Painting Scene Recognition Using Homogenous Shapes

Razvan George Condorovici*, Corneliu Florea, and Constantin Vertan

The Image Processing and Analysis Laboratory, LAPI
University "Politehnica" of Bucharest,
Bucharest, Romania
rcondorovici@alpha.imag.pub.ro, corneliu.florea@upb.ro,
constantin.vertan@upb.ro

Abstract. This paper addresses the problem of semantic analysis of paintings by automatic detection of the represented scene type. The solution comes as an incipient effort to fill the gap already stated in the literature between the low level computational analysis and the high level semantic dependent human analysis of paintings. Inspired by the way humans perceive art, we first decompose the image in homogenous regions, follow by a step of region merging, in order to obtain a painting description by the extraction of perceptual features of the dominant objects within the scene. These features are used in a classification process that discriminates among 5 possible scene types on a database of 500 paintings.

Keywords: scene analysis, scene classification, perceptual segmentation, paintings

1 Introduction

Since ancient times there was a close connection between human kind evolution and the form of art it produced. With the entering into the digital era and fostered by the growth of computer usage in daily life, there can be seen considerable efforts of creating automatic systems for facilitating a better understanding of art. As noted in the reviews of Stork et al. [18] and Cornelis et al. [4] such topics cover a very wide range, from systems used for high quality and accuracy digitization of paintings to image analysis and diagnostics or virtual restoration, color rejuvenation, pigment analysis, brush stroke analysis, lightning incidence, craquelure analysis or painting authentication and classification.

In most of the cases, adapting classical image processing techniques for art understanding offered good results, proving that computers are indeed able to

* This work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the European Social Fund Financial Agreements POSDRU/107/1.5/S/76903

help humans, experts or amateurs, to better comprehend art. However, art is fundamentally about humans and the way humans see and feel art is a major component in this understanding process. Following psycho-visually experiments, Ramachandran [15] concluded that the key for understanding the art perception are the human perceptual processes, rather than the analysis of the aesthetic properties, noting that *"the painter does not paint with his eyes, but with his brain"*. Subsequently, many works integrated the human perception principles and often concluded that a semantical interpretation of the painting is highly needed [8], [10], [16], [19]. The here proposed method follows the same line.

Regarding the analysis level, Wallraven et al. [19] observed that there are three possibilities: a low level of pictorial information (consisting in technique, thickness of brush strokes, type of painting material, color composition), a middle level about the information content (specific objects, type of painting or subject) and a high-level information containing background data (historical events or artist and period in general). While the first and third level are generally reserved for art experts, the mid-level content information is the most accessible for non-expert art viewers. By contrast, most of the computer aided painting analysis solutions known in the literature focus on the low-level features, as they are more suitable to be modelled with classical image processing techniques and the works focusing on mid-level features are rather scarce.

In this category we note the efforts in creating a system for mid-level painting analysis by Carneiro et al. [2]. They propose a solution for detecting the painting theme from a set of 27 types and for identifying human bodies in a database containing 988 grayscale images of paintings by exploring the accuracy of different state of the art inductive and transductive photographic image annotation methods. Yet, following [15], paintings are more perceptually constructed, when compared with natural photographs, which are more chaotic, thus motivating a perceptual approach on the scene analysis topic. Our current work is situated in the same mid-level, as it focuses on detecting the painting subject from five possible categories (portrait, nude, landscape, cityscape and still life) relying on a perceptual approach.

While we note the lack of relevant state of the art in painting scene analysis, nevertheless we stress that the literature holds many attempts of detecting the scene type in general photographic images. Lazebnik et al. [13] proposed a pyramidal decomposition of the scene in sub-regions described through histograms of local features. Oliva et al. [14] developed a spatial envelope model of perceptual features, constructing the now state of the art GIST feature image description for a detection rate of $\approx 90\%$ when classifying natural images from 4 classes. More recently, an unsupervised classification of 13 types of natural scenes, based on an hierarchical Bayesian model using codebooks, was proposed by Fei-Fei and Perona [9], reporting a detection rate of 76%. We note that, as a general direction, the query image is described by features that are salient with respect to a multi-resolution pyramidal, global decomposition, which, again, is different from the way humans perceive things.

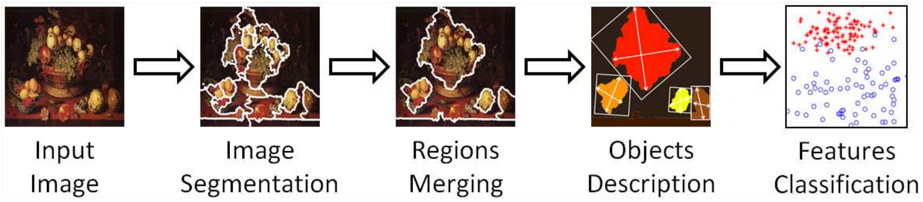


Fig. 1. Algorithm's overview

The overview of the proposed painting scene detection algorithm is presented in section 2. The database and the classification method are discussed in section 3, while experiments and results, as well as a comparison with state of the art are presented in section 4. Finally, the last section is dedicated to conclusions.

2 Algorithm Description

The proposed solution is based on the Gestalt [20] visual perception principles. The Gestalt (shape) theory states that the human eye perceives objects in their entirety before detecting their individual parts. In other words, when watching a scene, or a painting, as in this paper, the component objects are immediately identified and have a greater meaning than point-level details such as paint, canvas, brush stroke or subregion of the objects. In order to model this behavior, our solution starts with a segmentation process that will offer the basic parts of the images. These components will be further grouped according to Gestalt principles in more meaningful regions, called *objects*. Further on, the objects are described by a set of perceptual features that will be used in detecting the scene type. The algorithm's overview can be seen in Figure 1.

2.1 Image Segmentation

In order to obtain the basic components of a scene, a segmentation process is performed on the input image. Considering the fact that we aim to mimic the Gestalt perception principles, the segmentation algorithm used is Normalized Cuts [17] as it was developed based on the Gestalt theory.

The segmentation solutions in the graph cut category are seen as a graph partitioning problem, where each pixel is a node of the graph that is interconnected with spatially neighboring pixels and the weight of the link (and respectively the cut cost) is given by a similarity measure of pixels intensities and by the quantity of edges in between. The normalized cut criterion considers the total dissimilarity between different groups normalized by the similarity within the group in what one may perceive as an adaptation of the Fisher discriminant. In our solution, we have used a derivation of the original Normalized Cut. This segmentation, proposed by Cour et al [5], optimizes the initial method by considering a multi-scale spectral approach that exploits the isolation of the segmentation cues used

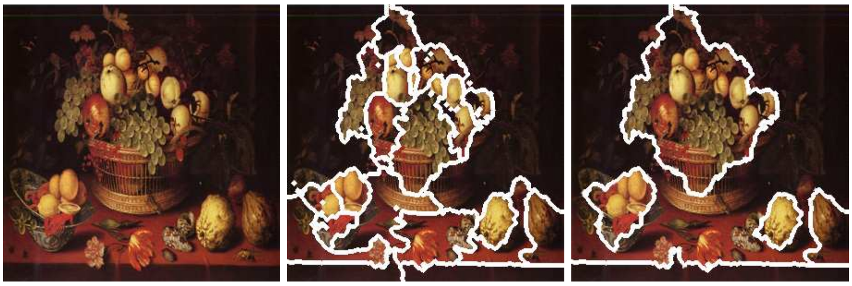


Fig. 2. Example of object detections (*a*) Original input image. (*b*) Segmented image (*c*) Image segmentation after regions merging

to detect coherent regions within faint boundaries. The algorithm segments the input image in a given number of clusters that will be discussed later.

2.2 Perceptual Regions Merging

Considering the fact that the Normalized Cuts segmentation is based on Gestalt perception principles, the segmentation output should be consistent with the way the human eye perceives the scene. However, our experiments showed that this is not always true, especially due to the over-segmentation phenomenon caused by the usage of a higher number of classes compared to the number of objects in the scene. In order to overpass this issue we have introduced an intermediary step that merge adjacent regions according to the Gestalt principles. The merging algorithm is:

- **Repeat:**
  1. Compute the Region Adjacency Graph. This graph will denote all pairs of adjacent regions;
  2. For each pair of neighboring regions compute the compatibility features, that will be described later;
  3. Compute the compatibility matrix by evaluating whether two regions should be merged. As discussed further, four scores are computed and the merging decision is taken by a Naive Bayes classifier;
  4. Merge all regions marked as belonging to the same object;
- **Until** no regions were merged in step 4 or the maximum iteration number was achieved

The Gestalt theory states several principles that are followed by the human brain when it perceives different regions as belonging to the same object [20]:

- *Similarity*: Regions that share visual characteristics like shape, size, color, texture will be seen as belonging to the same object;
- *Proximity*: Objects or shapes that are close to one another appear to form groups;

- *Closure*: The effect of suggesting a visual continuity between sets of elements which do not actually touch each other or the phenomenon of seeing complete figures even when part of the information is missing;
- *Continuity*: The effect of continuing shapes beyond their ending points, to meet up with other shapes or the edge of the picture plane.

In order to mimic the behavior of the human brain we have developed a set of compatibility features that models the Gestalt principles in order to assess whether two adjacent regions should be merged.

The *similarity* between two regions was modelled through the difference of the mean color of the two regions.

The *closure* aspect of the Gestalt theory was modelled through a measure derived from the spatial segmentation component of the JSEG segmentation algorithm [7]. Considering a region R_1, the sum of distances from each of the region's points to the weighting center of mass is computed and denoted by J_{R1}. Having two regions, R_1 and R_2, the JSEG measure is computed for both regions, as well as for the merged region $R_1 \cup R_2$. The final closure measure, μ_{close}, is computed as:

$$\mu_{close} = \frac{J_{R1} + J_{R2}}{J_{R1 \cup R2}} \tag{1}$$

The *continuity* property is far more complex and difficult to model, as also observed by Zlatoff et al. [22] as it involves more high level concepts such as direction or shape. Thus, at this stage we have chosen to model only one aspect of the continuity principle by computing the percentage of common border between two adjacent regions. This way, if for example a region is encapsulated by a larger one, the percentage will be maximum, indicating that most likely the human eye will tend to group those two regions.

With regards to *proximity*, this principle is automatically taken into consideration through the fact that two regions are considered for merging only if they share a border.

Having computed the compatibility features, the next step is *to decide* whether two adjacent regions should be merged into the same object. For this step we have used a Naive Bayes classifier, for the ease of bias prior classes probability. This bias was needed because the usage of equiprobable classes lead to an over-merging phenomenon. For example, in the simple case of three regions, R_1, R_2 and R_3, let's say that R_1 and R_2, as well as R_2 and R_3 were marked as belonging to the same objects, but R_1 and R_3 were marked as belonging to different objects. Due to the transitive property of the merging operation, the R_1 and R_3 regions are falsely merged into the same object. In order to avoid this, it was preferred to decide that two regions should be merged only if the decision was very clear. This was achieved by giving a much higher prior probability to the decision of not merging. The classifier was trained using more than 2000 positive and negative examples of merging and offered a detection rate of 72% on the training set.

An example of segmented image after the regions merging process can be see in Figure 2 c).

Table 1. Features that describe an object description.

Feature	Shape	Area	Perimeter	Color	Location
Size	8	1	1	3	9

2.3 Objects Description

The next step of the proposed solution is the description of the identified objects. Usually the segmented image, obtained according to the merging process, consists in an object corresponding to the background of the scene and one or more foreground objects. Due to the composition rules that are in general followed in paintings, the position and the general shape of the objects are consistent for a certain scene type. For example, in case of a portrait, usually there are maximal two foreground objects corresponding to the subject or to the subject's face and body, and a darker background object around them.

Considering this, a set of features is proposed to describe the objects composing the scene: object area, object perimeter, object shape, object mean color and object location. The area, perimeter and the mean color are self explanatory. The shape of the object is described using a 8 bins histogram of orientations computed on the object edge (the classical HoG [6], but restricted only to the outer object edges).

With regards to the *location* of the object, the most obvious choice would be the usage of the coordinates of the weighting center of mass. However, given that one of the most basic rules of composition is the "rule of thirds", we set a location descriptor based on this rule. The painting was divided in a 3×3 grid and for each of the resulting 9 rectangles the percentage of space occupied by the current object was taken as feature. The vector of 9 sub-unitary values is taken as the final feature indicating the object position in the image.

One of the challenges encountered was the fact that the *number of final objects* presented in the final image could vary in a quite large interval. We preferred to overcome this problem by considering a fixed number (N) of objects for all scenes. In this case, if the scene contains less objects than the chosen number, the features vector will be filled with zeros. Otherwise, only the first N objects will be considered.

This leads to another issue regarding the way the objects are *ranked.* This aspect is very important not only when it comes to choosing the first N objects to be described, but also because the same type of object from multiple paintings should have the same rank in the classification process. For example, if in a landscape painting the first object is the sky and the second one corresponds to the ground, the same order should be applied for all the other landscape paintings containing two objects. As will be shown in section 4, the best results were obtained when the objects were ranked by their *size.*

The complete set of variables in a feature vector describing an image is presented in table 1.

Table 2. Average Detection Rate (ADR) for various segmentation algorithms: Normalize Cuts – NormCut, K-Means , Mean Shift, Graph Cuts

Method	NormCut	K-Means	MeanShift	GraphCut
ADR[%]	67.4	63.0	55.9	62.3

3 Database and Recognition Scheme

The vector of features describing the resulted N objects within a painting is presented to a classifier in order to determine the scene type. For the classification process we have used a 10-fold validation scheme over a database containing 500 paintings out of 5 scene types: portrait, nude, landscape, cityscape, still life. When choosing the scene types we wanted to test the most encountered scene types in the art history and to ensure as well that both very different classes (like portrait and landscape) and very similar classes (like landscape and cityscape) are present. The images were taken from the Yorck Project Database [21] and the scene type was manually marked for establishing the ground truth.

The classifier was chosen by experiment and it turned out that a bagged ensemble of 25 decision trees proved to offer the best performance.

The criterion used for measuring performance is the average correct detection rate (ADR).

4 Results

4.1 Algorithm's Parameters

In order to study the performance of our solution, we have investigated the influence of various parameters or algorithmic blocks onto the overall performance.

Segmentation Method For the first stage of the solution we have also considered the usage of other widely used segmentation algorithms, but the non-perceptual solutions offered lower detection rates. The tested alternative segmentation solutions were K-Means, Mean Shift [3] and standard Graph Cuts [1], [12], leading to the detection rates presented in table 2.

Although the performance of the others methods were not very far behind, the chosen segmentation method (Normalized Cuts) offered the best results.

Number of clusters For the Normalized Cuts segmentation method we tested the usage of various initial numbers of clusters; the resulting average detection rates are showed in table 3. While a very small number of classes do produce enough separation and too many classes lead to over-segmentation, the best results are achieved with a medium number of classes. Because of the similarity in the average detection rates obtained for $N = 20$ and $N = 30$ segmentation clusters and taking into account the higher computation time necessary for $N = 30$ classes, we decided to continue the experiments using a $N = 20$ class segmentation.

Table 3. Variation of overall Average Detection Rate (ADR) with respect to the initial number of clusters considered in the Normalized Cut algorithm.

No of clusters	10	20	30	40
ADR[%]	64.6	67.4	67.4	64.7

Table 4. Average Detection Rate (ADR) achieved when considered various colorspaces

Color space	RGB	HSV	HMMD	Lab
ADR[%]	61.9	58.0	64.7	67.4

Color space We have also considered several color spaces (RGB, HSV, HMMD, Lab) for the segmentation process and for objects description. The achieved results are presented in table 4. We note that although the results are similar, perceptual color spaces provide more accurate results, which confirms Ramachandran [15] hypothesis.

Region merging As stated in chapter 2, for the merging stage of the algorithm, when deciding whether two regions should be merged or not, it is very important to control the prior probability of the two decisions. Therefore, a series of tests were performed in order to determine the best prior probabilities for the Bayes Classifier that establish whether two regions are merged. The tests showed that the detection rate varied from a minimum of 54.8% for equal prior probabilities to a maximum of 67.4% when the decision to keep the two regions apart had a probability of 0.8.

Features In order to assess the performance of our set of features models the Gestalt principles, we have implemented the set of features proposed by Zlatoff et al. [22], obtaining an average detection rate of 59.4%.

Furthermore, in order to assess the contribution of each group of features to the overall score, we independently removed each group from the object description and re-performed the classification. The results are presented in table 5. As the individual contribution of each set is small, we may conclude that features complement each other.

Object selection As previously described, once the main objects of the scene are detected, it is of great importance to find a consistent way of sorting them so that during the classification process corresponding objects from different scenes are situated at the same positions in features vector. In other words, for example in the case of a portrait, it is very important for the subject to be, the first object in the scene and the background the second. In order to achieve this order, we have tested several ordering criteria based on the features describing the objects; the results are presented in table 6. Although the sorting by size offered good results, we consider that this aspect might be a bottleneck in our solution, because often objects with similar sizes are detected in the scene and

Table 5. Contribution of each feature to the overall detection rate.

Feature group	Shape	Area	Perimeter	Color	Location
Contribution[%]	1.2	1.2	0.4	10.0	7.0

Table 6. Average Detection Rate (ADR) achieved when different criteria was chosen for selecting the representative object.

Sorting Criteria	Size	Intensity	Hue	Location
ADR[%]	67.4	55.4	46.9	55.4

miss-sorting them might introduce classification errors. We suggest here that some graph matching algorithms might offer better results.

Number of objects Another aspect related to the objects description that had to be assessed was the optimal number of objects to be kept in the final feature vector. Our tests showed that keeping the largest two objects from the scene offer good results, as can be seen in Table 7. These findings are consistent with the human perception mechanism because usually the two biggest objects from a scene are a clear cue for the scene type.

Classifier choice For the last stage of the proposed solution, several classifiers implemented in the open-source machine learning library Weka [11] and presented in Table 8 were considered. The fact that all tested classifiers had very similar performance leads us to believe that the features space is consistent and that the results would remain in the same range no matter what classifier will be used.

4.2 Scene Detection Accuracy and Comparison with Prior Art

Having set the parameters values, we have performed another set of tests in order to assess the overall performance of the proposed solution.

In order to compare our work with state of the art, we have described the same database using the GIST image descriptor [14] alone. Basically, GIST provides a global description of the scene, while our method follows Gestalt principle that states objects are more important for scene understanding. As can be shown in table 9 our algorithm outperforms the GIST solution for paintings scene classification. However, we consider that such a result is reasonable to expect only for painting analysis where human perception prevails to the nature randomness.

In order to assess how our solution performs for various number of possible scene types all possible combinations of the five scene types were tested. In Figure 3 a) the average, the lowest and the highest detection rates are presented for each possible number of classes. As expected, the overall detection rate decreases with the increase of possible scene types, but the results remain acceptable for all scene types.

Table 7. Average Detection Rate (ADR) when different number of objects were considered.

No. Objects	1	2	3	4	5
ADR[%]	66.8	67.4	66.4	65.6	62.6

Table 8. Average Detection Rate (ADR) for tested classifiers (Logistic Regression – LR, Multilayer Perceptron – MLP, Sequential minimal optimization – SMO, Bagged ensemble of tress – Ba, LogitBoost – LB, Naive Bayes – NB, Random Forrest – RF). Details regarding the implementation of the classifiers are to be found in [11] and references therein.

Classifiers	LR	MLP	SMO	**Ba**	LB	DT	RF
ADR [%]	60	61.8	62.2	67.4	60.4	54	59.8

Figure 3(b) shows the confusion matrix for the five scene types, where it can be seen that the lowest detection rate is obtained for the nude scenes that are sometimes confused with still life or portraits. Also, as it was expected, a slightly higher confusion occurs between cityscape and landscapes.

Figure 4 shows some examples of misclassified scenes. While for the cityscape and landscape examples a wrong classification might be expectable, for the nude example the justification might be not very visible at a first glance. However, this example of nude painting is not exactly a classical nude painting, compositionally speaking. Usually the nude paintings contained a skin-colored object in the middle of the paint, surrounded by a background. For the portrait painting the error's cause is much clear, the man's clothes having a color very similar to the skin's color.

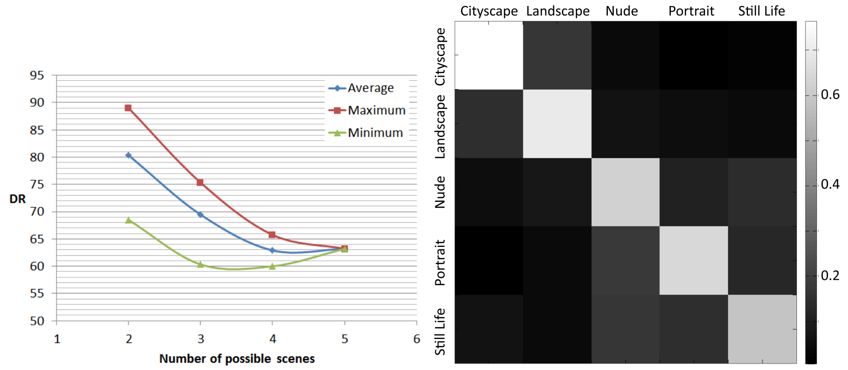
5 Conclusions and continuation

In this paper we have proposed a method for the automatic recognition of the scene type in digitized paintings. As art is dedicated to humans, we inspired our method from the Gestalt perception theory, stressing that objects are more important in scene identification than the overall description. We successfully validated the robustness of our method on a 5-scene, 500 images database.

As continuation paths, we need to investigate more thoroughly criteria for selecting the most representative object from a scene and to extend the testing, by increasing both the database and the number of considered scenes.

Table 9. Comparison with state of the art: average detection rate for the proposed algorithm and GIST solution.

Proposed Solution	67.4 %
GIST [14]	61.4 %

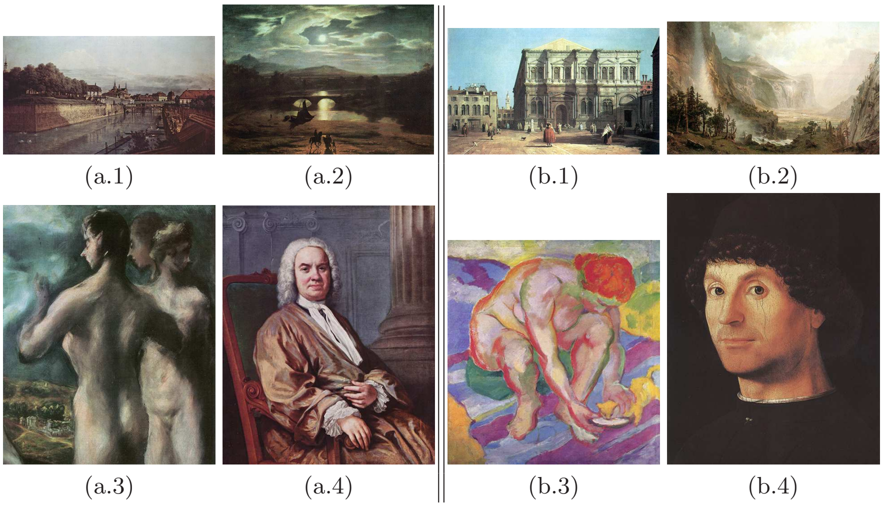


(a) (b)

Fig. 3. (a) Detection rates for various number of classes; (b) Confusion Matrix

References

1. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE transactions on PAMI 26(9), 1124–1137 (2004)
2. Carneiro, G., da Silva, N.P., Del Bue, A., Costeira, J.P.: Artistic image classification: an analysis on the PRINTART database. In: Proc. of ECCV, pp. 143–157. Springer (2012)
3. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. on PAMI 24(5), 603–619 (2002)
4. Cornelis, B., Dooms, A., Cornelis, J., Leen, F., Schelkens, P.: Digital painting analysis, at the cross section of engineering, mathematics and culture. In: Proc. of EUSIPCO. pp. 1254–1259 (2011)
5. Cour, T., Benezit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. In: Proc. of CVPR. vol. 2, pp. 1124–1131 (2005)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR. pp. 886–893 (2005)
7. Deng, Y., Manjunath, B.: Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. on PAMI 23, 800–810 (2001)
8. Durand, F.: An invitation to discuss computer depiction. In: Proc. International symposium on Non-photorealistic animation and rendering, NPAR. pp. 111–124. ACM (2002)
9. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. of CVPR. vol. 2, pp. 524–531 (2005)
10. Graham, D., Redies, C.: Statistical regularities in art: Relations with visual coding and perception. Vision Research 50(16), 1503 – 1509 (2010)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
12. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Trans. on PAMI 26(2), 147–159 (2004)



(a.1) (a.2) (b.1) (b.2)

(a.3) (a.4) (b.3) (b.4)

Fig. 4. (a) Examples of misclassified scenes. (a.1) cityscape painting classified as landscape (a.2) landscape classified as cityscape (a.3) nude classified as cityscape (a.4) portrait classified as nude. (b) Examples of correctly classified paintings: (b.1) cityscape (b.2) landscape (b.3) nude (b.4) portrait

13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of CVPR. vol. 2, pp. 2169–2178 (2006)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42(3), 145–175 (2001)
15. Ramachandran, V., Herstein, W.: The science of art: A neurological theory of aesthetic experience. Journal of Consciousness Studies 6, 15–51 (1999)
16. Shamir, L., Macura, T., Orlov, N., Eckley, D.M., Goldberg, I.G.: Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. ACM Transactions on Applied Perception 7(2), 1–17 (2010)
17. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. on PAMI 22(8), 888–905 (2000)
18. Stork, D.: Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature. In: Proc. of CAIP. pp. 9–24 (2009)
19. Wallraven, C., Fleming, R.W., Cunningham, D.W., Rigau, J., Feixas, M., Sbert, M.: Categorizing art: Comparing humans and computers. Computers & Graphics 33(4), 484–495 (2009)
20. Wertheimer, M.: Principles of perceptual organization. Readings in perception pp. 115–135 (1958)
21. Yorck, P.: The yorck project. `http://commons.wikimedia.org/wiki/Category:PD-Art_(Yorck_Project)` (dec 2012)
22. Zlatoff, N., Tellez, B., Baskurt, A.: Image understanding and scene models: a generic framework integrating domain knowledge and gestalt theory. In: Proc of ICIP. vol. 4, pp. 2355–2358 (2004)