Region Based Visual Object Categorization Using Segment Features and Polynomial Modeling

Huanzhang Fu, Alain Pujol, Emmanuel Dellandréa, and Liming Chen

LIRIS, UMR 5205 CNRS
Ecole Centrale de Lyon - 36 Av. Guy de Collongue - 69134 Ecully Cedex - France
{huanzhang.fu,alain.pujol,
emmanuel.dellandrea,liming.chen}@ec-lyon.fr

Abstract. This paper presents a novel approach for visual object classification. Based on Gestalt theory, we propose to extract features from coarse regions carrying visually significant information such as line segments and/or color and to include neighborhood information in them. We also introduce a new classification method based on the polynomial modeling of feature distribution which avoids some drawbacks of a popular approach, namely "bag of keypoints". Moreover we show that by separating features extracted from different sources in different "channels", which are then combined using a late fusion strategy, we can limit the impact of feature dimensionality and actually improve classification accuracy. Using this classifier, experiments reveal that our features lead to better results than the popular SIFT descriptors, but also that they 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.

1 Introduction

Generic visual object classification is one of the most challenging topics in computer vision. Indeed the number of real world object types as well as variations in view, imaging, lighting and occlusion pose serious problems in this domain. Furthermore we must add to this the difficulty 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 [1].

1.1 Related Work

Most works in the literature make use of a "bag of features" kind of approach [2,3] which tries to adapt the "bag-of-words" representation for text categorization to "Visual Object Categorization" (VOC) problem and has shown its effectiveness, obtaining the best performance in Pascal VOC contest [1]. These methods view images as an orderless distribution of local image features, typically using the popular SIFT features [4], extracted from salient image regions,

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 277–286, 2008.

called interest "points" [4,5,6] or more simply from points extracted using a grid [7]. The set of these local features is then characterized by a histogram of "visual keywords" from a visual vocabulary which is learned from the training set by a hard assignment (quantization) or a soft assignment through GMMs. These distributions can thus be compared to estimate the similarities between images and categorized through a machine learning process, for instance SVM.

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 this 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 is ruled by some Gestalt principles according to several studies on visual perception [8,9] 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 from well-known machine learning problems, namely the curse of dimensionality when increasing feature vector size which leads to exponential learning complexity as well as a small and biased training dataset, in particular with an imbalanced ratio of positives versus negative samples.

1.2 Our Approach

Our basic hypothesis is that effective visual object classification 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, especially 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 [10] 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, a *gestalt*. These grouping laws follow criterion such as spatial proximity, color similarity. They also highlight the interaction between regions.

We feel that the popular "bag of features" like approaches, while ignoring these principles, deprive themselves of meaningful information. Thus, instead of SIFT like local features, we propose some region-based meaningful features extracted from image regions with neighborhood information. These features result from perceptually significant "Gestalts" segmented according to some basic Gestalt grouping laws. Moreover, we propose using visually meaningful features, such as color and line segment based features which we extend to provide information from neighboring regions. Our goal is to compare the use of these region based

features with the popular SIFT features but also to check the efficiency of the combination of them to evaluate their complementarities.

We also propose a polynomial representation to model image feature distribution. Its interest is 3-fold. We can circumvent the difficulty of fixing the size of visual vocabulary, avoid the inaccurate assumption of Gaussian repartition of features and finally cope with a smaller number of feature vectors per image.

Finally, we also study and compare two fusion strategies, namely early fusion and late fusion. Experiments carried out show that our features can be combined with SIFT features using fusion strategy to provide better overall performance, especially in the case of using late fusion strategy.

The rest of the paper is organized as follows. In section 2 we describe our Gestalt-inspired region segmentation algorithm and the extraction of region-based features. Section 3 introduces our classification method while section 4 contains experimental results and discussion. Section 5 concludes the paper while depicting some future research directions.

2 Region Segmentation and Region-Based Features

In this section, we first introduce our Gestalt-inspired region segmentation scheme [11] and then our color and segment-based features extracted from the region map by our segmentation scheme.

2.1 Gestalt-Inspired Region Segmentation

The principle of our region segmentation algorithm is to segment an image into partial gestalts for further visual object recognition. These partial gestalts are significant regions within the image satisfying some grouping laws. The segmentation result should have some robustness against the color variation from one image to another. We thus made use of the following Gestalt basic grouping laws in our gestalt construction process: color constancy law, similarity law, vicinity

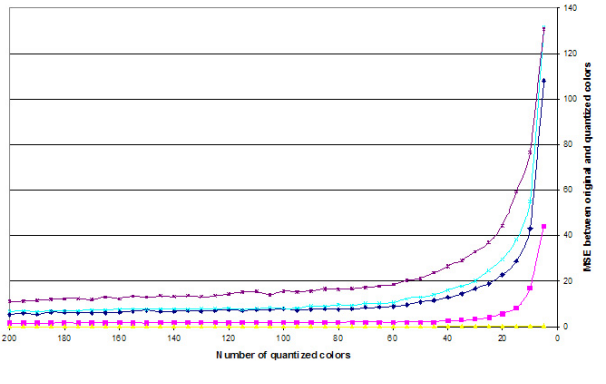

Fig. 1. Evolution of MSE between quantized colors and original colors

law and finally good continuation law. Because those laws are defined between regions and their context, at each step we assess the possibility to merge according to global information.

The segmentation is first based on color clustering making use of color constancy law but also it includes an extra post-processing step to ensure spatial consistency of the regions. In practice, our segmentation scheme is a 3-step process. First we filter the image for robustness to noise and reduce color depth by grouping perceptually similar colors. In the second step we use an iterative algorithm to determine a good color count which limits the quantization error. Indeed, quantization error can be measured by MSE between original and quantized color distribution as illustrated in Fig. 1, which clearly shows a sharp MSE rise when the number of color clusters is beyond a threshold. By performing several fast coarse clustering operations using Neural Gas algorithm [12], which is fast and less sensitive to initialization than its counterparts such as K-means, a target color cluster count is fixed at the number leading to a double MSE value of the one corresponding to 200 color clusters. We then use hierarchical ascendant clustering 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 [13]). Where n_i, μ_i, σ_i^2 are respectively the number of pixels, the average color and the variance of colors within region i.

$$D(R_1, R_2) = \frac{(n_1 + n_2)(\mu_1 - \mu_2)^2}{n_1\sigma_1^2 n_2\sigma_2^2} \quad . \tag{1}$$

Sample segmentation results on some Pascal challenge dataset images can be seen on Fig. 2. As we can see, we have original images at the first row having very different color distribution. The image on the left has very few colors, the middle one has very contrasted colors while the image on the right contains many colors. Our Gestalt-inspired segmentation algorithm has automatically adapted its segmentation process, producing significant partial gestalts.

Fig. 2. Sample Segmented Images

2.2 Region-Based Features

Two kinds of features are used in this work for the purpose of image classification: color features and segment-based features. Region based color features aim at capturing a coarse perception of partial gestalts. They are defined by color moments (mean, variance and skewness) [14] for each color channel in CIELch color space. CIELch was chosen in our work as variant of Lab color space as in (2), which best fits to the human perception [15]. These features are quite compact and have proven as efficient as a high dimension histogram [16].

$$L_{Lch} = L_{Lab} \qquad c = \sqrt{a^2 + b^2} \qquad h = \arctan \frac{b}{a} \ . \tag{2}$$

The segment-based features aim on the other hand at capturing some textural and geometrical properties of partial gestalts. They rely on a fast connective Hough transform (FCHT) [17] that quickly detects segments within an image. Once all segments identified by FCHT, they are distributed to the regions. Our segment-based features, a simple histogram combining length and orientation, are then computed as follows. We first divide lengths by the longest segment's length and then compute an average orientation so that all angles can be expressed with respect to it. We therefore obtain a feature that is invariant to translation, scale and 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 color and segment-based features are expressed at four 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 because on most images the fourth level will represent features extracted over the whole image.

3 Polynomial Image Modeling and Classification

We now turn to the problem of image modeling and classification. The basic problem is that the number of feature vectors as we propose to extract in the previous section can vary from one image to another, typically depending upon the number of segmented gestalts whereas machine-based learning schemes require a same data size for the classification. In the popular "bag-of-features" approach, one deals this problem by computing a kind of histogram for each image on the basis of a "visual vocabulary". This "visual vocabulary" is built using learning dataset either through a clustering algorithm on the set of feature vectors or the use of Gaussian Mixture Model (GMM). The drawbacks of such an approach are twofold. On the one hand, 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. On the other hand, when a GMM is used for a soft assignment, the number of parameters along with the number of Gaussians can quickly lead to the problem of "curse of dimensionality" [18].

Instead of building a "visual vocabulary", we propose here a simple polynomial modeling to characterize the visual content represented by the set of feature vectors extracted in the previous section. Moreover, we also propose to study several fusion strategies for a better image classification accuracy.

3.1 Polynomial Modeling Based Image Representation

In order to represent the visual content of an image characterized by a set of feature vectors, we propose to consider the distribution of values in each component of these feature vectors and to model such a distribution (histogram) by a simple polynomial. The coefficients of these polynomials will then be considered as the feature vector characterizing the visual content of an image.

The polynomial model for a given feature histogram is computed as follows. Given the set D of histogram values $D = \{(x_1, y_1), ..., (x_M, y_M)\}$ (M is the number of values), a polynomial $f(x)$ of degree N, described by its set of coefficients $P = \{p_1, p_2, ...p_{n+1}\}$, is computed to interpolate the data, by fitting $f(x_i)$ to y_i in a least squares sense. Thus, vector P can be used to characterize the distribution D. An example is given in Fig. 3. Once the distribution of each component from the feature set has been modeled thanks to a polynomial, a new image feature vector Q is produced by concatenating the coefficients of all polynomials.

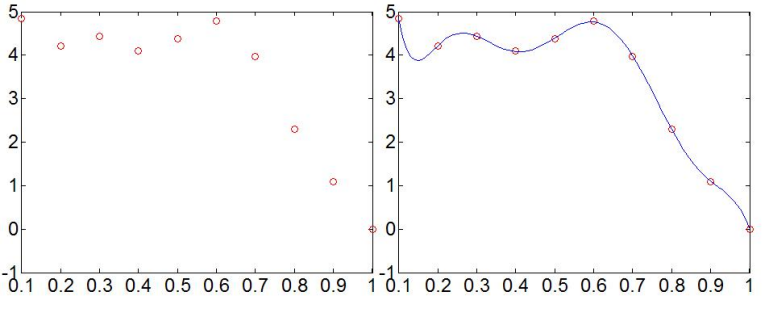

Fig. 3. (a) Histogram for one component of the image feature set. (b) A polynomial curve models the histogram in (a)

Assuming that our feature vector has L components and each component is modeled by a polynomial of degree N, then the vector Q has a dimension of $(N + 1) * L$, which generally ranges from hundreds to thousands. A vector of such high dimensionality used for classification can also lead to the "curse of dimensionality" [18]. Consequently, we further reduce the dimension using a feature selection method. While there exist several such methods in the literature [19], we have chosen the canonical discriminant analysis [20] as it is fast and enables an efficient feature dimension reduction by producing a new representation space which distinguishes the best the different classes. In most cases, we obtain

$K-1$ axes with the number of classes K. Thus with the help of this method, the overall feature vector Q becomes a much more simplified vector which is called in the subsequent *polynomial modeling based image representation.*

3.2 Classification and Fusion Strategies

Given an image to classify, we first characterize its visual content by extracting a set of feature vectors as proposed in section 2. However, other feature vectors, such as for instance SIFT, can also be used in our polynomial modeling and the subsequent classification process. Upon these feature vectors, our polynomial modeling based image representation is computed, leading to a single global feature vector for the input image. This new single global feature vector is then fed to a classifier beforehand trained or a set of such classifiers if a fusion strategy is applied, to judge whether this image contains a specified object. Any classifier, such as SVM or Neural networks, can be used for categorization of such an image representation. In our current experiments, we have chosen a simple multilayer perceptron for its ability to draw complex separating class border.

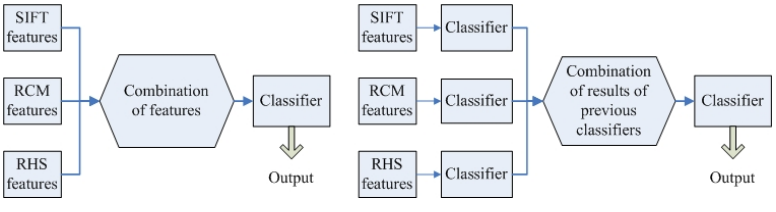

Fig. 4. (a) General scheme for early fusion (b) General scheme for late fusion

A fusion strategy can be applied when there exist several data sources in a classification process as for instance in video data analysis which can rely on both visual channel and audio one [21]. In our case, descriptors of different natures, such as SIFT, Region based Color Moments (RCM) and Region based Histogram of Segments (RHS), are extracted from an image. These descriptors can be considered as complementary modalities whose fusion can lead to a better accuracy in a classification process. Several fusion strategies can be applied: an early fusion is obtained when grouping all the features together and fed into a single classifier whereas a late fusion is to make use of "channels" with a separate classifier for each kind of features, the outputs of these classifiers being merged later [21] in a process similar to boosting [22]. Between these two strategies, there exist many intermediate strategies which consist in generating intermediate classes from different sources and to take a final decision based on these intermediate classes. In our current experiments, we have studied the two main strategies: early fusion and late fusion. Their schemes are illustrated respectively in Fig. 4(a) and 4(b).

4 Experimental Results

As a first experimental validation and evaluation of our approach, we have chosen 5 semantic representative classes namely airplane (248 images), bicycle (243 images), bus (186 images), horse (287 images) and person (2008 images) from the dataset of [1]. One versus all others multilayer perceptron was built for each class with a 4 fold-cross-validation. The structure of these perceptrons is composed of one hidden layer for all the tests, and the number of neurons in the hidden layer varies according to the number of inputs, and has 3 values: 5, 15, 2 for single channel, early fusion and late fusion respectively. Finally, the degree of polynomial for modeling the visual content of an image has been empirically set to 8.

Our region-based features, namely RCM and RHS, along with the popular SIFT features [23], were considered in our experiments. Early and late fusion strategies are also studied by interpreting these three types of features as different modalities. RCM and RHS are first merged by the strategies of Early Fusion and Late Fusion, noted as EF(RCM+RHS) and LF(RCM+RHS), and SIFT is combined afterwards, noted as EF(RCM+RHS+SIFT) and LF(RCM+RHS +SIFT). Here, we also use one versus all others multilayer perceptron in the two

Table 1. Results of 5 classes for object categorization

Classification rate	Plane	Bicycle	Bus	Horse	Person
SIFT	65,00%	55,21%	60,75%	65,49%	58,94%
RCM	72,69%	61,57%	67,90%	65,84%	62,77%
RHS	76,60%	61,98%	66,13%	62,59%	63,54%
EF(RCM+RHS)	80,34%	63,97%	70,75%	65,63%	65,17%
EF(RCM+RHS+SIFT)	81,47%	64,63%	69,30%	66,43%	65,50%
LF(RCM+RHS)	82,02%	70,95%	91,99%	79,65%	66,74%
LF(RCM+RHS+SIFT)	85,21%	72,73%	92,74%	81,54%	69,41%
Recall rate	**Plane**	**Bicycle**	**Bus**	**Horse**	**Person**
SIFT	68,66%	57,90%	62,58%	71,57%	60,93%
RCM	73,45%	64,10%	68,06%	66,53%	67,27%
RHS	76,55%	68,32%	71,61%	67,09%	69,11%
EF(RCM+RHS)	80,17%	65,67%	70,86%	67,09%	68,42%
EF(RCM+RHS+SIFT)	81,43%	66,67%	70,54%	70,10%	68,57%
LF(RCM+RHS)	84,20%	73,86%	89,35%	79,48%	70,01%
LF(RCM+RHS+SIFT)	85,38%	74,86%	89,78%	83,89%	72,89%
Precision rate	**Plane**	**Bicycle**	**Bus**	**Horse**	**Person**
SIFT	63,98%	54,90%	60,37%	63,76%	58,60%
RCM	72,35%	60,98%	67,85%	65,56%	61,72%
RHS	76,62%	60,60%	64,53%	61,49%	62,08%
EF(RCM+RHS)	80,44%	63,47%	70,71%	65,13%	64,24%
EF(RCM+RHS+SIFT)	81,50%	64,02%	68,84%	65,25%	64,61%
LF(RCM+RHS)	80,68%	69,77%	94,32%	79,71%	65,72%
LF(RCM+RHS+SIFT)	85,09%	71,77%	95,43%	80,08%	68,14%

fusion strategies. The performance of our experiments was evaluated through three classical rates, namely classification rate, recall rate and precision rate. The detailed results are presented in Table 1.

In Table 1, experimented classifiers can be categorized into 3 classes: Single Channel (SC) which means make use of only one kind of features, Early Fusion (EF) and Late Fusion(LF). As we can see, our region-based features, RCM and RHS, with an improvement of 5 points in average, perform better than SIFT features. These results tend to show the effectiveness of our RCM and RHS features using the polynomial modeling based image representation. Between RCM and RHS, we find that RHS is slightly better after comparing all 3 rates and RCM tends to favor negative side. Now focusing on EF and LF, we can note that the best classification rates are obtained when the 3 channels are merged using LF strategy which performs much better than SC and EF. The classes of bus and horse, for instance, record a classification rate increase by about 22 points and 15 points respectively compared to the second higher rate obtained in EF. This result seems to suggest that the three different channels carry complementary visual contents and their fusion helps to improve the final classification accuracy. Another reason might be that EF may have more chances to result in a conflict between different features which blur the boundary between classes. It can also explain that why EF performs only slightly better than SC and much worse than LF.

5 Conclusion and Perspectives

We have proposed in this paper a novel approach for visual object categorization, using polynomial modeling based image representation with new region based features. Two different fusion strategies, early and late, have been considered to merge information from different "channels" represented by the different types of features. Results have shown that good performance can be achieved with our approach and that our segment features carry information which is complementary to SIFT features, especially when merging feature channels according to a late fusion strategy. Encouraged by these first experimental results, we are currently experimenting our approach on the whole database in [1]. Among other perspectives under way, we are optimizing the quantification technique used for the construction of feature histograms as well as the choice of the degree of polynomial for modeling. Concerning features, we will evaluate richer representations of our region based features (e.g.: coocurrence matrix, ...) and also test more channels such as region shape and texture.

References

1. The PASCAL Visual Object Classes Challenge 2007 Results (2007), `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/`
2. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)

3. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints. International Journal of Computer Vision 66(3) (2006)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
5. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30(2), 79–116 (1998)
6. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60(1), 63–86 (2004)
7. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), pp. 524–531 (2005)
8. Kaniza, G.: Grammatica del vedere. Il Mulino (1997)
9. Wertheirmer, M.: Untersuchungen zur lehre der gestalt ii. Psychologische Forschung 4, 301–350 (1923)
10. Desolneux, A., Moisan, L., Morel, J.: From Gestalt Theory to Image Analysis: A Probabilistic Approach. Springer, Heidelberg (2008)
11. Pujol, A., Chen, L.: Coarse adaptive color image segmentation for visual object classification. In: Proceedings of the 15th International Conference on Systems, Signals and Image Processing (2008)
12. Martinetz, T., Schulten, K.: A "neural-gas" network learns topologies. Artificial Neural Networks I, 397–402 (1991)
13. Zhu, S.C., Yuille, A.: Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 18(9), 884–900 (1996)
14. Stricker, M.A., Orengo, M.: Similarity of color images. In: Storage and Retrieval for Image and Video Databases (SPIE), pp. 381–392 (1995)
15. Trémeau, A., Fernandez-Maloigne, C., Bonton, P.: Digital Color Imaging - From acquisition to Processing. Dunod (in French) (January 2004)
16. Deng, Y., Manjunath, B., Kenney, C., Moore, M., Shin, H.: An efficient color representation for image retrieval. IEEE Transactions on Image Processing 10(1), 140–147 (2001)
17. Ardabilian, M., Chen, L.: A new line extraction algorithm: Fast connective hough transform. In: Proceedings of PRIP 2001, p. 127 (2001)
18. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
19. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Oxford University Press, Oxford (2007)
20. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936)
21. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: MULTIMEDIA 2005: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 399–402. ACM, New York (2005)
22. Freund, Y., Schapire, R.: A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
23. Nowozin, S.: Libsift - scale-invariant feature transform implementation (2005), `http://user.cs.tu-berlin.de/~nowozin/libsift/`