Impact of Hierarchical Structures in Image Categorization Systems

Merza Klaghstan[1], Ronny Haensch[2], David Coquil[1] and Olaf Hellwich[2]
[1] University of Passau, Department of Distributed and Multimedia Information Systems
email: merza.klaghstan@fim.uni-passau.de, david.coquil@uni-passau.de
[2] Berlin Insitute of Technology, Department of Computer Vision
email: r.haensch@tu-berlin.de, olaf.hellwich@tu-berlin.de

***Abstract*— Image categorization refers to the process of assigning images to a number of predefined categories. The difficulty of problem solving is proportional to the number of categories the system addresses. This paper proposes an image categorization system, and studies the impact of dividing the categorization problem into smaller problems in a hierarchical structure. We compare examples solved with and without the proposed approach, to conclude its pros and cons.**

***Keywords*— Image categorization, Hierarchical structure, Subproblem, Features extraction, Dimensionality reduction, Learning.**

I. Introduction

With the massive growth of the size of image databases, categorization systems are becoming more important for scientific, commercial and personal purposes. Categorization is considered to be a very effective preprocessing step for image retrieval systems and search engines [1], in addition to many other potential applications such as photo archiving, driver assistance, autonomous robots and interactive games. Solving the categorization problem for a small number of categories with specific features and classifiers can be achieved relatively easy with satisfying results (see Section II), but the problem becomes more complicated for wider applications with a big number of targeted categories. Expressing categories in a in a hierarchical structure enables to divide the problem into smaller problems and deal with them independently, following divide-and-conquer paradigm. This paper, derived from [2], aims at first to investigate the operating subsystems of an image categorization system, along with state-of-the-art examples, and second to describe their implementation into a fully running example scenario, to compare results in different conditions.

The remaining of this paper is organized as follows. Section 2 lists related works. Section 3 overviews the categorization system and its underlying parts. Section 4 explains the testing environment used for the example in Section 5, which presents a categorization problem and analyzes the obtained results with and without the hierarchical approach. Finally, a conclusion is given in Section 6.

II. Related Work

Vailaya et al. [3] considered a hierarchical classification of vacation images, using fixed features and Bayesian binary classifiers. They used a hierarchical approach to overcome the limitation of binary classifiers, and apply them in a multi-class application. More generally, much work has been done to understand high-level semantics from images using low-level features. *Szummer and Picard* [4] proposed algorithms for indoor-outdoor scene categorization using a combination of color and texture features and a K-nearest-neighbour (K-NN) classifier. *Oliva and Torralba* [5] introduced a new concept for classification using spatial envelope properties to model the shape of a scene as a whole, and tested this using a support vector machine (SVM) classifier. *Xiao et al.* [6] understood the limitation of databases and categories in scene categorization applications. Therefore, they proposed an extensive Scene UNderstanding (SUN) image database. *Quattoni and Torralba* [7] tried to resolve a challenging classification problem in their work to recognize indoor scenes.

III. Categorization System Overview

In an image categorization system, smaller parts operate independently of each other, in order to deliver the full categorization functionality. Categorization process, as shown in Fig. 1, is split into two phases; learning and testing.

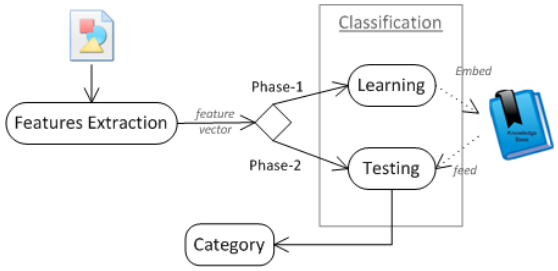

Fig. 1: Categorization process

Feature extraction is first applied to the input images in order to represent them using feature vectors. Dimensionality reduction techniques are needed for some features to compress the esulting high-dimensional feature vectors. Then, feature vectors are fed into a classifier, to build the knowledge-base at the learning phase, or decide for the corresponding category at the later testing phase.

978-1-4673-2584-4/12/$31.00 ©2012 IEEE

A. Features Extraction

Features extraction of an image changes its visual representation to a numerical form called "Feature Vector," in an N-dimensional feature space [8].
Features are grouped according to their attributes, examples are color vs. shape features and global vs. local features. In order to cover different aspects of features, we investigated color histogram, color moments, Gist and SIFT.

1) Color Histogram: Color histogram represents an image by a feature vector, whose elements (bins) characterize colors distribution in that image [9]. We implemented color histogram in the preferred HSV color space, since it separates the color components HS from the luminance component V [10], and empirically decided for the number of used bins as 24:16:8 for H, S and V channels respectively [2].

2) Moments: Moments quantitatively measure the shape of a set of points, by uniquely characterizing their probability distribution [9]. *Vailaya et al.* [3] used the first- and second-order moments in the LUV color space as color features. In a single color channel, the first moment is defined as the mean of all pixels, and the second central moment is the variance. To allow local and spatial properties to improve classification, it's suggested to divide the image into subblocks and compute the moments separately within them [4][3]. It's empirically shown that using 4x4 subblocks delivers better results [2]. Hence, the resulting feature vector will be 96-dimensional: 2 moments, per 16 subblocks, among 3 channels (LUV).

3) Gist: Gist refers to the amount of important semantic information, that an observer can keep in mind after quickly reviewing a scene [5]. The implementation of Gist comes in the form of Gabor-like scene filters, adapted to different orientations and scales [11]. In the original implementation of Gist, default values for the tunable variables are suggested as: 3 scales with 8 orientations for the first scale, 8 for the second and 4 for the third, applied onto 4x4 windows, in 3 color channels. Therefore, the resulting feature vector has the size $4 \times 4 \times (8 + 8 + 4) \times 3 = 960$, which is reduced using PCA (see Section III-B.2).

4) Scale Invariant Feature Transform: SIFT is a method to characterize an image with a set of feature vectors built upon local keypoints. The keypoints are defined as points of the image that are stable at different scales [12][13]. The resulting feature vector has $128 \cdot M$ values, where M is the number of keypoints. Since M differs from an image to another, the condition of fixed-size feature vectors is not fulfilled. Therefore, SIFT's output has to be manipulated. *Quelhas et al.* [14] proposed to assign each local descriptor to one quantizing cluster of a limited set of clusters, which is constructed by K-means clustering algorithm (see Section III-B.3).

B. Dimensionality Reduction

1) Features Selection: If there were a number of implemented features, for a given categorization problem one feature or a subset of features may be sufficient to solve that problem. However, determining a given feature's discriminative power is difficult [1]. This is approached in our work using direct estimation. This process encompasses running the system among learning and testing phases, using different features each time to directly estimate their accuracies, and selecting the feature with the highest one.

2) Principle Component Analysis: PCA is an approach to identify patterns in data, and represent them using their principal components, which results in reducing the number of dimensions without much loss of information [15]. It extracts the eigenvectors and eigenvalues from the covariance matrix of the input data, and then selects only the first R important principal components (eigenvectors). The value for R is determined according to the distribution of eigenvalues [16], so that their cumulative energy is above a certain threshold (80-90% of the total energy).

3) K-Means Clustering: K-means is an algorithm to divide N samples into K groups or clusters, according to their attributes. Clustering is done by minimizing the distortion between every data sample and its corresponding cluster centroid, where distortion is given by squared euclidean distances [17].

C. Learning

Categorization systems must learn from a number of training samples, to create a knowlege-base and then use it to resolve new unknown testing samples [8]. We used the K-Nearest-Neighbors algorithm as our standard method for testing in a supervised manner.

1) K-Nearest-Neighbor: KNN is a nonparametric classification technique, which does not depend on the data distribution, but instead works directly in the feature space [18][8]. The training phase is summarized by extracting feature vectors from known image-class pairs, and building the feature space. Then, the decision rule is built by picking the most frequent category among the K nearest neighbors. A good value of K can only be selected using statistical measurements. In our work, $K = 5$ proved to deliver the best results.

IV. Testing Environment

We implemented all of the previously described elements, and embedded them into a fully functioning categorization system, using a hierarchical approach to build categories' structure, split into different levels and subproblems.

A. Structure of Target Application

In order to sufficiently test the system, a structure is defined consisting of 3 levels and 12 leaf-nodes as final categories. Images are classified first as *Indoor* or *Outdoor*. *Outdoor* images are classified then as *Nature* (finally as *Forest*, *Beach* or *Desert*) or *Man-made* (finally as *Tower*, *Facade* or *Street*). On the other side, *Indoor* images are further classified as *Extended* (finally as *Subway*, *Corridor* or *Airport*) or *Closed* (finally as *Hall*, *Livingroom* or *Bookstore*).

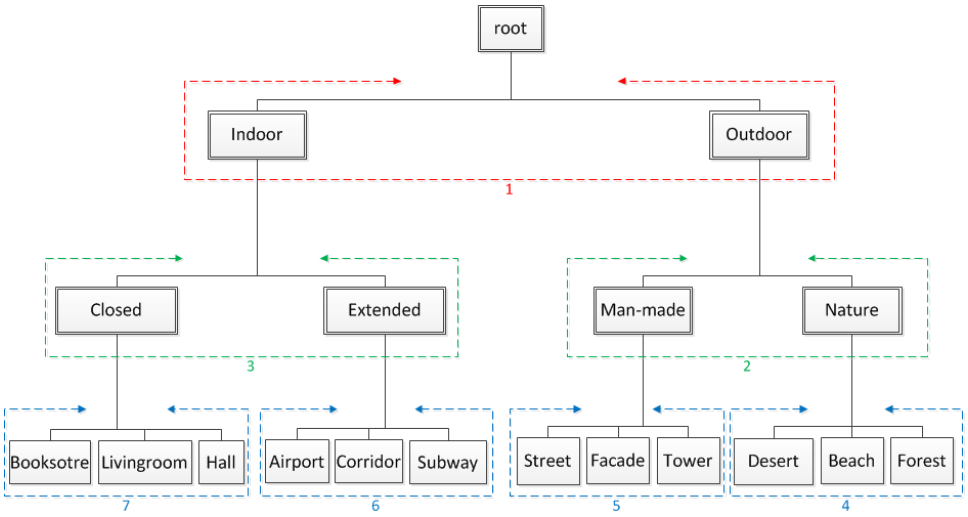

Fig. 2: Structure of target application

This structure, which is shown in Fig. 2, is inspired by the works of *Vailaya et al.* [3], *Oliva and Torralba* [5] and *Quattoni and Torralba* [7].

B. Image Database

The used image database consists of 2,400 images, distributed equally among the final 12 categories, 200 images pro category. All images are colored images, scaled to the size 256x256 pixels. Outdoor images are downloaded from [5], and indoor images are from [7]. Samples are given in Fig. 3.

C. Cross Validation

To reduce variability, experiments are done using k-fold cross validation [18], with $k = 4$ rounds, dividing the dataset of each category into 4 subsets, each with 50 images.

V. Experimentations

All results below are presented by the means of accuracies, where an accuracy is defined as the percentage ratio of correctly categorized images to the total number of images.

A. Separately Tested Subproblems

We start with simple categorization tasks, picking out subproblems-{4, 5, 6 and 7} from the structure in Fig. 2, and testing them separately each at once, as shown in Fig. 3a, 3b, 3c and 3d respectively.
Results of these problems are presented in Table 1.

Problem	**4**	**5**	**6**	**7**
Accuracy	87.4	87.5	74.7	72.8

Table 1: Selected features and percentage accuracies

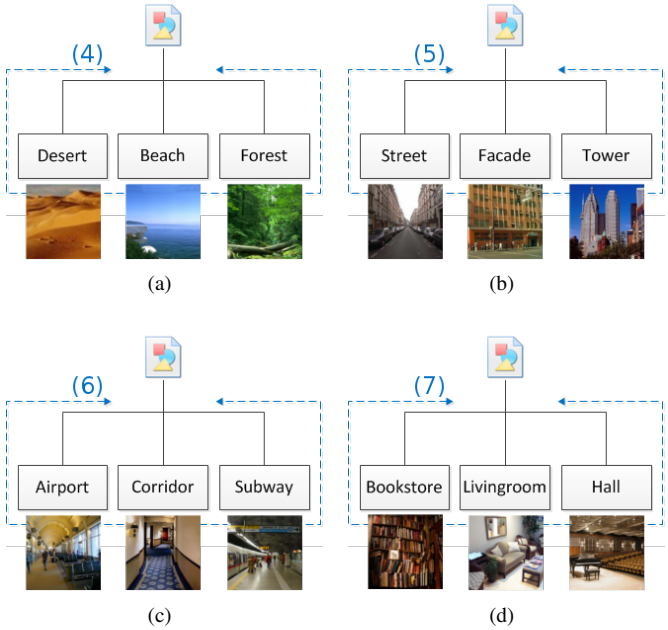

Fig. 3: Extracted subproblems from the main structure

B. Flat Structure

In this experiment, the final 12 categories are taken out of the hierarchical structure and set flat as direct children of the root node. Consequently, the number of problems is reduced from 7 to only 1, i.e. images must be directly classified as *Forest*, *Beach*, *Desert*, *Tower*, *Facade*, *Street*, *Subway*, *Corridor*, *Airport*, *Hall*, *Livingroom* or *Bookstore*. Final results are shown in Table 2 in 4 cross validation rounds and their corresponding average, using one feature at once. SIFT could not be tested because of its complexity when dealing with a large number of samples, which is derived from the complexity of the K-Means clustering algorithm.

	Accuracy (%)				
	R-1	R-2	R-3	R-4	Avg
HSV	37.2	41.7	41.7	41	**40.4**
LUV	37.8	41.3	36.8	38.8	**38.7**
Gist	67.8	64.2	65.7	65.8	**66**

Table 2: Results of a flat structure

C. Ordinary Full Test

Running direct estimation, the selected features for each of the 7 subproblems in the hierarchical structure are: (1:Gist), (2:Gist), (3:Gist), (4:Gist), (5:color moments), (6:Gist), (7:SIFT). Going through the 3 levels of categorization, obtained results are presented in Table 3.

		Accuracy (%)				
		R-1	R-2	R-3	R-4	Avg
K-NN	Level-1	86	86.3	88.5	89	**87.5**
	Level-2	77.7	78.3	81.5	80	**79.4**
	Final	61.8	62.3	65.2	63.3	**63.3**

Table 3: Full test results for the 3 levels of hierarchy

D. Analysis

As shown in Table 1, the system achieved good results with small categorization problems (here 3 categories). These results are to those obtained by original state-of-the-art works. The 87.4% accuracy of subproblem-4 is analogous to several results in [3], which range between 87.4-96.6%. Subproblem-5's 87.5% accuracy is comparable to the 83.7% of [5], and the accuracies of 74.7% and 72.8% for subproblems 6 and 7 are comparable to 63.2% in [7].
However, when the problem was extended to 12 categories at once, the results decreased significantly (Table 2).
Now, splitting the 12 categories into the structure shown in Figure 2 yielded better results (Table 3) when compared to HSV color histogram or LUV color moments in Table 2. Using Gist on the single flat problem yields a slightly better result than the hierarchical approach, because it fits the defined problem very well. However, this does not violate the general assumption that using a hierarchical structure gives the chance to select different features to fit better for the smaller subproblems. Furthermore, using a hierarchical structure prevents the clustering and learning algorithms from dealing with an extremely big number of samples in the feature space at once. This influences most significantly the algorithms that are non-linearly proportional to the number of samples, like K-means clustering for example. This fact practically saved us from using K-means clustering on samples from the 12 categories together, which would have consumed much time and memory space. Lastly, there might be applications, which are interested in categories in higher levels rather than final leaf categories. For example, *Nature vs. Man-made* rather *Beach vs. Street*, or *Indoor vs. Outdoor* rather *Forest vs. Bookstore*, and thus a hierarchical structure provides this possibility at better results than having all categories flat.
On the other hand, using a hierarchy has the drawback that when the structure grows vertically, this gives the chance for the error rate to increase accumulatively through the levels down. For example, the results in Table 3: the accuracy decreased from 87.5% to 79.4% and finally to 63.3% at the lower level.

VI. Conclusion

In this paper, we have shown the positive impact of the use of a hierarchical structure for image categorization. However, results could not be generalized because contradicted results were obtained. Additional work would be necessary to better prove or refute our assumption, by using broader examples, more features and more accurate classifiers. To this end, our system has the advantage that further improvements can be easily implemented due to its modular nature. Hence, once other components are implemented, they can be easily integrated into the system, as for example in [2] for example, in which different classifiers are used.

References

[1] A. Vailaya, A. Jain, and H. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition*, vol. 31, no. 12, pp. 1921–1935, 1998.
[2] M. Klaghstan, "Scene categorization, a hierarchical approach," 2011. [http://blog.merza-k.com/scene-categorization].
[3] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing," *Image Processing, IEEE Transactions on*, vol. 10, no. 1, pp. 117–130, 2001.
[4] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pp. 42–51, IEEE, 1998.
[5] A. Oliva and A. Torralba, "Modeling the shape of the scene," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001. [http://people.csail.mit.edu/torralba/code/spatialenvelope/].
[6] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," 2010.
[7] A. Quattoni and A. Torralba, "Recognizing indoor scenes," 2009. [http://web.mit.edu/torralba/www/indoor.html].
[8] A. Pinz, "Object categorization," *Foundations and Trends® in Computer Graphics and Vision*, vol. 1, no. 4, pp. 255–353, 2005.
[9] S. Sergyan, "Color content-based image classification," in *Proceedings of the 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics*, pp. 25–26, Citeseer, 2007.
[10] O. Chapelle, P. Haffner, and V. Vapnik, "Svms for histogram-based image classification," *IEEE transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1065, 1999.
[11] V. Prasad and J. Domke, "Gabor filter visualization," tech. rep., Technical Report, University of Maryland, 2005.
[12] D. Lowe, "Object recognition from local scale-invariant features," in *iccv*, p. 1150, Published by the IEEE Computer Society, 1999.
[13] U. Sinha, "Scale invariant feature transform," 2010.
[14] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 883–890, IEEE, 2005.
[15] L. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 52, 2002.
[16] I. Jolliffe, "Principal component analysis," 2002.
[17] K. Teknomo, "K-means clustering tutorial," *Medicine*, vol. 100, no. 4, p. 3.
[18] R. Duda, P. Hart, and D. Stork, *Pattern classification*, vol. 2. wiley New York:, 2001.