Semantic-based image analysis with the goal of assisting artistic creation

Pilar Rosado[1], Ferran Reverter[2], Eva Figueras[1], and Miquel Planas[1]

[1] Fine Arts Faculty, University of Barcelona, Spain,
pilarrosado@ub.edu,
[2] Statistics Department, University of Barcelona, Spain.

Abstract. We have approached the difficulties of automatic cataloguing of images on which the conception and design of sculptor M. Planas artistic production are based. In order to build up a visual vocabulary for basing image description on, we followed a procedure similar to the method Bag-of-Words (BOW). We have implemented a probabilistic latent semantic analysis (PLSA) that detects underlying topics in images. Whole image collection was clustered into different types that describe aesthetic preferences of the artist. The outcomes are promising, the described cataloguing method may provide new viewpoints for the artist in future works.

Keywords: Artificial vision, automated image cataloguing, Bag-of-Visual terms, probabilistic latent semantic analysis.

1 Introduction

Artists are image generators, since they constantly produce them in their creative process. Recently, an increasing community of researchers in computer vision, pattern recognition, image processing and art history have developed rigorous computer methods for addressing an increasing number of problems in the history of art [1]. In our case, the images used by the sculptor Miquel Planas [2] constitute an essential part of his creative process and he presents them as a large document collection on which subsequent work will be based, especially in the field of sculpture. The possibility was outlined that a study could be started to enable the creation of a system of image grouping and classification, not only with the aim of cataloguing them, but also in order to obtain new values, qualities and common characteristics of the compared images. In view of the fact that the piece of research of such characteristics was proposed within the field of fine arts, it was suggested that its results could be extrapolated to any kind of actions centred on creation, in which the comparison among images would be the main characteristic, with a view to create applications aimed at learning, knowledge acquisition and image research. This study extends previous work in assessing the performance of SIFT descriptors, BOW representation and spatial pyramid matching for automatic analysis of images that are the basis of the ideation and designing of art work [3].

2 Methodology

Image Representation In order to build up a visual vocabulary for basing image description on, we followed a procedure similar to the one used in automatic text analysis. The method is known as the "Bag-of-Words" (BOW) model because every document is represented as a distribution of frequencies of the words in the text, without considering the syntactic relationships among them. In the sphere of images we will refer to "Bag-of-Visual Terms" (BOV) representations. This approach consists in analysing images as a group of regions, describing only their appearance without taking into account their spatial structure. The BOV representation is built up based on the automatic extraction and quantization of local descriptors and it has proven to be one of the best techniques for solving different tasks in the field of computer vision. The BOV representation was first implemented [4] to develop an expert system specializing in image recognition. Figure 1 summarizes the process to be followed in order to obtain the BOV representation of the images of a collection [5, 6]. This representation of an image does not contain information concerning the spatial relationships among visual words, in the same way that the BOW representation removes the information relating to word order in documents. To overcome the limitations of the BOV approach we have implemented a method using pyramid histograms that configure an increasingly fine grid sequence over the image and conducts a BOV type analysis in each of the grids, finally obtaining a weighted sum of the number of matches that occur in each resolution level of the pyramid [7].

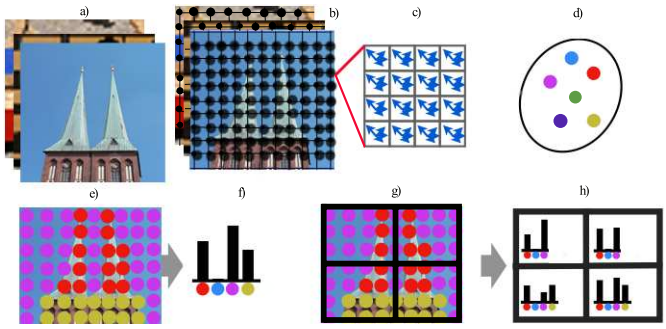

Fig. 1. BOW representation scheme. a) Collection of images b) A grid is defined over the images. c) Image descriptors are calculated. d) Descriptors are quantized in M clusters, which will define a visual vocabulary of M visual words. e) Once the vocabulary is available, the descriptors of each image are assigned to the nearest visual word. f) In order to obtain the BOV representation of a given image, the frequency of each visual word in the image is calculated. g),h) Sequence of grids on the image to draw histograms pyramid in order to take into account the spatial relationship between visual words.

Representation of latent aspects The BOV representation is easy to construct. However, it has two drawbacks: polysemy - a single visual word may represent different scene contents - and synonymy - various visual words may characterize the same image content. As a partial solution to the previous drawbacks, we have found the probabilistic latent semantic analysis (PLSA), a methodology originated from text mining [8]. Extending PLSA to image analysis involves considering images as documents in a visual vocabulary established by means of a quantization process, as previously mentioned. The PLSA will detect categories of objects, formal patterns, within the images in such way that an image which contains different types of objects is modeled as a mixture of subjects. We have at our disposal an image collection $D = \{d_1, \ldots, d_N\}$ and a vocabulary of visual words $V = \{v_1, \ldots, v_M\}$. We can summarize the observations in a $N \times M$ table of frequencies $n(d_i, v_j)$, where $n(d_i, v_j)$ indicates how frequently the visual word v_j occurs in image d_i. The PLSA is a generative statistical model which associates a latent variable $z_l \in \{z_1, \ldots, z_K\}$ with each observation, an observation being understood as the occurrence of a visual word in a given image. These variables, normally known as aspects, are used to obtain a model of joint probability based on the images and visual words, defined as:

$$P(d_i, v_j) = P(d_i) \sum_{k=1}^{K} P\big(v_j|z_k\big) P\big(z_k|d_i\big)$$

where $P(d_i)$ represents the probability of d_i , $P\big(v_j|z_k\big)$ represents the conditional probability of a specific visual word conditioned on the latent aspect z_k , and $P\big(z_k|d_i\big)$ represents the image-specific conditional probability. PLSA introduces a concept of conditional independence, according to which it is assumed that the occurrence of a visual word v_j is independent of the image d_i in which it appears, given an aspect z_k. The estimation of probabilities of the PLSA model is performed by means of the maximum likelihood principle, using the image collection $D = \{d_1, \ldots, d_N\}$. The optimization is carried out using the EM algorithm [9]. The EM algorithm alternates two steps. In step E, the a posteriori probabilities are calculated for the latent aspects based on current estimations of the probabilities of the model; in step M, the probabilities of the model are updated by maximizing the so-called "expected complete data log-likelihood":

– **Step E**

$$P\big(z_k|d_i, v_j\big) = \frac{P\big(v_j|z_k\big) P\big(z_k|d_i\big)}{\sum_{l=1}^{K} P\big(v_j|z_l\big) P\big(z_l|d_i\big)}$$

– **Step M**

$$P\big(v_j|z_k\big) = \frac{\sum_{i=1}^{N} n(d_i, v_j) P\big(z_k|d_i, v_j\big)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, v_m) P\big(z_k|d_i, v_m\big)}$$

$$P\big(z_k|d_i\big) = \frac{\sum_{j=1}^{M} n(d_i, v_j) P\big(z_k|d_i, v_j\big)}{n(d_i)}, \quad n(d_i) = \sum_{j=1}^{M} n(d_i, v_j)$$

Steps E and M alternate repeatedly until a certain condition of termination is achieved. The iterative process is started by assigning random values to the set of probabilities $P(z_k|d_i)$ and $P(v_j|z_k)$. As a result of the previous process, a new representation is obtained for the images of the collection based on the aspect distribution,

$$(P(z_1|d_i), ..., P(z_K|d_i))$$

Actually, it is also possible to determine the aspect distribution for any image d that does not belong to the initial collection [10, 11]. Using the previously described EM algorithm again will suffice, although in this case, in step M, only probabilities $P(z_k|d)$ will be updated, while probabilities $P(v_j|z_k)$, independent of the image and estimated from the collection in the learning stage, will remain fixed. Although the image representation based on aspects can be used as input for a scene classifier, we will focus on the use of said representation for image organization or ranking based on the distribution of underlying aspects. Given an aspect z, the images can be arranged according to the values of $P(z|d)$ thus, once the values of $P(z_k|d)$, $k = 1, ..., K$, for a given image d are estimated, we can arrange them in order and obtain an objective measure of the association between the image and every single aspect. As a result, we will associate the image with the aspect of higher probability. Based on this methodology, we have been able to analyse the image collection and determine underlying aspects by means of which the whole collection can be catalogued. The process has been carried out by means of scripts written in MATLAB, 2013a version (The MathWorks) (8.1.0.604). SIFT local descriptors [12] and the vocabulary of visual words have been implemented by means of functions available from the open code library VLFeat, 0.9.16 version [13]. The PLSA has been implemented by using functions developed by the authors themselves.

3 Results

The initial sample of our study is made up of 2,846 photographic images taken by the artist himself. It is a set of images, most of which were taken outside and from different angles and details (including fragments and special features which can be photographed as abstract and/or textured elements). The size of the images taken by the artist is between 480 x 480 pixels and 1400 x 1400 pixels, but the process rescales the images that exceed this size to 480 pixels. Taking into account the type of sample and the number of images, we have tested the system by setting the number of aspects to 10. For the outcome assessment of this set we will mainly take into consideration images categorized into a certain aspect with a probability equal to or exceeding 0.6. Computational evaluation is performed in greyscale. In light of these results, we perceive that the sample consists of two very distinct image types; a type of photographs which shows a single highly prominent aspect (we shall call them low entropy images), and another one which shows several simultaneously associated aspects (we shall call them high entropy images). To distinguish these two typologies, we have used the Shannon entropy index [14]. The PLSA methodology provides a distribution

of aspect probability in the images, that is, for a given image d, there is a vector of probabilities:

$$\big(P\big(z_1|d\big), P\big(z_2|d\big), ..., P\big(z_K|d\big)\big)$$

hence, we can calculate the Shannon Entropy index of image d by means of

$$H(d) = -\sum_{i=1}^{K} P\big(z_i|d\big) \log P\big(z_i|d\big)$$

Thus, an image which is associated with a single aspect, that is, an image with a probability vector with all values of zero except for a one, will have a minimum entropy value equal to $H(d) = 0$; on the other hand, an image which is equally associated with all the aspects, that is, with a probability vector of $1/10$ in each component, will have a maximum entropy value equal to $H(d) = 2.3026$. The theoretical entropy ranges in respect of 10 aspects would be from 0 to 2.3026. Those noticed in our sample practically range from 0 to 2.17. The images which show high entropy are those which have been associated in an equiprobable manner with each and every aspect. It is decided to select the images with an entropic value of above 1.4 and restart the search for aspects in this new sample made up of $1,482$ images. Thus, an attempt is being made to have the system establish new relationships between more visually complex images. Again the whole process of generating local descriptors and visual vocabulary is repeated, and thus an attempt is made to make the system be able to establish new relationships between visually complex images, resulting in new latent aspects different from the 10 first ones. The test is successful and different set of 10 aspects are generated on the new sample. In total, the system is able to categorize the total images analyzed in 20 groups, see Figures 2 and 3.

4 Conclusions

In view of the results obtained, we can conclude that, given a set of a large number of images, the system would enable a formal pre-selection by grouping them in a more objective manner. Artificial vision is not totally subject to or conditioned by human conceptual perception. The described cataloguing method may introduce new relationships, new sets which could provide the artist with new indications or viewpoints for future works. Thus, the results obtained in the different experiments previously described have enabled Miquel Planas to reapproach the photographic work carried out so far. In this respect, the artist's personal vocabulary will also be enriched by the new relationships obtained from the programmed classifications, as shown in the examples detailed in this study. The outcomes obtained will provide new ideas and nuances, which will directly benefit the final creative work. Thus, a creator's vocabulary can also be established, growing as the system is fed with new images. This will be highly useful in the creative, analytical, taxonomic and pedagogic process of the artwork.

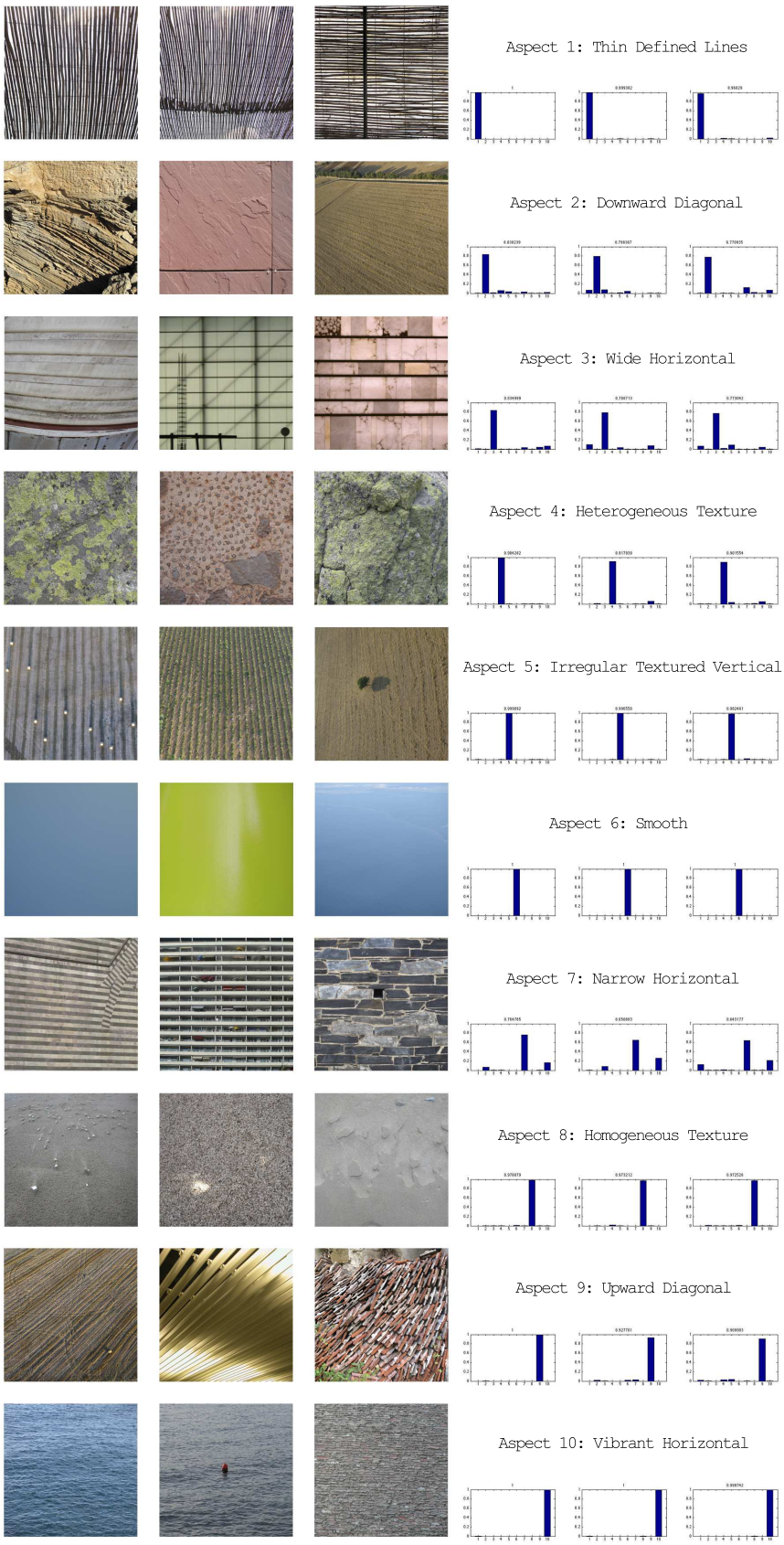

Fig. 2. Images and histogram of aspects obtained on the set of low entropy images.

Fig. 3. Images and histogram of aspects obtained on the set of high entropy images.

References

1. Stork, D.G. Computer image analysis of paintings and drawings: An introduction to the literature. Proc. of the Image processing for Artist Identification Workshop, van Gogh Museum (2008)
2. Planas, M. A.: Miquel Planas. http://www.miquelplanas.eu (2014)
3. Reverter, F., Rosado, P., Figueras, E., Planas, M.A. Artistic ideation based on computer vision methods", Journal of Theoretical and Applied Computer Science, Vol. 6, No. 2, 72-78 (2012)
4. Willamowski, J., Arregui, D., Csurka, G., Dance, C., Fan, L. Categorizing nine visual classes using local appearance descriptors. In Proceedings of LAVS Workshop, in ICPR'04, Cambridge (2004)
5. Lazebnik, S., Schmid, C. , Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 2169-2178. doi.ieeecomputersociety.org/10.1109/CVPR.2006.68 (2006)
6. Fei-Fei, L. , Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In Proc. CVPR (2005)
7. Grauman, K. , Darrel, T.: The pyramid match kernel: Discriminative classification with sets of image features. In Proceedings of IEEE International Conference on Computer Vision (ICCV) Beijing (2005)
8. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42,177-196 (2001)
9. Dempster, A. P., Laird, N. M., & Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. B, 39, 1-38 (1977)
10. Bosch, A., Zisserman, A., & Munoz, X.: Scene classification via PLSA. In Proceedings of the European Conference on Computer Vision, Graz, Austria (2006)
11. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars & Van Gool, L.: Modeling scenes with local descriptors and latent aspects. Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), 1, 883-890. doi: 10.1109/ICCV. 2005. 152 (2005)
12. Lowe, D. G.: Distinctive Image Features from Scale Invariant Keypoints. Int. Journal of Computer Vision, 60, 2, 91-110 (2004)
13. Vedaldi, A & Fulkerson, B.: VLFeat - An open and portable library of computer vision algorithms. Retrieved from http://www.vlfeat.org (2008)
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory (2on ed.). New Jersey: John Wiley & Sons (2006)