A Multi-Feature Optimization Approach to Object-Based Image Classification *

Qianni Zhang and Ebroul Izquierdo

Queen Mary, University of London,
Mile End Road, E1 4NS, London, UK
{qianni.zhang, ebroul.izquierdo}@elec.qmul.ac.uk

Abstract. This paper proposes a novel approach for the construction and use of multi-feature spaces in image classification. The proposed technique combines low-level descriptors and defines suitable metrics. It aims at representing and measuring similarity between semantically meaningful objects within the defined multi-feature space. The approach finds the best linear combination of predefined visual descriptor metrics using a Multi-Objective Optimization technique. The obtained metric is then used to fuse multiple non-linear descriptors is be achieved and applied in image classification.

1 Introduction

Content-based image retrieval uses descriptors derived from low-level image features and user relevance feedback to successively find pictures in a database according to a predefined metric in the descriptor space. These approaches rely on low-level analysis for the inference and classification process [1]. For this reason, the retrieval output often has little in common with high-level classification as expected by human observer.

Though low-level feature extraction algorithms are well-studied and able to capture important patterns in visual information [2], the bridge between automatic classification using such low-level primitives and higher level concepts remains an open problem. This challenge is referred to as 'the semantic gap' [3].

In this paper the problem of semantic image classification using multiple descriptors is considered. The emphasis is on single objects rather than on the whole scene depicted in the image. However segmentation is not assumed, since segmenting an image into single object is almost as challenging as the semantic gap problem itself. To deal with objects in images, small image blocks of regular size are considered. This paper focuses on devising an approach for combining the low-level descriptors and finding suitable metrics to represent and measure similarity between semantic objects. It is argued that semantic objects cannot be described by single low-level descriptors and metrics. Their nature is complex and requires a suitable combination of descriptors and multi-feature metric

* *The work leading to this paper was partially supported by the European Commission under contracts FP6-001765 aceMedia and FP6-027026 K-Space.*

spaces. But most low-level visual descriptors show non-linear behaviours and their direct combination may become meaningless. Some approaches to combine them have been suggested, like combining descriptor distances by reducing the metric combination to a single selected by a Boolean decision model and application of weighted linear merging of distances where the weights are accumulated from learned examples [4, 5]. A method to measure "visualness" of concepts in introduced in [6]. It performs probabilistic region selection for labeled images and computes an entropy measure of "visualness". In another approach user query is first classified into one of the predefined categories and the retrieval results with query-class associated weights are then aggregated by learning from the development data [7].

The idea of this paper is different from these methods and others in the literature. We propose to combine descriptors and optimize their metrics by analyzing the underlying patterns of low-level visual primitives in the training set. Then the classification of images is done based on the obtained metric. The proposed strategy is based on a Multi-Objective Optimization (MOO) technique, in particular the Pareto Archived Evolution Strategy (PAES) is adopted in this paper as optimization algorithm [8, 9, 10].

The paper is organized as follows: section 2 describes the strategy for block-based concept modeling and feature extraction. Section 3 introduces the proposed technique for building metric in a multi-feature space. Experimental evaluation of the proposed approach is presented in section 4 and the paper closes with conclusions and future work in section 5.

2 A Block-Based Approach to Visual Concept Modeling Using Low-level Primitives

Usually users are interested in finding single semantically meaningful objects rather than global descriptions of whole scenes such as landscapes, cityscapes, sunsets, or other elements make up the scenes. However current object segmentation technologies are not yet powerful enough to distinguish areas of an object from noisy areas like other objects or backgrounds when multiple objects are overlapping or when the object consists of several parts that are visually very different. Instead of doing segmentations, images can be regarded as mosaics of small building blocks as objects representation. In most cases these building blocks do not represent semantic concepts. But, small blocks of semantic objects should have certain similarity in their visual patterns. In this paper these blocks are referred as 'elementary building blocks', and some of these blocks that, according to professional user's subjective judgment, could best represent the visual patterns of a concept are chosen as 'representative building blocks'. Sets of these representative blocks can be used for visual pattern analysis and extraction. Having a small but very representative set of representative building elements for each semantic concept at hand, a suitable descriptor and its metric in a multiple feature space is sought by using the proposed method described in Section 3.

2.1 Object-based approach

In the proposed approach the each image is split into 8x8 blocks of regular size. Among the database of the elementary blocks, a professional user is required to select a set of best representative examples for a concept. Several visual features of the example set are analyzed and used as the training set of finding the most suitable metric space that combines them. An examples of choosing representative elementary building blocks of semantic concepts from one image is illustrated in Fig. 1.

2.2 Feature extraction and metric definition in multi-feature space

The primitives used by the proposed analysis are selected from the visual descriptors including MPEG-7 Colour Layout (CLD), Colour Structure (CSD), Dominant Colour (DCD), and Edge Histogram (EHD) [11]. Two texture features are also used as low-level primitives: Texture feature based on Gabor Filters (GF) [12] and Grey Level Co-occurrence Matrix (GLCM) [13]. Additionally, to emphasis invariance to saturation, Hue-Saturation-Value (HSV) [14] color system is also considered.

As mentioned before since most low-level visual descriptors are complex and non-linear, they cannot be combined directly. Thus in this paper a combination of distances with certain metric is used as a similarity measurement. In the very first stage, a step to measure the primitive distances of blocks within different descriptor spaces is conducted. A distance function or metric is defined as

$$d = dist(v_1, v_2) \tag{1}$$

and varies for different descriptors, where v_1, v_2 are the feature vectors for a particular descriptor.

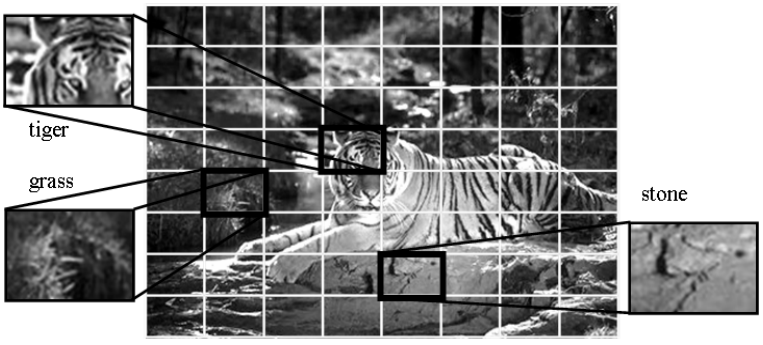

Fig. 1. An image consisting of complex objects is split into elementary building blocks representing single objects

Suppost n descriptors are considered, v_j is the j_{th} descriptor used, $j = [1, n]$. To combine the distance calculated for each elementary block, the most straightforward candidate of possible metrics in the multi-feature space is the linear combination of the distances defined for the descriptors:

$$D(V_1, V_2, A) = \sum_{j=1}^{n} \alpha_j d_j(v_j^{(1)}, v_j^{(2)}) \tag{2}$$

where D is the sum of a set of distance function as defined (1), and it measures the distance between two sets of feature vectors in a multi-feature space. A is the set of weighting coefficients α we are seeking to optimize. For the specific case given by (2), the optimality problem is regarded in the sense of both concept representation and discrimination power.

The approach to estimate a metric in the underlying multi-feature space relies on comparing different descriptors. Unfortunately, in most cases comparing these functions becomes meaningless. To ensure minimum comparability requirement all distances are normalized using simple *Min-Max Normalization*. This transforms the distance output into the range $[0, 1]$ by applying:

$$d_{j(new)} = \frac{d_j - min_j}{max_j - min_j} \tag{3}$$

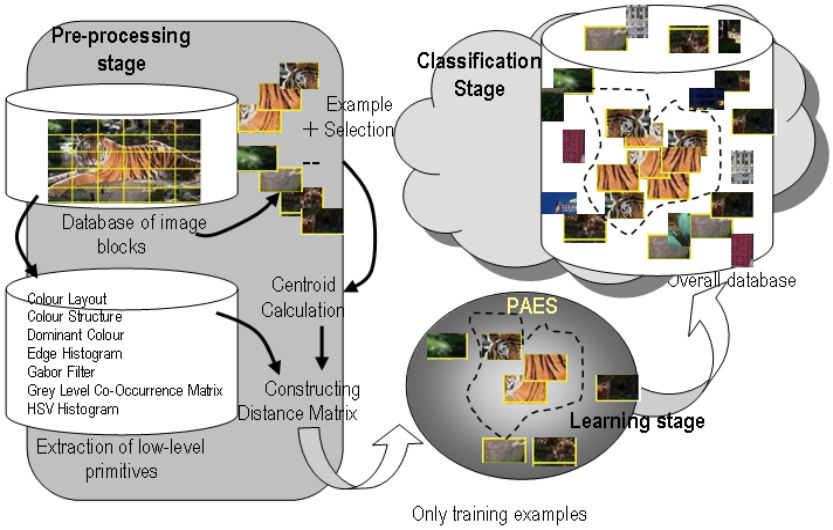

Fig. 2. The overall procedure in a scenario of searching for images of 'tiger' using proposed approach

3 Image Classification in Multi-feature Space

The overall procedure of the proposed approach can be divided into three stages: the pre-processing stage, the learning stage and the classification stage. This is illustrated in Fig. 2. The pre-processing stage includes steps of splitting images into blocks and low-level feature extraction. Besides, in this stage some representative block are selected, their centroid is calculated, and a distance matrix is constructed. The learning stage uses PAES to define a multi-dimensional similarity metric space. The classification stage uses the learned metric to classify blocks in database.

3.1 The pre-processing stage

Initially all images are split into blocks and the visual features of the blocks are extracted. Given a semantic concept that the user would like to retrieve, the first step of the proposed approach is to build up the training group of 'representative building blocks'. It is required for a visual representative element group to be able to represent the nature that the objects of a concept have in common. Besides, it is also required that it possesses the discriminating power of the concept from noise of unrelated elements. Therefore in this paper two types of the representative example are selected and both of the two types are combined in a training set. The first type of representatives is the most relevant examples to a concept in common understanding and they are referred as 'positive examples', while the second are 'negative examples' which, intuitively, should consist of blocks that are visually close to the concept but do not represent the targeting concept.

Let $S = \{s^{(i)} | i = 1, ..., m\}$ be the training set of elementary building blocks containing m elememts in total. For n low-level descriptors, a $m \times n$ matrix is formed in which each element is a descriptor vector. The centroid for each descriptor is calculated by finding the block with the minimal sum of distances to all other blocks in S. All the centroids across different descriptors form a particular set of vectors $\bar{V} = \{\bar{v_1}, \bar{v_2}, ..., \bar{v_n}\}$, in which $\bar{v_j}$ is the centroid vector for all the vectors of the j_{th} descriptor used.

In general $\bar{V}$ does not necessarily represent a specific block of S. Taking $\bar{V}$ as an anchor, for a given concept representing an object the following distance matrix can be constructed:

$$\begin{matrix} d_1^{(1)} & d_2^{(1)} & \dots & d_n^{(1)} \\ d_1^{(2)} & d_2^{(2)} & & d_n^{(2)} \\ \vdots & & \ddots & \\ d_1^{(m)} & d_2^{(m)} & & d_n^{(m)} \end{matrix} \tag{4}$$

is built. In (3) each row contains distances of different descriptors estimated for the same block, while each column display distances for the same descriptor for all blocks.

3.2 Learning weighting factors for the multi-feature metric

Semantic objects can be more accurately described by a mixture of low-level descriptors than by single ones. However, this leads to the difficult question about how descriptors can be mixed and what is the "optimal" contribution of each feature. In a realistic scenario, an approach based on optimizing a single objective function will not lead to acceptable results because of the complex nature of semantic objects. Often for semantically similar objects, their visual primitives are not similar. Even worse, in many cases different low-level visual features contradict each other. To illustrate the conflicting nature of the objective function presented in (2), an example is considered in Fig **??**.

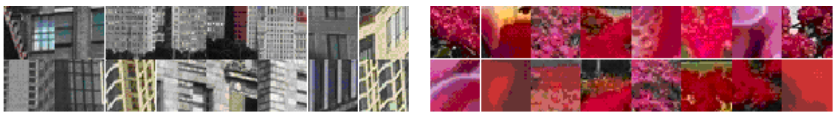

Fig. 3. Examples of image blocks for 'Building' group (left) and 'Flower' group (right)

Fig **??** shows Examples of image blocks for 'Building' group (left) and 'Flower' group (right). Blocks selected from images containing buildings ('Building' group) and blocks containing red flower ('Flower' group) are shown in this figure. Considering the 'Flower' group and its intrinsic concept (flower), a colour descriptor identifies blocks in which the red colour is dominant. That is, if single objective optimization is used, say for two descriptors, CLD and EHD, a weight of 1 will be assigned to CLD while a weight of 0 will be assigned to EHD. The retrieval process will mark predominantly red blocks in the database as "Flowers". In this case the edges or textures of the flowers, which also strongly contribute to the semantic concept, will be fully ignored. On the other hand, for the 'Building' group, using a single objective optimization for CLD and EHD, the EHD will dominate the similarity estimation while the CLD or other colour and texture features that are also important for the concept will again be neglected. In order to include all the descriptive characters contained in every single representative building block, each of them should be considered as an objective function in the optimization problem. However, when optimizing a set of contradicting objective functions, usually there is not unique solution achieving an optimum for all objectives at the same time. The solution of the problem at hand is closely related to multiple decision making strategy in which simultaneous optimization of multiple objectives is sought.

In this paper the Pareto Archived Evolution Strategy (PAES) [8] is adopted to optimize the combination metrics. PAES is an evolution strategy employing local search but using a reference archive of previously found solutions to identify the approximate dominance ranking of the current and candidate solution vectors. This will produce a set of set of Pareto Optimal Solutions. Unfortunately

none of these pareto-optimal solutions can be identified as better than others without any further consideration, so a second step is required: a higher-level decision-making involving further considerations to choose a single solution.

In this paper the second step is based on finding the minimum solution of

$$F = \frac{sum\ of\ objective\ functions\ of\ all\ positive\ examples}{sum\ of\ objective\ functions\ of\ all\ negative\ examples} \tag{5}$$

considering the that small sums of weighted distances of positive examples means better gathering of all positive points while big sums of weighted distances of negative examples means sparseness of the negative points, which is just the target we are seeking to achieve.

An intensive study as well as comparison with other algorithms in MOO is done in [8]. As a result PAES is a capable multi-objective optimizer across the problems tested.

The problem of finding the suitable metric consists of finding the optimal set of weighting factors α, where optimality is regarded in the sense of both concept representation and discrimination power. This optimization problem can be tackled by minimizing or maximizing one or several objective functions as in (2).

For a given semantic concept and its according distance matrix (3), the optimization is then performed on the set of objective functions like (2):

$$\bar{D}(V, \bar{V}, A) = \begin{cases} D_1(V_1, \bar{V}, A) \\ D_2(V_2, \bar{V}, A) \\ \vdots \\ D_m(V_m, \bar{V}, A) \end{cases} \tag{6}$$

In (5) $\bar{D}$ is the set of objective functions $\{D_i, i = 1, ...m\}$, D_i is the distance vector of the i_{th} block, and A is the collection of weighting factors. The optimal solution is to find the $A = \{a_j | j = 1, ..., n\}$ by which the objectives of positive examples in $\bar{D}$ reaches their minimal values while the objectives of negative examples in $\bar{D}$ reaches their maximal values, subject to constraint $\sum_{j=1}^{n} a_j = 1$. This set of weighting factors is assumed to be the metric that represents the symbolic nature of the concept within a multi visual feature space.

According to the different kinds of examples the way of optimizing the objective functions are different. The optimization process is to simultaneously minimizing the objective functions from the positive representative group and maximizing the objective functions from the negative representative group.

3.3 Classification: The Minimum (Mean) Distance Classifier

The Minimum (Mean) Distance Classifier (MDC) is utilized in this paper for classification within the obtained multi-feature metric space. The reason of choosing it as the classifier is that it is simple to implement and works well when the distance between means is large compared to the spread of each class. What's more,

because of its simplicity, it is easy and safe to be transformed into any desired non-linear high-dimensional multi-feature space. Some more intelligent classifier may also be used in future but for now they are avoided despite their various appealing characters, due to the uncertainty of their behaviours when adapted into the transformed metric space.

MDC is a special case of classifiers based on discriminant functions. It is usually applied in linear space, but in this paper, it is adapted to the metric space which combines several non-linear similarity functions of descriptors by the linear weighted function obtained from the PAES algorithm.

The centroid vector as we described in Section 2.2 from the positive examples is used as the mean of the positive class, and is referred later as $\bar{V}_+$. On the other hand a $\bar{V}_-$ that can be obtained by using the same method from the negative examples is used as the mean of the negative class. Using the obtained metric, say $A = \{a_j | j = 1, ..., n\}$, write the distance functions to the two mean values as:

$$D_{i+}(V_+^{(i)}, A) = \sum_{j=1}^{n} \alpha_j d_{j+} \tag{7}$$

$$D_{i-}(V_-^{(i)}, A) = \sum_{j=1}^{n} \alpha_j d_{j-} \tag{8}$$

$D_+^{(i)}$ is the distance vector of the i_{th} block to $\bar{s}_+$, and vice versa; while D_{i+} is the similarity estimation of the i_{th} block in the obtained metric space, and vice versa. The decision boundary which separates the positive class from the negative class is given by:

$$D_{i+}(V_+^{(i)}, A) - D_{i-}(V_-^{(i)}, A) = \delta \tag{9}$$

where δ is a variable that is usually 0 but can be changed for different concepts to fit different requirements.

4 Experimental Evaluation

As stated before current approaches from the conventional literature combine multiple features for image classification using a different model. Usually, classification is done applying single descriptors and fusion of results is performed after the initial mono-feature classification. This makes it difficult to compare our approach with relevant ones from the literature. Even a simple analysis of the final results, for the sake of comparison, is not feasible due to the lack of common test sets. A more critical fact rendering a fair comparative study almost impossible is that software implementation of previously reported approaches is not available and it is not trivial to implement them using only the reported algorithmic steps. For these reasons in this paper only a set of experiments using each single descriptor have also been performed for comparison with the proposed approach.

4.1 Experimental Setup

The test data contains 700 images selected from 'Corel' dataset. The images are labeled manually on 5 predefined concepts as ground truth. The concepts are "building" (141), "cloud" (264), "grass" (279), "lion" (100), and "tiger" (100). The numbers in brackets after concepts are the numbers of images containing the concepts in the test dataset according to ground truth.

As the propose approach classifies images based on their elementary building blocks, we argue that if a block of an image is classified as relevant to a concept, then the image itself is judged as similar to the concept. As a retrieving performance evaluation the MDC is used as a classifier.

4.2 Experimental Results and Evaluation

A group of 10 positive representative blocks and 10 negative representative blocks are manually selected to represent each concept by professional user. For each group (concept) a distance matrix (5) has been computed with the selected blocks and the 7 descriptors used resulting in sets of 7 weighting factors. Using the 7 weighting factors as a combination metric, the accuracy of relevant images classified by the MDC classifier is shown in Table 2, as the first column in each row. The experiments using each single descriptor are also shown in Table 2 for comparison and evaluation.

As it can be observed from Table 2, among the 5 groups of experiments, in the experiments for "cloud", "grass", "lion" and "tiger", the approach using the obtained metric outperforms the approached using any single one of the 7 descriptors. Only in the "building" group the single descriptor EHD slightly outperforms the proposed approach.

Table 1. The accuracy of image classification using obtained metric

%	Obtained metric	CLS	CSC	DCD	EHD	GF	GLCM	HSV
building	70	48	24	20	74	40	38	42
cloud	79	76	70	38	68	28	34	78
grass	92	92	86	28	82	64	88	88
lion	88	50	36	16	50	24	40	66
tiger	60	2	46	7	14	26	34	57

However the results show that the proposed approach using positive and negative representative blocks is generally better than the retrieval based on single descriptors. Even though in some cases specific single descriptors are dominant for a concept, the result from proposed approach is very close to it.

5 Conclusion and Future Work

A technique to estimate optimal linear combinations of predefined metrics by applying a Multi-Objective Optimization is presented. The core strategy uses MOO to optimize the metric in multi-feature space. The proposed approach has been tested for the classification of objects in images. A more comprehensive evaluation of the proposed technique and additional improvements of the method are being undertaken. Immediate work includes adopting more intelligent classifier which can employ the obtained multi-feature metric, as well as extension and evaluation with several other low-level descriptors. Future work will focus on non-linear combinations of descriptors and metrics.

References

1. J. R. Smith and S.Chang. "Visualseek: a fully automated content-based image query system". Proceedings of ACM Multimedia 96, pages 87–98, Boston MA USA, 1996.
2. S.-E Chang and T Sikora, A. Purl, "Overview of the MPEG-7 Standard", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 688-695, 2001.
3. J. O'Reilly, ContentEengineering. Electronics Communications Engineering Journal, vol. 14, No. 4, Aug. 2002.
4. H. Eidenberger and C. Breiteneder: "Macro-level Similarity Measurement in ViZir". 2002.
5. Q. Tian, Y. Wu, and T. S. Huang: "Combine User Defined Region-Of-Interest and Spatial Layout for Image Retrieval".IEEE ICIP'2000, pp. 746-749, Vol. 3
6. K. Yanai and K. Barnard, "Image Region Entropy: A Measure of Visualness of Web Images Associated with One Concept". Proc. ACM Multimedia 2005, pp.419-422.
7. R. Yan, J, Yang and A. G. Hauptmann, "Learning QueryClass Dependent Weights in Automatic Video Retrieval". Proc. ACM Multiemdia 2004, pp.548-555.
8. R.E. Steuer: "Multiple Criteria Optimization: Theory, Computation, and Application". New York: Wiley 1986.
9. J. Knowles and D. Corne: "Approximating the Non-dominated front using the Pareto Archived Evolution Strategy". 1999.
10. J. Knowles and D. Corne, "Properties of an adaptive archiving algorithm for storing nondominated vectors", 2002.
11. N. O' Connor and E. Cooke , Le Borgne H., Blighe M., Adamek T. "The aceToolbox: Lowe-Level AudioVisual Feature Extraction for Retrieval and Classification". Proc. of EWIMT'05, Nov. 2005.
12. B.S. Manjunath and W.T. Ma, "Texture features for browsing and retrieval of image data," IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837-842, August 1996
13. M. Tuceryan and A. K. Jain, Texture Analysis. The Handbook of Pattern Recognition and Computer Vision (2nd Edition), pp. 207-248, World Scien-tific Publishing Co., 1998.
14. M. J. A. Swain and D. H. A. Ballard, "Color indexing" International Journal of Computer Vision, vol. 7, pp. 11-32, 1991.