A New Class of Learnable Detectors for Categorisation

Jiri Matas and Karel Zimmermann

Center for Machine Perception,
Faculty of Electrotechnical Engineering,
Czech Technical University in Prague

Abstract. A new class of image-level detectors that can be adapted by machine learning techniques to detect parts of objects from a given category is proposed. A classifier (e.g. neural network or adaboost trained classifier) within the detector selects a relevant subset of extremal regions, i.e. regions that are connected components of a thresholded image. Properties of extremal regions render the detector very robust to illumination change. Robustness to viewpoint change is achieved by using invariant descriptors and/or by modeling shape variations by the classifier.

The approach is brought to bear on three problems: text detection, face segmentation and leopard skin detection. High detection rates were obtained for unconstrained (i.e. brightness, affine and font invariant) text detection (92%) with a reasonable false positive rate.

The time-complexity of the detection is approximately linear in the number of pixels and a non-optimized implementation runs at about 1 frame per second for a 640×480 image on a high-end PC.

1 Introduction

Methods relying on correspondences of local affine or scale covariant regions have furthered research in a number of areas of computer vision including object recognition [15, 8, 13, 5], wide-baseline stereo [14, 17, 4, 10, 11], tracking [3], categorisation [16, 1, 7] and texture recognition [6]. As a first step, the cited approaches detect a set of transformation-covariant regions that are stable both under illumination variations and local geometric transformations (either similarity or affine) induced by a viewpoint change. The detectors are generic and they have been shown to perform well in a wide range of environments.

Fig. 1. Text detection based on category-specific extremal regions

In categorisation, the problem we focus on, state-of-the-art approaches represent categories as probabilistic configurations of classified transformation-covariant regions [2, 16, 7, 12]. The (soft) classification of the

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 541–550, 2005.

transformation-covariant regions into components (parts) is based on rules learned in a training stage. The region detectors used in categorisation are generic, e.g. the salient regions of Kadir and Brady in the categorisation systems of Fergus et al.[2] and Fei-Fei et al.[1] or the affine-invariant interest points of Mikolajczyk and Schmid[11] and MSER regions [10] in the VideoGoogle system of Sivic and Zisserman [16].

As a main contribution of the paper, a new class of machine learnable category-specific detectors of covariant regions is presented. Machine learning techniques have been applied in the context of categorisation to find a representation of the configuration [16, 18] and to train classifiers for recognition of regions — components of the configuration [2, 16]. In this paper, machine learning is newly introduced to the image processing level i.e. it becomes part of the design of a category-specific detector. The benefits of learning at the detector level are demonstrated on two classical categorisation problems: text detection in images and licence plate recognition.

The proposed category-specific class of detectors is trained to select a relevant subset of extremal regions. A robust category-specific detector of extremal regions can be implemented as follows. Enumerate all extremal regions, compute efficiently a description of each region and classify the region as relevant or irrelevant for the given category. In a learning stage, the classifier is trained on examples of regions – components of objects from a given class. Such detection algorithm is efficient only if features (descriptors) for each region are computed in constant time. We show there is a sufficiently discriminative class of 'incrementally computable' features on extremal regions satisfying this requirement.

The proposed detector is robust to many image transformations. The affine invariance is achieved in reasonable scale by learning. The partial occlusion robustness, depicted in Figure 5, is caused by decomposition of the object to small individually detectable regions. The illumination invariance is demonstrated in Figure 1. Two images of a scene with different contrast levels are shown. A class-specific detector of character-like regions (the arrow and the pound sign are not in the training set) processed the two images. An object belonging to a 'text' class is defined as a (approximately) linear configuration of more than one character-like extremal regions. The hand-written text is detected even in the extremely low contrast image at the bottom of Fig. 1.

2 Category-Specific Extremal Region Detection

Our objective is to select from the set of extremal regions those with shape belonging to a given category. The model of the category is acquired in a separate training stage. Let us assume for the moment that the learning stage produced a classifier that, with some error, is able to assign to each extremal region one of two labels: 'interesting', i.e. is a component of our category, or 'non-interesting' otherwise. The detection of category-specific extremal regions can be then arranged as three interleaved steps: (1) generate a new extremal

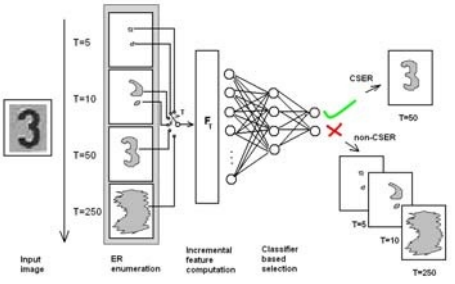

Fig. 2. The detection is implemented as interleaved enumeration of extremal regions, computation of incremental features and classification

region, (2) describe the region and (3) classify it. The interleaved computation is schematically depicted in Figure 2.

Extremal regions are connected components of an image binarised at a certain threshold. More formally, an extremal region r is a contiguous set of pixels such that for all pixels $p \in r$ and all pixels q from the outer boundary ∂r of region r either $I(p) < I(q)$ or $I(p) > I(q)$ holds. In [10], it is shown that extremal regions can be enumerated simply by sorting all pixels by intensity either in increasing or decreasing order and marking the pixels in the image in the order. Connected components of the marked pixels are the extremal regions. The connected component structure is effectively maintained by the union-find algorithm.

In this process, exactly one new extremal region is formed by marking one pixel in the image. It is either a region consisting of a single pixel (a local extremum, a region formed by a merge of regions connected by the marked pixel, or a region that consisting of union of an existing region and the marked pixels. It is clear from this view of the algorithm that there are at most as many extremal regions as there are pixels in the image. The process of enumeration of extremal regions is nearly linear in the number of pixels[1] and runs at approximately 10 frames per second on 2.5 GHz PC for a 700×500 image.

To avoid making the complexity of the detection process quadratic in the number of image pixels, the computation of region description must not involve all of its pixels. Fortunately, a large class of descriptors can be computed incrementally in constant time even in the case of a merge of two or more extremal regions (the other two situations are special cases). Importantly, combinations of incrementally computable features include affine and scale invariants. Incrementally computable features are analysed in Section 3.

The final step of the CSER detection, the selector of category-specific regions, is implemented as a simple neural network trained on examples of regions - components of the category of interest. The neural network selects rel-

[1] The (negligibly) non-linear term is hidden in the "maintenance of connected component structure".

evant regions in constant time. The overall process of marking a pixel, recalculating descriptors and classifying is thus constant time. The choice of neural network is arbitrary and any other classifier such as SVM or AdaBoost could replace it.

3 Incrementally Computable Region Descriptors

In the CSER detection process the descriptors of a connected component that evolves have to be computed. The evolution has two forms: growing and merging of regions. It is easy to see that if we can compute the description of a union $r_1 \cup r_2$ of two regions r_1 and r_2 then we can compute it in each step of the evolution (we use $r_.$ to identify both the region and its set of pixels). The following problem arises: what image features computed on the union of the regions can be obtained in constant time from some characterisation g of r_1 and r_2?

For example, let us suppose that we want to know the second central moment of the merged region. It is known that the second central moment (moment of inertia) can be computed in constant time from the first and second (non-central) moments and first and second (non-central) moments can be updated in the merge operation in constant time. A region descriptor (feature) ϕ will be called *incrementally computable* if the following three functions exists: a characterising function $g : 2^{Z^2} \to \mathcal{R}^m$, a characterisation update function $f : (\mathcal{R}^m, \mathcal{R}^m) \to \mathcal{R}^m$, and a feature computation function $\phi : \mathcal{R}^m \to \mathcal{R}^n$, where m is constant, n is the dimension of the feature and Z^2 is the image domain.

For each region, the characterising function g returns the information necessary for computing feature ϕ in a real vector of dimension m. The dimension m of the characteristic vector depends on the feature, but is independent of region size. Given the characterisation returned by g, the n-dimensional feature of interest (region descriptor) is returned by ϕ. Function f computes the characterisation of the merged region given the characterisation of the regions r_1, r_2. For efficiency reasons, we are looking for features with the smallest characterisation dimension m^*. An incremental feature is a triplet of functions (g^*, f^*, ϕ^*) defined as

$$g^* = \arg\min_g \{\dim(g(2^{Z^2}))\} \text{ subject to } \phi(g(r_1 \cup r_2)) = \phi(f(g(r_1), g(r_2))).$$

Example 1. Minimum intensity I of all pixels in a region is an incrementally computable feature with dimension $m^* = 1$. Given regions r_1 and r_2 with pixels $r_1^i \in r_1, r_2^j \in r_2$, the description of the union regions r_1, r_2 is

$$\phi(g(r_1 \cup r_2)) = \underbrace{1}_{\phi} . \underbrace{\min}_{f} \{ \underbrace{\min_{r_1^i \in r_1} I(r_1^i)}_{g(r_1)}, \underbrace{\min_{r_2^j \in r_2} I(r_2^j)}_{g(r_2)} \}$$

Example 2. The center of gravity ($m^* = 2$) of a union of regions r_1, r_2 with pixels r_1^i, r_2^j for $i = 1...k_1, j = 1...k_2$ is

$$\phi(g(r_1 \cup r_2)) = \underbrace{\frac{1}{k_1 + k_2}}_{\phi} \left(\underbrace{\sum_{i=1}^{k_1} r_1^i}_{g(r_1)} \underbrace{+}_{f} \underbrace{\sum_{j=1}^{k_2} r_2^j}_{g(r_2)} \right).$$

In this paper we use the following incrementally computable features: *normalized central algebraic moments* with $m^* \sim (k)^2$ where k is an moment order (calculation based on algebraic moments), *compactness* with $m^* = 2$ (using the area and the border), *Euler number* of a region with $m^* = 2$, *Entropy of cumulative histogram* with $m^* = 2$. Features that we are not able to compute incrementally are e.g. the number convexities and the area of convex hull.

4 Experiments - Applications and Properties of CSER Detection

4.1 Text Detection and Properties of CSER

$\theta \backslash \phi$	0^o	15^o	30^o	45^o
0^o	2.6	2.8	2.8	3.0
10^o	3.2	3.2	3.2	3.8
20^o	3.2	3.6	4.0	7.8
30^o	7.6	8.4	15.2	26.5

Fig. 3. (a) False negative rate (missed characters) as a function of viewing angles ϕ (elevation) , θ (azimuth); in percentage points

The favorable properties (e.g. bright and affine invariance or speed) of CSER detector are demonstrated in this experiment. We have decided for text detection problem only of one font to present mentioned properties of the detector.

The category of texts is modeled as a linear constellation of CSERs. The feed-forward neural network for CSER selection was trained by a standard back-propagation algorithm on approximately 1600 characters semi-automatically segmented from about 250 images acquired in unconstrained conditions. The region descriptor was formed by scale-normalised algebraic moments of the characteristic function up the fourth order, compactness and entropy of the intensity values. Intentionally, we did not restrict the features to be either rotation or affine invariant and let the neural network with 15 hidden nodes to model feature variability.

The detection of text proceeds in two steps. First, relevant CSER selected as described above. Second, linear configurations of regions are found by Hough transform. We impose two constraints on the configurations: the CSER regions must be formed from more than three regions and the regions involved must have a similar height.

Detection Rate. On an independent test set of 70 unconstrained images of scenes the method achieved 98% detection rate with a false positive appearing in approximately 1 in 20 images.

Speed. The detection time is proportional to the number of pixels. For a 2.5 GHz PC the processing took 1.1 seconds for a 640×480 image and 0.25 seconds for 320×240 image.

Robustness to viewpoint change was indirectly tested by the large variations in the test data where scales of texts differed by a factor of 25 (character 'heights' ranged from approximately 7-8 to 150 pixels) and were viewed both frontally and at acute angles. We also performed systematic evaluation of the CSER detector. Images of texts were warped to simulate a view from a certain point on the viewsphere. The false negative rates for the CSER detector (missed character percentages) with approximately 10 false positive regions per background image are shown in Table 3a). The CSER detector is stable for almost the whole tested range. Even the 27% false negative at the 30^o-45^o elevation-azimuth means that three quarters of characters on the text sign are detected which gives high probability of text detection.

Robustness to illumination change was evaluated in a synthetic experiment. Intensity of images taken in daylight was multiplied by a factor ranging from 0.02 to 1. As shown in Figure 4b, the false negative (left) and false positive (right) rates were unchanged for both the detector of CSER (bottom) and whole text signs (top) in the $(0.1, 1)$ range! For the 0.1 intensity attenuation, the image has at most 25 intensity levels, but thresholds still exist that separate the CSERs. The experiment also suggests that the interleaving of extremal region enumeration, description and classification cannot be simply replaced by detection of MSERs followed by MSER description and classification.

Robustness to occlusion is a consequence of modeling the object as a configuration of local components. Occlusion of some components does not imply the object is not detected.

Independence of internal classifier is presented by implementation of different classifier. The ROC characteristic in the Fig.4a shows the comparison of achieved results with built-in neural network and adaboost classifiers. Considering that detector is originally designed as filter we can see that for acceptable rate of FP bigger then 15% adaboost provides lower false negative rate then neural network. In the other hand, achieved results renders detector to be able to work alone (without any post-processing considering only linear constellation constraints) and for a such application neural network brings better results in the interval of $FP \leq FN$.

4.2 Text Detection in Unconstrained Conditions

We applied the CSER to the problem of text detection in images for which standard datasets are available. We used part of the ICDAR03 text detection competition set maintained by Simon Lucas at the University of Essex [9].

An object from the 'text category' was modeled as an approximately linear configuration of at least three 'text-like' CSERs. The neural network selector

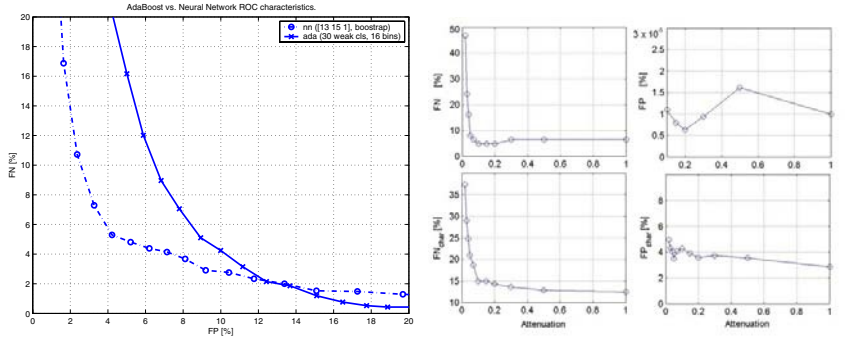

Fig. 4. (a)The ROC characteristic of the proposed detector comparing results with built-in neural network and adaboost classifiers.(b) Text detection in images with attenuated intensity

Fig. 5. Text detection results

was trained on examples from the 54 images from Essex (the `ifsofa` subset) and 200 images of licence plates. Compared to the preceding experiment, the neural network (again with 15 hidden nodes) has to select CSER corresponding to letters of much higher variability (different fonts, both handwritten and printed characters).

The text detector was tested on 150 images from the `ryoungt` subset of the Essex data. The false negative rate (missed text) was 8% and 0.45 false positives were detected per image. No post-filtering of the result with an OCR method was applied to reduced false positives. Examples of text detection on the ICDAR Essex set are shown in Figures 1 (top) and 5 (top row).

Further informal experiments were carried out to test insensitivity to lighting (Figure 1, center and bottom) and occlusion (Figure 5, bottom right). The image in the bottom left of Figure 5 includes two texts that have different scales and contrast; both are detected.

4.3 Face Detection

Extremal regions can be defined in colour images with respect to any ordering of RGB values. Intensity is just a particular scalar function that orders RGB values. This section describes and experiment, where human faces are detected as category-specific extremal regions. In this case, the scalar function is the likelihood ratio $\lambda(RGB) = P(RGB|skin)/P(RGB|non-skin)$. The assumption is that for a face there exists a threshold θ on the likelihood ratio λ separating the face from the background. As the enumeration of extremal region with respect to skin likelihood ratio $\lambda(RGB)$ proceeds, descriptors of the connected component are passed on to a classifier trained to detect face-like regions.

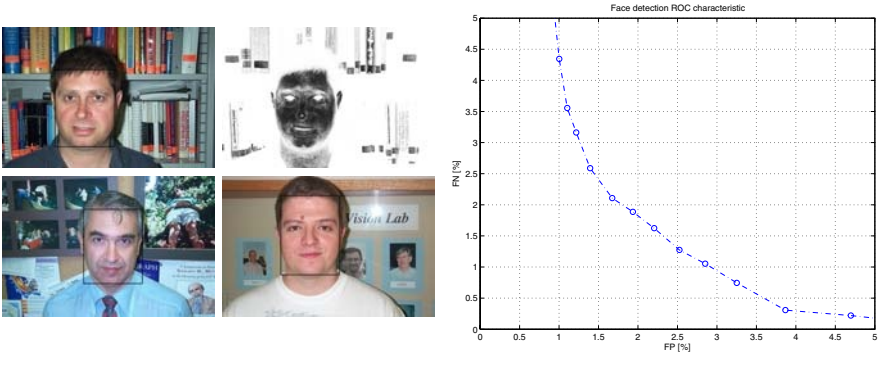

Fig. 6. Face detection:(a) Results and thresholding in the direction of skin probability, (b) ROC characteristic

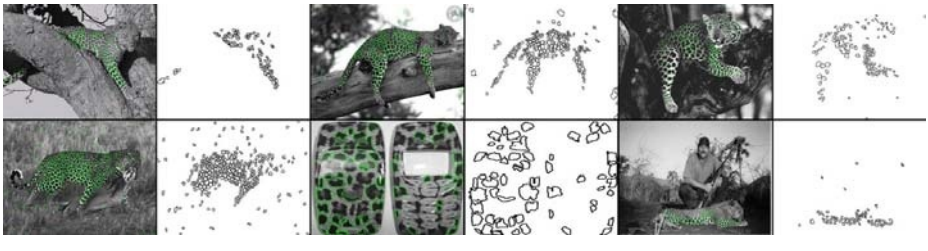

Fig. 7. Leopard skin detection; (b) sample results

The results on Caltech Human face (front) dataset are summarized by ROC characteristic in Fig.6, where false positive rate is normalized with respect to all extremal regions in the image. In this ROC curve at 99.5% detection rate, only 3.5% of windows have to be verified. The results present detector as rapid region pre-selector , i.e. weak classifier with false negative rate close to zero.

4.4 Leopard Skin Detection

The experiment on leopard skin detection shows whether CSERs can support detection of objects from the given category. We did not attempt to model the complex and flexible spatial configuration. The neural network was trained on spots from only four images. The spot-specific CSER detector than processed a number of images from the WWW. Sample results are shown in the Fig. 7. The density of CSER is high in the leopard skin area (skin-like area in the case of the mobile phone) and low elsewhere. The result suggest that learned CSER may be useful in viewpoint-independent texture detection.

5 Conclusions

We presented a new class of detectors that can be adapted by machine learning methods to detect parts of objects from a given category. The detector selects a category-relevant subset of extremal regions. Properties of extremal regions render the detector very robust to illumination change. Robustness to viewpoint change can be achieved by using invariant descriptors and/or by modeling shape variations by the classifier.

The detector was tested in three different tasks (e.g. text detection, face segmentation or texture detection) with successful results. The task of text detection presents affine and brightness invariance, the experiment of face detection introduces the ability of detector to process color images by thresholding in the learnable direction in RGB space and texture detection experiment demonstrates variability of the proposed detector.

The method can only detect regions that are extremal regions. It is not clear whether this is a significant limitation. Certainly many objects can be recognised from suitably locally threshold images, i.e. from extremal regions. Also note that different extremal sets can be defined by ordering pixels according to totally ordered quantities other than intensity, e.g. saturation. Efficiency of the method requires that the CSER are selected on the basis of incrementally computable features. This restriction can be overcome by viewing the interleaved classifier as a fast pre-selector in a cascaded (sequential) classification system.

Acknowledgements

Karel Zimmermann was supported by The Czech Academy of Sciences Multi-Cam project 1ET101210407 and Jiri Matas was supported by The European Commission under COSPAL project IST-004176.

References

1. L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV03*, pages 1134–1141, 2003.

2. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*, pages II: 264–271, 2003.
3. V. Ferrari, T. Tuytelaars, and L. Van Gool. Real-time affine region tracking and coplanar grouping. In *CVPR*, pages II:226–233, 2001.
4. V. Ferrari, T. Tuytelaars, and L. Van Gool. Wide-baseline multiple-view correspondences. In *CVPR*, 2003.
5. T. Kadir and M. Brady. Saliency, scale and image description. *IJCV01*, 45(2):83–105, 2001.
6. S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *ICCV03*, pages 649–655, 2003.
7. B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR03*, pages II: 409–415, 2003.
8. D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.
9. S. Lucas. Icdar03 text detection competition datasets. In *http://algoval.essex.ac.uk/icdar/Datasets.html*, 2003.
10. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC02*, volume 1, pages 384–393, London, UK, 2002.
11. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV01*, pages 525–531, 2001.
12. G. Mori, S. Belongie, and J. Malik. Shape contexts enable efficient retrieval of similar shapes. In *CVPR01*, pages I:723–730, 2001.
13. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *BMVC*, volume 1, pages 113–122, London, UK, 2002.
14. P. Pritchett and A. Zisserman. Matching and reconstruction from widely separated views. In *SMILE98*, 1998.
15. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, 1997.
16. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV03*, pages 1470–1477, 2003.
17. T. Tuytelaars and L. van Gool. Content-based image retrieval based on local affinely invariant regions. In *VIIS*, pages 493–500, 1999.
18. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV00*, pages I: 18–32, 2000.