Empirical Analysis of Descriptor Spaces and Metrics for Image Classification

D. Djordjevic, E. Izquierdo

Electronic Engineering Department
Queen Mary, Mile End Road, University of London, UK
E-mail: \{divna.djordjevic, ebroul.izquierdo\}@elec.qmul.ac.uk

Abstract The objective of this work is to devise an approach for measuring similarity of low-level content features in multi-descriptor spaces. Since no single descriptor is able to represent all the properties and patterns encapsulated in natural images, the combination of several descriptors appears to be a sensible strategy to increase their discrimination power and classification properties. We are interested in combining descriptors using weights according to their discriminative properties for given semantic concepts. In an attempt to better understand the discriminatory properties of relevant descriptors a thorough empirical and statistical analysis of their basic properties is reported in this paper. The goal is to capture the behavior of descriptors by considering the particular syntax and appropriate distance metric depending on the syntax of each descriptor.

1 Introduction

During the last decade numerous image descriptors for classification, annotation and retrieval have been investigated and developed. Some of them reflect global or local pixel color and intensity distributions in images. Others capture the spatial patterns of pixels building edges and texture [1], [2] and [3]. More elaborated descriptors in computer vision are based on interest points and appropriate feature extraction from their corresponding neighborhoods [4]. A very important set of descriptors specifically designed for image annotation and retrieval was defined and standardized in MPEG-7 [1]. However, the analysis of most image descriptors defined so far, have been confined to single primitives and independently from the features and properties of other descriptors. Since no single descriptor is able to represent all the properties and patterns encapsulated in natural images, the combination of several descriptors appears to be a sensible strategy to increase their discrimination power and classification properties. Though the idea of combining descriptors has been addressed for years in of pattern recognition, most methods developed so far are limited to specific low-level vision tasks. In the more challenging context of natural image classification, the combination of image descriptors has not been well studied and the underlying problems are not well understood. Some existing approaches are based on low-level descriptors and feature distance weighting. The reasoning behind this idea is that the feature providing the most compact clustering of relevant images and separation of relevant and irrelevant image should be weighted the most [5]. Other computer vision approaches do not use global descriptors, since they are mainly looking at object recognition and detection of objects in very specific application scenarios. These approaches use affine invariant interest point detectors and descriptors invariant to geometric and illumination variations [6]. The third group of approaches avoids the problem of descriptor amalgamation by applying single descriptors separately and combining the results after parallel classifications for each descriptor. An example of such approach is given by the PicSOM system [7]. Contrary to these approaches we are interested in combining descriptors using weights according to their discriminative properties for given semantic concepts. In an attempt to better understand the discriminatory properties of relevant descriptors a thorough empirical and statistical analysis of their basic properties is reported in this paper. The goal is to capture the behavior of descriptors by considering their particular syntax and appropriate distance metric models for each particular. A clustering approach that follows these rules in features spaces is then introduced and used to estimate the performances of each descriptor. We use the discriminative property of the descriptors to combine them in a multi-descriptor framework. For a given classification scenario, the discriminative property of each descriptor is defined as the ratio between examples retrieved as truly relevant and all returned examples. The paper is organized as follows: In section II a
subjective evaluation of descriptors is presented. A classification framework with the discriminative properties of each feature descriptor is analyzed and presented in section III. In the final section conclusions and future work are outlined.

2 Evaluation of Descriptor Performance, Descriptor Syntax and Similarity Measures

In this section a simple visual inspection of the selected descriptors is conducted. Though this approach may appear simplistic, it is not. Visual inspection is probably the only weapon we have to correctly assess how low-level descriptors represent human perception of visual information. The goal is to empirically evaluate how close low-level features reflect human perception of images. Clearly, a complete link between low-level features and human perception is the breakthrough needed to bridge the semantic gap. Though this is not the objective of this work, better understanding of low-level descriptor behavior is needed as a basis to approach the more challenging problem of automatic image recognition. The difficulty of the problem arises from the generic and intrinsically different nature of the descriptors considered. They are usually formed using a number of different algorithms. Descriptors have individually specific syntax depending on the nature of the values obtained and the procedure for generating them. The underlying consequence is that different descriptors ‘live’ in completely different feature space with their own metrics and statistical behavior. Low-level descriptors try to capture the low-level commonality of image content and are used to infer semantic meanings. They represent different numerical characteristics in an effort to complement each other. For this reason, their combination would help to make a leap over the semantic gap. However, here lies the problem. Though their extraction, representation, statistical behavior and similarity measures are designed, as much as possible, to simulate human perception, they do not naturally and straightforwardly mix into a meaningful combination. Since a visual feature space described with colour, texture, edges and shape does not follow any strictly defined mathematical guidelines, statistical properties are an important source of information that help us to build up image descriptions. Conventional low-level descriptors are extracted automatically for the whole database. In the sequel, the MPEG-7 Colour Layout (CLD), Edge Histogram (EHD), Colour Structure (CSD) descriptors, are considered. Additionally texture feature based on Gabor Filters (GF) is also used in the analysis [3]. To stress invariance to saturation (shades of the same colour), Hue-Saturation-Value (HSV), the colour system based descriptor introduced in [2] is considered. In an attempt to simplify the analysis, images are split into small blocks. The motivation for this is twofold: Firstly, the better images represent low-level patterns and descriptions, the easier is to recognize the image content. If in an experimental setting the semantic gap problem is reversed and the human is forced to think as the machine, then the learning process becomes more transparent. Secondly, images are too complex to be understood at once by a computer programme. Clearly, analyzing small image building blocks would be much easier. Semantic objects and complete images can be regarded as mosaics of small building blocks. In Fig. 1 low-level descriptors for three image blocks elements of object or background are considered. As it can be seen, for different descriptors there is a degree of overlapping between the features of the object of interest and the background. The Color Layout descriptor (CLD) has an obvious pick in the plotted chrominance of the colour blue and a noticeable shift of this pick for blocks constituting the object.

From the plots of the Edge Histogram Descriptor (EHD) it becomes obvious that a common difference in distribution of edges is present and might serve as a good indication on how to differentiate edges and no edges. There are no edges present in the block presenting the clear sky, contrary to the rest of image blocks with high edge presence. Distance between descriptors of the same class is estimated using the metric recommended by the MPEG-7 standard and suggestions based on the nature of generating algorithms. Similarity measures vary in form and meaning for different descriptors, Table 1.

For this as the basic measure of similarity in the multi-feature space we use a linear combination of individual distances:

\[
D(x, y) = \sum_{j=1}^{M} \alpha_i \cdot d_i(x_{ij}, y_{ij}) | j = 1, M_i
\]
<table>
<thead>
<tr>
<th>Table 1 Similarity measure functions for visual features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Similarity measure functions</strong> ( \nu_1, \nu_2 )</td>
</tr>
<tr>
<td><strong>CLD</strong> [\text{Sum of weighted L1 distances on 3 subsets of the feature.}]</td>
</tr>
<tr>
<td><strong>CSD</strong> [\text{Normalized version of L1 distance, } d_{CSD}]</td>
</tr>
<tr>
<td><strong>HSV</strong> [\text{Histogram intersection of normalized histograms.}]</td>
</tr>
<tr>
<td><strong>EHD</strong> [\text{Sum of L1 distances over the original features, global and semi-global histograms.}]</td>
</tr>
<tr>
<td><strong>GF</strong> [\text{Sum of weighted L1 distances, } a(i)\text{ normalizing standard deviation of appropriate feature components.}]</td>
</tr>
</tbody>
</table>

\( M_i \) is the maximum number of coefficient that correspond to that particular descriptor, and \( M \) is the number of descriptors; Where \( \alpha_i \) represents functions inducing constant positive weighting factors, for each of the metric distances.

**Lemma 1** Consider a finite number of metric spaces \( (X_i, d_i) \), \( 1 \leq i \leq n \), and let set \( X \) be a Cartesian product of individual sets \( X_i \), \( \prod_{i=1}^{n} X_i \). Points in the set \( X \) are \( x = (x_1, \ldots, x_n) \) and \( y = (y_1, \ldots, y_n) \) with \( x_i, y_i \in X_i \), then a new metric on the set \( X \) is defined as in Eq.(1).

Axioms of self-identity, symmetry and triangular inequality for a linear combination of weighted distance metrics are satisfied. The expression Eq.(1) is therefore a metric and can be used to merge the descriptors together.

The problem at hand is that now when using the distance functions we obtain scalar value representing a level of similarity for that low-level primitive, but how to combine them since the values and ranges are particular for each descriptor. We approach this problem from a statistical point of view trying to estimate the distribution of distances, for a set of 12000 images, that is 774400 descriptor. We approach this problem from a statistical point of view trying to estimate the distribution of distances, for a set of 12000 images, that is 774400 descriptor. Using Monte-Carlo method we obtain three random sets each being a fraction of the total number of blocks and we try to observe properties of completely random distances and perform estimate the distributions and their averages, per descriptor Fig. 2. More then 17 million elements are used to build up each distribution of distances.

Each of the distribution can be closely fitted with a mixture of Gaussian, where \( d \) is the variable representing distance:

\[
p_j(d) = \sum_{i=1}^{N_j} \alpha_{ij} \cdot \exp(-\frac{(d - \mu_{ij})^2}{\gamma_{ij}}) \quad j = 1, M
\]

\( N_j \) is the number of used Gaussians in the mixture, \( j \) is the index of used descriptor. Hence, we obtained all parameters \( \{\alpha_{ij}, \mu_{ij}, \gamma_{ij}\} \) from the experimental plot by fitting the distribution function of distances for each descriptor. For different descriptors as CLD, CSD, EHD, HSV,GF for a close fit of curves presented in Fig. 2 we obtain different number of mixed Gaussians as \( (5, 7, 4, 3, 3) \), respectively. We use this probability fits in the next section to weight differently distances in multi-feature spaces.

**3 Discrimination Powers of Descriptors in a Classification Framework**

Understanding low-level features and how they behave, helps us to link the machine-extracted features to a generic classification approach to infer semantic meanings. A meaningful evaluation of automatic clustering of semantic blocks is a difficult and tiresome task. Even when ground truths are available the problem is complex and limited by subjectivity. In this section results of experiments conducted for three high level semantic categories are reported: ‘elephants’, ‘tigers’ and ‘horses’. Out of 6400 blocks per concept, e.g. ‘elephants’, 969 blocks were declared members of the representative set. The rest of the blocks were declared to be a part of the background set. In the same way the two remaining categories were labeled to build the ground truth data.

Particular attention is paid to the syntax inherent to all descriptors analyzed in this paper, and given in Table 1. It is worth to note that the underlying \( n \)-dimensional feature spaces and associated metrics are not always Euclidian spaces. In most cases they are not even vector spaces. This fact has important consequences in the selection of the clustering technique. For instance, the conventional k-means approach becomes meaningless since...
it is designed for Euclidian spaces. Clearly, k-means requires the estimation of an ‘average vector’. This is not possible without a suitable sum and scalar multiplication in the underlying feature space. However, K-Medoid clustering can be used in a more generic model without the restrictions imposed by techniques designed for Euclidean space. It is applicable to arbitrary objects and distance functions. It is also less sensitive to noisy data and outliers than the k-means. The drawbacks are the high computational cost and low efficiency for large databases. For this reason a variations of the basic K-median approach PAM (Partitioning Around Medoids) was used in this work [8].

In Table 2 selected results of the clustering for individual descriptors of 6400 block elements per image category are presented. The classification accuracy indicates potentially good discriminative property of object oriented visual features since some individual descriptors show a tendency to discriminate relevant blocks, better then other.

Therefore in Table 3 we have classification accuracy with k-medoids for a multi-feature space, using distance given in Eq. (1), denoted as ‘Mix’. And the accuracy for clustering with a similarity measure given with the same equation but with the weighting values obtained ahead, and denoted as ‘Mixp’.

\[
\alpha_j = 1/p_j(d)|j = 1, M
\]

For all concept the mixture of descriptors, multi-feature space performs better then individual spaces, either just as summation of normalized distances ‘Mix’, a weighed version ‘Mixω’, or including the frequency histogram of a distance ‘Mixp’. Though the values for mixing parameters are probably not optimal they suggest ways to generate a more effective mixed similarity metric in a multi-descriptor space.

### Table 2: Accuracy (%) in separation properties of clusters when K-medoids is used with specific distances appropriate to the syntax of the descriptors. For 5 individual descriptors and three concepts

<table>
<thead>
<tr>
<th></th>
<th>CLD</th>
<th>CSD</th>
<th>HSV</th>
<th>EHD</th>
<th>GF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elephant</td>
<td>59.78</td>
<td>47.4</td>
<td>46.37</td>
<td>52.34</td>
<td>55.05</td>
</tr>
<tr>
<td>Tiger</td>
<td>66.54</td>
<td>56.88</td>
<td>51.46</td>
<td>49.39</td>
<td>62.32</td>
</tr>
<tr>
<td>Horse</td>
<td>42.56</td>
<td>66.84</td>
<td>75.71</td>
<td>43.65</td>
<td>67.51</td>
</tr>
</tbody>
</table>

### Table 3: Accuracy (%) in separation properties of clusters when K-medoids is used with specific distances appropriate to the syntax of the descriptors. For three combinations of 5 descriptors with Eq.(1) and three concepts

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>Mixω</th>
<th>Mixp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elephant</td>
<td>60.44</td>
<td>55.78</td>
<td>56.92</td>
</tr>
<tr>
<td>Tiger</td>
<td>68.63</td>
<td><strong>68.58</strong></td>
<td>58.72</td>
</tr>
<tr>
<td>Horse</td>
<td>79.93</td>
<td>55.90</td>
<td><strong>84.23</strong></td>
</tr>
</tbody>
</table>

### 4 Conclusions

Underlying low-level properties of elementary building blocks for semantic concepts were investigated. A clustering approached based on descriptor specific syntax was used according to underlying properties of considered feature spaces and their specific descriptors. The goal was to investigate the possibility of descriptor combination in a more suitable metric space and to use the appropriate distances. Future work will deal with investigating better ways for optimization techniques for obtaining effective similarity measures that would correctly represent semantic concepts.

**Acknowledgement:** This research was partially supported by the European COST292 Action and the aca-Media project FP6-001765.

### References