OPTIMIZING METRICS COMBINING LOW-LEVEL VISUAL DESCRIPTORS FOR IMAGE ANNOTATION AND RETRIEVAL

Qianni Zhang and Ebroul Izquierdo

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

ABSTRACT

An object oriented approach for key-word based image annotation and classification is presented. It considers combinations of low-level descriptors and suitable metrics to represent and measure similarity between semantically meaningful objects. The objective is to obtain "optimal" metrics based on a linear combination of single metrics and descriptors in a multi-feature space. The proposed approach estimates an optimal linear combination of predefined metrics by applying a Multi-Objective Optimization technique based on a Pareto Archived Evolution Strategy. The proposed approach has been evaluated and tested for annotation of objects in images.

1. INTRODUCTION

Recognizing objects and semantic concepts in images has become a major research topic in image processing. It is the pillar for automatic annotation and retrieval of visual information and in a more generic context it is the base for high-level image understanding. Fully automatic analysis for accurate image annotation with high-level semantic concepts is not feasible today. Though, low-level feature extraction algorithms are well-understood and able to capture subtle differences between colours, statistic and deterministic textures, the global colour layout and distribution in images, etc, the bridge between the automatic classification of such low-level primitives and higher concepts remains an open problem. This problem is referred to as 'the semantic gap' and commonly defined as *the discrepancy between low-level features or content descriptors that can be computed automatically by current machines and algorithms, and the richness, and subjectivity of semantics in high-level human interpretations of audiovisual media*. To bridge this gap is a challenge that has captured great attention from researchers in computer vision, pattern recognition, image processing and other related fields, evidencing the difficulty and importance of such technology and the fact that the problem remains unsolved. The challenge offers the possibility of adding the audiovisual dimension to well-established text databases to build truly multimedia enabled information retrieval.

Much related work on image indexing and retrieval has focused on the definition of low-level descriptors and the generation of metrics in the descriptor space [1], [2]. These techniques are aimed at defining image signatures using primitives extracted from the content, e.g., pixel patterns and dynamics in image and video or sampling patterns in audio signals. They are extremely useful in some generic image classification tasks or when a query by example is considered. However, if the aim is to annotate single objects in complex images using semantic words or sentences, two profound challenges become evident: how to deal with the subjective interpretation of images by different users under different conditions; and how to link a semantic-based concept with low-level metadata. The first problem originates from the fact that perceptual similarity is user and context dependent. It can be tackled by constraining the annotation to a given set of words in a well-defined taxonomy or ontology. The second challenge is a synonym of the semantic gap. To tackle it the machine needs to learn associations between complex combinations of low-level patterns and concepts. Consequently, complex combinations of features building high-dimensional and heterogeneous feature spaces need to be considered.

The objective of this paper is twofold: to present an object oriented approach for key-word annotation and to consider combinations of low-level descriptors and suitable metrics to represent and measure similarity between objects present in images. The rational behind the first objective is that real-world applications require annotations at object level rather than global descriptions of whole scenes such as landscapes, cityscapes, sunsets, etc [3], [4]. Here the emphasis is on single objects rather than on the whole scene depicted in the image, without however assuming segmentation, since we argue that segmenting an image into single semantically meaningful object is as challenging as almost equivalent to the semantic gap problem. Consequently, to deal with objects a very simple approach

---

This work was partially supported by the European Commission under contract FP6-001765 aceMedia.

is taken based on what we call *elementary building elements of images* or small image blocks of regular size. The proposed technique was inspired by three observations: users are mostly interested in finding objects in images and do not care the surroundings in picture; elementary building elements are closer to low-level descriptions than whole objects or complete pictures; and objects are made up of these elementary building elements.

The second objective is motivated by the fact that semantic objects cannot be described by single low-level descriptors and metrics. Their nature is usually complex and requires a suitable combination of descriptors and metrics in multi-feature metric spaces. To find an optimal combination of low-level primitives into single multi-feature descriptors is very difficult since the interaction of different descriptors has not been sufficiently studied and comprehended so far. Most low-level visual descriptors show non-linear behaviors and their direct combination may easily become meaningless. Some approaches to combine them have been suggested in the past. For instance, combining descriptor distances by reducing the metric combination to a single one selected according to a Boolean decision model [5]. Using weighted linear merging of distances in which the weights are accumulated from learned examples [6].

The focus of this paper is to obtain "optimal" metrics based on a linear combination of single metrics and descriptors in a multi-feature space. The aim is to classify or identify elementary building elements belonging to semantic meaningful objects using such complex metrics in the multi-feature space. The proposed approach estimates an optimal linear combination of single metrics by applying a Multi-Objective Optimization (MMO) technique [8] based on a Pareto Archived Evolution Strategy [9].

The paper is organized as follows: section 2 describes a strategy for modeling and construction of visual concepts as well as descriptor extraction and corresponding metrics. Section 3 introduces the proposed technique for metric optimized 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. VISUAL CONCEPT MODELING USING LOW-LEVEL PRIMITIVES

Most annotation and retrieval approaches from the literature have dealt with either whole images or segmented regions. When complete images are used, annotation and retrieval engines fail to produce satisfactory results specially if the user is interested in finding objects in images regardless other scene elements that make up the whole picture. If segmentation is used the presence of noisy regions and over-segmentation is unavoidable. This leads to confusing and inadequate retrieval results. Consequently, on the one hand we need to consider the fact that even the best image segmentation techniques cannot reliably extract meaningful semantic objects and thus it is not reasonable to assume object segmentation in a retrieval system. On the other hand, without segmentation most learning machines fail to capture the specific key-word representing the single object the user is interested in. To alleviate this problem in this paper we consider small blocks of regular size. The strategy consists of classifying these elementary building elements using a suitable combination of low-level features and a previously annotated training data set. Once elementary building elements are classified the annotation process may appear to have been finessed since the annotated blocks are assumed to belong to semantic objects of the same class.

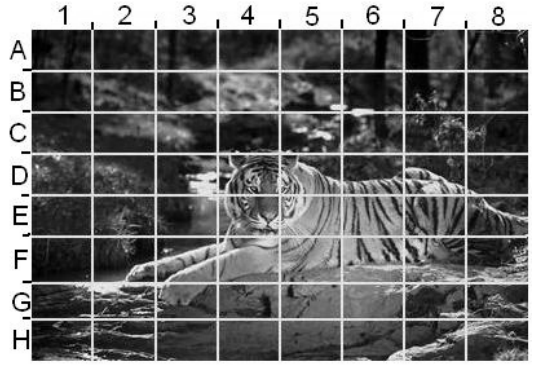

Figure 1: Elementary building elements of objects.

As shown in Figure1 splitting it into small blocks bring us closer to the objects and to some extent to the low-level descriptions than can be calculated automatically by a machine. It appears than a block based approach can be used to classify objects without relying on segmentation. Having a small but very representative set of elementary building elements for each semantic concept at hand, a suitable descriptor and metric in a multi descriptor space can be sought. It is expected that such descriptor and metric possess high discrimination power and thus can be used to automatically identify and annotate other elementary blocks belonging to semantic objects in the whole database. The strategy to build such multi-descriptor and metric is the core of this paper and will be elaborated in the next section.

2.1 Feature extraction

It is assumed that the whole database consists of elementary building elements created by simply splitting the images into 4x4 blocks of regular size and a small number of representative blocks has been manually labeled and can be used as training set. It is also assumed that conventional low-level descriptors can be extracted automatically for the whole database. For the experimental evaluation described in this paper the MPEG-7 feature extractor developed within the *aceMedia* project was used [7]. In particular the MPEG-7 *Colour Layout* (CLD), *Dominant Colour* (DCD), and *Edge Histogram* (EHD) descriptors and *Grey Level Co-occurrence Matrix* (GLCM) descriptor were used as low-level primitives. The distance between descriptors of the same class is estimated using the metric recommended by the MPEG-7 standard. A distance function or metric is

defined as $d = dist(v_1, v_2)$ and vary for different descriptors. Since in some cases this similarity measure is not a metric in mathematical sense, the term distance function is used.

2.2 Distance normalization

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 the minimum comparability requirement all distances are normalized using a simple *Min-Max Normalization*, which transforms the distance output into the range [0, 1] by applying:

$$d_{ij(new)} = (d_{ij} - \min_j)/(\max_j - \min_j) \quad (1)$$

After the pre-processing steps described in this section, a normalized matrix (2) is constructed. The search for the suitable metric combining the multiple low-level primitives will be conducted based on the matrix.

3. COMBINING DISTANCES IN A MULTIPLE FEATURE SPACE

Let be $S = \{s^{(i)} \mid i = 1,...,m\}$ the training set of elementary building elements. 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 in different descriptors form a particular vector $s^* = \{v_1^{(k1^*)}, v_2^{(k2^*)}, ..., v_n^{(kn^*)}\}$. Observe that in general s* does not necessarily represent a specific block of *S*. Taking s* as anchor, a distance matrix can be constructed according to (2):

$$d_j^i = dist(v_j^{(k^*)}, v_j^{(i)}) . \quad (2)$$

Thus, for a given concept representing an object the matrix:

$$\begin{matrix} d_1^{(1)} & d_2^{(1)} & \cdots & d_n^{(1)} \\ d_1^{(2)} & d_2^{(2)} & & d_n^{(2)} \\ \vdots & & \ddots & d_4^{(3)} \\ d_1^{(m)} & d_2^{(m)} & \cdots & d_n^{(m)} \end{matrix} \quad (3)$$

is built. Observe that in (3) each row contains distances of different descriptors estimated for the same block, while each column display distances for the same descriptor for different blocks. The most straightforward candidate of possible metrics in the multi-feature space is the linear combination of the distances defined for single descriptors:

$$M(\mathrm{A}, D) = \alpha_1 d_1 + \alpha_2 d_2 + \alpha_3 d_3 + ... \quad (4)$$

Here A is the set of weighting factors and *D* the set of distance functions for single descriptors. The problem consists of finding the optimal set of weighting factors α, where optimality is regarded in the sense of both concept representation and discrimination power.

The solution of the problem at hand is close related to decision making strategies in which simultaneous optimization of multiple objectives is sought. Contrasting single-objective optimization in which the best design solution corresponding to the optimum value of a single objective function is sought, multi-objective optimization usually involves conflicting objectives subject to a set of constraints. To illustrate the conflicting nature of the objective functions in (4), let's consider the example shown in Figure 2 where two groups of image blocks are shown. The first group contains blocks extracted from natural images of flowers with predominant red colour and is called the "Flower" group. The second group was generated synthetically using a mixture of colours and very dominant horizontal edges and is called the "Hori" group.

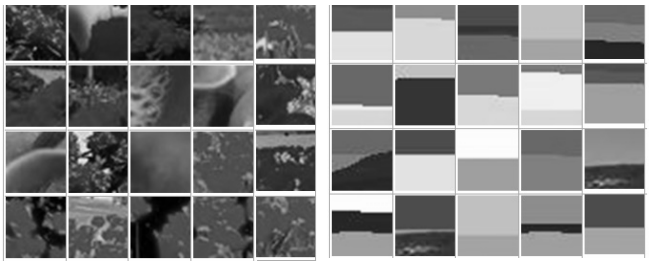

Figure 2: Selected image blocks. 'Flower' group (left) and 'Hori' group (right).

Considering the 'Flower' group and its intrinsic concept (flower), a colour descriptor will identify blocks in which the red colour is dominant. The annotation (or retrieval) process will mark predominantly red blocks in the database as "flowers". In this case the edges of the flowers, which strongly contribute to the semantic concept, will be fully ignored. On the other hand if an edge descriptor is used the result will be blocks of very varied colours featuring clear edges and not representing flowers. The obtained results are conflicting from the point of view of semantic similarity as understood by the human visual system. Moreover, in a more realistic scenario semantic concepts are complex and can be better described using a mixture of single descriptors. In the "flower" case the colour descriptor will be important but the edges will strongly contribute to a final positive classification.

To optimize (4) there is no single solution achieving optimum for all objectives at the same time. The interaction between different objectives leads to a set of compromised solutions, largely known as the *pareto-optimal solutions* or *pareto front* [8]. Since none of these pareto-optimal solutions can be identified as better than others without any further consideration, the goal is to find representative pareto-optimal solutions. Once the pareto-optimal solutions are found, a second processing step is required: a higher-level decision-making involving further considerations to choose a single solution. In this paper the *Pareto Archived Evolution Strategy* (PAES) [9] is adopted to optimize (4). The second processing step is based on finding the minimum sum of all objective solutions considering the that small sums of distances means better gathering of all points, which is just the target we are seeking to achieve. The proposed strategy to optimize (4) is outlined in the remaining of this section.

For a given semantic concept the distance matrix (3) is first estimated. The optimization of (4) is then performed by applying PAES on S according to:

- Original condition: $\alpha_i = 1/n$, where $j = 1, \ldots, n$;
- Consider: $M(\mathrm{A}, D) = \begin{cases} m_1(\mathrm{A}, D^{(1)}) \\ m_2(\mathrm{A}, D^{(2)}) \\ .... \\ m_m(\mathrm{A}, D^{(m)}) \end{cases}$

  where A is the collection of decision variables, and $D^{(i)}$ is the i_{th} distance vector;
- The optimal solution is to find the minimal value of M and its $\mathrm{A} = \{\alpha_j \mid j = 1,..., n\}$, subject to constraint $\sum_{j \in (1,n)} \alpha_j = 1$.

4. EXPERIMENTAL EVALUATION

Results obtained with the four concepts "*grass*", "*cloud*", "*building*" and "*lion*" are reported in this section. Twenty elementary building blocks were manually selected to represent each concept. The K-Nearest Neighbourhood (KNN) Search is employed as a simple binary classification to test the introduced metric. Using the distance matrix (3) defined with the 20 blocks and 4 descriptors (CLD, EHD, DCD, GLCM), multi-objective functions of 4 variables, producing 4 weights for the concepts optimization is performed. Thus, 4 weighting factors were estimated for each concept. Figure 3 shows the weights returned by the MOO algorithm.

	CLD	EHD	DCD	GLCM
grass	0.8686	0.0770	0.0592	0.0603
cloud	0.5647	0.3591	0.0612	0.0171
building	0.2524	0.4079	0.0180	0.3755
lion	0.0629	0.5944	0.0222	0.3349

Table 1: Weights obtained for the concepts grass (1), cloud (2), building (3) and lion (4).

To assess the discrimination power of the estimated distances in the underlying 4-feature space, a generalization process over a subset of the "Corel" dataset was conducted. The subset contains 450 images and was labeled manually to be used as ground truth. The distances between the centroid and all elementary building elements (blocks) for the 450 images were estimated. Element blocks within the K nearest neighborhoods are classified as relevant.

	grass	cloud	building	lion
Precision of proposed metric	**95%**	**75%**	**80%**	**40%**
Precision of CLD only	80%	45%	55%	15%
Precision of DCD only	45%	55%	25%	5%
Precision of EHD only	55%	50%	70%	20%
Precision of GLCM only	70%	45%	75%	15%

Table 2: Precision values of retrieved images of four concepts

Table 1 displays the obtained precision value for each concept using the optimized metric in multi-feature space and the original single metrics of each descriptor. As it can be observed that the results obtained from the combinational metric significantly outperform the retrieval performance using any of the single descriptors. For the concept 'lion' the precision of 40% is not good enough due to the complexity of the concept itself. But it is still much higher than the approach of using single descriptors.

Due to space limitations only numerical and graphical results of this work are presented. A more comprehensive analysis and experimental evaluation is in preparation [10].

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 Pareto Archived Evolution Strategy to optimize the metric in multi-feature space. The proposed approach has been tested for the annotation of objects in images. A more comprehensive evaluation of the proposed technique and additional improvements of the method are being undertaken and will be reported shortly [10]. Immediate work includes the combination of multi-objective optimization and relevance feed-back, 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]. S.-E Chang, 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.*

[2]. Aleksandra Mojsilovic: A computational model for color naming and describing color composition of images. *IEEE Transactions on Image Progressing 14(5): 690-699 (2005)*

[3]. 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.*

[4]. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom et al. "Query by Image and Video Content: The QBIC System". *IEEE Computer, 28(9):23--32, September 1995.*

[5]. H. Eidenberger, C. Breiteneder: "Macro-level Similarity Measurement in ViZir". *2002.*

[6]. 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*

[7]. O' Connor N., Cooke E., Le Borgne H., Blighe M., Adamek T. "The aceToolbox: Lowe-Level AudioVisual Feature Extraction for Retrieval and Classification". *Proc. of EWIMT'05, Nov. 2005.*

[8]. Steuer, R.E.: "Multiple Criteria Optimization: Theory, Computation, and Application". *New York: Wiley 1986.*

[9]. J. Knowles, D. Corne: "Approximating the Non-dominated front using the Pareto Archived Evolution Strategy". *1999.*

[10] Q. Zhang, E. Izquierdo, "A New Approach For Image Retrieval In A Multi-Feature Space", *In preparation.*