A New Class of Learnable Detectors for Categorisation

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 neural network 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 modelling shape variations by the classifier.

The approach is brought to bear on three problems: license plate detection, text segmentation and leopard skin detection. High detection rates were obtained for both license plate detection (98%) and text detection (92%). In the license plate experiment, test views included 25-fold change of scale and views from acute angles.

The time-complexity of the detection is approximately linear in the number of pixels in the input image and a non-optimized implementation runs at about 1 frame per second for a 640x480 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, 9, 13, 6], wide-baseline stereo [14, 17, 5, 10, 11], tracking [4], categorisation [16, 2, 8] and texture recognition [7]. 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.

In categorisation, the problem we focus on, state-of-the-art approaches represent categories as probabilistic configurations of classified transformation-covariant regions [3, 16, 8, 12]. The (soft) classification of the 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.[3] and Fei-Fei et al.[2] or affine-invariant interest points[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, 19] and to train classifiers for recognition of regions — components of the configuration [3, 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 license plate recognition.

The proposed category-specific class of detectors is trained to select a relevant subset of extremal regions. The set of extremal regions is the union of connected components of binary images obtained by thresholding (the union is over all threshold levels). The selected subsets inherit properties of the set of extremal regions that support robust and invariant detection [10]: the set is closed under monotonic transformation of intensity and under coordinate transformations that are diffeomorphisms (a class that includes homography and affine transformation). Moreover, the number of extremal regions is not grater than the number of pixels in the image and an efficient algorithms exist for their enumeration.

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.

Figure 1: Text detection based on category-specific extremal regions.

In the literature, one particular subset of extremal regions, the maximally stable extremal regions (MSER), has been used for object recognition and categorisation [16, 13]. MSERs are extremal regions that stay virtually unchanged (i.e. 'are stable') over a range of thresholds. Roughly speaking, MSERs are connected components separated from the rest of the image by a range of intensities. In the presented work, the requirement of stability over a range of thresholds is removed and relevant extremal regions are selected on the basis of their *shape*. Therefore a *single* threshold separating a region of a given shape is sufficient to detect a category-specific extremal region (CSER). As a consequence, the set of CSERs is virtually unaffected by severe illumination changes. The property is demonstrated in Figure 1. Three 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 three 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.

The rest of the paper is organised as follows. First, the structure of the algorithm for category-specific extremal region detection is presented. We show that CSERs are efficiently selected by interleaving enumeration of extremal regions and classification of

their incrementally computable features. The class of incrementally computable features is studied next, necessary conditions for the class are found and examples of such features are given (Section 2.1). We then apply the method to two well know problems of text detection and license plate recognition (Section 3). The flexibility of the framework is tested on an unrelated 'toy' problem - detection of leopard skin. The paper is summarised in Section 4.

2 Category-specific extremal region detection

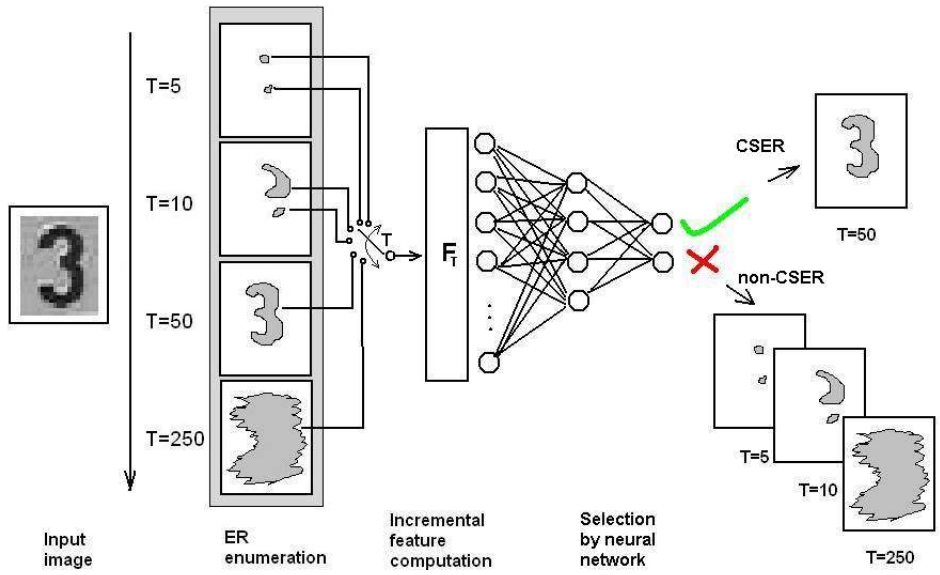


Figure 2: The detection is implemented as interleaved enumeration of extremal regions, computation of incremental features and classification.

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. it 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 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 pixel. 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 2.1.

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 relevant 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.

At this point it is interesting to compare the proposed CSER detection process with the seminal face detection method of Viola and Jones [18]. Viola and Jones use cascaded AdaBoost to classify (module sub-sampling) every rectangular window of a predetermined size using features computed in constant time from the integral image. There are strong analogies. In both cases, the number of classifications made is equal to the number of pixels in the image (which is equal both to the number of rectangular windows of fixed size and the number of extremal regions). In both cases, features describing the classified regions are computed in constant time. In the Viola-Jones approach the assumption is that the object from the category (faces) are well represented in a rectangular window. In our case, the assumption is that the category of interest has components that are extremal regions. The difference in the adopted classifier is superficial. The CSER can be characterised by an AdaBoost-trained cascaded classifier instead of the adopted neural network.

2.1 Incrementally Computable Region Descriptors

In the CSER detection process, we are given two or more disjoint regions r_1 and r_2. By marking a pixel in the image, these regions merge to form a new extremal region. The new region is the union of $r_1 \cup r_2$ (we use $r_\cdot$ 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

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

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 $g : 2^{Z^2} \to \mathscr{R}^m$, a characterisation update function $f : (\mathscr{R}^m, \mathscr{R}^m) \to \mathscr{R}^m$, and a feature computation function $\phi : \mathscr{R}^m \to \mathscr{R}^n$, 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.

3 Experiments - Applications of CSER detection

3.1 License plate detection

At least in constrained conditions, license plate detection, as demonstrated e.g. by the London congestion charge system, is more an engineering than a research problem. Here

we demonstrate that an unconstrained license plate detector is developed easily (and without ad hoc techniques) using CSERs. By 'unconstrained license plate detector' we mean viewpoint and illumination independent and robust to occlusion.

The category of license plates is modelled as a linear constellation of CSERs. Information about the rectangular shape of the place as a whole is not exploited. 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 to 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. Counterexamples were obtained by ten rounds of bootstrapping. In each round, the CSER detector processed 250 training images and the false positives served as negative examples in the next round of training.

The detection of license plates proceeds by in two steps. First, relevant CSERs are selected as described in Section 2. Second, linear configurations of regions are found by Hough transform. We impose two constraints on the configurations: it must be formed from more than three regions and the pixels in the regions involved must have a similar maximum distance from the baseline of the linear configuration.

Detection Rate. On an independent test set of 70 unconstrained images of scenes with license plates the method achieved 98% detection rate with a false positive appearing in approximately 1 in 20 images. Example of detected license plates and the type of data processed are shown in Figure 3.

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 license plates differed by a factor of 25 (character 'heights' ranged from approximately 7-8 to 150 pixels) and the plates were viewed both frontally and at acute angles, see Figure 3. We also performed systematic evaluation of the CSER detector. Images of license plates were warped (see Figure 4b) to simulate a view from a certain point on the viewsphere. The false negative rates for the CSER detector (missed character percentages) are shown in Table 4a. The CSER detector is stable for almost the whole tested range. Even the 27% false negative rate at the 30^o-45^o elevation-azimuth means that approximately three quarters of characters on the license plate are detected on average - the probability of detecting the whole plate is still high.

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 5, the false negative (left) and false positive (right) rates were unchanged for both the detector of CSER (bottom) and whole license plates (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 as demonstrated in Figure 3b is a consequence of modelling the object as a configuration of local component. Occlusion of some components does not imply the object is not detected.

Figure 3: License plate detection in unconstrained conditions.

$\theta \backslash \phi$	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

(a)

(b)

Figure 4: (a) False negative rate (missed CSER on license plates) as a function of viewing angles ϕ (elevation) , θ (azimuth); in percentage points. (b) An Example of a synthetically warped license plate to ϕ, θ equal to $(0^o, 0^o), (0^o, 45^o), (30^o, 0^o)$ and $(30^o, 45^o)$.

3.2 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 [1].

An object from the 'text category' was modelled as an approximately linear configuration of at least three 'text-like' CSERs. The neural network selector was trained on examples from the 54 images from Essex (the `ifsofa` subset) and 200 images of license plates. Compared to the license plate 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. Most of the false positives appeared in areas of repetitive image structure with character-like regions (e.g. closely spaced windows or I-shaped parts of fences). No post-filtering of the result with an OCR method was applied to reduce false positives. Given the linear configuration, it is easy to compensate for local affine distortion and apply standard OCR techniques. Examples of text detection on the ICDAR Essex set are shown

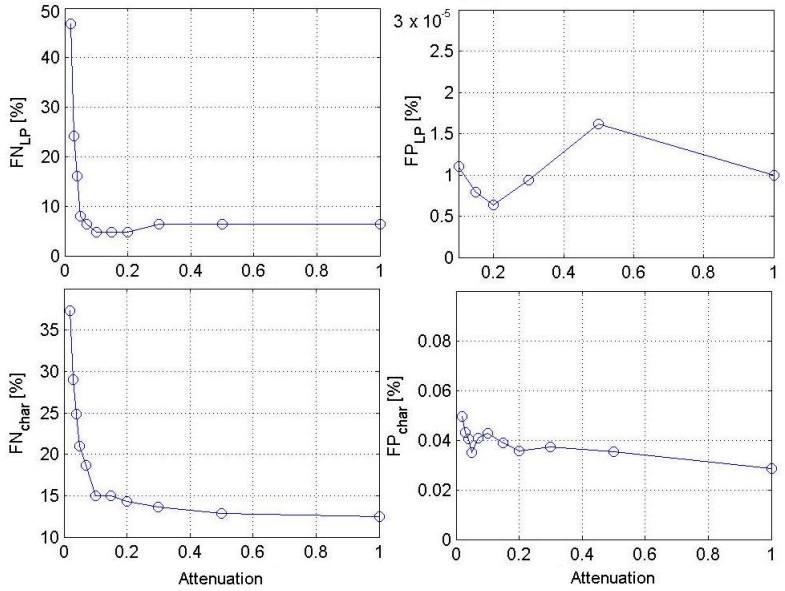


Figure 5: License plate detection in images with attenuated intensity.

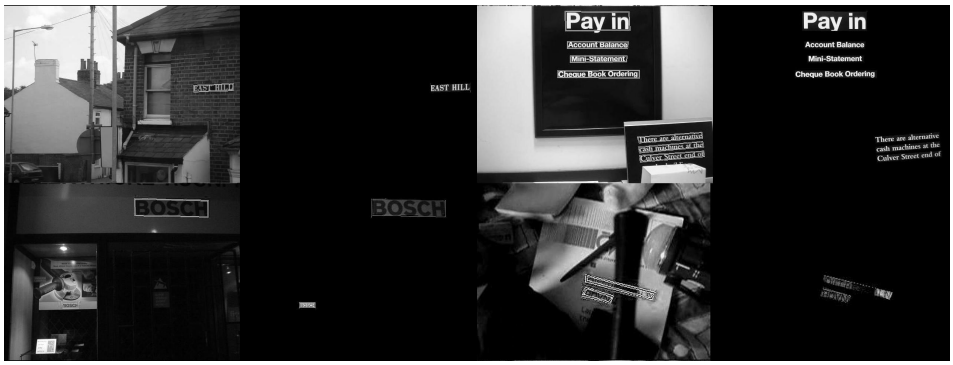
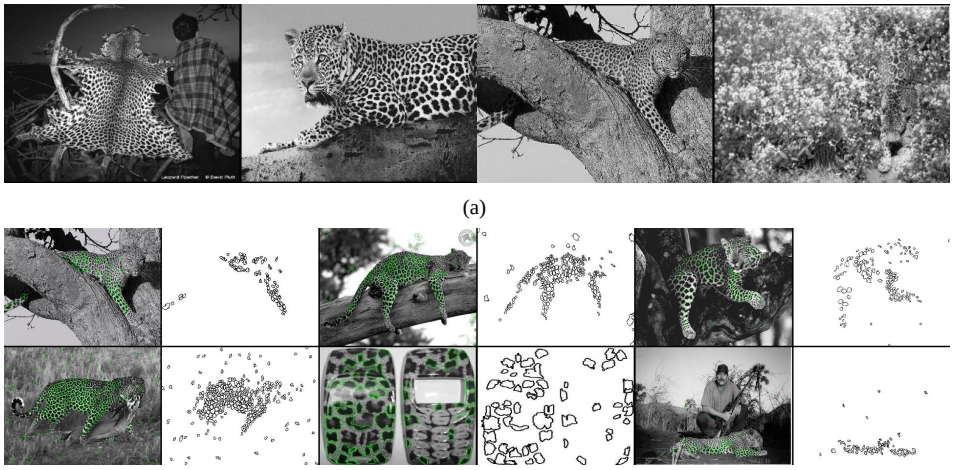


Figure 6: Text detection results

in Figures 1 (top) and 6 (top row).

Further informal experiments were carried out to test insensitivity to lighting (Figure 1, center and bottom) and occlusion (Figure 6, bottom right). The image in the bottom left of Figure 6 includes two texts that have different scales and contrast; both are detected.



(a)

(b)

Figure 7: Leopard skin detection; (a) the training set and (b) sample results.

3.3 Leopard skin detection

The experiment on leopard skin detection shows whether CSERs can support detection of objects from this category. We did not attempt to model the complex and flexible spatial configuration. The neural network was trained on spots from only four images (Fig. 7, top row). The spot-specific CSER detector than processed a number of images from the WWW. Sample results are shown in the bottom row of 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 suggests that learned CSER may be useful in viewpoint-independent texture detection.

4 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. The approach was tested on three problems: license plate detection, text segmentation and leopard skin detection. High detection rates were obtained for both license plate detection (98%) and text detection (92%). In the license plate experiment, test views included 25-fold change of scale and views form acute angles.

The time-complexity of the detection is approximately linear in the number of pixel and the current implementation [2] runs at about 1 frame per second for a 640*x*480 image on a high-end PC.

The method can only detect subset of extremal regions. It is not clear whether this is a significant limitation. Certainly many objects (e.g. faces) can be recognised from suitably

[2] For a straightforward code without careful optimization.

locally thresholded 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 CSERs 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.

References

[1] ICDAR03 text detection datasets. In `http://algoval.essex.ac.uk/icdar/Datasets.html`.

[2] 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.

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*, pages II: 264–271, 2003.

[4] V. Ferrari, T. Tuytelaars, and L. Van Gool. Real-time affine region tracking and coplanar grouping. In *CVPR*, pages II:226–233, 2001.

[5] V. Ferrari, T. Tuytelaars, and L. Van Gool. Wide-baseline multiple-view correspondences. In *CVPR*, 2003.

[6] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV01*, 45(2):83–105, 2001.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *ICCV03*, pages 649–655, 2003.

[8] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR03*, pages II: 409–415, 2003.

[9] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

[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, 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, 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] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.

[19] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV00*, pages I: 18–32, 2000.