A Novel Structural-Description Approach For Image Retrieval

Christoph Rasche, Constantin Vertan
Laboratorul de Analiza si Prelucrarea Imaginilor
Universitatea Politehnica din Bucuresti
Bucuresti 061071, RO
rasche15@gmail.com, cvertan@alpha.imag.pub.ro

Abstract

We tested our image classification methodology in the photo-annotation task of the ImageCLEF competition [Nowak, 2010] using a visual-only approach performing automated labeling. Our labeling process consisted of three phases: 1) feature extraction using color histogramming and using a novel method of structural description, that was exploited in a statistical manner only; 2) classification using Linear Discriminant (LD) or Average-Retrieval Rank (ARR) methods that provided the confidence (scalar) values, which were then thresholded to obtain the binary values; 3) eliminating labels (setting binary values to 0) on the testing set thereby exploiting the calculated joint-probabilities for pairs of concepts from the training set. The results show that our present system performs better on 'whole-image' labels than on object labels.

1 Introduction

The main novelty of the presented approach is the use of a decomposition of structure as introduced in Rasche [Rasche, 2010]. The decomposition output is particularly suited to represent the geometry of contours and the geometry of their relations (pairs or clusters of contours), but it is applied here only in a statstical form for reason of simplicity [Rasche, 2011], together with a color histogramming approach as described in Vertan et al. [Vertan and Boujemaa, 2000b]. This statistical classification has already been shown to be useful for video indexing [Ionescu et al., 2010].

Looking at the provided photo annotations we realized that the spatial size of the annotated object or scene can vary substantially in reference to the image size: an annotation can describe either the image content as a whole and is thus suitable for (semantic) image classification, or it can describe a part of a scene (e.g isolated objects) and is thus rather suited for object-detection systems. A clear distinction between whole or part annotation is difficult of course, but is in our opinion desirable to better exploit the annotations, e.g. by providing a scalar value denoting the size of the object (1=whole image, 0.3=part/object covering ca. one third of the image). A typical recognition system is specialized for one process, either image classification or object detection. Our methodology is geared toward image classification and therefore is limitedly useful for 'part' annotations.

2 Method

2.1 Feature Extraction

Color and texture characterization The classical histogram image content description approach was further refined by the classification of the image pixels in several classes, according to a local attribute (such as the edge strength). We can easily imagine a classification in three classes, consisting of pixels characterized by a small, medium and high edge strength. The number of classes is thus related to the number of quantization level of the pixel attribute's. At the limit, since every pixel has acquired a supplementary, highly relevant characteristic, we can easily imagine a one pixel per class approach, which will certainly provide a very accurate description of the image, but will require a very important size.

In order to keep the balance between the histogram size and the discrimination between pixels we propose to adaptively weight the contribution of each pixel of the image into the color distribution [Vertan and Boujemaa, 2000b]. This individual weighting allows a finer distinction between pixels having the same color and the construction of a weighted histogram that accounts both color distribution and statistical non-uniformity measures. Thus, we will use a modified histogram, defined as:

$$h(\mathbf{c}) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} w(i,j) \delta \left(f(i,j) - \mathbf{c} \right), \quad \forall \mathbf{c} \in C \tag{1}$$

In the equation above $w(i,j)$ is the weighting coefficient of the color at spatial position (i,j). We may notice that, since $w(i,j)$ must be a scalar, we cannot use any color statistics (which are necessarily vector triples).

Intuitively the accounting within the color distribution of some local measures of each pixel could be considered as a way of integrating both color and texture, provided that the local measure have a textural background. The Laplacian-weighted histograms [Vertan and Boujemaa, 2000b], [Vertan and Boujemaa, 2000a] are defined as:

$$\tilde{h}(\mathbf{c}) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta \left(f(i,j) - \mathbf{c} \right) \frac{1}{1 + \Delta^2(i,j)}, \quad \forall \mathbf{c} \in C, \text{ or} \tag{2}$$

$$\tilde{h}(\mathbf{c}) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta \left(f(i,j) - \mathbf{c} \right) \Delta^2(i,j), \quad \forall \mathbf{c} \in C \tag{3}$$

The relation (2) emphasizes the weight of pixels that belong to constant (uniform) regions: their Laplacian is very small, so they sum with an unitary weight; the pixels placed on the edges are characterized by an important Laplacian and thus their contribution to the corresponding $\mathbf{c}$ bin is very small. This behavior is thought to reduce the influence of the uncertain colors, situated at the border between different objects and is derived from the gray-scale image case of choosing the segmentation thresholds as the minima of the histogram. The relation

from (3) corresponds to a dual behavior, counting the colors proportionally to their edge strength.

Colors are uniformly quantized with 6 bins per RGB color component, yielding a 216 components feature vector per image.

Structure characterization Images were downsampled to a maximum size of 300 pixels for any side length (width or height) to decrease computation time. The structural processing started with contour extraction ([Canny, 1986]) at 4 different scales (sigma=1,2,3 and 5). Contours were then partitioned and represented as described in (Rasche 2010) leading to 7 geometric and 5 appearance parameters for each contour segment (arc, 'wiggliness', curvature, circularity, edginess, symmetry, contrast, 'fuzziness'). Contour segments are then paired and clustered leading to another 58 parameters describing various distance measurements (between segments end and center points) and structural biases (degree of parallelism, T feature, L feature,...), see [Rasche, 2010] for details. For each parameter a 10-bin histogram is generated; the histograms are then concatenated to form a single vector of 700 dimensions. The average processing time for structural processing is ca. 40 seconds per image on a 2.6 GHz machine.

- Integration: The color and structural parameters are then concatenated to a single image vector with 916 dimensions (700 structural and 216 color parameters).

2.2 Classification

- LDA: A Linear Discriminant Analysis was applied to train a one-versus-all classifier for each of the 93 concepts (on the 8000 training images). This resulted in an average number of 24.9 labels per photo, more than twice as much as the average number of labels per training image (12.0). The posterior values of the classifier are provided as confidence values.

- ARR: The concepts for any test image are assigned based on a weighted average retrieval rank (ARR) of all training images retrieved following the query by example with the said test image. The binary concepts are obtained by a concept-adaptive threshold; the concept thresholds are computed based on the training image set annotations under the assumption that the test image database is statistically similar to the train image database.

2.3 Label Elimination

Because the LDA method (see above) returned a much larger proportion of labels for the testing set (24.9 labels/image) than for the training set (12.0), we attempted to reduce the number of labels by eliminating unlikely labels based on the joint-probabilities observed in the training set. Within the training set, we determined which pairs appeared as mutual exclusive (joint probability equal 0). If a testing image contained a pair of labels that are mutual exclusive in the training set, then the one label (of the pair) was eliminated that showed

a lower posterior value (obtained from the LDA classifier) in reference to the entire distribution of posterior values for each concept. After label elimination, the average number of labels was lower by ca. 6 labels, see last column in table number 1.

2.4 Runs

All runs contained the same structural preprocessing, but differed in their choice of color processing; thus, each run was tested with 916 parameters. The runs differed also in the choice of the classifier method and whether label elimination was used.

Table 1. Runs. Structure information was used in all runs. After the LDA (runs 3 to 5), the average number of labels in the testing set was 24.9 labels without label elimination. RGB: equation 1; Laplinv: equation 2; Lapl: equation 3.

Run #	Color	Classifier	Label-Elim	No. of Labels
1	Laplinv	ARR	-	(12.1)
2	Lapl	ARR	-	(12.1)
3	Laplinv	LDA	yes	17.3
4	Lapl	LDA	yes	16.7
5	RGB	LDA	yes	18.5

3 Results

3.1 Comparison of our runs

Runs 1 and 2 performed substantially better than the other three runs, for the MAP (Mean Average Precision; concept-based) and the F-ex measure (F-measure; example-based). For the OS-fcs (Ontology Score with Flickr Context Similarity costmap; example-based), runs 1 and 2 also performed better but not as distinctively as for the other two measures:

Table 2. Performance

Run #	MAP	F-ex	OS-fcs
1	0.259	0.531	0.562
2	0.258	0.529	0.562
3	0.206	0.408	0.513
4	0.182	0.366	0.487
5	0.201	0.414	0.520

Thus, ARR classification without label elimination outperforms LDA classification and label-elimination, but whether the performance difference is due to choice of classifer (LDA) or the attempt to eliminate labels can not be determined.

The detailed results per concept are shown in figure 1. Concepts with low average precision are skateboard, horse, cat, fish, etc. and tend to be objects, some of them likely part of the image only. Concepts with high precision are neutral illumination, no visual season, no blur, no persons etc. and tend to be whole-image annotations. This is what we roughly expected.

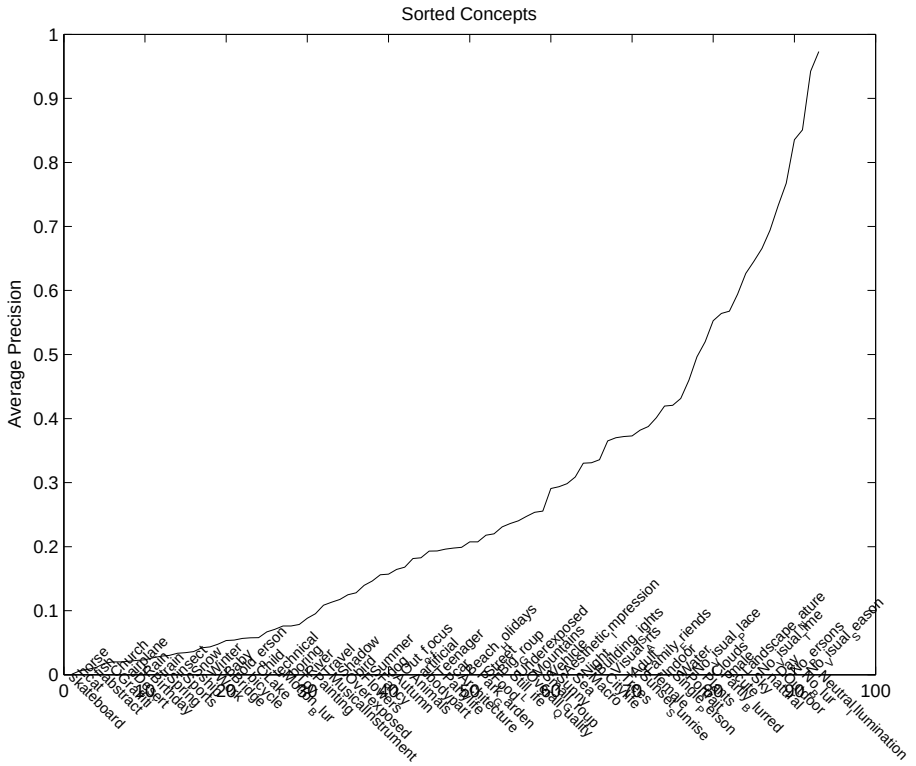

Fig. 1. Average performance for individual categories (sorted). Labels are alternatively tilted to improve readability (from left: skateboard, horse, cat,...)

3.2 Comparison to other groups

In comparison to other classification systems, our best results (of run no. 1) rank as 22nd out of 46 for the first two measures (MAP and F-ex) and 15th for the ontology-score measure (OS-fcs).

4 Discussion

Although we applied our structural decomposition in statistical manner only and on down-scaled image resolutions, it achieved already a performance comparable to other approaches. We do not expect a much better performance if the full image resolution was employed; rather, the long-term improvement lies in exploiting the individual contours and their relations, for which a proper learning algorithm needs to be developed [Rasche, 2011]. The fact that image size is not crucial for the extraction of semantic meaning - at least not for a human observer - was well pointed out by Torralba and co-workers [Torralba et al., 2008]. A quick solution to improve the present performance of our system could be to merge it with one of the appearance-based methods [Shotton et al., 2008,Heitz et al., 2009].

That label elimination did not lead to a significant improvement was unexpected, indeed it may have been even detrimental. But we still think that a proper exploitation of the joint probabilities can lead to a better performance.

References

[Canny, 1986] Canny, J. (1986). A computational approach to edge-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

[Heitz et al., 2009] Heitz, G., Elidan, G., Packer, B., and Koller, D. (2009). Shape-based object localization for descriptive classification. *International Journal of Computer Vision*, 84:40–62.

[Ionescu et al., 2010] Ionescu, B., Rasche, C., Vertan, C., and Lambert, P. (2010). A contour-color-action approach to automatic classification of several common video genres. In *AMR 8th International Workshop on Adaptive Multimedia Retrieval. Linz, Austria*.

[Nowak, 2010] Nowak, S. Huiskes, M. (2010). New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *In the Working Notes of CLEF 2010*.

[Rasche, 2010] Rasche, C. (2010). An approach to the parameterization of structure for fast categorization. *International Journal of Computer Vision*, 87:337–356.

[Rasche, 2011] Rasche, C. (2011). Contour groupings and their description for structural recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Under Review.

[Shotton et al., 2008] Shotton, J., Blake, A., and Cipolla, R. (2008). Multi-scale categorical object recognition using contour fragments. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 30(7):1270–1281.

[Torralba et al., 2008] Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2008*, 30(11):1958–1970.

[Vertan and Boujemaa, 2000a] Vertan, C. and Boujemaa, N. (2000a). Spatially constrained color distributions for image indexing. In *Proc. of CGIP 2000*, pages 261–265, Saint Etienne, France.

[Vertan and Boujemaa, 2000b] Vertan, C. and Boujemaa, N. (2000b). Upgrading color distributions for image retrieval: Can we do better ? In Laurini, R., editor, *Advances in Visual Information Systems*, volume 1929 of *Lectures Notes in Computer Science LNCS*, chapter , pages 178–188. Springer Verlag, Berlin, Germany.