SOFT CATEGORIZATION AND ANNOTATION OF IMAGES WITH RADIAL BASIS FUNCTION NETWORKS

Moreno Carullo, Elisabetta Binaghi and Ignazio Gallo
Università degli Studi dell'Insubria, via Ravasi 2, Varese, Italy

Keywords: Content-based image retrieval, Image categorization, Image annotation, Soft classification, Neural networks.

Abstract: This work focuses on fast approaches for image retrieval and classification by employing simple features to build image signatures. For this purpose a neural model for soft classification and automatic image annotation is proposed. The salient aspects of this solution are: a) the employment of a Radial Basis Function Network built on top of an image retrieval distance metric b) a soft learning strategy for annotation handling. Experiments have been conducted on a subset of the Corel image dataset for evaluation and comparative analysis.

1 INTRODUCTION

The growing demand for digital visual data in many applications related to scientific, commercial and cultural contexts has aroused a significant interest in Content-Based Image Retrieval (CBIR) aimed at defining methods to archive, query and retrieve these data based on their content (Datta et al., 2007; Smeulders et al., 2000).

The problem of filling the gap between visual, low level similarity and abstract semantic similarity is even more complicated when dealing with general-purpose, broad content image databases, such as Internet image archives, because of the large size of the database, the heterogeneity of the recorded scenes and the imaging techniques employed (Li and Wang, 2008). Usually these are databases where images are annotated with semantic labels, enabling the user to specify the query through a natural language description of the visual concepts of interest. These aspects combined with the cost of manual image annotation, have generated significant interest in the problem of automatically extracting high level semantic descriptors from images.

The problem can be addressed by different approaches (Datta et al., 2007; Smeulders et al., 2000). Early methods followed the geometrical approaches focusing directly on an explicit definition of a similarity function, powerful enough to represent high level meaning. In this direction strategies aimed at updating the query or optimizing the similarity function, thanks to the user annotations, have been proposed. Recent studies confirm that a single similarity measure can barely produce robust semantically meaningful ranking of images (Datta et al., 2007). An alternative approach which in some sense circumvents the problem, is the use of automated machine learning techniques able to induce, and then implicitly define from a set of already classified/annotated images, semantically meaningful similarity functions with which to categorize, ranking and annotate images (Datta et al., 2007; Vailaya et al., 2001).

Classification methods can be divided into two major branches: generative modeling and discriminative approaches (Bishop, 1996). In generative modeling, the searched category is modeled as a density probability function and the Bayes formula is then used to compute the posterior (Li and Wang, 2008). Discriminative modeling approaches are more direct in finding classification boundaries (Chen and Wang, 2004; Shotton et al., 2008).

Early works in the image categorization field make use of global color and texture histograms (Swain and Ballard, 1990). Recent works that try to exploit local features include the bag-of-features (Chen and Wang, 2004) where learning models are applied to collection of local features and pyramidal approaches (Grauman and Darrell, 2005; Lazebnik et al., 2006) where geometric description of the scene is accomplished. New trends also include approximated segmentation techniques are also applied to obtain good results with stricter time constraints (Li and Wang, 2008). All recent works can be roughly divided into approaches that prefer fast and

Carullo M., Binaghi E. and Gallo I. (2009).
SOFT CATEGORIZATION AND ANNOTATION OF IMAGES WITH RADIAL BASIS FUNCTION NETWORKS.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 309-314
DOI: 10.5220/0001785203090314

simple techniques for general purpose images and approaches with deeper and more costly techniques for image segmentation and object recognition.

The present work focuses on discriminative modeling for categorization and annotation tasks on large, content varied image databases. The main contribution of our study is to investigate whether these tasks can be approached successfully using the approximation capability of a novel supervised neural learning technique based on Radial Basis Function Network (RBFN). A salient aspect of the proposed solution is the integration within the RBFN of the Earth Movers Distance (EMD) which has been recognized as a useful similarity metric in information retrieval (Rubner et al., 2000).

In the context of automated image annotation the neural model is considered a soft classifier to better represent the inherent vagueness and imprecision with which images are annotated by users. The output of the neural classifier, for a given image signature provided in input, usually interpreted as crisp class assignment according to the winner-takes-all rule, must be softened here, considering the values of the output neurons directly as gradual relevance of the corresponding class/annotation to the image.

The overall strategy was experimentally evaluated using the Corel database subset used in (Chen and Wang, 2004). Several experiments have been conceived and conducted to quantify and compare the contribution of the different solutions adopted.

2 RBFN-BASED LEARNING FOR IMAGE CATEGORIZATION AND ANNOTATION

The present work focuses on the learning task within a CBIR strategy. However, to make the work self-contained, the important pre-processing phase concerning visual signature extraction is derived from previous works.

Section 2.1 describes the strategy adopted for visual signature extraction, while sections 2.2 and 2.3 explain the learning model and the annotation strategy.

2.1 Extraction of Visual Signature

This phase is crucial for the ability of the learning model to understand and predict concepts and categories. Proceeding from solutions adopted by Li and Wang (Li and Wang, 2008) a signature extraction technique for generic images is adopted, computationally easy but powerful enough to solve real world problems.

To build the signature a set of two feature $F = \{f_1, f_2\}$ where $f_1 = color$ and $f_2 = texture$ is considered. Each signature feature f_i is built on a set of vectors extracted from the image, one for each pixel. Vectors are then grouped together into a set of centroids $v_{j,k},\ k = 1,\ldots,K$ with the K-Means clustering method (with fixed K), and for each centroid $v_{j,k}$ a weight $w_{j,k}$ is computed to express the relevance of related pixels.

For the color feature the LUV color space components for each pixel are considered, while the Daubechies 4 wavelet transform (Daubechies, 1992) is employed as a texture descriptor. The texture descriptor is computed on the L-plane of the image, considering the LH, HL and HH planes to form the set of vectors that are in turn clustered.

Each image I_i is thus represented by its signature $\gamma_i \in \Gamma$ and is formed by features $\beta_{i,j},\ \ j = 1,\ldots,|F|$ where each feature $\beta_{i,j} = \{(v_{i,1}, w_{i,1}), \ldots, (v_{i,K}, w_{i,K}))\}$ is a discrete distribution.

The clustering phase for the extraction of the discrete distributions is a mean to summarize images by dividing them into regions with similar feature vectors. Several strategies can be adopted to exploit the differences among different types of images in the collection: in (Li and Wang, 2008) an adaptive meta-clustering method based on K-Means is used. A common and simpler alternative is to use the K-Means algorithm with fixed K. This strategy is adopted for the proposed solution.

2.2 The Soft Classifier

The present work addresses the main problem of semantic categorization and annotation of images with a machine learning approach. In particular, we chose adopted the RBFN model introduced by (Moody and Darken, 1989) for its proven training speed and robustness on classification and regression tasks. These capabilities are especially suitable for the inherent vagueness related to categorization and annotation within the CBIR context.

RBFNs have a single hidden layer of processing units with local, restricted activation domains: a Gaussian function is commonly used, but any other locally-tunable functions can be used. They were introduced as a neural network evolution of exact interpolation (Moody and Darken, 1989), and have been shown to have the universal approximation property (Hartman et al., 1990). As outlined in (Jain et al., 2000), the RBFN main advantages are that the classification function is non-linear, the model may produce confidence values and it may be robust to outliers; its drawbacks are the potential sensitivity to input parameters, and potential overtraining sensitivity.

The need to learn and predict on signature objects instead of regular vector patterns requires the standard Euclidean distance within the Gaussian activation units and the first-level K-Means clustering to be substituted with a distance tailored to discrete distributions. Considering the previous works on appropriate metrics for CBIR and CBIR systems making use of such metrics (Lv et al., 2006; Almeida et al., 2008) we selected the EMD - Earth Mover's Distance as the image distance metric within the RBFN model (Rubner et al., 2000).

The network is structured as a regular RBFN and its non-linear function $f : \Gamma \to \mathbb{R}^C$ maps the signature space to the categories space as a result of the learning phase on the training set $TrS = \{(\gamma_1, \mathbf{y}_1), \ldots, (\gamma_N, \mathbf{y}_n)\}$, where γ_i is a signature and $\mathbf{y}_i \in \mathbb{R}^C$ is the vector whose j-th component is the soft membership truth for the the j-th annotation.

The network is structured as follows:

1. a first level of M Gaussian Processing units $\phi_i : \Gamma \to \mathbb{R}^C$.

$$\phi_i(\gamma) = \exp(-\mathrm{emd}(\gamma, \gamma_i)/\sigma_i) \qquad (1)$$

   where $\mathrm{emd}(\gamma, \gamma_i)$ is the mean EMD over all signature features between the signature given as argument in ϕ_i and γ_i is the centroid signature for processing unit ϕ_i.

2. a second level of C linear weights $\mathbf{w}_i = \{w_{i,1}, \ldots, w_{i,C}\}$ connect each first level unit with each output unit.

3. the two levels are then linearly combined to build the model function f:

$$o_c(\gamma) = \sum_{i=1}^{M} \phi_i(\gamma) \cdot w_{i,c} \qquad (2)$$

$$f(\gamma) = \{o_1(\gamma), \ldots, o_C(\gamma)\} \qquad (3)$$

Following (Moody and Darken, 1989), the training scheme is two-phased: one is unsupervised and decides values for $\gamma_i,\ i = 1, \ldots, M$ while the other solves a linear problem to find values for $\mathbf{w}_i,\ i = 1, \ldots, M$.

1. the first phase finds suitable centroid signatures $\gamma_i,\ i = 1, \ldots, M$ by running an EMD-based iterative K-Means clustering algorithm with $k = M$. Then the p-means heuristic (Moody and Darken, 1989) is applied to compute the processing unit spreads $\sigma_i,\ i = 1, \ldots, M$.

2. the second phase computes $\mathbf{w}_i,\ i = 1, \ldots, M$. This phase is supervised and therefore the training set is considered; the objective is to minimize the difference between predicted output and truth by Least Mean Squares, computed through the pseudoinverse.

   (a) Φ is a $N \times M$ matrix where $\Phi_{i,j} = \phi_j(\hat{\gamma}_i)$

   (b) W is a $M \times C$ matrix where $W_{i,j} = w_{i,j}$

   (c) T is a $N \times C$ matrix where $T_i = \hat{\mathbf{y}}_i$

   the minimization problem to solve is $\Phi W = T$ and thus $W = \Phi^{\dagger} T$, where $\Phi^{\dagger}$ is the pseudoinverse.

The model has therefore two user parameters:

1. the number M of first level local processing units
2. the number p of the p-means heuristic, used to determine the spread of first level processing units.

2.3 Annotations and Categories

The visual content of an image can be described with words that have an accepted meaning. Be $A = \{a_1, \ldots, a_{|A|}\}$ the global dictionary of known annotations, the process of annotating each image I_i results in a set of weights $A_i = \{\alpha_1, \ldots, \alpha_{|A|}\}$ with $\alpha_j \in [0;1]$ and positive values of α_j are set for annotations a_j that belong to the image I_i.

A soft classification framework can be set up by teaching the model the annotation weights as the expected output for a given image; the training and test sets elements $(\gamma_i, \mathbf{y}_i)$ are such that $\mathbf{y}_i = \{\alpha_{i,1}, \ldots, \alpha_{i,|A|}\}$.

The RBFN output $\hat{\mathbf{y}} \in \mathbb{R}^C$ for a given image signature γ describes the level of confidence for each annotation, and can be used to predict the set of annotations. This can be addressed by considering only elements whose output units are activated with values higher than a threshold parameter $\varepsilon \in [0;1]$.

The elicitation strategy of annotation weights α_j is manifold. Considering real-world scenarios where users interact with the system by providing examples of tagged images, it is easy to imagine a simple graphical user interface where each annotation can be given a weight by adjusting its "visual size" just like a geometrical shape can be within a painting program. In simpler scenarios where only annotations can be taught and learned, the expected output $\mathbf{y}_i$ can be such that all components are equal to

$$\frac{1}{|\{a_{i,j} | a_{i,j} \text{ is an annotation of } I_i\}|} \qquad (4)$$

assuming that images with fewer annotations probably have stronger and clearer membership in respect to the annotation set.

Table 1: Error matrix of the hard classification analysis over the five runs, with User Accuracy (UA) and Producer Accuracy (PA) for each category.

-	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	Tot U	UA
ω_1	193	7	18	9	0	7	0	0	7	12	253	76.28 %
ω_2	4	157	17	2	0	2	3	0	46	0	231	67.97 %
ω_3	10	23	151	10	0	12	2	0	13	3	224	67.41 %
ω_4	0	11	13	210	0	0	0	0	5	4	243	86.42 %
ω_5	0	0	0	0	250	6	0	0	0	3	259	96.53 %
ω_6	18	14	15	3	0	204	0	1	8	6	269	75.84 %
ω_7	1	1	21	0	0	0	231	1	4	8	267	86.52 %
ω_8	4	4	0	1	0	13	10	247	5	3	287	86.06 %
ω_9	6	31	10	10	0	4	0	0	162	6	229	70.74 %
ω_{10}	14	2	5	5	0	2	4	1	0	205	238	86.13 %
Tot P	250	250	250	250	250	250	250	250	250	250	-	-
PA	77.20 %	62.80 %	60.40 %	84.00 %	100.00 %	81.60 %	92.40 %	98.80 %	64.80 %	82.00 %	-	-
Total accuracy: 80.4000 % (2010 hit, 490 miss, 2500 total)												

The global set of annotations A can grow unexpectedly when users are allowed to add their own new words. Its size can be kept under control by grouping clusters of elements into a single high-level annotation. In scenarios where automated tagging is used as a basic suggestion for the user, we expect that the most relevant elements are presented to the user, minimizing the presence of words with the same visual semantic.

3 EXPERIMENTS

The experimental analysis aims at assessing the performance of the proposed approach as an automated image annotation method. As shown in section 2 the overall process relies on the ability of the underlying machine learning model to predict a soft membership of a given set of conceptual classes. To better isolate the contribution of the learning model and of the annotations management task, the experiments were divided in two parts: first the proposed model is assessed as a hard classifier of images, while the second part considers the soft classification and automatic annotation capabilities.

For both the hard and soft experiments, the K-Means clustering technique used for signature building is employed with $K = 5$ as a result of trial and error phase, also taking into account reasonable computational times.

3.1 Hard Classification Analysis

For the hard classification analysis the Corel database subset used in (Chen and Wang, 2004) is considered. This dataset [1] is composed of 1000 small JPEG images divided into 10 categories: African people and villages (ω_1), Beach (ω_2), Historical buildings (ω_3), Buses (ω_4), Dinosaurs (ω_5), Elephants (ω_6), Flowers (ω_7), Horses (ω_8), Mountains and glaciers (ω_9) and Food (ω_{10}).

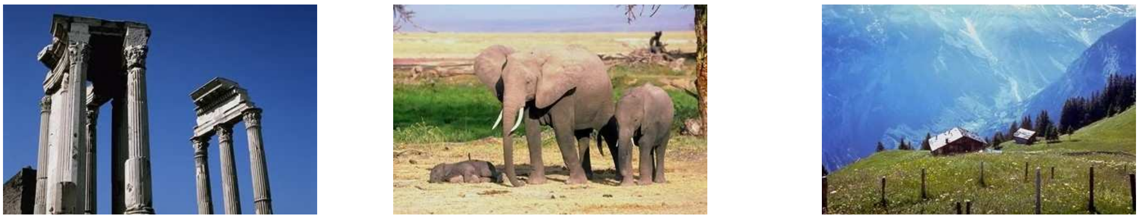
The set of images within each category is randomly split into two subsets of 50 elements to form the training and test set. Each experiment is repeated five times and the average overall accuracy (OA) is then reported as the main evaluation metric. When available, the number of processing units (NPU) of the learning model is reported. A complete error matrix (Congalton, 1991) over the five random runs is presented in table 1.

Two image categorization models proposed in (Chen and Wang, 2004) and (Andrews et al., 2003) respectively are considered to compare our method: MI-SVM, an extension of the standard Support Vector Machines model to the multiple-instance learning paradigm and DD-SVM, which aims at improving the MI-SVM by going beyond the single-prototype bag model.

We also compare the performance of a RBFN using a 125 bins LUV histogram (R-Hist) with that obtained by SVM employing the same image representation technique (HistSVM). Results are from (Chen and Wang, 2004). Experimental results of the overall accuracy are reported in table 2.

The HistSVM and R-Hist figures confirm that Radial Basis Function Networks can be employed in image catgorization tasks with similar performance to SVM models. The performance of the proposed R-EMD proves that a standard RFBN training technique combined with EMD-based radial basis functions and K-means can compete with more complex models based on the multiple instance framework.

[1] Dataset labels are now available at http://john.cs.olemiss.edu/~ychen/ddsvm.html. Images can be downloaded from http://wang.ist.psu.edu/docs/related.shtml



building, monument, sky — animal, elephant, tree, vegetation, grass, sky — mountain, building, vegetation, grass, tree

Figure 1: Sample images from the dataset used for soft classification and annotation analysis.

Table 2: Hard classification results of the proposed approach (R-EMD). Overall Accuracy (OA) with 95% confidence interval is reported; when available the number of Processing Units is also presented.

Model	*OA%* - [conf.int]	*NPU*
R-EMD	80.40 - [77.80 – 82.60]	100
R-EMD	77.52 - [74.91 – 80.13]	50
R-Hist	71.16 - [68.32 – 73.99]	100
R-Hist	67.88 - [64.96 – 70.80]	50
DD-SVM	81.5 - [78.5 – 84.5]	n.d.
MI-SVM	74.7 - [74.1 – 75.3]	n.d.
HistSVM	66.7 - [64.5 – 68.9]	n.d.

3.2 Soft Classification and Annotation Analysis

To investigate the performance of the model for annotation purposes, an annotated image dataset was needed. The absence of a widely accepted benchmark dataset in the CBIR research area lead us to put together a subset of the Corel images found in (Chen and Wang, 2004)[2] and adding proper annotations.

A set of 29 annotations $A = \{$*animal*, *beach*, *boat*, *building*, *cloth*, *cloud*, *decoration*, *desert*, *elephant*, *face*, *flower*, *forest*, *grass*, *horse*, *lake*, *monument*, *mountain*, *palace*, *person*, *river*, *rock*, *sand*, *sea*, *sky*, *snow*, *street*, *tree*, *vegetation*, *water*$\}$ is used to annotate 573 images, some examples are provided in figure 1. Annotations defined on images are converted to soft memberships as explained in section 2.3 by considering uniform weights as suggested in (4).

The whole image dataset is randomly split into two parts - for the training and test sets. The model is then trained and the neural network's output is evaluated within the soft paradigm as suggested in (Binaghi et al., 1999), considering the OA descriptive measure of the fuzzy error matrix. This evaluation metric isolates the behavior of the RBFN model without considering the threshold parameter ε. The annotation process is then evaluated considering the well-known Information Retrieval metrics Precision (P), Recall (R) and F-Measure (Frakes and Baeza-Yates, 1992) (F1) with the micro-average approach. These metrics, in particular F-Measure, describe the user-perceived performance of the system. All experiments are repeated five times and the average OA, Precision, Recall and F-Measure are reported.

Table 3: Soft classification and automated annotation results.

Model	*F.OA%*	*P%*	*R%*	*F1%*	*NPU*
R-EMD	48.44	64.43	55.95	57.23	20
R-EMD	53.59	66.56	63.99	62.49	50
Random	*n.a.*	14.62	52.27	20.69	*n.a.*

The model is evaluated fixing the threshold parameter $\varepsilon = 0.1$ and annotation performance is compared to a random annotator that selects a random number of tags from the available ones. This assesses the overall utility of the method with a lower bound method.

The Fuzzy OA (F.OA) shows that the model can learn soft memberships reasonably. The model was not supposed to behave perfectly with respect to this metric, and in addition the vagueness of learned and evaluated data makes Fuzzy OA behave differently from conventional, crisp OA.

The F1 score obtained shows the utility of the model over a completely random approach, by delivering an average 62.49% of correct annotations over the expected ones. Looking into the the F1 in detail, the Precision and Recall figures show that the major impact provided by the model is found in making the set of suggested tags more precise, or in other words small enough to contain the set of expected annotations.

[2] The dataset is available at `http://www.dicom.uninsubria.it/~moreno.carullo/cbir/datasets.html`

4 CONCLUSIONS

This work presented and evaluated a Radial Basis Function Network based approach to image categorization and annotation. Experimental analysis confirms that the proposed solution can be employed for both categorization and annotation tasks with encouraging results. The proposed soft classification approach seems promising and adequate for the management of intrinsic uncertainty of user-provided annotations. Future works involve the investigation of the performance on larger datasets with more images and annotations to assess the impact on the model's behavior.

REFERENCES

Almeida, J., Rocha, A., Torres, R., and Goldenstein, S. (2008). Making colors worth more than a thousand words. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1180–1186, New York, NY, USA. ACM.

Andrews, S., Tsochantaridis, I., and Hofmann, T. (2003). Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press.

Binaghi, E., Brivio, P. A., Ghezzi, P., and Rampini, A. (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recogn. Lett.*, 20(9):935–948.

Bishop, C. M. (1996). *Neural networks for pattern recognition*. Oxford University Press, Oxford, UK.

Chen, Y. and Wang, J. Z. (2004). Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939.

Congalton, R. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1):35–46.

Datta, R., Joshi, D., Li, J., James, and Wang, Z. (2007). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 39.

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Frakes, W. B. and Baeza-Yates, R. A., editors (1992). *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465.

Hartman, E., Keeler, J. D., and Kowalski, J. M. (1990). Layered neural networks with gaussian hidden units as universal approximations. *Neural Comput.*, 2(2):210–215.

Jain, A., Duin, R., and J.Mao (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA. IEEE Computer Society.

Li, J. and Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6).

Lv, Q., Josephson, W., Wang, Z., Charikar, M., and Li, K. (2006). Ferret: a toolkit for content-based similarity search of feature-rich data. In *EuroSys '06: Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, pages 317–330, New York, NY, USA. ACM.

Moody, J. E. and Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121.

Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *Semantic Texton Forests for Image Categorization and Segmentation*.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380.

Swain, M. and Ballard, D. (1990). Indexing via color histograms. *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 390–393.

Vailaya, A., Member, A., Figueiredo, M. A. T., Jain, A. K., Zhang, H.-J., and Member, S. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130.