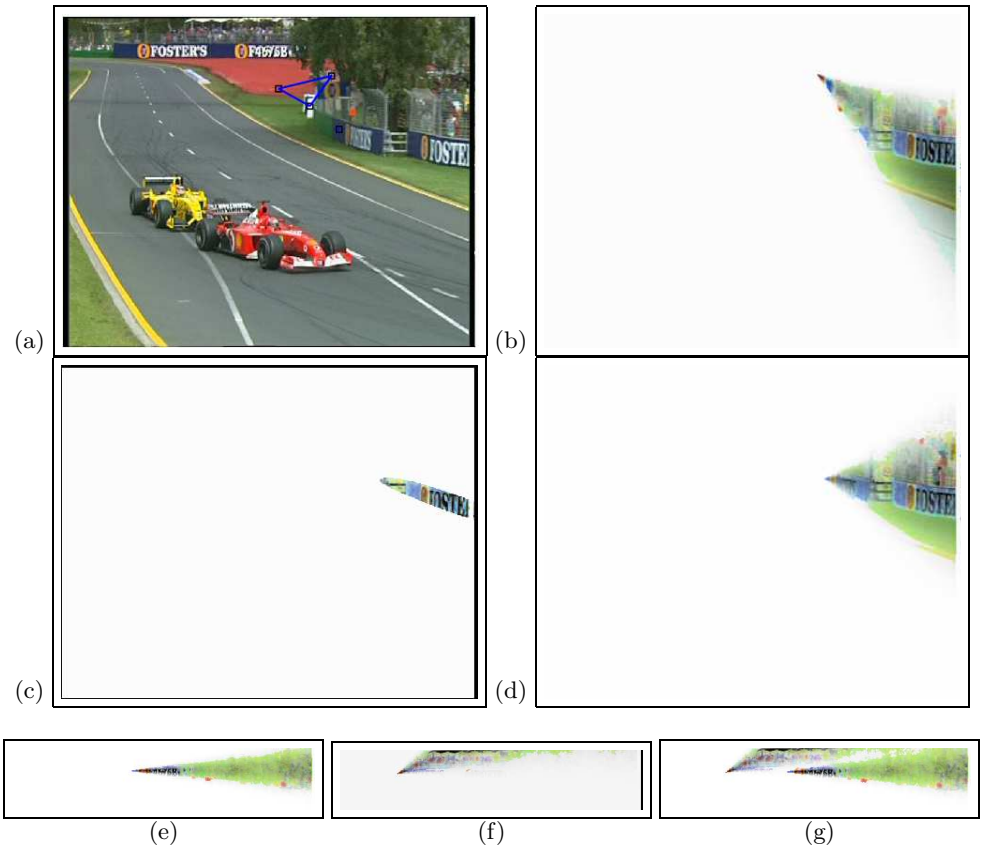


Fig. 6. Recursive contextual cueing and spatial attention for object detection. (a) Original frame with extracted landmark configurations. (b,d) Confidence maps derived from 2 individual landmarks. (c) Confidence map after Bayesian integration, depicting confidence beyond the threshold of $\Theta = 0.9$. (e) Accurate and (f) inaccurate predictive search regions, (g) integrated confidence map contains target. Since the target is represented in the fused confidence map, landmark 1 and 2 would impose object hits. In contrast, single landmark evaluation of 1 and 2 would produce one erroneous result.

Extensive analysis %	CCA analysis %	unprocessed image part (CCA) %
73,1	46,1	36,7

Table 1. Performance analysis of contextual cueing for attentive detection of objects of interest. Using extensive image analysis, ca. 73,1 % of the image had to be analyzed in order to detect a logo. In contrast, using CCA (contextual cueing for attention) analysis, only 46, 1% of the complete image had to be processed, resulting in a gain of 36,7 % of unprocessed image parts. CCA does not only provide impressive gains in speedup, but also in the statistically estimated accuracy of object detection as illustrated in Fig. 7b.

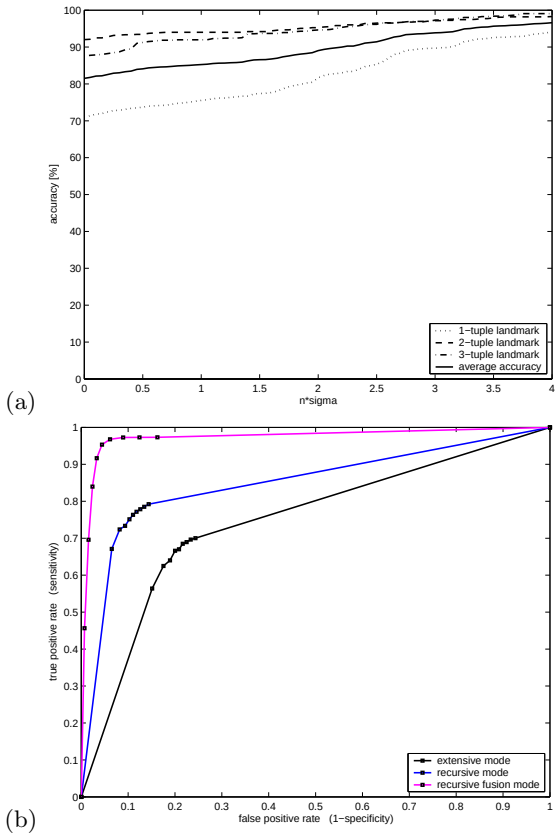


Fig. 7. Performance evaluation of the contextual cueing system. (a) Tolerated error and associated percentage of predictions (%) within this interval (for n-tuples of landmarks: point=1-tuple, dotdashed=2-tuple, dashed=3-tuple landmarks, line=avg.). (b) Receiver operator characteristic (ROC) curve demonstrating the high capabilities for object detection understanding the contextual cueing in terms of a detection system.

6 Conclusions

Context information contributes in several aspects to robust object detection from video. This work presents a predictive framework to focus attention on detection events instead of extensively searching the complete video frame for objects of interest.

Firstly, the probabilistic recognition of scenes from a landmark based description of the scene context are the innovative components that enable both rapid, predictable, and robust determination of relevant search regions. Secondly, grouping of local features can be rapidly applied and yields improved results. Additional computing derives the *context from the geometry of local features* which has been demonstrated to dramatically improve object recognition. Further ex-

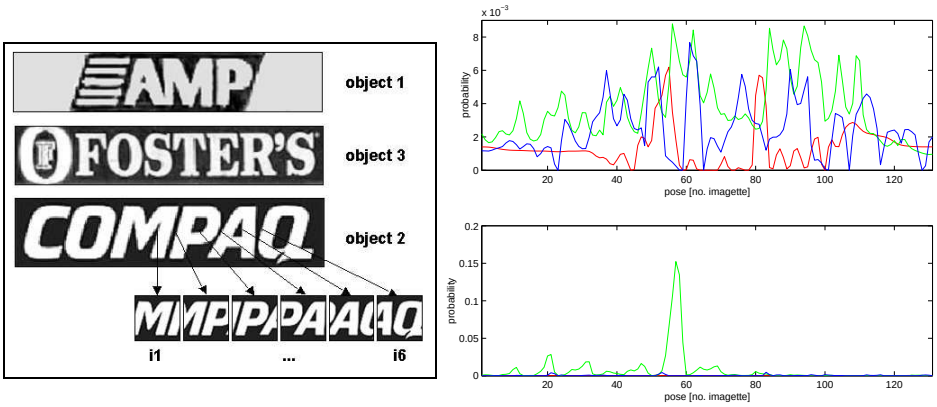
periments on contextual cueing demonstrate that prediction of object events from landmark based scene context can decisively determine an efficient focus of attention that would permit to save a substantial amount of computational resources from extensive processing.

Future work will focus on the extraction of local context from scene information in order to predict the future locations of detection events. We will consider the temporal context in the occurrence of landmark configurations and therefore most probably improve the landmark based scene recognition, together with the prediction performance as well.

References

1. J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2):52–60, 2002.
2. S. Becker. Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation*, 11(2):347–374, 1999.
3. S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using em and its applications to content-based image retrieval. In *Proc. International Conference on Computer Vision*, pages 675–682. Bombay, India, 1998.
4. I. Biederman, R.J. Mezzanotte, and J.C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177, 1982.
5. M.M. Chun and Y. Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36:28–71, 1998.
6. J. L. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual components for context aware computing. In *Proc. 4th International Conference on Ubiquitous Computing*, 2002.
7. A. K. Dey. Understanding and using context. In *Proc. 3rd International Conference on Ubiquitous Computing*, 2001.
8. B. A. Draper. Learning control strategies for object recognition. In K. Ikeuchi and M. Veloso, editors, *Symbolic Visual Learning*, chapter 3, pages 49–76. Oxford University Press, New York, 1997.
9. G.H. Granlund and A. Moe. Unrestricted recognition of 3-D objects using multi-level triplet invariants. In *Proc. Cognitive Vision Workshop*, Zürich, Switzerland, September 2002.
10. J. Hoffmann and W. Kunde. Location-specific target expectancies in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 25:1127–1141, 1999.
11. A. Hollingworth and J. Henderson. Does consistent scene context facilitate object perception. *Journal of Experimental Psychology: General*, 127:398–415, 1998.
12. A. Hollingworth and J.M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113–136, 2002.
13. M. Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 86(5):905–921, 1998.
14. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

15. M.H. Kalos and P.A. Whitlock. *Monte Carlo Methods, I: Basics*. John Wiley & Sons, New York, NY, 1986.
16. A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.
17. M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.
18. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conference*, 2002.
19. A. Oliva and P.G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176–210, 2000.
20. L. Paletta, A. Goyal, and C. Greindl. Selective visual attention in object detection processes. In *Proc. Applications of Artificial Neural Networks in Image Processing VIII*. SPIE Electronic Imaging, Santa Clara, CA, in print, 2003.
21. L. Paletta and C. Greindl. Context based object detection from video. In *Proc. International Conference on Computer Vision Systems*, pages 502–512. Graz, Austria, 2003.
22. L. Paletta and A. Pinz. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31(1-2):71–86, 2000.
23. L. Paletta, M. Prantl, and A. Pinz. Learning temporal context in active object recognition using Bayesian analysis. In *Proc. International Conference on Pattern Recognition*, pages 695–699, 2000.
24. F. Pelisson, D. Hall, O. Riff, and J.L. Crowley. Brand identification using Gaussian derivative histograms. In *Proc. International Conference on Computer Vision Systems*, pages 492–501. Graz, Austria, 2003.
25. F. Sadjadi. *Automatic Target Recognition XII*. Proc. of SPIE Vol. 4726, Aerosense 2002, Orlando, FL, 2002.
26. B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, pages 31–50, 2000.
27. C. Schmid. A structured probabilistic model for recognition. In *Proc. IEEE International Conference on Computer Vision*, 1999.
28. C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. International Conference on Computer Vision*, pages 872–877. San Jose, Puerto Rico, 1996.
29. R. Sims and G. Dudek. Learning visual landmarks for pose estimation. In *Proc. International Conference on Robotics and Automation*, Detroit, MI, May 1999.
30. Y. Takeuchi and M. Hebert. Finding images of landmarks in video sequences. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.
31. A. Torralba and P. Sinha. Statistical context priming for object detection. In *Proc. IEEE International Confernce on Computer Vision*, 2001.
32. P. Viola and M. Jones. Rapid object detection using a bossted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.



(a) (b)

Fig. 8. (a) The logo object set and associated pattern test sequence. (b) Probability distribution on pose hypotheses w.r.t. all 3 logo objects from a single imagette interpretation (*top*) and after the 5th fusion of local evidences (*bottom*).

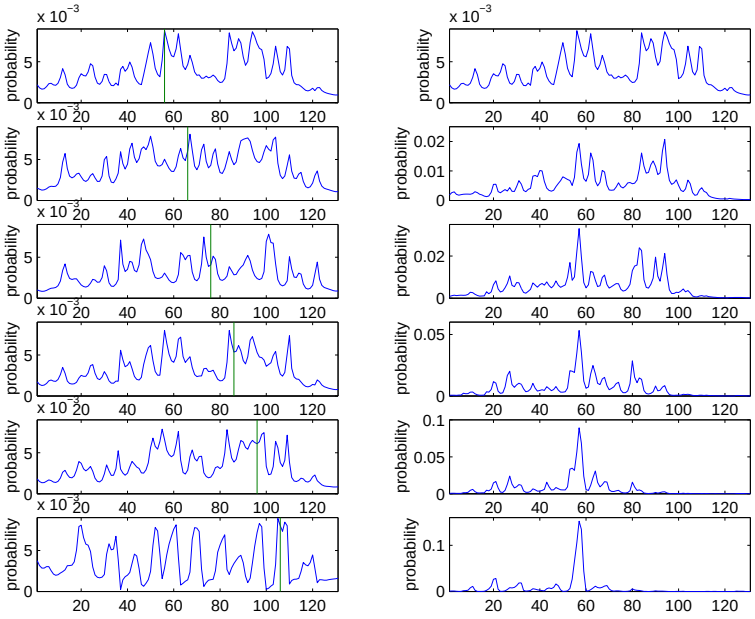


Fig. 9. *Left:* Probability distributions over pose hypotheses (imagette pose no.1-131 within logo) from individual test patterns no. 1-6, from top to bottom. *Right:* Corresponding fusion results using spatal context from geometry illustrating fusion steps no.1-5.