An Unified Framework Based on $p$-Norm for Feature Aggregation in Content-Based Image Retrieval

Jun Zhang and Lei Ye
School of Computer Science and Software Engineering
University of Wollongong
Wollongong NSW 2522 Australia
{jz484, lei}@uow.edu.au

Abstract

Feature aggregation is a critical technique in content-based image retrieval systems that employ multiple visual features to characterize image content. In this paper, the $p$-norm is introduced to feature aggregation that provides a framework to unify various previous feature aggregation schemes such as linear combination, Euclidean distance, Boolean logic and decision fusion schemes in which previous schemes are instances. Some insights of the mechanism of how various aggregation schemes work are discussed through the effects of model parameters in the unified framework. Experiments show that performances vary over feature aggregation schemes that necessitate an unified framework in order to optimize the retrieval performance according to individual queries and user query concept. Revealing experimental results conducted with IAPR TC-12 ImageCLEF2006 benchmark collection that contains over 20,000 photographic images are presented and discussed.

1. Introduction

With the explosively growing amount of information made available in digital form, the information retrieval plays a more and more important role in work and daily life. Image retrieval is an important area of information retrieval. Traditional keyword-based image retrieval makes use of the annotations of images to search for images. In this paradigm, image retrieval is a form of text information retrieval. Content-based image retrieval (CBIR) addresses another problem of searching and ranking images based on their visual similarity, in many cases with a query that is expressed by an example image. The state-of-art technology is to characterize image content using visual features and the similarity is measured with the feature distances. Each feature extracted from images characterize certain aspect of image content. Multiple features are necessarily employed to provide an adequate description of image content in order for a CBIR system to retrieve relevant images. In CBIR systems using visual features, the relevance is defined as visual similarity of image content that is in turn specified by various visual features. However, it is an challenging problem to measure the image similarity from various individual feature similarities as different features are not compatible in the sense that are defined in different spaces. The distances of different feature vectors are not therefore directly comparable with each other. Research in feature aggregation is aimed to addressing this problem.

Some efforts have been reported to provide working solutions. In the context of relevance feedback, linear combination of feature distances is one of the first methods [9, 3]. To treat the feature distance array as a vector, Euclidean distance is used to measure the aggregated similarity of multiple features in [10, 1]. There are some systems such as MARS [8] and BlobWorld [2] attempting to address this problem using the Boolean logic. To overcome the limit of traditional Boolean logic, decision fusion scheme using fuzzy logic is introduced in [5]. These efforts have achieved certain success in their applications. The results reported individually are obtained in different conditions and the implication of formulae in various schemes and their mechanism are yet to be further explored.

In this paper, we introduce the $p$-norm into feature aggregation that provides a framework to unify various previous feature aggregation schemes such as linear combination, Euclidean distance, Boolean logic and decision fusion schemes in which previous schemes are instances. Some insights of the mechanism of how various aggregation schemes work are discussed through the effects of model parameters in the unified framework. Different feature aggregation schemes effectively emphasize either the difference or the commonness of features in the query image and
images in the collection, which may reflect the underlying design philosophy of various CBIR systems. The observation on our extensive experiments suggests that feature aggregation schemes need to adapt to queries to satisfy user query concept in various queries, which necessitates an unified feature aggregation framework.

In Section 2, we introduce $p$-norm into the feature aggregation and the unified framework. The mechanism of various schemes are also discussed in the context of the framework through the effects of model parameters. In Section 3, we describe our experiments and present some revealing results. We conclude with a brief discussion of our work and some future work that may be inspired from the work presented in this paper.

2. Feature Aggregation in CBIR

In this section, we will discuss the feature aggregation problem, introduce the $p$-norm as an aggregation measure and propose an unified framework based on $p$-norm. It is shown that previous feature aggregation schemes are instances of the framework.

2.1. Feature Aggregation

In CBIR systems, images are retrieved according to the relevance of content of images in an image collection and that of the query image. The content of images is characterized by visual features such as visual descriptors suggested in MPEG7 visual tools [11, 6]. The relevance of image content is in turn defined as the similarity of visual features measured by the distance of visual descriptors. In contrast to early work in CBIR that has been focused on selecting a good feature to characterize the image content, recent research recognizes that each visual feature describes one aspect of image content and multiple features are necessary to adequately characterize the content of images. Various features are extracted from the query image and their similarity measured by distances to those of images in the collection are calculated. In CBIR systems employing multiple features, the relevant images are ranked according to an aggregated similarity of multiple feature descriptors, as shown in Fig.1, where $x_i, (i = 1, 2, ..., n)$ stands for the $i^{th}$ feature distance between the query image and an image in the collection. The performance of the retrieval is largely dependant on a sensible feature aggregation scheme and how they will contribute to the process of measuring the relevance of image content for a given query. Ideally, the contribution of individual features in feature aggregation should correspond to its significance in describing the query concept of specific queries, which varies from query to query.

Previous work on feature aggregation has proposed a few schemes. In the context of relevance feedback, a linear combination of various features were used [9, 3]. The Euclidean distance is also proposed [10, 1] to measure the aggregated similarity of various features. Those two schemes treat the feature aggregation problem in the vector space. In [8, 2], the problem is formulated as a Boolean logic. Effectively, it measures the content similarity using one of the features selected by an aggregation strategy expressed with logic operations. To further extend the Boolean model, [5] introduced the decision fusion formulated based on fuzzy logic to extend AND and OR operations in Boolean logic.

2.2. Motivation of the Work

The $p$-norm is a more general measure and widely used to induce distance measures in normed spaces. We will introduce $p$-norm to CBIR as a measure for feature aggregation. Various feature aggregation schemes are evaluated in the same conditions using IAPR TC-12 ImageCLEF2006 benchmark collection that contains an adequate amount of photographic images. Experiments show that the performances of various schemes vary over query images and
scheme parameters. It is desirable that the feature aggregation scheme could be adaptive to queries to optimize the overall retrieval performance. With the introduction of $p$-norm as the aggregation measure, it is found that previous schemes such as linear combination, Euclidean distance, Boolean logic and decision fusion are special cases of the $p$-norm based aggregation, that will be discussed in the following subsection. The $p$-norm based feature aggregation can provide an unified framework.

2.3. Unified Feature Aggregation Framework

We use the T-S compensatory operation introduced in [7] to formulate the feature aggregation. T-S compensatory operation is expressed as

$$C(x_1, x_2, ... x_n) = (1 - \gamma) \cdot T(x_1, x_2, ... x_n) + \gamma \cdot S(x_1, x_2, ... x_n),$$

where $T$ and $S$ represent a t-norm and a t-conorm, respectively and $C$ is the aggregated image similarity. The t-norm operator is a generalization of the aggregation similar to logical AND in Boolean logic and t-conorm operator similar to that of the logical OR. The parameter $\gamma$ plays a role in balancing the two compensatory operations. $T$ operation emphasizes more different features in images while $S$ operation more similar features.

In this paper, we introduce $p$-norm into the $T$ and $S$ operations as

$$T_p(x_1, x_2, ... x_n) = 1 - \left( \frac{\sum_{i=1}^{n} (1 - x_i)^p}{n} \right)^{1/p}, \quad 1 \leq p \leq \infty,$$

and

$$S_p(x_1, x_2, ... x_n) = \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p}, \quad 1 \leq p \leq \infty. \quad (3)$$

With Eqs.1, 2 and 3, the feature aggregation can be formulated as

$$C_p(x_1, x_2, ... x_n) = (1 - \gamma) \left( 1 - \left( \frac{\sum_{i=1}^{n} (1 - x_i)^p}{n} \right)^{1/p} \right) + \gamma \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p}, \quad 1 \leq p \leq \infty. \quad (4)$$

Eq.4 formulates an unified feature aggregation framework in the sense that previous feature aggregation schemes become instances of this framework.

Let $p = 1$, Eq.4 becomes

$$C_p(x_1, x_2, ... x_n) = \frac{\sum_{i=1}^{n} x_i}{n}. \quad (5)$$

Eq.5 represents the feature aggregation scheme using linear combination of individual feature distances. In this case, all feature distances contribute equally to the similarity measurement of image content.

Let $p = 2$ and $\gamma = 1$, Eq.4 becomes

$$C_p(x_1, x_2, ... x_n) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}. \quad (6)$$

Eq.6 represents feature aggregation scheme using the Euclidean distance. With $\gamma = 1$, Euclidean distance measures the image similarity by the more different features in images as the result of the companding effect of the squared feature distances.

Let $p = \infty$, Eq.4 becomes

$$C_p(x_1, x_2, ... x_n) = (1 - \gamma) \cdot \min(x_1, x_2, ..., x_n) + \gamma \cdot \max(x_1, x_2, ..., x_n). \quad (7)$$

Eq.7 represents feature aggregation scheme using decision fusion introduced in [5].

Furthermore, let $\gamma = 0$ in Eq.7, it becomes

$$C_p(x_1, x_2, ... x_n) = \max(x_1, x_2, ..., x_n). \quad (8)$$

Eq.8 represents a special case of the feature aggregation using Boolean logical AND to select the most different feature as the content description for the image similarity measurement, in which case the most different feature in the image will determine the similarity of the image. Similarly, let $\gamma = 1$ in Eq.7, it becomes

$$C_p(x_1, x_2, ... x_n) = \min(x_1, x_2, ..., x_n). \quad (9)$$

Eq.9 represents a special case of the feature aggregation using Boolean logical OR to select the most similar feature as the content description for the image similarity measurement, in which case the most similar feature in images will determine the similarity of images.

For $\gamma = 0$ or 1, Eq.4 becomes

$$C_p(x_1, x_2, ... x_n) = \left( 1 - \left( \frac{\sum_{i=1}^{n} (1 - x_i)^p}{n} \right)^{1/p} \right), \quad 1 \leq p \leq \infty, \quad (10)$$

or

$$C_p(x_1, x_2, ... x_n) = \left( \frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{1/p}, \quad 1 \leq p \leq \infty. \quad (11)$$

They are two forms of extended Boolean schemes.

For any $0 < \gamma < 1$, it reflects a balancing strategy in the system to emphasize more on either the difference or commonness of features in images. In the case of a larger $\gamma$ that emphasizes the difference, the CBIR system will look
Figure 2: Relationships of various feature aggregation schemes in the unified feature aggregation framework

for images that have all features of the images similar to all features of the query image while in the case of a smaller \( \gamma \) that emphasizes the commonness, the CBIR system will look for images that have at least some of the features of the images similar to some of the features of the query image.

For all \( p \neq \infty \), all features contribute to the similarity measurement of image content and their significance of differences is determined by the order of \( p \)-norm.

Fig. 2 summarizes the relationships of various feature aggregation schemes in the unified feature aggregation framework.

3. Experimental Results

In Section 2, we introduced the \( p \)-norm to feature aggregation that provides an unified framework. In this section, we will present experimental results of a comparative study on various feature aggregation schemes under the unified framework. Previous schemes were reported with different implementations and experiments were conducted in different image collections with different queries. This study provides a fair evaluation of various schemes under the same conditions that provides a basis for selecting feature aggregation schemes, which is useful for scheme adaptation. The effects of \( p \) and \( \gamma \) on the retrieval performance are investigated as well.

3.1. The System

An experiment system is implemented to evaluate the performance of various feature aggregation schemes. The following steps are executed in the system.

**Step 1:** Extract the features of query image in real time.

**Step 2:** Compute the distances between query image and database image based on features using the functions recommended by MPEG-7.

**Step 3:** The feature similarity is defined as the feature distance normalized by an exponential membership function given in [5] as

\[
s_{ij} = e^{-\frac{d_{ij}}{\alpha_j}},
\]

where \( d_{ij} \) is the distance between \( j \)-th feature of the query image and \( j \)-th feature of the \( i \)-th image in the collection, and \( \alpha_j \) is a normalization factor. \( \alpha_j \) is calculated as

\[
\alpha_j = \frac{\text{median}(D_j) - \log(0.5)}{D_j},
\]

where \( D_j = \{d_{1j}, d_{2j}, ..., d_{mj}\} \) is the distance array of the \( j \)-th feature between query image and \( m \) images in the collection.

**Step 4:** With the notation of feature similarity in Eq. 12, the aggregated image similarity \( S_i \), Eq. 4, of the query image and the \( i \)-th image in the collection is reformed as

\[
S_i = C(s_{i1}, s_{i2}, s_{i3}, s_{i4}) = (1 - \gamma) \left( 1 - \left( \frac{\sum_{k=1}^{4} (1-s_{ik})^p}{4} \right)^{1/p} \right) + \gamma \left( \frac{\sum_{k=1}^{4} s_{ik}^p}{4} \right)^{1/p}, 1 \leq p \leq \infty.
\]

Four standardized MPEG-7 visual descriptors [11] are used in the system including the Dominant Color Descriptor (DCD), Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD) and the Homogeneous Texture Descriptor (HTD).

3.2. The Experiments

The IAPR TC-12 benchmark image collection (ImageCLEF2006) [4] is used in the experiments. It contains over 20,000 photographic images. We examined the queries and their ground truth sets defined in the CLEF Cross-language Image Track 2006 and they are deemed not suitable for use directly in our experiments as they are defined for combined keyword and content-based retrieval systems. To evaluate content-based retrieval only, we selected one example image from each query set and adapted the corresponding ground truth set based on visual similarity and ignored the text annotations of all queries and image annotations in the
collection. This resulted in 20 queries and their corresponding ground truth sets. Each ground truth set consists of about 40 ground truth images.

Experiments are conducted with varying \( p \) and \( \gamma \). As discussed in Section 2, typical \( p \) and \( \gamma \) are set to instantiate the previous feature aggregation schemes along with general values of \( p \) and \( \gamma \). \( p = 1 \) and \( p = \infty \) are set for linear combination (Linear) and decision fusion schemes (Fusion), respectively, including \( \gamma = 0 \) or 1 for basic Boolean logic schemes; \( p = 2 \) and \( \gamma = 1 \) are set for Euclidean distance scheme (Euclidean). \( p \)-norm schemes (\( p \)-norm) of \( p = 3, 4, 5 \) with \( \gamma \) from 0 to 1 in a 0.1 increment are evaluated as well.

Average precision-recall over 20 queries is used to measure the retrieval performance, as defined as

\[
\text{precision} = \frac{FG(k)}{k}, \tag{15}
\]

and

\[
\text{recall} = \frac{FG(k)}{NG}, \tag{16}
\]

where \( k \) is the number of retrieved images, \( FG(k) \) is the number of matches after \( k \) image retrieved and \( NG \) is the number of ground truth images.

3.3. The Results

In this subsection, we report and discuss the revealing results of typical schemes in the unified framework based on \( p \)-norm.

Fig.3 depicts the results of typical \( p \) and \( \gamma \) that represent previous proposed schemes. It shows that the linear combination and Euclidean schemes have a similar performance and both of them are superior to the decision fusion scheme with an advantage of about 10 percent. It is worth while to point out that in the decision fusion scheme, effectively only the most different and similar features, which change from query to query, have contributed to the image similarity measurement as a result of the minimal and maximal aggregation. The schemes that take into account of all features in image similarity measurement perform better.

Experiments show that the value of \( \gamma \) has different effects on schemes of different values of \( p \). Figs.4, 5 and 6 depict the results of various schemes of different values of \( p \) over the range of \( \gamma \), respectively. For \( p = 2 \), there are limited effects of \( \gamma \) on the performance while for \( p = 5 \) and \( p = \infty \), the effects of \( \gamma \) on the performance are dramatic. When \( \text{recall} < 0.5 \), the performances are diverse while \( \text{recall} > 0.5 \), they converge. The average performances could be as different as about 15 percent and the performance difference at certain recall rate could be as high as about 40 percent. The \( \gamma \) dramatically affects the performance at lower recall rates therefore they are most obvious.
to the users as these images are highly ranked. It is interesting to note that the performances of schemes with larger values of \( p \) can approach each other with tuned values of \( \gamma \) as shown in Fig.7 where \( [p = 2 \& \gamma = 1], [p = 3 \& \gamma = 0.8], [p = 4 \& \gamma = 0.7] \) and \( [p = 5 \& \gamma = 0.8] \).

To observe the difference of performances manifested in the ranked retrieval results, we present some retrieval results. Fig.8 are top ranked 10 retrieved images for the linear combination and decision fusion schemes for the query named "Scenes of Footballers in Action". The first image at the top-left position is the query image. There are 4 irrelevant images appearing in the top 10 retrieved images in the decision fusion scheme.

With tuned values of \( \gamma \), the effects of \( \gamma \) can alter the performance of the schemes. Figs.9 and 10 are the top ranked 10 images for the linear combination for the queries named "People on Surfboards" and "Drawing in Rock", with comparisons to the decision fusion scheme and \( p \)-norm scheme when \( p = 3 \), respectively.
4. Conclusions and Future Work

The feature aggregation in content-based image retrieval using multiple visual features is a challenging problem as various feature distances are not directly comparable with each other. Previous work treated this problem using either a vector model or a logic model. In this paper, we introduced the $p$-norm for the feature aggregation and proposed an unified framework in which previous schemes are instances of the framework. Some insights of the mechanism of how various aggregation schemes work are discussed through the effects of model parameters $p$ and $\gamma$ in the unified framework. Extensive experiments were performed to evaluate various schemes under the same conditions with IAPR TC-12 benchmark image collection (ImageCLEF2006) that contains an adequate amount of photographic images along with its defined challenging queries. Revealing experimental results were presented and discussed in this paper. The proposed framework is not feature-aware so that it is general enough to apply to any possible features and to be built into layers of any content-based image retrieval systems. Experiments showed that it is necessary to adapt the feature aggregation schemes to the queries to achieve a better performance. By unifying various feature aggregation schemes, the framework provides a foundation for such work.

As discussed in the paper, different feature aggregation schemes effectively emphasize either the difference or the commonness of features in the query image and images in the collection. Therefore their performance is largely dependent on the query and query concepts of the users or what the users expect the system to look for. The future research in content-based image retrieval should address the problems of how to analyse and map the user query concept to the feature aggregation requirements and how to adapt the feature aggregation scheme, or model parameters in the proposed unified framework, to individual queries.

References


