Visual object categorization based on the fusion of region and local features

Catégorisation d’objets visuels par une approche reposant sur la fusion de descripteurs de régions et locaux

Huanzhang Fu — Alain Pujol
— Emmanuel Dellandréa — Liming Chen

Université de Lyon
Ecole Centrale de Lyon
LIRIS, CNRS, UMR5205, F-69134, France
{huanzhang.fu,alain.pujol,emmanuel.dellandrea,liming.chen}@ec-lyon.fr

RÉSUMÉ. Ce papier présente une nouvelle approche pour la catégorisation d’objets visuels en utilisant des descripteurs de régions et une modélisation de l’image basée sur des mesures statistiques. Nos descripteurs de régions sont extraits de régions éparse obtenues par un algorithme de segmentation inspiré de la théorie de Gestalt, et capturant des informations visuelles importantes telles que les segments et les couleurs. La modélisation du contenu visuel d’une image repose sur des mesures statistiques caractérisant la distribution des descripteurs de régions, évitant la principale difficulté de l’approche populaire “bag-of-local features” qui nécessite de fixer la taille d’un vocabulaire visuel. Plusieurs processus de classification, intégrant des techniques de sélection de descripteurs (par exemple, ACP ou Adaboost) et des stratégies de fusion, sont également mis en œuvre et comparés. Nos expérimentations sur un sous-ensemble de la base d’images Pascal VOC ont montré qu’en séparant les descripteurs extraits de différentes sources dans différents “canaux”, puis en les combinant en utilisant une fusion précoce, les performances de classification pouvaient être améliorées. En outre, nos résultats expérimentaux démontrent que nos descripteurs de régions peuvent être combinés avec les descripteurs SIFT pour améliorer la classification, ce qui suggère que nos caractéristiques portent une information complémentaire à celle des descripteurs SIFT.
ABSTRACT. This paper presents a novel approach for visual object categorization using region based features and statistical measures based image modeling. Our region-based features are extracted from coarse regions obtained by the Gestalt theory inspired region segmentation algorithm and they capture visually significant information such as segments and colors. The modeling of the visual content of an image relies upon some statistical measures over sparse region-based features, thus avoiding the major difficulty of the popular “bag-of-local features” approach which needs to fix a visual vocabulary size. Several classification schemes, including feature selection techniques (e.g. PCA or Adaboost) and fusion strategies, are also implemented and compared. Experimented on a subset of Pascal VOC dataset, we show that by separating features extracted from different sources in different “channels”, and then to combine them using an early fusion, we can actually improve classification performance. Moreover, experimental results demonstrate that our region-based features can be combined with SIFT features to reinforce performance, suggesting that our features managed to extract information which is complementary to the one of SIFT features.

MOTS-CLÉS : Catégorisation d’objets visuels, Descripteurs de régions, Stratégies de fusion, Sélection de descripteurs

KEYWORDS: Visual object categorization, Region based features, Fusion strategies, Feature selection

1. Introduction

Generic Visual object categorization aims at predicting whether at least one or several objects of some given classes are present within an image. This is a really challenging problem in the computer vision field because of the number of real world object types which need to be discriminated, as well as variations in viewpoint, scale, imaging, lighting and occlusion. To all these difficulties we also need to add the one induced by intra-class variations, typical of semantic classes of everyday objects. As such it has attracted a lot of attention in the past years [EGW+08].
1.1. Related Work

Recently, the “bag of features” kind of approach [DWF+04, PDCB06], which tries to adapt the “bag-of-words” representation for text categorization to “Visual Object Categorization” (VOC) problem, has been widely used and has shown its effectiveness in PASCAL VOC challenge [EGW+08]. These approaches describe an image as a bag of discrete “visual words”, where the histogram containing the number of occurrences of each visual word is used for further object categorization. A visual vocabulary including all these visual words is beforehand learned from the training set by a hard assignment (quantization) or a soft assignment through GMMs using local image features, for instance the popular SIFT feature [Low04]. They can be sparsely extracted from salient image regions, called interest “points” [Lin98, MS04] or more densely from points extracted using several grids at multiple scales [LP05]. A classifier, e.g. SVM, can then be used for learning and classifying the images modeled by their histogram of “visual words”.

Although the “bag-of-local features” approach has achieved the best performance in the last Pascal VOC contests, the overall performance, with an average precision around 60% over 20 classes achieved by the best classifier, is still far from real application-oriented requirements. In particular, the size of visual vocabulary which is the basis of “bag-of-local features” approach is hard to be fixed as there are no evident similar concepts in images as compared to a textual document. The basic problem is that the “bag-of-local features” approach, while adapting the best practice from text categorization, does not necessarily correspond to a human visual perception process which seems to be ruled by some Gestalt principles according to several studies on visual perception [Kan97, Wer23] and supposed to perform a holistic analysis combined with a local one through a fusion process. Moreover, the schemes so far proposed in the literature for automatic generic visual object classification also suffer the problem of a small and biased training dataset, in particular with an unbalanced ratio of positives versus negative samples.
1.2. Our Approach

Our basic hypothesis is that effective visual object classification or detection should be inspired by some basic human image interpretation principles. In this paper we propose overcoming the shortfalls of the popular “bag-of-local features” approach and make use of some basic principles from the Gestalt theory, in particular the well known Gestalt laws of Perceptual Organization which suggest both the grouping of pixels into homogeneous regions as well as the interaction between regions.

Desolneux et al. have given in [DMM08] a comprehensive introduction to Gestalt theory in an image analysis perspective. Gestalt theory starts with the assumption of active grouping laws in visual perception which recursively cluster basic primitives into a new, larger visual object, called gestalt. These grouping laws follow criteria such as spatial proximity, color similarity. These laws also highlight the interaction between regions. This interaction is also confirmed by Navon [Nav77] who showed the preponderance of global perception over local perception. Following these basic Gestalt perception laws, our approach also claims that an effective description of the visual content of an image needs to model the partial gestalts and their interactions. We feel that lacking these principles, the popular “bag of features” approaches [DWF+04, PDCB06] deprive themselves of meaningful information. One exception is the work of Barnard et al. [BDG+03] which is a region-based approach where regions are labeled with probable categories. However, they don’t take into account the interaction among regions.

In our approach, we propose to make use of some region-based meaningful features extracted from visual regions with neighborhood information. These region based features result from perceptually significant “Gestalts” segmented according to some basic Gestalt grouping laws. As the regions resulted from a segmentation process may not be consistent with object boundaries, individual regions are not labeled as did Barnard et al. [BDG+03]. Regions produce a feature vector which is supposed to have no meaning on its own but that can contribute to one or more classes. Regarding features, we propose using visually meaningful
features, such as color and line segment based features which we will extend to provide information from neighboring regions. In this paper, not only we compare the use of these region based features with popular SIFT features but also we check the efficiency of the combination of our features with SIFT ones.

A set of statistical measures is then used to model the distribution of image feature vectors. Its interest is three-fold. First, we circumvent the difficulty of fixing the size of visual vocabulary; secondly, we avoid the inaccurate assumption of Gaussian distribution of feature vectors and thirdly we can cope with a small number of feature vectors per image as it is the case with our region-based features. Using such a statistical measures-based visual content modeling, we experiment several classification schemes, including for instance the cascade of classifiers and Adaboost feature selection, on a subset of Pascal VOC 2007 dataset. Furthermore, we also compare two fusion strategies, namely early fusion strategy by grouping all the features together and fed into a single classifier, and late fusion strategy which makes use of “channels” with a separate classifier for each kind of features, the outputs of these classifiers being merged later [SWS05] in a process similar to boosting [FS99]. Experiments carried out on a subset of Pascal VOC dataset show that our region-based features not only perform adequately, but also can be combined with SIFT features to provide better overall performance. Moreover, the fusion of the different channels by an early fusion strategy performs better than late fusion strategy.

The rest of the paper is organized as follows. Section 2 describes our region segmentation algorithm and the visual features we extract. Our classification scheme is introduced in section 3 while section 4 gives respectively implementation details and experimental results. Section 5 concludes this work and evokes some future research prospects.

2. Region Segmentation and Region based Features

In this section, we first introduced our Gestalt-inspired region segmentation scheme [PC08] and then the color and segment based features we extract from the region map given by our segmentation scheme.
2.1. Region Segmentation Scheme

As we have seen in the previous section, studies on human perception strongly hint at a region based approach. On the other hand, introducing region segmentation brings about a host of new problems regarding segmentation robustness and accuracy. Thus, while this approach suits human perception better, we have no guarantees its benefits will overcome its drawbacks. In our approach we specifically designed a robust region segmentation method that aims at automatically producing coarse regions from which we can consistently extract feature vectors [PC08]. We will now briefly describe the outline of the algorithm.

The principle of our region segmentation algorithm is to segment an image into partial gestalts for further visual object recognition. We thus made use of the following Gestalt basic grouping laws in our gestalt construction process: The color constancy law stating that connected regions where color does not vary strongly are unified; the similarity law leading to group similar objects into higher scale object; the vicinity law suggesting grouping close primitives with respect to the others; and finally good continuation law saying that reconstructed amodal object, i.e. partially perceived physical structure which is reconstructed through understanding, should be as homogenous as possible. Because those laws are defined between regions and their context, at each step we assess the possibility to merge regions according to global information.

The algorithm is based on color clustering but also includes an extra post-processing step to ensure spatial consistency of the regions. In order to apply previously mentioned Gestalt laws, we defined a 3-step process: first we filter the image and reduce color depth, then we perform adaptive determination of the number of clusters and cluster color data and finally we perform spatial processing to split unconnected clusters and merge smaller regions.

Images are first filtered for robustness to noise; colors are then quantified by following a first, fast color reduction scheme using an accumulator array in CIELab color space to agglomerate colors that are perceptually similar. In the second step, we use an iterative algorithm to determine a good color count which limits the quantization error. Indeed,
quantization error measured by MSE between original and quantized colors evolves as per Fig. 1 according to the number of clusters.

This clearly shows a threshold cluster number under which quantization MSE begins to rise sharply. By performing several fast coarse clustering operations using Neural Gas algorithm [MS91], which is fast and less sensitive to initialization than its counterparts such as K-means, we are able to compute the corresponding MSE values and generate a target cluster count. We then use hierarchical ascendant clustering which is more accurate but much slower thus executed only once in our case, to achieve segmentation. The third step consists in splitting spatially unconnected regions, merging similar regions and constraining segmentation coarseness. Merging of similar regions is achieved through the use of the squared Fisher’s distance as (1) (used for a similar task in [ZY96]), where \( n_i, \mu_i, \sigma_i^2 \) are respectively the number of pixels, the average color and the variance of colors within region \( i \). This distance still stays independent towards image dynamics as it involves intra-cluster distance vs. inter-cluster distances. Finally, regions which are too small to provide significant features are discarded.

\[
D(R_1, R_2) = \frac{(n_1 + n_2)(\mu_1 - \mu_2)^2}{n_1\sigma_1^2 n_2\sigma_2^2}.
\]  
(1)
With this algorithm we obtain consistent coarse regions that can be used for our classification system. Sample segmentation results on Pascal challenge dataset images are given in Fig. 2. As we can see, our Gestalt-inspired segmentation algorithm has automatically adapted its segmentation process to the color depth of the images, producing significant partial gestalts.

2.2. Feature Extraction

In order to represent the information carried by regions, we make use of two kinds of features: color features and segment features. Region based color features aim at capturing a coarse perception of partial gestalts, in the form of color moments (mean, variance and skewness) [SO95] for each color channel. These features are quite compact and have proven as efficient as a high dimension histogram [DMK+01]. Various color spaces were experimented for the computation of these features and best results were achieved in the CIELch color space which is derived from CIELab as in (2) and best fits to the human perception [TFMB04].

\[
L_{Lch} = L_{Lab} \quad c = \sqrt{a^2 + b^2} \quad h = \arctan \frac{b}{a}.
\]  

(2)

The segment features aim at capturing some textual and geometrical properties of partial gestalts. We thus developed segment based features
relying on a fast connective Hough transform [AC01] that performed well in global image classification [PC07] and more specifically provided more significant information than gradient based features. These features are relevant regarding our approach of following human visual interpretation as, most of the time, there are few segments within a region but, on the other hand, they represent features that stand out visually and their simple presence is significant.

The principle of our segment based feature extractor is the following. As for any other Hough transform, we start from an edge map of the processed image. Because we wish to avoid problems related to edge thickness, we use a Canny Edge Detector [Can86] to process our image in order to ensure a one pixel thickness for our edge map. For an edge point on the edge map, we examine its neighborhood identified by its relative angular position \((r, \theta)\): each direction \(\theta\) is processed while a connected edge is found at distance \(r + 1\), which gives us a list of segments by orientation for this edge point. Once we have this list, we store the longest segment and remove it from the edge map. To avoid hindering intersecting segment detection, we use two separate edge maps: one for segment source point detection and one for connected points detection. Removed segments are only removed from the source point map, which avoids detecting the same segment twice while preserving intersecting segments. These segment features are extracted once for the whole image.

During this extraction step, we can build a map from image coordinates to the corresponding segments. Therefore, we can quickly detect segments within a region. For validation purposes, our “segment” shape features are a simple histogram combining length and orientation. In order to obtain scale invariant features, we normalize lengths by dividing them by the longest segment length. We then obtain rotation invariance by computing an average orientation in order to have a stable average and by expressing all angles with respect to this average direction. We therefore obtain a feature that is invariant to translation, scale as well as rotation. The size of the histograms was experimentally determined and set to 6 bins for orientation and 4 for length.

Finally, in order to include neighborhood information, our region based features (color moments and Hough segment features), are expres-
sed at four different levels: original region, region + neighbors, region + neighbors + neighbor’s neighbors, etc. Those levels are concatenated in the final feature vector. This is a basic way to integrate spatial relationship but also to include global information in each feature vector. On most images, the fourth level will represent features extracted over the whole image.

3. Classification Process

This section mainly focuses on the classification processes we have developed using our region based features presented in the previous section, combined with popular SIFT features. The image representation thanks to statistical measures is first presented and then classification frameworks are described.

3.1. Statistical Measures based Image Representation

Once having extracted a set of feature vectors from an image, an efficient characterization of the visual content represented by this information needs to be elaborated. A simple approach would be to concatenate these feature vectors to build a huge single vector. However, the number of feature vectors as we proposed to extract in the previous section can vary from one image to another, typically depending upon the number of segmented gestalts. Since machine-based learning schemes require input data to have a constant size, a possibility is to model the distribution of feature vectors and to use the parameters of this distribution as new features for the classification. The popular “bag-of-features” approach follows this strategy: the distribution of original features is modeled thanks to a histogram for each image on the basis of a “visual vocabulary”, which can be built either by using a clustering algorithm or by using a parametric distribution such as Gaussian Mixture Models (GMM). The drawback of such an approach is obvious: the optimal size of this visual vocabulary is hard to be fixed as there is no easy intuitive counterpart in image compared to keywords in text document. Regarding GMM as an example, if the number of Gaussians is too small then it can’t supply enough normal distributions for a large amount of
diversified feature vectors to be modeled, while a too high number of Gaussians suffers from an insufficient number of feature vectors to optimize the parameters of the model. Therefore, we propose to solve this problem by modeling the feature distribution thanks to statistical measures.

The basic idea of this image representation is to model the distribution of values in each component of these feature vectors by descriptive statistical measures and then to concatenate these statistical measures into one new single feature vector that will characterize the visual content of an image and will be used for object categorization in the next step.

<table>
<thead>
<tr>
<th>Name of statistics</th>
<th>Description or formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic average</td>
<td>$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$</td>
</tr>
<tr>
<td>Harmonic mean</td>
<td>$m = n/ \sum_{i=1}^{n} \frac{1}{x_i}$</td>
</tr>
<tr>
<td>Trimmed mean</td>
<td>mean of $X$ excluding the highest and lowest 10% of observations</td>
</tr>
<tr>
<td>Range</td>
<td>max($X$)-min($X$)</td>
</tr>
<tr>
<td>Mean absolute deviation</td>
<td>$y = \frac{1}{n} \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$s = \left( \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2}$</td>
</tr>
<tr>
<td>Percentiles</td>
<td>quantiles of $X$ with orders that are multiples of 0.25 (5 values obtained in the interval [0, 1])</td>
</tr>
</tbody>
</table>

Totally 12 statistical measures have been used to describe the distribution of data in each component, among which stands the number of zeros. Indeed, due to the computation process of our visual features as well as the one of SIFT features, the data contains a high number of zeros that may disturb the computation of the data distribution. Thus, this information is carried in the feature called “number of zeros” and then zeros are removed from the new data that is characterized by the remaining 11 statistical measures which mainly belong to 3 groups : 1, Measures of central tendency to locate a distribution of data along an appropriate scale ; 2, Measures of dispersion to find out how spread out
3.1. Statistical Measures

The data values; 3, Percentiles to provide information about the shape of data as well as its location and spread. A detailed presentation of these 11 statistical measures is in Table 1, where \( X = \{x_i\}, i = 1...n \) is a set of observations for one component.

Once the distribution of each component of the feature set has been modeled thanks to the statistical measures, a new image feature vector \( Q \) is produced by concatenating these statistical measures for all components, which we call statistical measures based image representation. Assuming that we have \( L \) components in a feature set, then the vector \( Q \) has a dimension of \( 12 \times L \), which ranges generally from hundreds to thousands. A vector of such high dimensionality used for classification may lead to the well-known “curse of dimensionality” [Bel61] problem. Consequently, the dimension of these vectors should be reduced. Numerous feature selection methods are conceivable [SIL07]. The ones we considered are discussed in the next subsection.

3.2. Classification Schemes

The classification process, in the context of visual object categorization, aims at predicting whether at least one or several objects of some given classes are present within an image. The elaboration of such classification schemes is generally empirical as its efficiency will depend on numerous factors such as the nature of visual features used to carry the information in images, the high dimensionality of the distribution of these features and the complexity of the frontiers between classes in the feature space. Thus, we present in this section several classification schemes presenting a general overview of conceivable classification techniques that will be further evaluated for visual object categorization purposes.

Given an image to classify, we first detect points of interest or regions from which the visual features are extracted. These features are then transformed to form a new feature vector through statistical measures based image representation using the method introduced in the previous subsection. Finally, this new feature vector will pass through the classifier beforehand trained or pass through a set of such classifiers, according to the fusion strategy that will be presented in the next sub-
section, to judge whether this image contains or not a given object. In this procedure, two particular problems should be taken into consideration. The first one is that a biased dataset (usually there are much more negative examples than positive ones) is often envisaged in the training stage, especially when a one-versus-all strategy is used for multi-class classification. However, this unbalanced dataset generally degrades the performance of a classifier as the training set has to be as representative as possible. As a result, we envisage three principal ways to address this problem: 1, the simplest one is to construct a balanced dataset using only a subset of negative examples through sub-sampling (by random for example); 2, a series of classifiers is built up according to a cascade technique, all classifiers having at their disposal balanced dataset created using different samplings; 3, “weak” side of the dataset is compensated by giving it a higher weight during the training. The second issue is the feature selection which aims at choosing the most useful features in order to avoid the potential “curse of dimensionality”. In our study, 4 different solutions are considered in order to evaluate their respective efficiency for our image categorization problem: 1, no feature selection method is used; 2, a canonical discriminant analysis [Fis36] is used; 3, a principal component analysis [Pea01, Jol02] is used; 4; an adaboost algorithm is used for feature selection [FS99, FS97, SB04]. A brief introduction of all these techniques is given in the following paragraphs.

**Balanced classifier**: In this case, a subset of negative examples is chosen through random sampling. Its size is equal to the one of the positive example set.

**Cascade of classifiers**: This is a series of balanced classifiers in each of which the positive examples are always the same whereas the negative examples are composed of the false positives of the previous balanced classifier and new added negative examples until the two sides reach a new balance (see Fig. 3). The process terminates when there is no new negative example left. The final score is the sum of the scores given by each balanced classifier.

**Biased classifier**: This corresponds to a single global classifier which is trained using all available examples. However, in order to handle the unbalanced effect of the dataset, different weights are gi-
ven to the positive and negative samples. As weight values are classifier and dataset dependent, they are determined experimentally.

**Principal component analysis (PCA)**: It is a simple, widely-used and non-parametric method for extracting relevant information from confusing dataset. With minimal additional effort PCA provides a roadmap for how to reduce a complex dataset to a lower dimension to reveal the sometimes hidden, simplified structure that often underlie it, that is to say it transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal component [Pea01, Jol02, Shl05].

**Canonical discriminant analysis (CDA)**: It is a quick algorithm which allows reducing the dimension by producing a new representation space which distinguishes the best the different classes. Its principle is to produce a series of uncorrelated discriminating variables, in order to have individuals in the same class projected on these axes as close as possible and individuals from different classes as distant as possible. In most cases, we obtain $K - 1$ axes where $K$ is the number of classes.

**Adaboost feature selection (ADA)**: The adaboost algorithm, first introduced by Freund and Schapire [FS97], calls a weak classifier repeatedly in a series of rounds $t = 1...T$. For each round, the weak classifier is forced to focus on the examples incorrectly classified by the previous weak classifier through increasing the weights for these hard examples. Finally, a strong classifier can be created by linearly combi-
ning these weak classifiers [SB04, Sch02, VJ01]. In our approach, we use adaboost as a feature selection method since each weak classifier can also be seen as a selected single feature which best separates positive and negative examples. Thus, after $T$ rounds, the best $T$ features for the classification have been selected, and they can feed other classifiers, such as SVM or Neural Networks.

3.3. Fusion Strategies

Fusion strategy is usually used in multimedia data analysis. Indeed, generally three modalities have to be handled in videos, namely the auditory, the textual, and the visual modality. Thus, a fusion step is necessary to combine the results of the analysis of these modalities considered independently in a first step [SWS05]. The same idea can be employed in visual object categorization, since, in order to extract a visual information as exhaustive as possible, different types of features from the same image can be computed to form several information streams such as SIFT, Region based Color Moments (RCM) and Region based Histogram of Segments (RHS) in our case. These streams need to be fused in order to elaborate a single decision from several sources of information. This fusion of different types of features can follow several strategies: an early fusion is obtained when grouping all the features together in order to build a single feature vector that will feed the classifier whereas a late fusion makes use of “channels” with a separate classifier for each kind of features, the outputs of these classifiers being merged later [SWS05] in a process similar to boosting [FS99]. Between these two strategies, numerous intermediate strategies can be conceivable which consist in generating intermediate classes from different sources and to take a final decision based on these intermediate classes. In our experiment, we evaluate the two main strategies: early fusion and late fusion. Their schemes are illustrated respectively in Fig. 4(a) and Fig. 4(b).
4. Experiments

4.1. Implementation

We have used in our experiments the database of PASCAL challenge 2007 [EGW+07]. This database consists in 20 object categories and contains 2501 images taken in real world provided for training, 2510 for validation and 4952 for testing. For the purpose of evaluating our classification approaches, we have chosen 5 semantic representative classes namely airplane (238 images for training), bicycle (243 images for training), bus (186 images for training), horse (287 images for training) and person (2008 images for training). Some image samples for these 5 classes are given in Fig. 5.

Concerning the classifier, we have chosen the Support Vector Machine (SVM) [CV05, CST00] for its high ability in solving the small dataset, nonlinear and high dimensional pattern recognition problems.
As far as we know, the choice of kernel and its parameter optimization are two crucial aspects for object categorization using SVM. According to [CL01], 3 reasons encouraged us to use the Radial Basis Function (RBF) kernel whose formula is in (3). The first reason is that the RBF kernel has similar performances as the linear kernel [KL03] or the sigmoid kernel [LL03] for certain parameters. Secondly, its small number of hyperparameters facilitates the following parameter optimization task. Finally, it has less numerical difficulties.

\[ K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad \gamma > 0 \]  

(3)

We have performed SVM parameter optimization thanks to a grid search using 4-fold cross-validation technique in order to find out the best-fit group of parameters \((C, \gamma)\), where \(C > 0\) is the penalty parameter of the error term. This parameter also offers us the possibility to construct a biased classifier mentioned in 3.2 by giving different weights on \(C\) for the positive and negative side. A good estimation of the weights has been obtained according to (4) through several preliminary experiments, where \(w_{\text{pos}}\) and \(w_{\text{neg}}\) are the weights applying on \(C\) for the positive and negative side respectively, \(p\) and \(n\) are the number of positive and negative examples.

\[ w_{\text{pos}} = (p + n)/p \quad w_{\text{neg}} = (p + n)/n \]  

(4)

Finally, one-versus-all SVM classifiers have been built for each class in our evaluation. In order to save computation time, the 4 feature selection approaches presented in 3.2 will only apply to the balanced classifier, and the winner will go further with the cascade of classifiers and biased classifier.

4.2. Results

We present in this section the results of our experiments concerning the evaluation of the different classification schemes we presented in
the previous section for the visual object categorization purpose. Thus, three types of features are considered, namely SIFT (computed using the C# “libsift” implemented by Sebastian Nowozin [Now05] for their extraction), RCM and RHS. Moreover, 2 fusion strategies, early and late (noted respectively as EF and LF), have been evaluated together with the 4 feature selection approaches (noted as NON when no feature selection is used, PCA, CAD and ADA) using these feature sets with the balanced classifier, in order to evaluate their efficiency in our case of visual object categorization. We have chosen the Average Precision (AP) as a measure of classification efficiency, which represents the average of precisions over the entire range of recall. A good score of AP requires both high recall and high precision, that it is particularly interesting for classification problems. Finally, the number of features for 3 channels SIFT, RCM and RHS is respectively 1536, 432 and 1152 after the modeling by statistical measures without feature selection, which is the case in NON. ADA selects the best 50% of the original features in NON sorted according to adaboost feature selection scheme for all the 3 channels. However, PCA and CDA would greatly reduce this number to about a few tens.

Tableau 2 – Results for 5 representative classes using the combinations of 2 fusion strategies and 4 feature selection approaches with a balanced classifier.

<table>
<thead>
<tr>
<th></th>
<th>Plane</th>
<th>Bicycle</th>
<th>Bus</th>
<th>Horse</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF</td>
<td>0.4085</td>
<td>0.1933</td>
<td>0.1920</td>
<td>0.3302</td>
<td>0.7224</td>
</tr>
<tr>
<td>EF</td>
<td><strong>0.4225</strong></td>
<td><strong>0.2522</strong></td>
<td><strong>0.2806</strong></td>
<td>0.3859</td>
<td><strong>0.7499</strong></td>
</tr>
<tr>
<td>LF</td>
<td>0.4051</td>
<td>0.1347</td>
<td>0.1085</td>
<td>0.1923</td>
<td>0.7079</td>
</tr>
<tr>
<td>EF</td>
<td><strong>0.3739</strong></td>
<td>0.2104</td>
<td>0.2150</td>
<td>0.2249</td>
<td>0.7246</td>
</tr>
<tr>
<td>LF</td>
<td>0.1988</td>
<td>0.0771</td>
<td>0.0576</td>
<td>0.0968</td>
<td>0.5494</td>
</tr>
<tr>
<td>EF</td>
<td>0.1878</td>
<td>0.0891</td>
<td>0.0540</td>
<td>0.2194</td>
<td>0.5449</td>
</tr>
<tr>
<td>LF</td>
<td>0.3479</td>
<td>0.1872</td>
<td>0.0954</td>
<td><strong>0.4035</strong></td>
<td>0.6946</td>
</tr>
<tr>
<td>EF</td>
<td>0.4153</td>
<td>0.2374</td>
<td>0.2229</td>
<td>0.3732</td>
<td>0.7364</td>
</tr>
</tbody>
</table>

From Table 2, which shows the results for 5 representative classes using the combinations of 2 fusion strategies and 4 feature selection approaches with a balanced classifier, we can see that NON generally per-
forms best among all the 4 feature selection approaches, even if results of ADA are somewhat comparable. However, PCA and CDA seriously hurt the performance in our case. Considering the number of features in different feature selection approaches as well, we found that the approaches that have a huge number of features (for example, EF_NON has 1536+432+1152=3120 features) generally perform better than the ones having a small number of features. This fact is probably due to the blur of boundary between classes when realizing the transformations of PCA and CDA. Then focusing on LF and EF, the results show that EF performs better than the second fusion strategy. One of the reasons might be the good ability of SVM in solving high dimensional problems so that it benefits EF in which all the features are merged to form a long feature vector. This conclusion is also consistent to the fact observed previously when comparing different feature selection approaches. As a result, early fusion together with no feature selection will be applied on the cascade of classifiers and biased classifier, whose results are listed in Table 3.

Tableau 3 – Results for 5 representative classes using early fusion with balanced classifiers, cascades of classifiers and biased classifiers.

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>Plane</th>
<th>Bicycle</th>
<th>Bus</th>
<th>Horse</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF_Balanced</td>
<td>0.4225</td>
<td>0.2522</td>
<td>0.2806</td>
<td>0.3859</td>
<td>0.7499</td>
<td></td>
</tr>
<tr>
<td>EF_Cascade</td>
<td>0.5038</td>
<td>0.2868</td>
<td>0.3032</td>
<td>0.4527</td>
<td>0.7499</td>
<td></td>
</tr>
<tr>
<td>EF_Biased</td>
<td><strong>0.5170</strong></td>
<td><strong>0.3507</strong></td>
<td><strong>0.3176</strong></td>
<td><strong>0.5853</strong></td>
<td><strong>0.7554</strong></td>
<td></td>
</tr>
</tbody>
</table>

In Table 3, EF_Cascade and EF_Biased get an AP much higher than EF_Balanced for all the classes. An increasing of 13% to 51% can be observed between EF_Biased and EF_Balanced, depending on the class except “person” in which only 1.41% augmentation has been observed. An explanation consists in the fact that persons appear in almost all the training images so that the training set of EF_Balanced doesn’t differ very much form the other two. Until now, we have got the best results using EF_Biased, which are also comparable to the results in [EGW+07].

The improvement recorded between single channels and early fusion in Table 4 means that our region based features managed to extract in-
Tableau 4 – Results for 5 representative classes between single channels (SIFT, RCM, RHS) and early fusion with biased classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Plane</th>
<th>Bicycle</th>
<th>Bus</th>
<th>Horse</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>0.4019</td>
<td>0.2120</td>
<td>0.1812</td>
<td>0.3515</td>
<td>0.6524</td>
</tr>
<tr>
<td>RCM</td>
<td>0.4428</td>
<td>0.2093</td>
<td>0.1744</td>
<td>0.4267</td>
<td>0.6514</td>
</tr>
<tr>
<td>RHS</td>
<td>0.3074</td>
<td>0.1788</td>
<td>0.2127</td>
<td>0.2975</td>
<td>0.6564</td>
</tr>
<tr>
<td>EF</td>
<td>0.5170</td>
<td>0.3507</td>
<td>0.3176</td>
<td>0.5853</td>
<td>0.7554</td>
</tr>
</tbody>
</table>

formation which is complementary to the one of SIFT features so that the fusion of these single channels helps to improve the classifier performance. Among single channels, their performances are more or less the same using statistical measures based image representation, but vary significantly from one class to another.

5. Conclusion

We have presented in this paper an evaluation of different classification schemes leading to the proposition of a novel approach for visual object categorization, using statistical measures based image representation with new region based features. Thus, an evaluation of several feature selection schemes and classifier construction techniques facing unbalanced dataset has been carried out. Moreover, two concurrent fusion strategies, early and late fusion, have been considered in order to merge information from different “channels” represented by the different types of visual features extracted from images. Experiments performed on PASCAL 2007 database have shown that a good classification accuracy can be achieved with the image representation we propose and that our region based features carry information that is complementary to the one of SIFT features, especially when merging feature channels according to an early fusion strategy. In our future work, we envisage to explore two directions for improving our method. On the one hand, the efficiency of shape and texture features will be evaluated for our purpose of visual object categorization. On the other hand, based on the conclusions drawn in this paper, a hierarchical architecture for
classifiers will be envisaged to better handle the complex distribution of images from the different classes in the feature space.

Références


ANNEXE POUR LA FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE
PAPIER
DE LEUR ARTICLE

1. ARTICLE POUR LA REVUE :
   Studia Informatica Universalis.

2. AUTEURS :
   Huanzhang Fu — Alain Pujol
   — Emmanuel Dellandréa — Liming Chen

3. TITRE DE L’ARTICLE :
   Visual object categorization based on the fusion of region
   and local features

4. TITRE ABRG POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   VOC

5. DATE DE CETTE VERSION :
   10 octobre 2010

6. COORDONNES DES AUTEURS :
   – adresse postale :
     Université de Lyon
     Ecole Centrale de Lyon
     LIRIS, CNRS, UMR5205, F-69134, France
     \{huanzhang.fu,alain.pujol,emmanuel.dellandrea,liming.chen\}@ec-lyon.fr
   – tphone : 04 72 18 64 59
   – tlcopie : 04 72 18 64 43
   – e-mail : huanzhang.fu@ec-lyon.fr

7. LOGICIEL UTILIS POUR LA PRPARATION DE CET ARTICLE :
   \LaTeX{}, avec le fichier de style studia-Hermann.cls,
   version 1.2 du 03/12/2007.