Segmentation-based multi-class semantic object detection

Remi Vieux, Jenny Benois-Pineau, Jean-Philippe Domenger, Achille Braquelaire

Noname manuscript No.
(will be inserted by the editor)

Segmentation-based Multi-Class Semantic Object Detection

Remi VIEUX · Jenny BENOIS-PINEAU ·
Jean-Philippe DOMENGER · Achille BRAQUELAIRE

Received: date / Accepted: date

Abstract In this paper we study the problem of the detection of semantic objects from known categories in images. Unlike existing techniques which operate at the pixel or at a patch level for recognition, we propose to rely on the categorization of image segments. Recent work has highlighted that image segments provide a sound support for visual object class recognition. In this work, we use image segments as primitives to extract robust features and train detection models for a predefined set of categories. Several segmentation algorithms are benchmarked and their performances for segment recognition are compared. We then propose two methods for enhancing the segments classification, one based on the fusion of the classification results obtained with the different segmentations, the other one based on the optimization of the global labelling by correcting local ambiguities between neighbor segments. We use as a benchmark the Microsoft MSRC-21 image database and show that our method competes with the current state-of-the-art.

Keywords Object Detection, Segmentation, Relaxation Labelling, Late Fusion, SVM

1 Introduction

The object detection and categorization aims to extract two kinds of information from the images: which objects are present in the image and where exactly are those objects located. The detection is closely linked to the task of object segmentation. One of the difficulties of object segmentation is to extract complex, highly structured objects composed of regions visually dissimilar. On the other hand, the recognition of objects is affected by several problems: low quality of images, partial occlusions, visually heterogeneous object categories, cluttered background, ... The discrepancy between the low level features that can be extracted from the images and the high level semantic interpretation of images is known as the semantic gap [28]. Many research efforts have been carried in the task of global image recognition such as scene classification [29,13,35], but object recognition remains a challenge. This task is even more ambitious when multiple objects are present in the images.

Remi Vieux
LaBRI CNRS UMR 5800
Tel.: +33540003880
E-mail: firstname.lastname@labri.fr

One of the earliest work closely linked to the object detection and recognition topic was proposed by Duygulu *et al.* [9]. In this work, the authors address the recognition problem as the process of attaching words to image segments considering the task as a translation between one language (English words) to another (visual words, or *blobs*). A mapping between the keywords and the visual blobs is performed using a method based on Expectation Maximization. The rest of the literature [15,26,32,12,2] noticeably differs from the original work by Duygulu *et al.* in the sense that the models built try to exploit the maximum of information that can be extracted from the image: not only low level features (color, texture, *etc.*), but also local contextual relationships between pixels or image segments, location and even global relevance estimates. He *et al.* [15] proposed a pixel-wise labelling into a finite set of labels using a multiscale conditional random field formulation(mCRF). The mCRF combines the output of a local pixel-wise classification, a representation of the local geometric relationships between the objects and global label features into a probabilistic framework, trained using supervised labeled image data. Shotton *et al.* [26] introduced a new approach for learning a discriminative model of object classes based on texture, layout and context information. Their model is based on *texture-layout filters*, a feature which jointly model patterns of texture and their spatial layout. Texture layout filters are combined with lower level image features (color, location) into a Conditional Random Field to provide a high-level discriminative model. They also discuss an efficient learning method of the CRF parameters based on boosting and piecewise training method. Verbeek and Triggs [32] point out that Markov Random Field and aspect models such as Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation are complementary methods which attempt to improve the coherence of the labellin of image patches. MRF provides crisper local labelling by exploiting neighbourhood-level couplings while PLSA and LDA use global relevance estimates. They studied two spatial extensions of the aspect models, one based on a forest of minimal spanning trees and the other one on a regular 8-neighbor MRF. Galleguillos *et al.* [12] have shown that introducing contextual information about the co-occurrences and the relative location of image regions with local appearance-based features improves the global labellin. Athanasiadis *et al.* [2] define a framework for simultaneous image segmentation and object labellin operating at the semantic level. They represent the contextual information as an ontological taxonomy of the set of possible semantic labels and employ fuzzy algebra to adjust the labelling of the regions given by region growing segmentation algorithms.

Our approach has been inspired by the notion of superpixels in image segmentation given by Ren and Malik [24]. As opposed to pixels which are an unnatural arbitrary grid sampling of an image into a very large number of entities, superpixels provide a larger, locally homogeneous and coherent regions that preserve most of the structure necessary for accurate segmentation. We take up the idea that an image over-segmentation is sufficient to perform efficient recognition of small objects or object parts. The visual descriptors extracted at the region level are statistically more robust to noise than descriptors extracted at the pixel level. We will show that a Support Vector Machine classifier trained in an appropriate descriptor space for representing the segments already yields reasonable performances. A similar approach of using image segments was proposed by Hoeim [16] for the recovery of geometry context from single images. Yang *et al.* integrated appearance based classification of meanshift patches, a bag-of-keypoints model and global shape based Elliptical Fourier descriptors within a unified framework for object-based segmentation. Gould *et al.* proposed to combine appearance-based features computed on superpixels patches with relative location priors in a two stage classification process [14]. Malisiewicz and Efros have

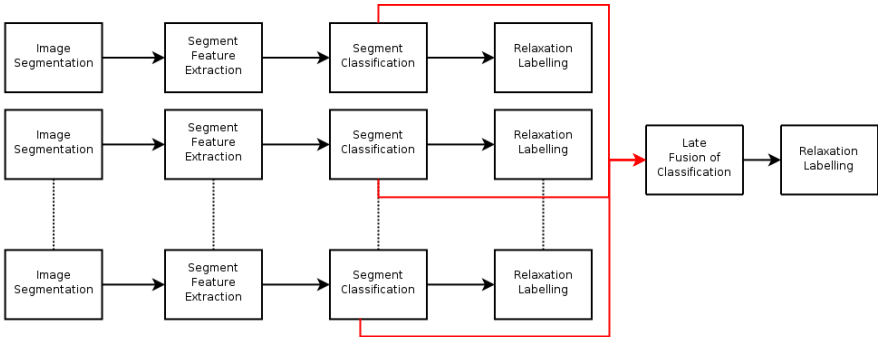


Fig. 1 Object detection and categorization framework.

Fig. 2 MSRC-21 class database, images with their corresponding pixel-wise labelling (best viewed in color).

also shown that using image segmentation was efficient to improve the spatial support for object detection and recognition [18].

In 1976, Rosenfeld *et al.* have formulated a process to reduce the ambiguities in object identification based on the relationships between the objects [25]. They proposed an ambiguity-reduction process by iterated parallel relaxation operations known as Relaxation Labelling. This approach was widely studied by the communauty in the context of graph matching for object recognition in videos [6]. We show that relaxation labelling improves the global coherence of the segment classification and provides cleaner, regularized object extraction. Our framework for object detection and categorization is depicted in figure 1. Each row going from image segmentation to relaxation labelling describes the different steps to achieve semantic object extraction using a single segmentation method. The image is first segmented, feature descriptors are extracted on the segments and the classification of the segment into one of the semantic categories is performed based on these descriptors. The initial classification is further refined with the relaxation labelling algorithm. As the figure shows, it is possible to run the framework using multiple segmentation algorithms as the initial preprocessing step. When multiple segmentations and detection are available, it is possible to combine them using a late fusion schema, which is depicted by the red arrows going from the several segment classification boxes to the late fusion box. Again, it is possible to apply the relaxation labelling at the very end on the combined results as a last regularization step.

In this paper, we provide a detailed contribution for each step of the framework. In order to gain insights about the impact of image segmentation on the segment recognition performances, we applied our framework using four segmentation algorithms which will be introduced in section 2. In section 3, we introduce an efficient low level feature representation of the image segments. We propose to model the segments using color, texture and

keypoint-based descriptors. For the latter, we propose a new way of sampling the keypoints locations to obtain a robust keypoint-based description of the segments. In the same section, we discuss the segment classification method focusing on how to obtain a multi-class probability estimates for each segment using Support Vector Machines (SVM) classifiers. Then, we present our methods for improving the recognition results: the relaxation labelling algorithm is described in section 4, while the late fusion strategies are introduced in section 5.

In section 6 we perform the quantitative evaluation of our approach and compare it with the current state-of-the-art. The experiments are carried using the MSRC-21 class database, a collection of 591 pictures of 21 object categories taken under different viewpoint and illumination conditions. Each image has a corresponding pixel-wise segmentation ground truth enabling training and evaluation (see figure 2).

2 Image segmentation

There exists a tremendous quantity of image segmentation approaches [11,21]. Our goal in this paper is not to propose a new segmentation algorithm, but to study the influence of the segmentation as a preprocessing step in the object detection and recognition pipeline. Hence, we apply our framework using four segmentation algorithms, each of them providing segmentations of different granularity (*i.e.* in terms of the number of segments produced) and accuracy (*i.e.* segments boundaries located at the actual object boundaries).

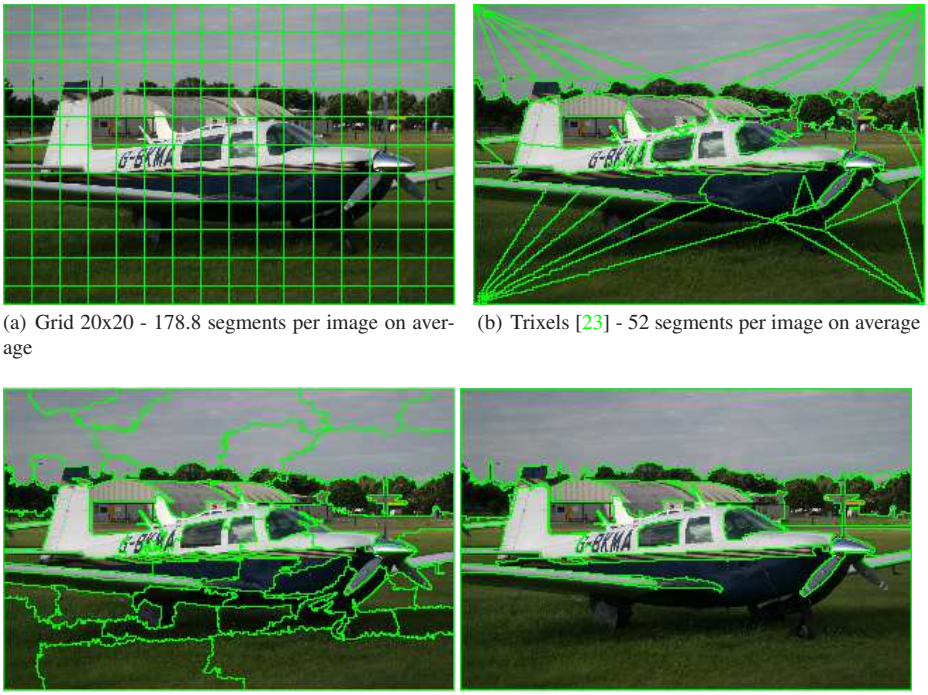
The first segmentation is a simple partitioning of the image into regular square patches of size 20×20 pixels. The size of the patches was chosen following the approach of Verbeek and Triggs [32]. This produce an average of 178 segments per image using the MSRC dataset, and the segmentation accuracy is low. An example is given in figure 3(a). In the following we will refer to this segmentation as the grid segmentation, or gridseg.

The second segmentation algorithm is the dual contour/region segmentation of Prasad and Skourikine [23](figure 3(b)). The approach consists in an edge detection step followed by a Constrained Delaunay Triangulation (CDT) to close the region contours. The CDT ensures that the segments should not cross the actual objects boundaries, assuming that the edge detection is accurate. However, false edges are likely to appear in highly textured areas hence producing a high number of small triangular segments. On the other hand, low contrast edges can be missed and some segments can overlap between different objects (as shown in figure 3(b) below the wing of the airplane). The average number of segments produced by this method is 52, and the segmentation accuracy is slightly better than the grid. We will refer to this segmentation as the triangle segmentation.

The third segmentation was obtained using the synergistic segmentation of Christoudias *et al.* [7] (figure 3(c)). The algorithm is an improvement of the mean-shift color segmentation [8] with the addition of an edge confidence parameter [19] in the feature space. The introduction of the edge confidence makes the segmentation more robust for low contrast region boundaries than the traditional mean-shift segmentation. We used the EDISON software package[1] with the default parameters value, and obtained an average of 45 highly accurate segments. We will refer to this segmentation as the edison segmentation.

The last segmentation algorithm is the one proposed by Felzenszwalb and Huttenlocher [10] (figure 3(d)). It is based on the detection of boundaries between regions using a graph-based representation of the image. This algorithm is fast and has the ability to preserve detail

[1] http://www.caip.rutgers.edu/riul/research/code/EDISON/index.html



(a) Grid 20x20 - 178.8 segments per image on average

(b) Trixels [23] - 52 segments per image on average

(c) Synergistic Segmentation [7] - 45.3 segments per image on average

(d) Graphseg [10] - 21.1 segments per image on average

Fig. 3 Examples of the different segmentation methods (best viewed in color).

in low variability regions while ignoring details in high variability regions. Again, we kept the default parameters value for segmentation and obtained an average of 21 segments per image. Since the number of segments is much lower than with the other algorithm, it is likely that some segments overlap different objects. However, the accuracy of the segmentation is usually good. We will refer to this segmentation as the graph-based segmentation or graphseg.

3 Region Classification

3.1 Description Space

The description space has to be rich enough to enable the computation of models for the recognition of image segments. We have based our description space on the combination of colour, texture and keypoint-based representation of the segment content. The shape of the segments is not considered as it is not robust with respect to the segmentation process, in the case of partial occlusions or in the case of complex objects made of different parts, and is also irrelevant for some object classes (*e.g.* sky, grass, water). We based the computation of the feature vectors of segments on the *bag of visual words* approach that was popularized

by Sivic and Zisserman [27]: for each descriptor, we compute a codebook (or dictionary) of representative elements using k-means++ [1] clustering. Then, the descriptors computed on the image segment are matched with their closest representative in the codebook. The segment feature vector is the statistical distribution of the segment descriptors according to the codebook.

The content of a segment is described by a color feature vector $\mathbf{x}_\mathfrak{c} = \{x_{\mathfrak{c}1}, x_{\mathfrak{c}2}, \ldots, x_{\mathfrak{c}|\mathfrak{C}|}\}$ where $|\mathfrak{C}|$ is the dimensionality of the color codebook and $x_{\mathfrak{c}i}$ is the statistical frequency of the i^{th} codework computed on the pixels belonging to the segment. We fixed the size of the color codebook to $|\mathfrak{C}| = 100$ representative colors computed on the training images in the HSV color space.

We compute a texture feature vector $\mathbf{x}_\mathfrak{t} = \{x_{\mathfrak{t}1}, x_{\mathfrak{t}2}, \ldots, x_{\mathfrak{t}|\mathfrak{T}|}\}$ following the statistical approach to material classification using image patch exemplars by Varma and Zisserman [31]. The texture at a pixel is considered as a vector of grey levels of a 3x3 neighborhood centered on this pixel. We extract the texture around each pixel belonging to the segment, and match them to a texture codebook of size $|\mathfrak{T}| = 100$, computed on the training images.

Finally, we compute a *keypoint-based* feature vector $\mathbf{x}_\mathfrak{k} = \{x_{\mathfrak{k}1}, x_{\mathfrak{k}2}, \ldots, x_{\mathfrak{k}|\mathfrak{K}|}\}$. The keypoint based representation of image content is now undoubtedly one of the most effective approaches for the detection of semantic concepts [17]. The keypoint-based representation relies on two steps, the localization of interest point and the computation of a feature descriptor located at these interest points. In our case, relying on the location of interest points introduces the potential drawback that some segments might not yield enough interest points for a statistically robust description, not to say might not yield any interest point at all. However, Nowak *et al.* [20] have shown that random sampling of interest points gives equal or better classifiers than sophisticated multiscale interest point detectors for a moderate to large number of samples. Rather than random sampling the interest points, we sample them on the segments boundaries, which provides a good combination of the advantages of random and interest-point based sampling. Indeed, the salient point in images are most likely to occur on strong edges or corners, which correspond the objects boundaries extracted with an accurate segmentation. At the same time, it is possible to compute much more feature descriptors on the segment boundaries to obtain a dense, statistically more robust description of the segment. We have used the SURF feature descriptors of Bay *et al.* [4] for their low computational burden yet efficient discriminative power. Again, the size of the codebook of SURF descriptors was fixed to $|\mathfrak{K}| = 100$, and the dictionnary was computed by clustering the SURF descriptors extracted on the training set images.

The overall feature vector for a segment that will be used to train the models is the concatenation of color, texture and keypoint-based feature vectors $\mathbf{x} = \{\mathbf{x}_\mathfrak{c} \vee \mathbf{x}_\mathfrak{t} \vee \mathbf{x}_\mathfrak{k}\}$. This is the simplest cases of early fusion of features.

3.2 Segments Classification

SVM have known an increasing popularity during the last 15 years since their introduction by Vapnik [30], due to their solid theoretical background, good practical results and their ability to deal with high dimensional data. To train an SVM model it is necessary to provide a representative set of training samples. We are working here on the problem of multi-class classification. Let us denote the training set $\mathscr{S} = \{(\mathbf{x}_k, \Lambda)|k = [1, \ldots, K]\}$, with $\mathbf{x}_k \in \mathbb{R}^n$ the feature vectors and $\Lambda \in \{\lambda_1, \ldots, \lambda_M\}$ the sample class. Given a set of training images we compute the training set $\mathscr{S}$ by assigning a label to every segments yielded by the segmentation algorithm. A segment is assigned label λ if more than 80% of its area is covered by

pixels belonging to the class λ according to the ground truth segmentation. If no class reach this threshold, the segment is assigned to the class **void**.

We aim to get a probability estimate of the segment labelling given the whole set of possible classes. SVM have been initially formulated as binary discriminative classifiers with decision function $\hat{f}(\mathbf{x})$ such that sign $\hat{f}(\mathbf{x})$ is used to predict the label of any test example $\mathbf{x}$. Platt [22] formulated an approximation of the posterior class probability $p(y=\lambda_l|y=\lambda_l or \lambda_m,\mathbf{x})$ by a fitting a sigmoid function (equation 1):

$$p(y=\lambda_l|y=\lambda_l \text{ or } \lambda_m,\mathbf{x}) \approx \frac{1}{1+\exp^{A\hat{f}+B}} \tag{1}$$

A and B are estimated by minimizing the negative log-likelihood function from the known training data and their decision values.

A binary SVM model is trained with each pair of classes. Given the binary probability estimates $r_{ij} = p(y=i|y=i \text{ or } j)$, Wu *et al.* formulate the multi-class probability estimates as the optimisation of equation 2 [34]:

$$\min_{\mathbf{p}} \frac{1}{2} \sum_{m=1}^{M} \sum_{l:l\neq m} (r_{lm}p_m - r_{ml}p_l)^2 \tag{2}$$

subject to

$$\sum_{m=1}^{M} p_m = 1, p_m \geq 0, \forall m \tag{3}$$

Hence, for each segment characterised by its feature vector $\mathbf{x}$, we obtain from the SVM model a multi-class probability estimate vector $\mathbf{p} = \{p_{\lambda_1},\ldots,p_{\lambda_M}\}, \sum_{m=1}^{M} p_{\lambda_m} = 1$ where $p\lambda_i = p(y=\lambda_i|\mathbf{x})$.

The whole process of binary SVM classification, binary probability estimates and multi-class probability estimates is implemented in the software package LibSVM [5].

4 Relaxation Labelling

The class probabilities p_λ estimated with SVMs are inferred using only local features characterizing the segment content. Each segment is classified independently. We now want to increase the coherence of the global image labelling taking into account the relationships between the segments. Let us denote $p_i(\lambda)$ the probability that segment i takes the label λ. Given the initial probabilities guess $p_i^{(0)}(\lambda)$, we can iteratively update the class probabilities of the segments using the relaxation labelling equations of Rosenfeld *et al.* [25] :

$$p_i^{(k+1)}(\lambda) = \frac{p_i^{(k)}(\lambda)[1+q_i^{(k)}(\lambda)]}{\sum_\lambda p_i^{(k)}(\lambda)[1+q_i^{(k)}(\lambda)]} \tag{4}$$

$$q_i^{(k)}(\lambda) = \sum_j d_{ij}\Big[\sum_{\lambda'} \eta_{ij}(\lambda,\lambda')p_j^{(k)}(\lambda')\Big] \tag{5}$$

$\eta_{ij}(\lambda,\lambda')$ is called the affinity coefficient between labels λ and λ', d_{ij} are coefficients satisfying $\sum_j d_{ij} = 1, \forall i$. The j denotes the set of segments interacting with segment i. We considered it as the set the direct neighbors of the segment. d_{ij} weights the contribution of the interactions between segments i and j. We set $d_{ij} = \rho(i,j)/\rho(i)$ with $\rho(i,j)$ the length of the common boundary between segments i and j and $\rho(i)$ the total length of the boundary

of segment i, so that segments sharing a large portion of boundary have a stronger influence on each other than segments sharing a small boundary. Note that $d_{ij} \neq d_{ji}$.

Rosenfeld suggests to compute the affinity coefficient $\eta_{ij}(\lambda,\lambda')$ as the correlation coefficient between the events that segment i has label λ and segment j has label λ' [25] (equation 6, 7).

$$cor(\lambda,\lambda') = \frac{cov(\lambda,\lambda')}{\sqrt{\left[(p(\lambda)-p^2(\lambda))(p(\lambda')-p^2(\lambda'))\right]}} \tag{6}$$

$$cov(\lambda,\lambda') = p(\lambda,\lambda') - p(\lambda)p(\lambda') \tag{7}$$

Hence $q_i^{(k)}(\lambda)$ gives the way we want to change the labelling probabilities $p_i^{(k)}(\lambda)$, increasing highly correlated pairs of labels and decreasing uncorrelated ones. We can easily estimate the covariances from the segmented training database, as depicted in equation 8.

$$p(\lambda) = \frac{|R_\lambda|}{|R|},\ p(\lambda,\lambda') = \frac{|\mathscr{R}_{\lambda,\lambda'}|}{|\mathscr{R}|} \tag{8}$$

with $|R_\lambda|$ the number of segments labeled λ, $|R|$ the total number of segments, $|\mathscr{R}_{\lambda,\lambda'}|$ the number of pairs of adjacent segments having respective labels λ and λ' and $|\mathscr{R}|$ the total number of pair of adjacent segments.

5 Late fusion of classifiers for segment labelling

In section 3, we saw how to obtain probability estimates of classes for each segment based on their local appearances. In section 2, we have introduced different segmentation algorithms for which the detection and recognition performances are evaluated in section 6. Hence, multiple segmentations with classified segments are available for the same image. A natural and seducing idea is to perform a late fusion of the classification results. Our fusion strategy is straightforward: first, we compute a new image segmentation mask from the the previously computed segmentations. This new segmentation is a partition of the image such that each segment S_i in this new segmentation is an intersection of the previous segments. An example of such a segmentation is shown in figure 4, where the segmentations of figure 3 are merged.

The segmentation hence obtained is more detailed than any of the previous ones, each segment having a one-to-many relationship with the original segments. Then, we assign the multi-class probabilities to the new segments. To do so, we combine the probabilities estimates obtained by the single segmentation detections using the max (equation 9), mean (equation 10) or multiplication (equation 11) operators.

$$p_{\lambda_m} = \frac{1}{Z}\max_{s\in S} p^s_{\lambda_m} \text{ with } p^s_{\lambda_m} \text{ the probability obtained using segmentation } s. \tag{9}$$

$$p_{\lambda_m} = \frac{1}{Z}\sum_{s\in S} p^s_{\lambda_m}/|S| \text{ with } |S| \text{ the number of segmentations/detections combined} \tag{10}$$

$$p_{\lambda_m} = \frac{1}{Z}\prod_{s\in S} p^s_{\lambda_m} \tag{11}$$

with Z a normalization factor to ensure that the probabilities sum to 1: $Z = \sum_{m=0}^{M} p_{\lambda_m}$

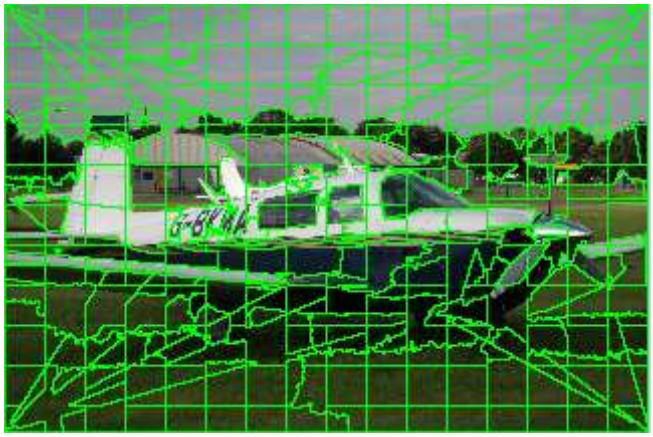


Fig. 4 Example of the fusion of the segmentations shown in figure 3. Very small regions (less than 50 pixels) are fused their biggest adjacent regions to avoid unnecessary isolated pixels.

This fusion strategy is different from the late fusion of SVM-Based classifiers for semantic indexing proposed by Ayache *et al.* [3]. In a classical late fusion schema, the single classification results are combined, *e.g.* using yet another classifier trained on the classification outputs of the single classifiers. Here, we take advantage of the probability estimates of the single classifiers and the final decision is a trade-off between the single estimates.

6 Experiments

The goal of this section is to assess the impact of the following parameters in the quality of the object detection and categorization:

- The influence of the quality of the segmentation algorithm
- The benefits of the neighborhood regularization method relaxation labelling
- The interest of late fusion of multiple classification outputs using different segmentation support

The experiments were conducted on the MSRC-21 class image database. We split the database in two halves for training and testing. Following the protocol of the other works [26,33,14,36], we ignored the **void** class during training and evaluation, and horse and mountain classes were also ignored due to insuffiencient amount of representative in the database. The reported results are the pixel-wise accuracy, that is the percentage of pixels correctly classified with respect to the total number of pixels in the segmentation ground truth.

6.1 Impact of the segmentation algorithm

Figure 5 shows the overall pixel-wise accuracy obtained using the different segmentation algorithms presented in section 2. The best results are achieved using Edison, which clearly indicates the benefits of an accurate segmentation for the recognition. This is emphasized by

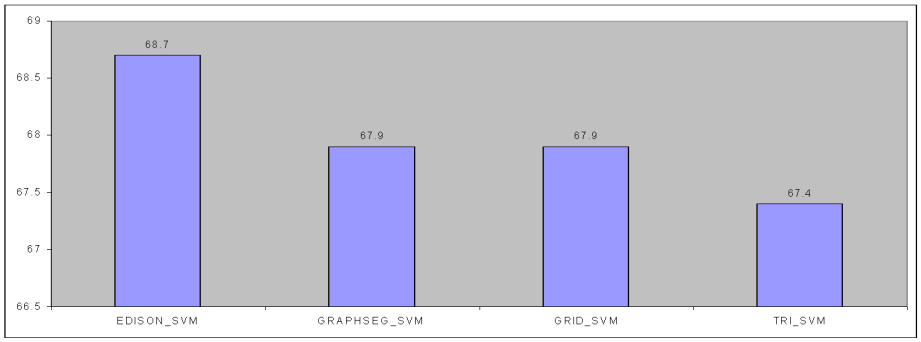


Fig. 5 Pixel-wise accuracy for the direct SVM segment classification obtained with the different segmentation algorithms.

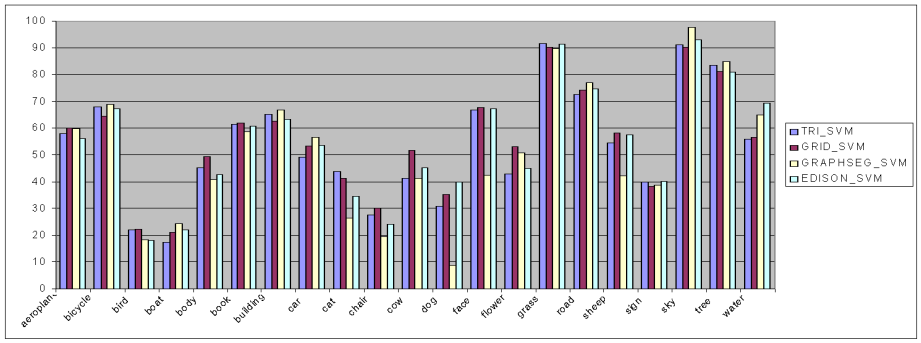


Fig. 6 Pixel-wise classification accuracy of direct segment classification using various segmentation algorithms (best viewed in color).

the fact that the results obtained by the triangle segmentation, despite the approximate same number of segments per image, are much lower. Graphseg and grid segmentation lead to the same overall pixel-wise accuracy, despite their very different spatial support (large accurate segments versus small, arbitrary patches). It seems that the very limited view of the objects provided by the small square patches is still sufficient for recognition. It is interesting to see the differences of results at the object class level, as shown in figure 6. We can see that no segmentation yields better results for all the classes, but significant differences among the classes appears. For example, we obtain much worse classification results for categories dog and sheep using the graph-based segmentation than using the other algorithms, while the best results are obtained with this segmentation for the categories sky, building and cars. It is hard to tell from these results if one algorithm is particularly suited for some kind of object classes. However, an interesting result to note is the fact that good performances are obtained for some categories (body cow and sheep) with the grid segmentation while we would have expected that accurate segmentation would help for their recognition. The results obtained with the graph-based segmentation seems to suffer a lack of robustness in the recognition, since for some classes the results are very low compared with the other segmentation. The difference in recognition among the classes is due to the fact that the *easiest* classes are the

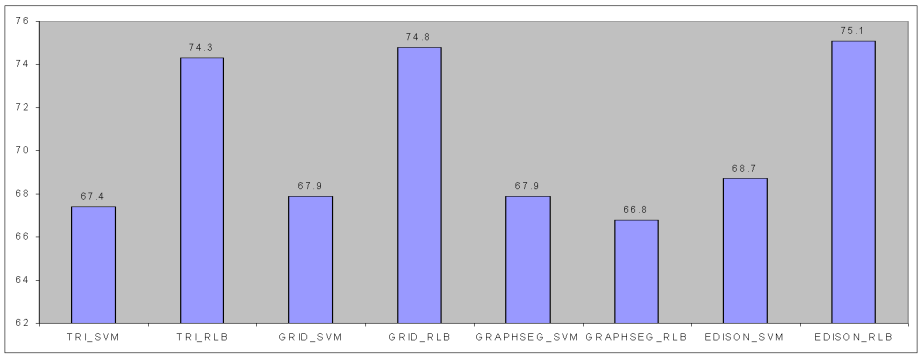


Fig. 7 Global pixel-wise classification accuracy of runs before and after relaxation labelling. Name_SVM are the results of direct segment classification, Name_RLB are the results obtained after the relaxation labelling.

one with the lowest visual variability (sky, grass) while the hardest classes are the ones with a high visual variability and fewer training examples.

From these results, we can draw the conclusion that each segmentation carry its own advantages that help to recognize some of the particular classes. The fact that the best results are achieved by edison contributes to the idea that an accurate segmentation improves the recognition, as was highlighted by other authors [14,16,18], Yet very different segmentation such as the high over-segmentation given by the grid and the coarser but accurate graph-based segmentation lead to the same recognition accuracy (overall). Nevertheless, in the case of single segmentation recognition, an over-segmentation should be favored since it appears to be more robust for the recognition of the different classes, despite the partial view of an object provided by the segments.

6.2 Impact of the relaxation labelling

Figure 7 presents the results obtained before and after relaxation labelling. The relaxation labelling step significantly increases the classification performances: 67.4 to 74.3 for triangle segmentation, 67.9 to 74.8 for grid segmentation, 68.7 to 75.1 for edison segmentation. Hence, we see that the regularization of the global labelling taking into account the relationships between neighboring regions is an efficient way of further improvement of the classification results. The probabilities estimates from the SVM model are sufficiently accurate to enable a convergence of the algorithm toward improved results. This is not the case, however, for the results obtained with graphseg, which are actually decreased after relaxation (66.9 to 66.8). Figure 8 presents the detailed classification results for the object classes before and after relaxation labelling. We note that, in fact, relaxation labelling is improving the results for most of the object classes over the direct classification results. However, for some classes which are the most frequent in the database (grass, sky), the results are decreased. Again, the large segments provided by the graph-based segmentation algorithm are a drawback for the relaxation labelling, since a wrong correction of a segment label has a stronger impact on the overall results.

Overall, the best classification performances are obtained using edison segmentation, followed by the grid, triangle and graphseg. The fact that the performances obtained with the grid are just slightly below the ones obtained by edison have to be contrasted. With a

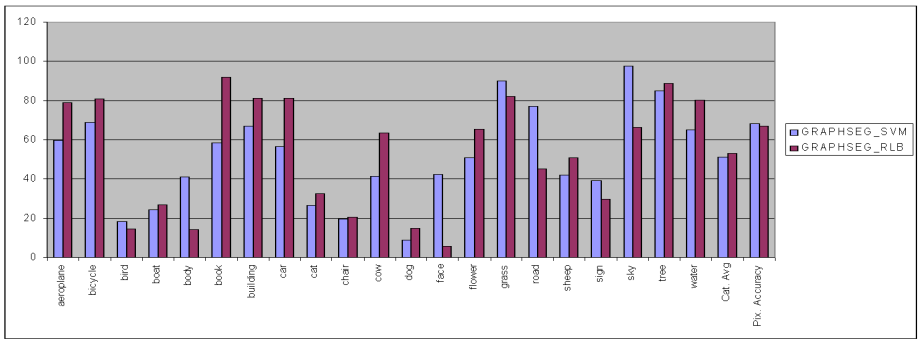


Fig. 8 Results of applying relaxation labelling to the segment classification obtained by graphseg on the different object classes.

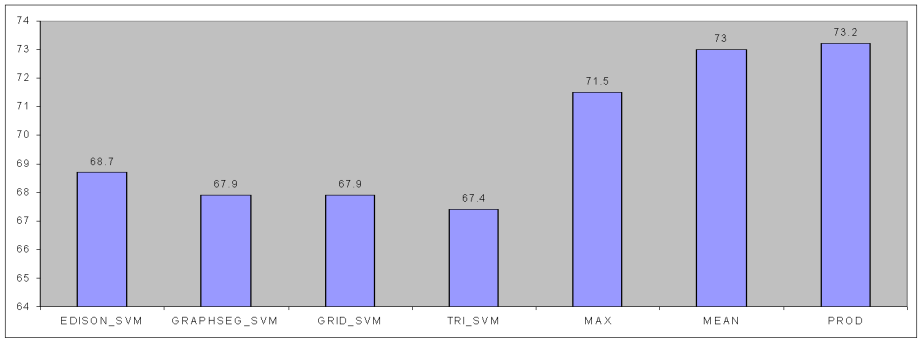


Fig. 9 Comparison of the classification accuracies obtained by the single segmentation detection results and by the combined segmentation with the different operators.

much lower number of segments per image, the processing time using edison segmentation is decreased both for SVM classification and for relaxation labelling.

6.3 Impact of the late fusion of multiple partitions

Figure 9 shows the classification accuracies obtained with the different merging strategies. The results obtained by any of the merging strategies outperformed the classification performances obtained using a single segmentation. The best results are achieved by multiplying the probabilities (73.2), then by taking the average probabilities (73) and taking the max (71.5). The multiplication operator strongly emphasize regions where each single detection agree on a common object class.

After merging the estimated probabilities, we are given for each new segment a multi-class probability estimate. We can then apply again the relaxation labelling, starting form these estimates.

Figure 10 shows the results obtained after relaxation labelling. Again, relaxation labelling improves the global coherency of the segments classification and has a positive impact on the overall classification results, even if the increase is not as significant as with

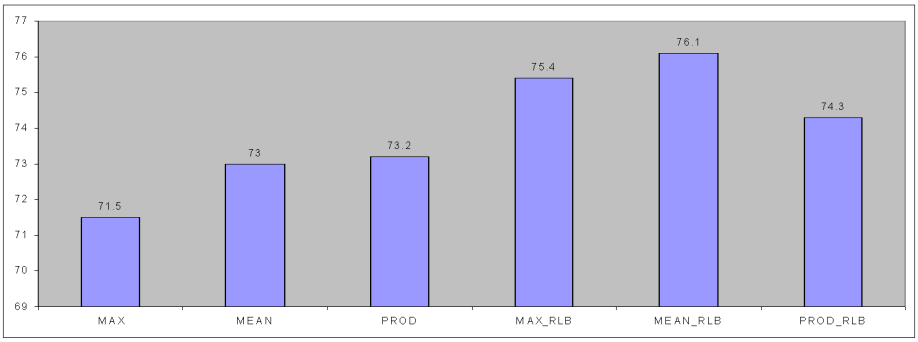


Fig. 10 Classification performances after the relaxation labelling step using the combined segmentation results as initialization.

the noisy, single segmentation detection results. It is interesting to see that the multiplied probabilities, which led to the best initialization did not lead to the best overall results after the relaxation labelling step. Multiplying the probabilities leads to an estimate with some regions already having very strong (*e.g.* not ambiguous) probabilities for some regions, which causes the relaxation to converge very fast without many changes. The max and mean operator allow more freedom for the convergence of the labelling, which in the end leads to better classification results. Overall, the best performances are obtained by the combination of the classifiers using the mean operator, followed by relaxation labelling which achieves 76.1 classification accuracy. This is a 1 point improvement over the best single segmentation detection results, 75.1 obtained with edison segmentation and relaxation.

Figure 11 shows the comparative results of our method with current state-of-the-art on the same dataset. Note that the results were not obtained using the same training and test datasets. Gould *et al.* [14] used 40% images for training, 60% for test, Shotton *et al.* [26], Verbeek and Triggs [33] split the dataset in two halves, Yang *et al* [36] reports the average results over a 5-fold cross validation. Our best method (merging multiple segmentation with mean operator) ranks second with for global pixel-wise accuracy with 76.1 behind Gould (76.5). The third one is our method using multiple segmentation and the max operator (75.4), then comes Yang *et al.* (75.1) and our method using edison segmentation (75.1).The main advantage of the method of Gould *et al.* is the introduction of the relative location priors between the object classes (a full 3-D spatial relationships between objects is inferred), which clearly helps the recognition as was also highlighted by Galleguillos *et al.* [12]. We model only the co-occurences between adjacent regions in the relaxation labelling. It will be interesting to model finer grain spatial relationships in the relaxation labelling, *e.g.* using different affinity coefficients for adjacent regions according to their spatial relationships such as above/below. The fact that the top 2 methods in the current state-of-the-art and ours rely on accurate image over-segmentation (superpixels for Gould *et al.* , mean shift patches for Yang *et al.* and our method) is again a plea for such kind of preprocessing. In figure **??**, we show the category-based results of our best method compared with the state-of-the-art. Example images are shown in figure 13 using single segmentation approach and in figure 14 using the late fusion results.

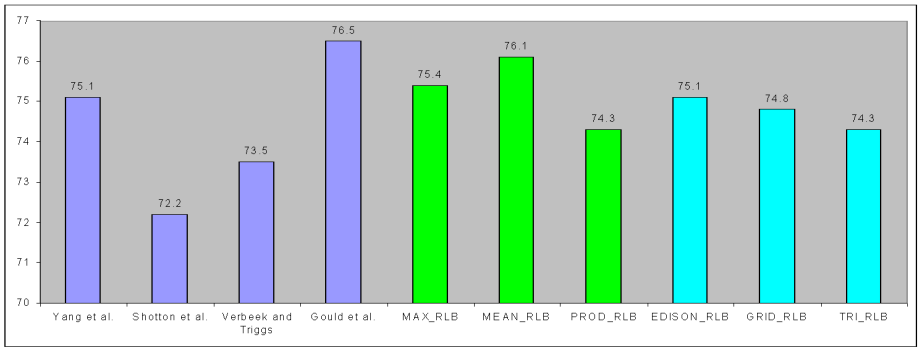


Fig. 11 Comparison with current state-of-the-art. Dark blue are results reported in the litterature, light green our results with late fusion, light blue our results with single segmentation approach.

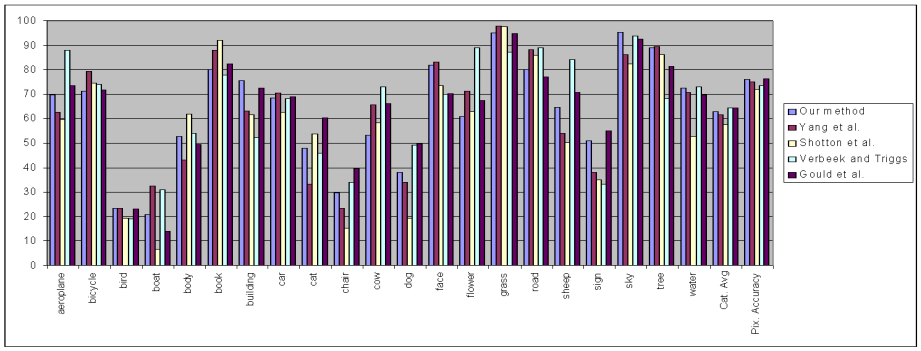


Fig. 12 Comparison with current state-of-the-art. Dark blue are results reported in the litterature, light green our results with late fusion, light blue our results with single segmentation approach.

7 Conclusion and perspectives

We have shown in this paper that the task of object detection and categorization can be successfully handled by the classification of the different image parts obtained by an initial image segmentation. Although trained using purely appearance based descriptors, the classification of segments already yields significantly good results. The combination of color, texture and keypoint-based descriptors is a powerful discriminative feature for the recognition of various object classes. In this work, we have kept the size of the codebook for the different descriptors low. However, there are evidences that a bigger codebook, at least for the keypoint-based descriptors, increases the classification results [17]. Our experiments have confirmed our assumptions that an accurate segmentation method as preprocessing step has a very positive impact on the semantic object extraction results, while saving some computational burden. As far as the detail of the initial segmentation is concerned, we have shown that in a single segmentation/recognition process, an accurate over-segmentation should be preferred.

The relaxation labelling algorithm of Rosenfeld *et al.* [25] improves greatly the coherence of the global image labelling, by resolving ambiguities of local adjacent segments. The

results obtained show an improvement over the direct recognition method for almost all the cases, and reach the current state-of-the-art obtained with more sophisticated method such as CRFs. The fact that this method is not optimal let the way open for future research on more global optimization that could improve the results even further.

Finally, we have proposed a new *late fusion* schema of the results of classifiers trained on the same initial feature space but using different segmentation algorithms as preprocessing. This is the first adaptation of late fusion approach in the image segmentation framework to our knowledge, and the results are promising. A direction of future research will be to apply our framework for object detection by combining several layers of coarse to fine segmentation methods, in order to capture and propagate recognition results over the different scales.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: SODA07, pp. 1027–1035 (2007) 5
2. Athanasiadis, T., Mylonas, P., Avrithis, Y., Kollias, S.: Semantic image segmentation and object labeling. IEEE transactions on circuits and systems for video technology **13**(3), 298–312 (2007) 2
3. Ayache, S., Qunot, G., Gensel, J.: Classifier fusion for svm-based multimedia semantic indexing. Lecture Notes in Computer Sciences **4425**, 494–504 (2007) 8
4. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding **110**, 346–359 (2008) 6
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm 7
6. Chevalier, F., Domenger, J.P., Benois-Pineau, J., Delest, M.: Retrieval of objects in video by similarity based on graph matching. Pattern Recognition Letter **28**, 939–949 (2007) 3
7. Christoudias, C., Georgescu, B., Meer, P.: Synergism in low level vision. In: 16th International Conference on Pattern Recognition, pp. 150–155 (2002) 4, 5
8. Comanicu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell. **24**, 603–619 (2002) 4
9. Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV'02, pp. 97 – 112 (2002) 2
10. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**, 167–181 (2004) 4, 5
11. Freixenet, J., Muoz, X., Raba, D., Mart, J., Cuf, X.: Yet another survey on image segmentation: Region and boundary information integration. In: ECCV'02 (2002) 4
12. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR'08. Anchorage, AK (2008) 2, 13
13. Gokalp, D., Aksoy, S.: Scene classification using bag-of-regions representations. In: CVPR'07, pp. 1–8 (2007) 1
14. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. International Journal of Computer Vision (2008) 2, 9, 10, 13
15. He, X., , Zemel, R.S., Carreira-Perpinan, M.: Multiscale conditional random fields for image labeling. In: CVPR'04, pp. 695–702 (2004) 2
16. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV'05 (2005) 2, 10
17. Jiang, Y.G., Yang, J., Ngo, C., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. IEEE Transactions on Multimedia **12**, 42–53 (2010) 6, 14
18. Malisiewicz, T., Efros, A.: Improving spatial support for objects via multiple segmentations. In: British Machine Vision Conference 2007 (2007) 2, 10
19. Meer, P., Georgescu, B.: Edge detection with embedded confidence. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(12), 1351–1365 (2001) 4
20. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV'06 (2006) 6
21. Pal, N., Pal, S.: A review on image segmentation. Pattern Recognition **26**, 1277–1294 (1993) 4
22. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74 (2000). URL http://citeseer.ist.psu.edu/platt99probabilistic.html 7

23. Prasad, L., Skourikhine, A.N.: Vectorized image segmentation via trixel agglomeration. Pattern Recognition **39**(4), 501–514 (2006) 4, 5
24. Ren, X., Malik, J.: Learning a classification model for segmentation. In: ICCV'03, vol. 1, pp. 10–17 (2003) 2
25. Rosenfeld, A., Hummel, R.A., Zucker, S.W.: Scene labeling by relaxation operations. IEEE Transactions on Systems, Man and Cybernetics **6**, 420–433 (1976) 3, 7, 8, 14
26. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout and context. International Journal of Computer Vision **81**(1), 2–23 (2009) 2, 9, 13
27. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV'03, vol. 2, pp. 1470–1477 (2003) 5
28. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12), 1349–1380 (2000) 1
29. *et. al*, C.S.F.: Columbia university/vireo-cityu/irit trecvid2008 high-level feature extraction and interactive video search. In: TRECVID'08 (2008). URL http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/columbia.pdf 1
30. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995) 6
31. Varma, M., Zisserman, A.: A statistical approach to material classification using image patch examplars. IEEE Transactions on Pattern Analysis and Machine Interlligence **31**, 2032–2047 (2008) 6
32. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: CVPR'07, pp. 1–8 (2007). URL http://lear.inrialpes.fr/pubs/2007/VT07 2, 4
33. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: CVPR'07 (2007) 9, 13
34. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5**, 975–1005 (2004) 7
35. Y. Peng Z. Yang, J.Y.L.C.H.L.J.Y.: Peking university at trecvid 2008: High level feature extraction. In: TRECVID'08, p. (on line). NIST (2008). URL http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/peking-university.pdf 1
36. Yang, L., Meer, P., Foran, D.J.: Multiple class segmentation using a unified framework over mean-shift patches. In: CVPR (2007) 9, 13

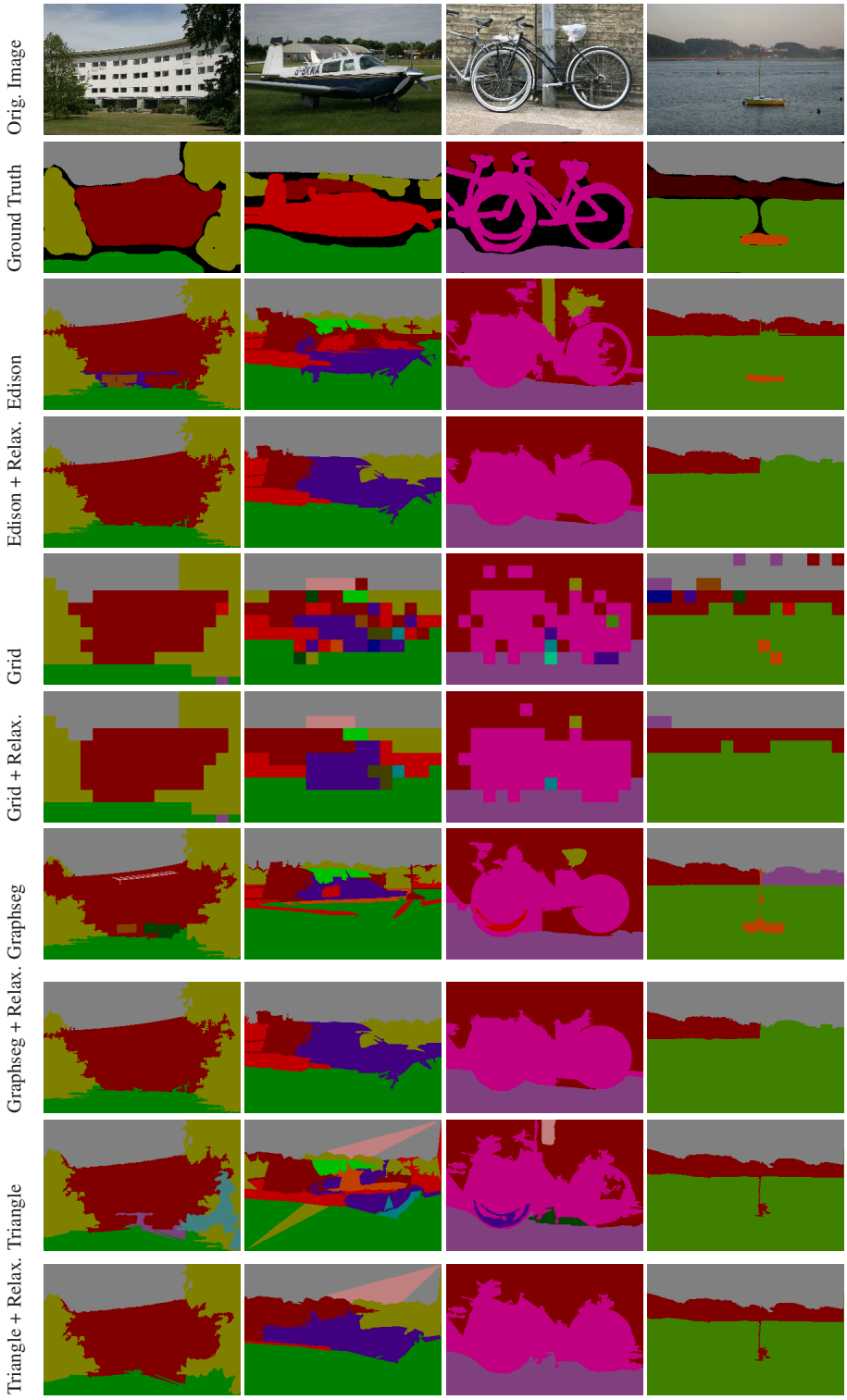


Fig. 13 Example of semantic object detection based on single segmentation (best viewed in color).

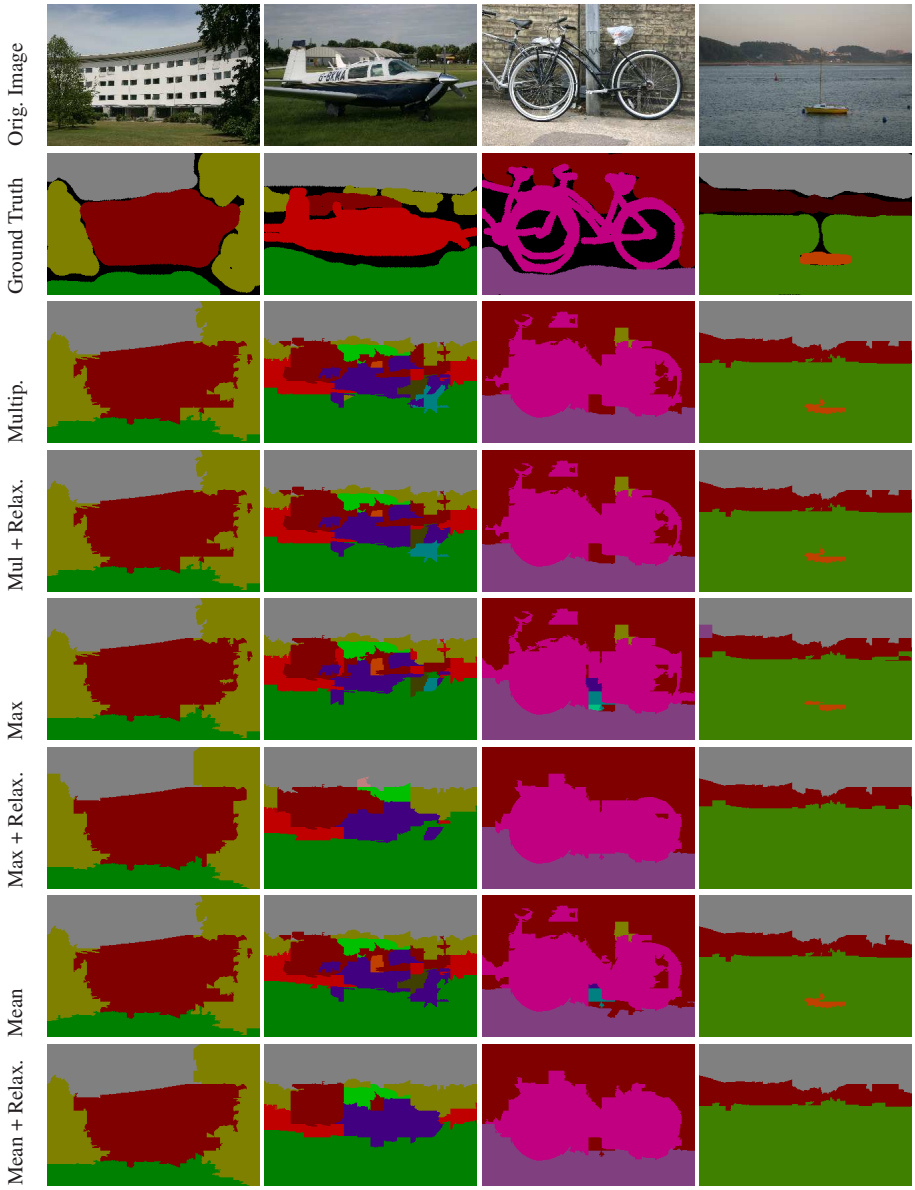


Fig. 14 Example of semantic object detection by late fusion of multiple segmentation detections (best viewed in color).