Journal of Theoretical and Applied Computer Science
ISSN 2299-2634

Vol. 6, No. 2, 2012, pp. 72–78
http://www.jtacs.org

Artistic ideation based on computer vision methods

Ferran Reverter, Pilar Rosado, Eva Figueras, Miquel Àngel Planas
University of Barcelona, Spain
{freverter,efigueras}@ub.edu, {prrforma,miquelplanas}@gmail.com

Abstract: *This paper analyzes the automatic classification of scenes that are the basis of the ideation and the designing of the sculptural production of an artist. The main purpose is to evaluate the performance of the Bag-of-Features methods, in the challenging task of categorizing scenes when scenes differ in semantics rather than the objects they contain. We have employed a kernel-based recognition method that works by computing rough geometric correspondence on a global scale using the pyramid matching scheme introduced by Lazebnik [7]. Results are promising, on average the score is about 70%. Experiments suggest that the automatic categorization of images based on computer vision methods can provide objective principles in cataloging images.*

Keywords: *bag-of-features, SIFT descriptors, pyramid match kernel, artistic ideation*

1. Introduction

Image representation is a very important element for image classification, annotation, segmentation or retrieval. Nearly all the methods in computer vision which deals with image content representation resort to features capable of representing image content in a compact way. Local features based representation can produce a versatile and robust image representation capable of representing global and local content at the same time. Describing an object or scene using local features computed at interest locations makes the description robust to partial occlusion and image transformation. This results from the local character of the features and their invariance to image transformations.

The bag-of-visterms (BOV) is an image representation built from automatically extracted and quantized local descriptors referred to as visterms in the remainder of this paper. The BOV representation, which is derived from these local features, has been shown to be one of the best image representations in several tasks.

The main objective of this study is assessing the performance of SIFT descriptors, BOV representation and spatial pyramid matching for automatic analysis of images that are the basis of the ideation and designing of art work. Additionally, we explore the capability of this kind of modelization to become useful for the production of software based art.

2. Image Representation and Matching

The BOV representation was first used [1] as an image representation for an object recognition system. In the BOV representation, local descriptors f_j are quantized into their respective visterms $v_i = Q(f_j)$ and used to represent the images from which they were extracted. The quantization process groups similar descriptors together, with the aim that the descriptors in each resulting group arises from local patterns with similar visual appearance. The number

of occurrences of each visterm in a given image is the elementary feature of the BOV representation. More precisely, the BOV representation is the histogram of the various visterms' occurrences.

To construct the BOV feature vector h from an image I four steps are required. In brief, local interest points are automatically detected in the image, then local descriptors are computed over the regions defined around those local interest points (certain applications may require that local descriptors may be computed on a dense grid over the image instead over local interest points). After this extraction step, the descriptors are quantized into visterms, and all occurrences of each visterm of the vocabulary are counted to build the BOV representation of the image.

2.1. Feature extraction

The BOV construction requires two main design decisions: the choice of local descriptors that we apply on our images to extract local features, and the choice of which method we use to obtain the visterms' vocabulary. Both of these choices can influence the resulting system's performance. Nevertheless BOV is a robust image representation, which retains its good performance over a large range of parameter choices.

For better discriminative power, we utilize higher dimensional features which are SIFT (Scale Invariant Feature Transform) descriptors introduced by [2]. The SIFT descriptor is a histogram based representation of the gradient orientations of the gray-scale image patch. In our study, SIFT descriptors are computed at points on a regular grid with spacing 8 pixels. At each grid point the descriptors are computed over circular support patches. Our decision to use a dense regular grid instead of interest points was based on the comparative evaluation of [3], who have shown that dense features work better for scene classification. Intuitively, a dense image description is necessary to capture uniform regions such as sky, calm water, or road surface. SIFT was also found to work best for the task of object classification [4] and [5].

2.2. Visual Vocabulary

In order to obtain a text-like representation, we quantize each local descriptor s into one of a discrete set $\mathcal{V}$ of visterms v according to a nearest neighbor rule:

$$s \mapsto Q(s) = v_i \leftrightarrow \text{dist}(s, v_i) \leq \text{dist}(s, v_j),$$

for all $j = 1, ..., M$, where M denotes the size of the visterm set.

We will call vocabulary the set $\mathcal{V}$ of all the visterms. The vocabulary construction is performed through clustering. More specifically, we apply the k-means algorithm to a set of local descriptors extracted from training images, and keep the means as visterms. We used the Euclidean distance in the clustering and quantization processes, and choose the number of clusters depending on the desired vocabulary size.

Finally, the BOV representation is constructed from local descriptors according to:

$$h(d) = \big(n(d, v_1), n(d, v_2), ..., n(d, v_M)\big)$$

with $n(d, v_i)$, $i = 1, ..., M$, denotes the number of occurrences of visterm v_i in image d. To classify an input image d represented either by the bag-of-visterms vector $h(d)$ we employed Support Vector Machines (SVMs).

This vector-space representation of an image contains no information about spatial relationships between visterms, in the same way the standard bag-of-words text representation removes the word ordering information.

For such whole-image categorization tasks, bag-of-features methods, which represents an image as an orderless collection of local features, have recently demonstrated impressive levels of performance. However, because these methods disregard all information about the spatial layout of the features, they have severely limited descriptive ability. In particular, they are incapable of capturing shape or of segmenting an object from its background.

2.3. Spatial matching scheme

To overcome the limitations of the bag-of-visterms approach, a spatial pyramid matching scheme was introduced in [8] and [7]. Informally, pyramid matching works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two points are said to match if they fall into the same cell of the grid; matches found at finer resolutions are weighted more heavily than matches found at coarser resolutions.

More specifically, let X and Y be two sets of vectors in a p-dimensional feature space. Let us construct a sequence of grids at resolutions $0, ..., L$ such that the grid at level ℓ has 2^ℓ cells along each dimension, for a total of $D = 2^{p\ell}$ cells. Let H_X^ℓ and H_Y^ℓ denote the histograms of X and Y at this resolution, so that $H_X^\ell(i)$ and $H_Y^\ell(i)$ are the numbers of points from X and Y that fall into the ith cell of the grid. Then the number of matches at level ℓ is given by the histogram intersection function:

$$\mathcal{I}(H_X^\ell, H_Y^\ell) = \sum_{i=1}^{D} \min\left(H_X^\ell(i), H_Y^\ell(i)\right).$$

With the aim of brevity, we will abbreviate $\mathcal{I}(H_X^\ell, H_Y^\ell) = \mathcal{I}^\ell$. Note that the number of matches found at level ℓ also includes all the matches found at the finer level $\ell + 1$. Therefore, the number of new matches found at level ℓ is given by $\mathcal{I}^\ell - \mathcal{I}^{\ell+1}$ for $\ell = 0, ..., L-1$. The weight associated with level ℓ is set to $\frac{1}{2^{L-\ell}}$, which is inversely proportional to cell width at that level. Intuitively, we want to penalize matches found in larger cells because they involve increasingly dissimilar features.

Putting all the pieces together, the pyramid match kernel [8] is defined by

$$\kappa^L(X, Y) = \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} L\frac{1}{2^{L-\ell+1}}\mathcal{I}^\ell.$$

As introduced in [8], a pyramid match kernel works with an orderless image representation. It allows for precise matching of two collections of features in a high dimensional appearance space, but discards all spatial information.

Lazebnik et all. [7] advocates an approach that has the advantage of maintaining continuity with the popular "visual vocabulary" paradigm. It performs pyramid matching in the two-dimensional image space, and uses traditional clustering techniques in feature space.

Specifically, we quantize all feature vectors into a set of M discrete types, visual terms, and make the simplifying assumption that only features of the same type can be matched to one another.

Each channel m gives us two sets of two-dimensional vectors, X_m and Y_m, representing the coordinates of features of type m found in the respective images. The final kernel is then the sum of the separate channel kernels:

$$K^L(X, Y) = \sum_{m=1}^{M} \kappa^L(X_m, Y_m). \tag{1}$$

This approach agrees with Bag-of-visterms, in fact, it reduces to a standard bag of features when L = 0.

Because the pyramid match kernel is simply a weighted sum of histogram intersections, and because $c\min(a, b) = \min(ca, cb)$ for positive numbers, we can implement (1) as a single histogram intersection of "long" vectors formed by concatenating the appropriately weighted histograms of all channels at all resolutions. For L levels and M channels, the resulting vector has dimensionality $M\sum_{\ell}^{L} 4^l = M\frac{1}{3}(4^{L+1} - 1)$.

In summary, both the histogram intersection and the pyramid match kernel are Mercer kernels [8]. Lazebnik et all. [7] extend the pyramid match kernel to the pyramid of histogram of visual terms. Bosch et al. [6] implement a pyramid of histograms of visual terms, inspired in the above spatial matching scheme, but using a gaussian like kernel. In this implementation the similarity between a pair of images I and J is computed using a kernel function between their pyramid of histograms of visual terms D_I and D_J, with appropriate weighting for each level of the pyramid:

$$K(D_I, D_J) = \exp\big\{\frac{1}{\beta}\sum_{l\in L} \alpha_l d_l(D_I, D_J)\big\}$$

where β is the average of $\sum_{l\in L} \alpha_l d_l(D_I, D_J)$ over the training data, α_l is the weight at which level l and d_l is the χ^2 distance [9] between D_I and D_J at pyramid level l computed using the normalized histograms at that level.

Spatial histograms could be used as image descriptors and fed to a linear SVM classifier. Linear SVMs are very fast to train, but also limited to use an inner product to compare descriptors. Vedaldi and Fulkerson [10] have shown that much better results can be obtained by computing an explicit feature map that emulates a non linear χ^2-kernel as a linear one.

3. Results

In this paper we propose to automatically analyze images from a database of photographs by Dr. M.A. Planas Rosselló (Professor of sculpture, University of Barcelona). The image resolution is 480×480. The database consists of 150 images previously classified in 5 categories: Central architecture (CA), Geometric stone (GS), Irregular stone (IS), Textured stone (TS) and Silhouettes (SI). These categories correspond to 5 different typologies identified in the photographic images from the database. Images are the basis for the ideation and design of an artist's sculptural work.

Figure 1 schematizes the steps in image analysis using a pyramid of histograms of visual terms. A dense grid of points is defined on the image, then local descriptors are computed over the regions defined around those points in the grid. After this extraction step, the descriptors are quantized into visual terms (visterms). Then, the image is represented by visterms, each

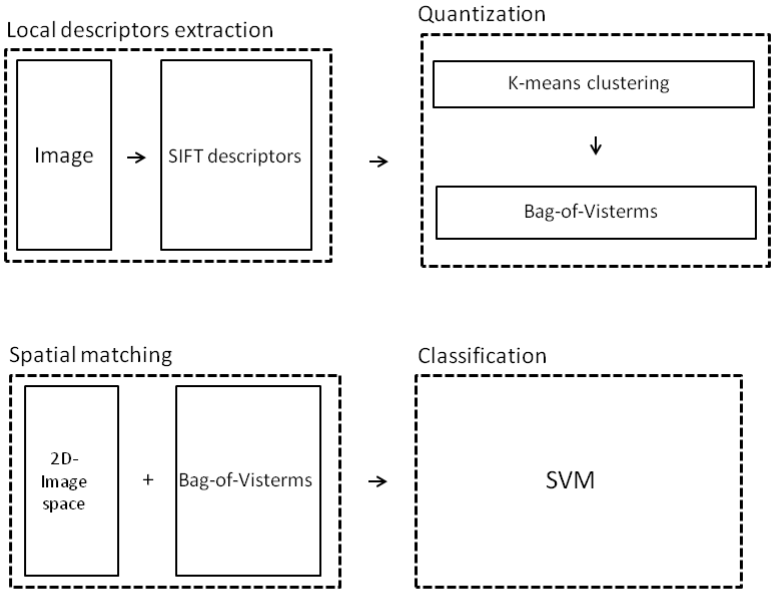

Figure 1. Summary of the steps involved in the process.

descriptor in the grid is replaced by the nearest visterm. Finally, a SVM classifier is trained employing a suitable kernel function for the pyramid of histograms of visterms.

Figure 2 shows a sample of the database in our experiment. We have analyzed a training dataset of 75 images, 15 images from each category. From this dataset we have built a vocabulary of 300 visterms. Then we have computed the pyramid of histograms of visual terms of each image. Finally, we compute the feature map associated with the χ^2-kernel and estimate the multiclass SVM classifier. Efficient code to compute our feature maps is available as part of the open source VLFeat library [11].

In order to assess the performance of the enabled methodology we classify a set of test images (75 images; 15 images from each category). The classification process is repeated 10 times, changing at random the training and test sets. Table 1 shows the mean and the standard error of the proportion of misclassification from each category. Central architecture and Silhouettes are the categories with a higher proportion of correct classification, 79% and 85% respectively. Subsequently, we find the categories Textured stone and Irregular stone with 61% and 57% of correct classification. Most classification errors in these categories are due to errors between both categories. The category Geometric stone has a lower proportion of correct classification, 41%. Most errors occur with the Irregular stone category.

4. Conclusions

The problem of classifying images based on the objects they contain constitutes an area of great activity in computer vision research. The set of methodologies currently available that addresses the problem of classifying images into categories is very efficient. In this work we have explored the behavior of bag-of-features techniques when faced with a database of images whose categories are determined by semantic aspects involved in the process of artistic ideation. We have shown that methods based on a bag of local descriptors and spatial pyramid

	CA	GS	IS	SI	TS
CA	0.79(0.04)	0(0)	0.04(0.003)	0.17(0.03)	0.01(0.007)
GS	0.14(0.02)	0.41(0.04)	0.31(0.05)	0.11(0.011)	0.03(0.011)
IS	0.01(0.01)	0.19(0.04)	0.57(0.02)	0(0)	0.23(0.011)
SI	0.11(0.01)	0.01(0.01)	0(0)	0.85(0.014)	0.03(0.011)
TS	0(0)	0.01(0.01)	0.37(0.011)	0(0)	0.61(0.017)

Table 1. True category in rows and Predicted category in columns. Categories are: Central architecture (CA), Geometric stone (GS), Irregular stone (IS), Silhouettes (SI) and Textured stone (TS). Cells in the table show the mean and the standard error, in brackets, of the proportion of misclassification.

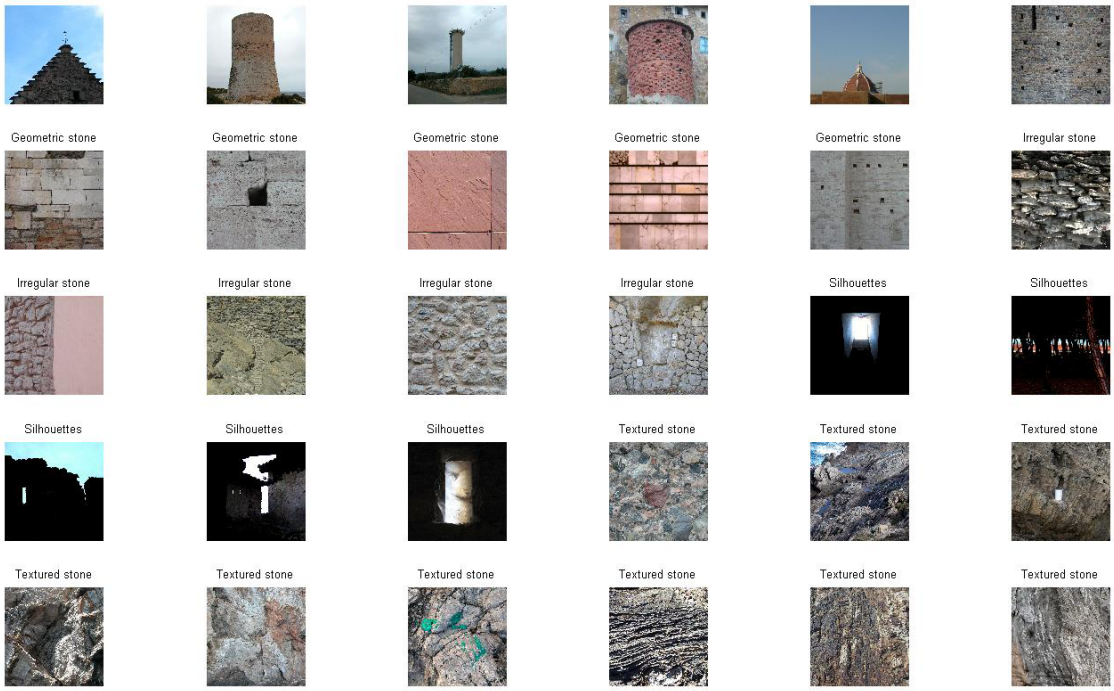

Figure 2. A sample from the dataset of images.

matching are adequate for the classification of images whose categories are based on semantic aspects. Experiments suggest that the automatic categorization of images based on computer vision methods can provide objective principles in cataloging images.

5. Acknowledgements

We would like to thank two anonymous reviewers for helpful comments on the manuscript. This work was partially funded by the University of Barcelona grant APPCSHUM 2011-2012.

References

[1] Willamowski, J., Arregui, D., Csurka, G., Dance, C- and Fan, L. 2004. Categorizing nine visual classes using local appearance descriptors. Proceedings of LAVS Workshop, in ICPR'04. Cambridge.

[2] Lowe, D. G. 2004 Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2),91:110.

[3] Fei-Fei, L and Perona, P. 2005. A Bayesian hierarchical model for learning natural scene categories. In Proceedings of CVPR.

[4] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. 2005. Discovering objects and their location in image collections. In Proceedings of IEEE International Conference on Computer Vision, Beijing.

[5] Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., and Gool, L. V. 2005. Modeling scenes with local descriptors and latent aspects. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Beijing.

[6] Bosch, A., Zisserman, A., Muñoz, X. 2007. Image Classification using Random Forests and Ferns. In Proceedings of IEEE International Conference on Computer Vision (ICCV).

[7] Lazebnik, S., Schmid, C., and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of CVPR.

[8] Grauman, K. and Darrel, T. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In Proceedings of IEEE International Conference on Computer Vision (ICCV), Beijing.

[9] Zhang,J., Marszaek,M., Lazebnik, C., and Schmid, S. 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision. DOI: 10.1007/s11263-006-9794-4.

[10] Vedaldi, A.,and Zisserman, A. 2010. Efficient Additive Kernels via Explicit Feature Maps. In Proceedings of CVPR.

[11] Vedaldi, A., and Fulkerson, B. 2008. VLFeat library (http://www.vlfeat.org/).