# A Comparison of Multiclass SVM Methods for Real World Natural Scenes

Can Demirkesen, Hocine Cherifi

# A Comparison of Multiclass SVM Methods for Real World Natural Scenes

Can Demirkesen[1] and Hocine Cherifi[1,2]

[1] Institue of Science and Engineering, Galatasaray University, Ortakoy, Istanbul, 34257, Turkey
candemirkesen@gmail.com
[2] Faculté des Science Mirande 9, Avenue Alain Savary, BP 47870 21078 Dijon, France
hocine.cherifi@u-bourgogne.fr

**Abstract.** Categorization of natural scene images into semantically meaningful categories is a challenging problem that requires usage of multiclass classification methods. Our objective in this work is to compare multiclass SVM classification strategies for this task. We compare the approaches where a multi-class classifier is constructed by combining several binary classifiers and the approaches that consider all classes at once. The first approach is generally termed as "divide-and-combine" and the second is known as "all-in-one". Our experimental results show that all-in-one SVM outperforms the other methods.

## 1 Introduction

Rapidly growing need for natural image classification challenges both image content representation studies and classification techniques. A performing multiclass classifier is needed when classifying natural scenes. From model-based methods to learning algorithms, there are many choices for an appropriate classifier. Among these, support vector machines (SVMs) appear to be a good candidate because of their ability to generalize in high-dimensional spaces without the need to add a prior knowledge. The appeal of SVMs is based on their strong connection to the underlying statistical learning theory. For several pattern classification applications [1][2], SVMs have been shown to provide better generalization performance than traditional techniques such as neural networks [3].

SVM is basically conceived for binary classification. The idea is to separate two classes by calculating the maximum margin hyperplane between the training examples. Several methods have been proposed to extend SVM in order to classify more than two classes because classification problems are mostly multi class. Image classification is naturally a multi class problem as well. Currently there are two major approaches for extending SVM to multiclass classification: (1) considering all data in a single optimization. (2) Combining several binary SVM classifiers; generally the first approach is called 'all-in-one' (AIO). and the second 'divide-and-combine' The main methods for divide-and-combine are One-Against-All (OAA), One-Against-One (OAO) and Directed Acyclic Graph (DAG). There is some work in the literature [1],[2] comparing these methods for classical datasets like iris, wine, glass, letter etc. In [2], OAA, OAO MaxWins (with majority voting), DAG and AIO are compared.

J. Blanc-Talon et al. (Eds.): ACIVS 2008, LNCS 5259, pp. 752–763, 2008.

The authors show that there is not one method that performs best for every dataset but that OAO MaxWins and DAG perform better with large number of classes. In [1], OAO MaxWins, OAA, DAG and Neural Networks are compared. The authors show that the methods have comparable performance on accuracy and error rate but that OAO and DAG need less time for training phases. This conclusions point out that one should compare these methods for a specific classification problem, in this case image classification, because the best method can depend on the problem at hand. In [4], OAO MaxWins, OAO-Pairwise Coupling, DAG and Neural Networks are compared for natural texture images like grass, leaves, brick etc. using a mixed color and texture representation. The authors conclude that OAO MaxWins and DAG have almost the same performance and they are both better then neural networks in terms of accuracy. In [5], OAO MaxWins, OAA, DAG, maximum likelihood and back propagation neural networks are compared for satellite images like water, construction, wood, bare soil etc. using topographical raster data for image representation. According to their results OAO with MaxWins majority voting is the most performing in terms of accuracy. In [6], AIO, neural networks, discriminant analysis and decision trees are compared for land cover images using random pixels for image representation. It is shown that AIO SVM outperforms other techniques in accuracy.

Despite all these studies there is not a fully complete comparison of multi-class SVM classification methods. The comparisons in [1] and [5] do not cover OAO pairwise coupling and AIO. In [4], OAA and AIO are missing. Even in the most complete comparison [2], OAO-Pairwise Coupling method is missing. In addition to our knowledge there is not one comparison of these methods for classification of real world natural scenes like forest, coast, mountain or city view categories and previous results show that performances are greatly influenced by nature of the data. Another criticism to existing work is that only one performance measure is used. Different performance criteria measure different tradeoffs in the predictions made by a classifier, and it is possible that a learning method performs well on one metric, but be suboptimal on other metrics. Because of this it is important to evaluate algorithms on a broad set of performance metrics.

In this paper, we evaluate and compare all of the multiclass SVM methods mentioned above on a challenging image database by following an experimental approach. We compare performance of the methods for natural image categorization task using global and local image representations. We perform an extensive evaluation using multiple performance measures. The rest of the paper is organized as follows: Section 2 introduces multiclass methods. In section 3 features used for image representation are briefly presented. Section 4 describes performance measures that we use in experimentations. Experimental results are given in section 5 and finally conclusion in section 6.

## 2 Multi Class SVM

### 2.1 Divide and Combine

Strategies described below can be applied to build N-class classifiers using binary SVM classifiers. They can be decomposed in two main steps: (1) classification, (2)

fusion. In classification step an instance x is classified by all of the binary classifiers. In the fusion step, classification outputs are combined together to provide a decision.

**One Against One SVM Classifiers**

One-Against-One (OAO) method involves N(N-1)/2 binary SVM classifiers. Each classifier is trained to separate each pair of classes. There are different strategies used to combine these binary classifiers. The main strategies widely used in literature are 'Pairwise Coupling' and a majority voting strategy which is called 'MaxWins'. When classifiers are combined through majority voting scheme, the class with maximal number of votes is the estimation. In pairwise coupling [3], a pairwise probability $p_{i,j}$ is obtained from each binary SVM output noted as $f_{i,j}(x)$.

$$p_{ij} = \frac{1}{2} f_{ij}(x) + 0.5 . \tag{1}$$

These pairwise probabilities are coupled into a common set of posterior probabilities $p_i$:

$$p_i = \frac{2}{N(N-1)} \sum_{j \neq i} p_{i,j} . \tag{2}$$

The decision function is given by:

$$c(x) = \arg\max_{1 \leq i \leq N} (p_i) . \tag{3}$$

**DAG SVM classifiers**

Directed Acyclic Graph (DAG) SVM is proposed by Platt et al [1]. Training is the same as the OAO using N(N-1)/2 binary SVMs. However, in the testing phase, a directed acyclic graph with N(N-1)/2 internal nodes and N leaves is used. Testing a sample starts at the root node and it moves to either left or right depending on the output value. Therefore, we go through a path before reaching a leaf node, which indicates the predicted class. An advantage of using a DAG is that its testing time is less than the OAO methods. An example DAG for three classes is given in Figure 1.

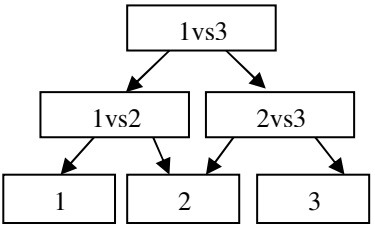


**Fig. 1.** Decision DAG for three classes. Each node is a binary classifier.

**One-Against-All SVM classifiers**

One-Against-All (OAA) is the most common and simplest approach [7]. It involves N binary SVM classifiers, one for each class. Each binary SVM is trained to separate

one class from the rest. The winning class is the one that corresponds to the SVM with highest output value i.e. the largest decision function value. This approach may suffer from error caused by markedly imbalanced training sets. The decision function for OAA is :

$$c(x) = \arg\max_{1 \le i \le N} f_i(x)$$

Where $f_i(x)$ is the output of the binary SVM classifier trained for class *i* against all the other classes.

### 2.2 All-in-One

There are certain limitations of the approaches that extend binary SVMs to multi-class problems. One of these limitations is that they do not consider the full problem directly. The one-against-all approach degrade the balance of the training sets (there are far more negative training examples in each binary classifier's training set), and the one-against-one method uses only information from the two classes that it works with. Each one-against-one classifier loses the information from all the remaining classes. All-in-one (AIO) is a more natural approach that considers the multi-class problem directly as a generalization of the binary classification algorithm The idea is similar to the OAA approach. It constructs *N* two-class rules where the *ith* function separates training vectors of the class *i* from the other vectors. Hence there are *N* decision functions but all are obtained by solving one problem [8]. The decision function is:

$$c(x) = \arg\max_{1 \le i \le N} (w_i x + b_i)$$

Where $w_i x + b_i$ is the hyper plane that separates the class *i* from the other classes. Note that for N=2, this formulation reduces to the binary SVM decision function.

## 3 Image Representation

Existing image categorization systems in the literature can be generally classified into two categories based on the underlying framework for image content representation. The first category segments the image into some meaningful components and uses them as semantic elements to characterize image content. The second category takes an image as a whole visual appearance and characterizes image contents by using image-based global visual features. These two approaches usually called as local and global have both shortcomings, for instance global approach do not take into account individual objects. And for local approach, image sub-blocks have little correspondence with global semantic of the image. Therefore, local and global representations are used together in order to overcome their shortcomings. The role of local and global information has also been studied by numerous experiments on human participants. These studies have shown that local part-based information and global information are processed separately by human visual system then integrated together [9].

In a previous work we evaluated a collection of local and global features used to represent color, texture, edge and spectral information for binary classification [10]. We have shown that texture leads to the highest classification accuracy as a local representation, while spectral information is the most performing global representation. We used co-occurrence matrices to characterize texture information and gist to characterize spectral information. Gist is a low dimensional representation of the scene structure based on the output of filters tuned to different orientations and scales [12]. To sum up the previous work, holistic spatial scene properties may be best estimated using spectral information and local information is best described by texture information. The combination of texture and gist improves classification performance in binary classification. The same representations can be adopted for multi class image classification.

## 4 Performance Measures

An important aspect of our study is the use of performance criteria to evaluate multi class SVM methods. Classification techniques are now used in many domains, and different performance metrics are appropriate for each domain. There exist numerous performance measures in the literature of image classification domain. For example Precision/Recall measures are used in information retrieval. The most widely used methods are correct classification rate [6], error rate [1], classification accuracy in percentage [4], [5] and ROC curves (sensitivity-specificity curves). A brief summary of the measures with the formulas are given in Table 1.

**Table 1.** Performance measures derived from the confusion matrix

| Measure | Formula |
|---|---|
| Precision | $TP/(TP+FP)$ |
| Recall | $TP/(TP+FN)$ |
| Specificity | $TN/(TN+FP)$ |
| F-measure | *2.Precision.Recall/(Precision+Recall)* |
| Accuracy | $(TP+TN)/(TP+TN+FP+FN)$ |
| Kappa | $\dfrac{(TP+TN)-(TP_{\exp ected}+TN_{\exp ected})}{(TP+TN+FP+FN)-(TP_{\exp ected}+TN_{\exp ected})}$ |

Let us consider that instances belong either to a positive class or to a negative class. The entries of a confusion matrix are true positives (TP) correctly classified positive instances, false positives (misclassified negatives), true negatives (correctly classified negatives) and false negatives (misclassified positives). Accuracy is the

simplest way to compare two confusion matrices because it is a measure that represents the whole classification not only one class prediction. That is the reason why accuracy is the most widely used measure. In an N classification problem Precision, recall, specificity and F-measure represent the performance of the prediction for only one class. To compute these measures one has to consider the class under investigation as positive and all the other classes as negative. Kappa statistic is used to compare the degree of consensus between raters. In this context it is used to measure the quality of classification. Like accuracy, kappa statistic can represent a confusion matrix with a single value. It varies in interval [-1, 1], 1 for perfect classification and -1 for a classifier that makes wrong decision systematically. Expected values of confusion matrix elements are obtained by:

$$TP_{\exp ected} = \frac{(TP+FN)\cdot(TP+FP)}{TP+TN+FP+FN}, \quad TN_{\exp ected} = \frac{(FP+TN)\cdot(TN+FN)}{TP+TN+FP+FN}$$

## 5 Experimental Results

### 5.1 Image Database

Our image database contains 8 categories of natural scenes: highway(260), streets(292), forest(328), open country(410), inside of cities(308), tall buildings(356), coast(360) and mountain(374) images (Numbers in brackets represent the size of each categories). The database provided by Oliva and Torralba was collected from a mixture of COREL images as well as personal photographs [11]. All images are colored and sized of 256x256 pixels. For each classification experiment 100 images of each category are reserved for test purpose and the remaining images are used as training set. Samples images for the 8 categories are given in Figure 2.

### 5.2 Choice of Modalities

We use our image database to generate two groups of images that contain both four classes. These groups are arranged in such a way that one group contains the four

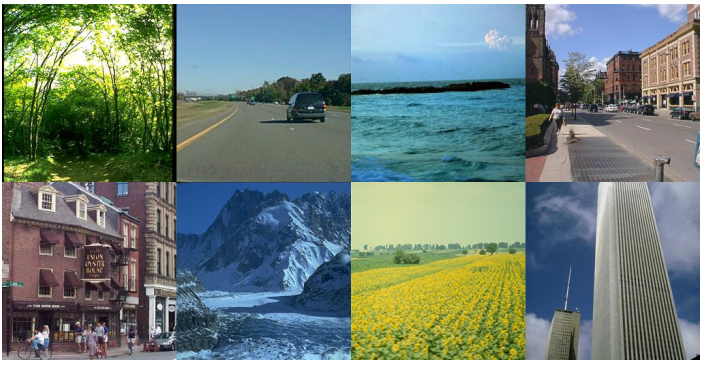

**Fig. 2.** Sample images of the database. From top left to bottom right: Forest, Highway, Coast, Street, Inside of city, Street, Mountain, Open Country, Tall building.

**Fig. 3.** Sample images from the least similar classes: Forest, highway, coast, street

**Fig. 4.** Sample images from the most similar classes: Inside of city, street, tall building, mountain

most similar classes and the other the four least similar ones. We use these two groups in the remaining experimentations to compare multiclass classification methods. We suppose that if two classes are similar then the binary classification performance for these classes is low and vice versa. In other words, similarity of two classes varies in the opposite way with binary classification accuracy of these classes. In order to obtain the groups of classes mentioned above we performed binary classifications between every possible pair of classes in our image database (Combination (8, 2) =28) based on a local texture feature that is obtained by extracting four attributes namely energy, entropy, homogeneity and inertia from gray level co-occurrence matrix. This feature is extracted from block of 64x64 pixels. We sorted the binary classification results by accuracy. Keeping in mind that 6 classifiers are needed to build a 4-class classifier the 6 best performing classifiers sufficient to construct a 4-class classifier are selected; these four classes are the most similar ones according to the feature that is used. Following the same procedure the four least similar classes are found. The four most similar classes are Inside of city, Street, Tall building and Mountain and the four least similar classes are Forest, Highway, Coast, Street (Figure 2 and Figure 4). This result is in accordance with ordering based on spectral signature as presented in [11]. For each category the spectral signature is obtained by averaging the power spectra of a few hundred images that belong to this category. The authors showed that spectral signature is very appropriate to discriminate the categories. Categories very close to each other exhibit similar spectral signatures while for more distant categories the shape of the spectral signatures is less similar.

### 5.3 Classification Based on Local Representation

Texture feature that is defined in 5.2 is extracted from images on blocks of 64x64 pixels. The most similar image classes are used for classification. Classification

**Table 2.** Classification results for the least similar classes using local image representation

| Methods | *MeanFmeasure* | *Accuracy* | *Kappa* |
|---|---|---|---|
| DAG | 0.767 | *0.767* | *0.690* |
| OAO-Pairwise Coupling | 0.774 | *0.775* | *0.700* |
| OAO-MaxWins | 0.779 | *0.780* | *0.706* |
| OAA | 0.783 | *0.785* | *0.713* |
| AIO | **0.789** | ***0.790*** | ***0.720*** |

results in terms of three performance measure are presented in Table 2. Mean F-measure is calculated by averaging F-measures of individual classes.

All three performance measures agree on the rank of the strategies under investigation. The results show that the methods are ascendant ordered as AIO, OAA, OAO-MaxWins, OAO-Pairwise Coupling and DAG. One should note that the best performing two strategies (AIO and OAA) have similar training phases. Remaining three methods use the exact same binary classifiers; MaxWins voting strategy is the winner of these last three methods. The most discriminative performance measure is Kappa statistic because it has a wider range comparing to the others.

**Table 3.** Classification results for the most similar classes using local representation

| Methods | *MeanFmeasure* | *Accuracy* | *Kappa* |
|---|---|---|---|
| DAG | 0.579 | 0.582 | 0.443 |
| OAO-Pairwise Coupling | 0.592 | 0.590 | 0.453 |
| OAO-MaxWins | **0.603** | 0.600 | 0.466 |
| OAA | 0.552 | 0.587 | 0.450 |
| AIO | 0.598 | ***0.605*** | ***0.473*** |

We performed the same experiment using the most similar classes; results are shown in Table 3. Accuracy and kappa statistic perfectly agree with the ordering AIO, OAO-MaxWins, OAO-Pairwise Coupling, OAA and DAG. According to F-measure the ordering is OAO-MaxWins, AIO, OAO-Pairwise Coupling, DAG and OAA. This is due to the surprisingly low recall values of Inside of city and mountain classes. The extreme F-measure values for these two classes decreased the mean F-measure value for OAA and AIO. Accuracy and kappa statistic has not been influenced by that as much as mean F-measure because both accuracy and kappa statistic are calculated in a

more global way to summarize the confusion matrix. Note that the overall performance decreased comparing to the previous experiment with the least similar classes.

### 5.4 Classification Based on Global Representation

The least similar classes have been classified using global representation; results are shown in Table 4. The methods are ordered by their performance as AIO, OAO-MaxWins, OAA, OAO-Pairwise Coupling and DAG for all three performance measure. An increase of performance is observed for all five methods comparing to the classification based on texture feature with the same classes. This shows that gist is more discriminative then texture feature. Kappa statistic is the most discriminative performance measure for this classification.

Classification results of the four most similar classes using global representation is shown in Table 5. Methods are ranked as: AIO, OAA, OAO-MaxWins, OAO-Pairwise Coupling and DAG for all the performance criteria. OAO-MaxWins, OAO-Pairwise Coupling and DAG that have the same binary classifiers are grouped together in performance rank. OAA and AIO made a second group with very similar results that can be explained by the similarity of their training phases. Overall performance in this experiment is better than the performances in classifications of the least similar

**Table 4.** Classification results for the least similar classes global representation

| Methods | *MeanFmeasure* | *Accuracy* | *Kappa* |
|---|---|---|---|
| DAG | 0.874 | 0.875 | 0.833 |
| OAO-Pairwise Coupling | 0.891 | 0.892 | 0.856 |
| OAO-MaxWins | 0.914 | 0.915 | 0.886 |
| OAA | 0.904 | 0.905 | 0.873 |
| AIO | **0.941** | ***0.942*** | ***0.923*** |

**Table 5.** Classification results for the most similar classes using global representation

| Methods | *MeanFmeasure* | *Accuracy* | *Kappa* |
|---|---|---|---|
| DAG | 0.773 | 0.775 | 0.700 |
| OAO-Pairwise Coupling | 0.816 | 0.817 | 0.756 |
| OAO-MaxWins | 0.819 | 0.820 | 0.760 |
| OAA | 0.845 | 0.847 | 0.796 |
| AIO | **0.858** | ***0.860*** | ***0.813*** |

**Table 6.** Correlation Coefficients between Performance Measures

| Class Similarity | Representation | Fm-Acc | Fm-Kappa | Acc-Kappa |
|---|---|---|---|---|
| Least similar | Local | 0.9983 | 0.9987 | 0.9995 |
| Most similar | Local | 0.6786 | 0.6581 | 0.9995 |
| Least similar | Global | 1.0000 | 0.9999 | 1.0000 |
| Most similar | Global | 0.9999 | 1.0000 | 1.0000 |

classes based on local representation which is an interesting conclusion that confirms the superiority of the discriminative power of global representation.

In order to evaluate the redundancy of the performance measures we calculated correlation coefficients. If two measures perfectly agree on every case it means that one of the measures is redundant. Correlation coefficients for different representation systems and for two extremes cases (the least and the most similar classes) are given in Table 6. Correlation coefficients vary in interval [-1, 1]. 1 is for perfect agreement between measures and -1 for disagreement. It is noted that accuracy and kappa statistic are perfectly correlated for each of the cases. So we can use only one of them as measure of performance. Kappa statistic is a better candidate because its range is wider than accuracy. There is perfect agreement between performance measures for all experiments except from the second experiment where this is due to the sensibility of mean F-measure to an extreme case.

**Table 7.** Overall Ranking of the multiclass strategies according to mean F measure and Kappa statistic A: Least similar images and local representation, B: Most similar images and local representation C: Least similar images and global representation D:Most similar images and global representation

| Methods | A | B | C | D | Total score A+C+D |
|---|---|---|---|---|---|
| AIO | 1+1 | 2+1 | 1+1 | 1+1 | 6 |
| OAO-MaxWins | 3+3 | 1+2 | 2+2 | 3+3 | 16 |
| OAA | 2+2 | 5+4 | 3+3 | 2+2 | 14 |
| OAO-Pairwise Coupling | 4+4 | 3+3 | 4+4 | 4+4 | 24 |
| DAG | 5+5 | 4+5 | 5+5 | 5+5 | 30 |

Table 7 shows the rank of methods under investigation according to mean F-measure and kappa statistic separated by '+'. For example, '2+3' means that the rank of a method is 2 according to f measure; 3 according to kappa statistic.

If we exclude the experiment involving the most similar classes with a local representation (B) one should note that performance measures perfectly agree on the rank of the methods without any exception. This is also reflected on the correlation coefficient table where the correlation values are near 1. If we rank the classification strategies according to the total score for these three cases the most performing method is AIO followed by OAA, OAO-MaxWins, OAO-Pairwise Coupling and DAG. Note that the total score of OAO-MaxWins and OAA are very close. If we rank the

strategies according to the total score using the four cases (A+B+C+D) the results are identical except for OAO-MaxWins and OAA which switch places. Nevertheless their scores are still very close.

To summarize we can say that AIO is always the most performing method in any situation followed by either OAO-MaxWins or OAA whose performance are very similar. One should note that there is a relation of ordering that remains unchanged for all of the four experiments and with agreement of all three performance measures without exception. OAO-MaxWins > OAO-Pairwise Coupling > DAG. It is an interesting result considering that these methods use the same binary SVM classifiers.

## 6 Conclusion

Results show that All-In-One method is the most performing SVM multiclass classification strategy for natural scene classification. This conclusion is confirmed with all four experiments performed on two separate groups of images using two different types of representation, one local, one global. We evaluated five methods, the first, fourth and last places of the ordering are very robust to image representation system and to class proximity. Our work show that the usage of multiple performance measures, multiple image representations and multiple test groups is not only preferable but also necessary because considering experiments separately leads to different performance orderings. An important result is that MaxWins majority voting is always the best performing strategy among the one-against-one strategies.

## References

1. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large Margin Dags for Multiclass Classification. Advances in Neural Information Processing Systems 12, 547–553 (2000)
2. Hsu, C., Lin, C.: A Comparison of Methods for Multi-class Support Vector Machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)
3. Kreßel, U.H.G.: Pairwise Classification and Support Vector Machines. In: Advances in Kernel Methods, pp. 255–268. MIT Press, Cambridge (1999)
4. Ren, J., Shen, Y., Ma, S., Guo, L.: Applying Multi-Class SVMs into Scene Image Classification. In: Proceedings of the 17th International Conference on Innovations in Applied Artificial Intelligence, pp. 924–934 (2004)
5. He, L., Kong, F., Shen, Z.: Multiclass SVM Based Land Cover Classification With Multi-Source Data. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3541–3545 (2005)
6. Foody, G.M., Mathur, A.: A Relative Evaluation of Multiclass Image Classification by Support Vector Machines. IEEE Transactions on Geoscience and Remote Sensing 42, 1335–1343 (2004)
7. Vapnik, V.: The Nature of Statistical Learning Theory. Wiley, New York (1998)
8. Weston, J., Watkins, C.: Support Vector Machines for Multi-Class Pattern Recognition. In: Proceedings of the Seventh European Symposium on Artificial Neural Networks (1999)
9. Vogel, J., Schwaninger, A., Wallraven, C., Bülthoff, H.: Categorization of Natural Scenes: Local vs. Global Information. In: Symposium on Applied Perception in Graphics and Visualization APGV, Boston, MA, USA (2006)

10. Demirkesen, C., Cherifi, H.: Local or Global Image Representation for Support Vector Machine Image Categorization. In: IEEE The 15th International Conference on Systems, Signals and Image Processing, IWSSIP (to appear, 2008)
11. Torralba, A., Oliva, A.: Statistic of Natural Image Categories. Network: Computation in Neural Systems 14, 391–412 (2003)
12. Oliva, A., Torralba, A.: Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research: Visual perception 155, 23–36 (2006)