Predicting Detection Events from Bayesian Scene Recognition*

Georg Ogris and Lucas Paletta

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

Abstract. This work is conceptually based on psychological findings in human perception that highlight the utility of scene interpretation in object detection processes. Objects of interest are embedded in their visual context, i.e., in visual events within their spatial neighborhood. The implication for a detection system is that early recognition of this environment might provide information to directly map to an object event. The original contribution of this work is to outline a detection system that gains prospective information out of rapid scene analysis in order to focus attention on estimated object locations. Scene recognition is outlined on the basis of rapid detection of triplet configurations of landmarks which determine the discriminability of a particular location within the scene. Formulating scene recognition in terms of posterior landmark interpretation enables a recursive integration of target predictions and hence a probabilistic representation for attention based object detection.

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 prediction schemes 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 (e.g., [6]). 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. Recent work on real-time interpretation therefore considers attentional mechanisms and cascaded systems [10] to coarsely analyze the complete video frame in a first step, reject irrelevant hypotheses, and iteratively apply increasingly complex classifiers with appropriate level of detail [13, 9].

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

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 1058–1065, 2003.

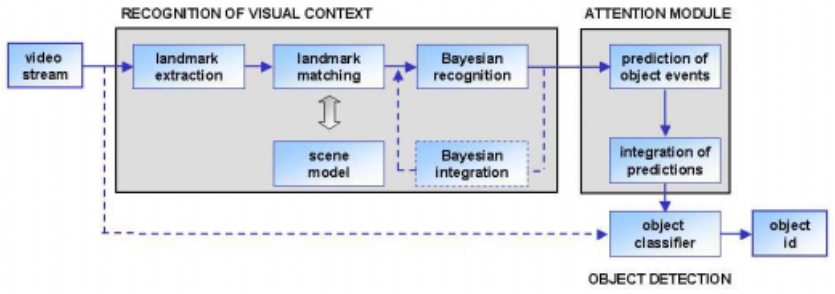

Fig. 1. Concept diagram. Landmarks are extracted from the scene and matched towards a primitive scene model. Bayesian recognition enables evidence integration over time and space. Attentive predictions on the location of embedded objects finally instantiate a complex object classifier that verifies or rejects the object hypotheses.

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 [3, 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*).

The original contribution of this work 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 (Fig. 1). The knowledge which is derived from a primitive, vectorized scene model is activated from simple and rapid feature extractions, i.e., landmarks, in order to operate only in those image regions where object detection events will most likely being encountered. The localization within an already modeled video scene is on the basis of a Bayesian prediction scheme and enables more elaborated detection schemes on previously attended objects (Fig. 1).

In the experimental results on logo detection in sport broadcasts we demonstrate a detailed analysis of a complete video test sequence and illustrate that landmark based prediction of detection events may actually provide results of attentional effects on the use of resources.

2 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 already been efficiently defined on features such as, colors and edges [12], local appearances [11], distinguished regions [7], etc. We apply an approach that rapidly extracts color and shape features but also considers the contrast of the extracted region with respect to the corresponding features of its local neighborhood. Landmarks can be combined into landmark configurations

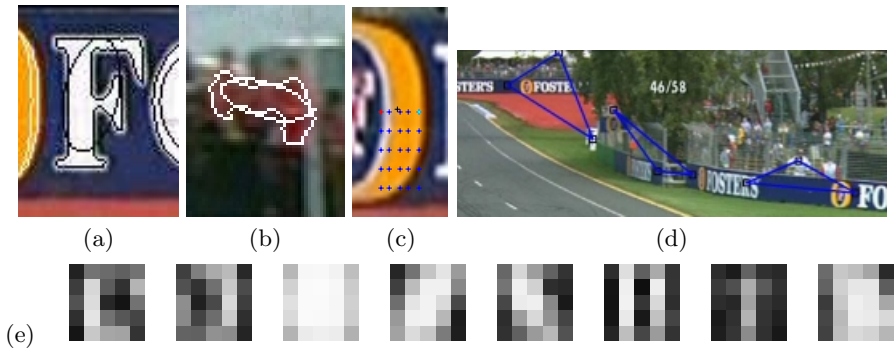

Fig. 2. Characteristic landmark features. (a,b,c) Color based regions for landmark definition, denoting the region border and the variance ellipsoid of the spatial distribution of region member pixels. (c) Class based extraction of shape: Sampling (crosses) of the local neighborhood of the region that result in a binary pattern received from color class interpretation of the pixels sampled therein.), attributed to class 4 in (e). (e) All 8 prototypical patterns of the learned Gaussian shape clusters, no. 1-8 from left to right.

(1-, 2-, and 3-tuples of landmarks) that carry even more distinctive power for positioning. Triplets of localized image properties own specific characteristics of scale invariance, ordering and topology [4] that make them attractive for landmark construction.

Landmarks are attributed to one out of a set of predefined color classes. The concept for color based segmentation includes region growing out of a local class seed and within predefined constraints on its region size (number of member pixels with reference to frame size) and the extension (ratio of the two principal components of the spatial distribution of region pixels) of a region. Additional constraints on region selection are based upon the contrast between region member pixels and neighboring pixels within a $k\sigma_{ij}^2$ environment, σ_{ij}^2 is the co-variance of the spatial distribution of the region pixels (Fig. 2). A shape parameter is then attributed from a projection of a local shape pattern within the region into the eigenspace (Fig. 2c,d), and applied via class features of prototypical shape patterns that have been clustered from training data (Fig. 2e). The orientation and size of the pattern is normalized with reference to the region, a binary pattern results then from positive (1, inside region) and negative (0, outside region) local responses.

Landmarks are not only defined by a singular region but also in terms of region configurations of n-tuples of landmarks (Fig. 3). Each *single 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 triplet *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 $\boldsymbol{\nu}_i$.

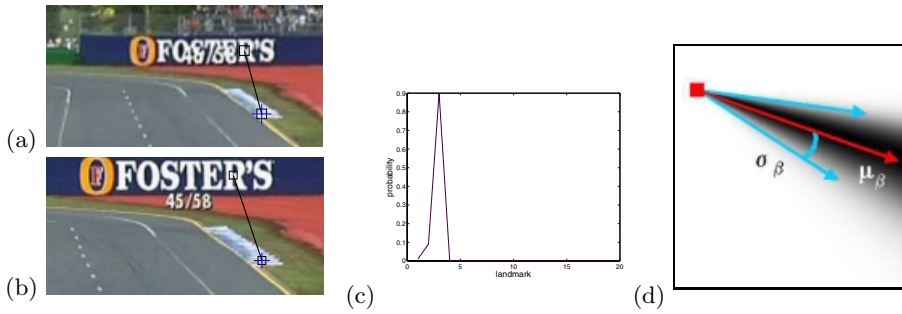

Fig. 3. (a,b) Triple configuration of landmarks in a sample video frame using the landmark extraction described in Section 2. (a) Landmark in test sequence and (b) matched landmark in training sequence. (c) associated $P(l_i|\boldsymbol{\lambda})$, (d) confidence map attributed to a landmark and corresponding directional prediction towards nearest object event.

3 Bayesian scene recognition

The goal of rapid scene modelling is to provide a simple and efficient encoding of the environment. It will serve the purpose of localization within a complete video sequence, with reference to physical identities l_i of landmarks (configurations). In our model, we omit any processing of the low level information such as it is performed in mosaicking. Instead, we pursue a framework of recognition and attribute each landmark feature sample $\boldsymbol{\lambda}$ to a physical landmark identity l_i and associated semantic blocks (frames) f_j in the reference (training) video sequence.

A primitive scene model is then 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

$$l^* = \arg\min_{l_i} ||\boldsymbol{\lambda}_t \quad \boldsymbol{\lambda}(l_i)||, \tag{1}$$

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,

$$P(l_i|\boldsymbol{\lambda}) = \frac{p(\boldsymbol{\lambda}|l)P(l_i)}{p(\boldsymbol{\lambda})} = \frac{p(\boldsymbol{\lambda}|l_i)\sum_{j=1}^{F} P(l_i|f_j)P(f_j)}{p(\boldsymbol{\lambda})}, \tag{2}$$

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.

In practice, we can simplify Eq. 2 using the mechanism of matching of individual landmarks $\boldsymbol{\lambda}$ as described above, since the number of all possible l_i's might be too large to be processed in real-time. Given a sample landmark vector $\boldsymbol{\lambda}_t$, we therefore select all landmark identities l_i with $||\boldsymbol{\lambda}_t \quad \boldsymbol{\lambda}(l_i)|| < \epsilon$, ϵ is a predefined threshold on the distance within a local neighborhood to the test vector $\boldsymbol{\lambda}_t$. We then apply Eq. 2 only on these identities, resulting in the probability distribution $P(l_i|\boldsymbol{\lambda})$ (Fig. 3e).

4 Recursive attention to detection events

Assuming that the scene has been repeatedly viewed and in a prevalent direction, each landmark configuration can be mapped to a pointer to a succeeding object event that has been extracted before using any highly accurate, computationally expensive object identification method. If the pointer is formulated in probabilistic terms, probabilistic inference can be applied to recursively estimate a target's location using evidences from multiple landmarks.

4.1 Prediction of object events

In the scene model, a directional information in terms of an angle interval $(\boldsymbol{\mu}_\beta \pm \sigma_\beta)$, is provided in which the object event is completely embedded; $\boldsymbol{\mu}_\beta$ is in the direction of the center of the predicted detection event, and $\pm\sigma_\beta$ designates an angle interval so that the detection event is completely embedded within (Fig. 3f). This interval $\pm\sigma_\beta$ defines the standard deviation with respect to a one-dimensional normal distribution, i.e., $N_{l_i}(\boldsymbol{\mu}_\beta, \sigma_\beta)$ (Fig. 4), that is defined normal to the vector originating from landmark l_i with angle $\boldsymbol{\mu}_\beta$. In total, these operations will define a probability density function (PDF) on the image, $p(\mathbf{x}(\beta)|\Omega, l_i)$, with image locations $\mathbf{x}$ carrying confidence information about the support for a local object detection event, out of the set of objects o_k, i.e., $o_k \in \Omega$, and in terms of a landmark specific *confidence map* (Fig. 3f, 4). This confidence map is an idealistic representation of the probabilistic information. However, in real-time implementations, Monte-Carlo sampling would be appropriate to approximate the estimated PDF.

4.2 Recursive contextual cueing to objects

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,

$$p(\mathbf{x}(\beta)|\Omega, l_1, l_2, \ldots, l_K) = \prod_{k=1}^{K} p(\mathbf{x}(\alpha)|\Omega, l_k), \tag{3}$$

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), weighting individual contributions according to the confidences given in Eq. 2 (Fig. 4).

5 Experimental results

The experiments were conducted on logo object detection in 'Formula One' broadcast videos. In particular, a video sequence of 71 frames (of 720×596 pixels, PAL) was used as training sequence and analyzed to generate 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, 4).

The region color information was clustered into 15 Gaussian unimodal kernels via expectation maximization (EM [2]). Shape patterns were clustered into 8 classes using PCA subspace information. The interpretation of this sequence resulted in 4011 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 under specific camera motion (e.g., 'left to right') so that events are always encountered from one direction, as it is the case in most sport broadcast captures.

Fig. 3 illustrates results for the localization of a 2-tuple landmark (c) in the test frame (d), (e) shows the associated probability distribution $P(l_i|\lambda)$, $\forall i = 1..L$. Via the localization of landmarks one can predict the successive detection event. Fig. 5a illustrates the error in degree per single prediction (avg. $2,6°$, std. $6,39°$). A resulting receiver operator characteristic (ROC) curve (Fig. 5b) interprets the contextual cueing method in terms of a classifier, leading to excellent results with respect to its object detection performance. Finally, the experiments demonstrate the gain in resources: Using extensive image analysis, ca. 73,1 % of the image has to be analyzed in order to detect a logo; using contextual cueing for attention, only $46,1\%$ of the image pixels had to be investigated. Considering the detection feature, we can even omit the application of complex object recognition methods that would have to be operating on simple regions of interest. More details are described in [8].

These results clearly demonstrate that we can achieve a promising level of accuracy in the predictions on search regions. Furthermore, Fig. 4 illustrates an improvement in performance using recursive cueing, by integrating the confidence maps from consecutive landmarks.

6 Conclusions

This work presents a probabilistic framework to focus attention on detection events instead of extensively searching the complete video frame. The probabilistic recognition of scenes from a landmark based description of the scene context, and the directional attention mechanism are the innovative components

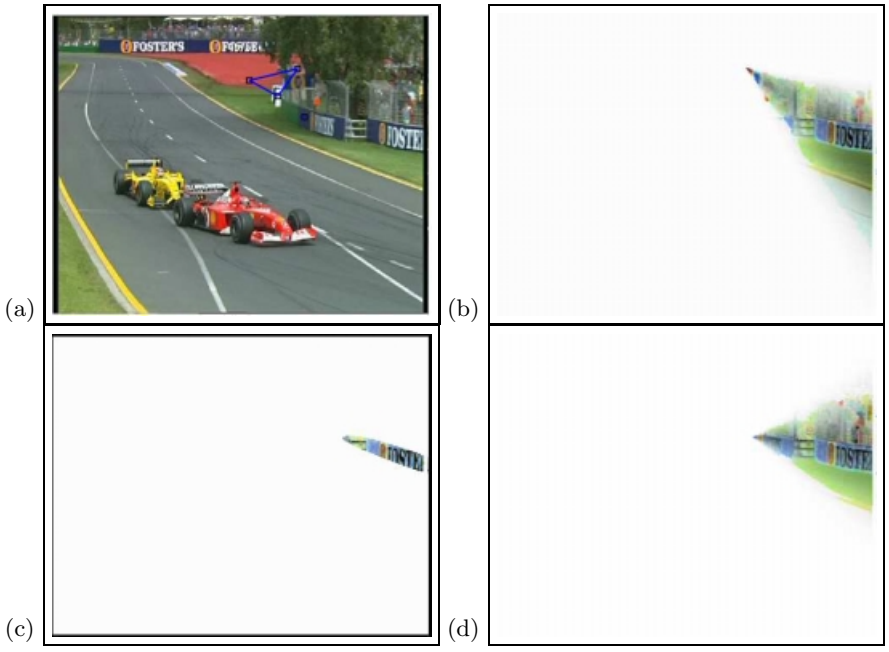

Fig. 4. Recursive contextual cueing and spatial attention for object detection. (a) Original frame with extracted landmark configurations. (b,d) Confidence maps derived from 2 individual landmarks. (c) Confidence map after Bayesian integration, depicting confidence beyond the threshold of $\Theta = 0.9$.

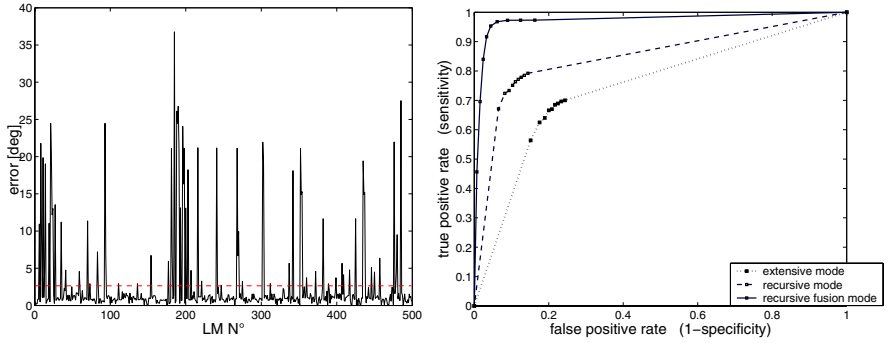

Fig. 5. Performance evaluation of the contextual cueing system. (a) Error in the directional predicition, $\boldsymbol{\mu}_\beta$, in degrees (mean $2,6^\circ$). (b) Receiver operator characteristic curve demonstrating the high capabilities for object detection understanding the contextual cueing in terms of a classifier.

that enable both rapid, predictable, and robust determination of relevant search regions.

The experiments 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.

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. 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.
3. I. Biederman, R.J. Mezzanotte, and J.C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177, 1982.
4. 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.
5. 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.
6. 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.
7. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conference*, 2002.
8. G. Ogris. *Attention from scene context for object detection in video.* MsThesis, Inst. of Digital Image Processing, Joanneum Research, Graz, Austria, 2003.
9. 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.
10. 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.
11. R. Sims and G. Dudek. Learning visual landmarks for pose estimation. In *Proc. International Conference on Robotics and Automation*, Detroit, MI, May 1999.
12. Y. Takeuchi and M. Hebert. Finding images of landmarks in video sequences. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.
13. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.