Series Feature Aggregation for Content-Based Image Retrieval

Jun Zhang and Lei Ye
School of Computer Science and Software Engineering,
University of Wollongong,
Wollongong NSW 2522 Australia

***Abstract*—Feature aggregation is a critical technique in content-based image retrieval systems that employ multiple visual features to characterize image content. One problem in feature aggregation is that image similarity in different feature spaces can not be directly comparable with each other. To address this problem, a new feature aggregation approach, series feature aggregation (SFA), is proposed in this paper. In contrast to merging incomparable feature distances in different feature spaces to get aggregated image similarity in the conventional feature aggregation approach, the series feature aggregation directly deal with images in each feature space to avoid comparing different feature distances. SFA is effectively filtering out irrelevant images using individual features in each stage and the remaining images are images that collectively described by all features. Experiments, conducted with IAPR TC-12 benchmark image collection (ImageCLEF2006) that contains over 20,000 photographic images and defined queries, have shown that SFA can outperform the parallel feature aggregation and linear distance combination schemes. Furthermore, SFA is able to retrieve more relevant images in top ranked outputs that brings better user experience in finding more relevant images quickly.**

I. 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 characterizes 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 [1], [2]. To treat the feature distance array as a vector, Euclidean distance is used to measure the aggregated similarity of multiple features in [3], [4]. There are some systems such as MARS [5] and BlobWorld [6] 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 [7]. These efforts have achieved certain success in their applications. However, the problem of how to measure the relevance of images using visual features is yet to be answered. The mechanism of how multiple individual visual features describe collectively the image content is still to be understood.

In the prior work, individual features are extracted independently from images and feature aggregation methods take into consideration of each feature by formulating the aggregated similarity as a combination of individual features in parallel. In other words, they are applied to rank the images at the same time.

In this paper, we propose a new feature aggregation approach, Series Feature Aggregation (SFA). SFA does not need to compare or aggregate distances from different feature spaces. SFA selects relevant images using features one by one in series from images highly ranked by the previous feature. Images are filtered out by each feature that does not describe the image content well. The remaining images are collectively well described by all features.

In Section II, we discuss the structure of feature aggregation. In Section III, we describe our experiments and present some revealing experimental 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.

II. Series Feature Aggregation

In this section, we will discuss the feature aggregation problem and propose a new approach, series feature aggregation. It is shown that SFA can avoid the difficult in merging different feature distances in different feature spaces that, in principle, are not comparable and their summation does not make any sense in describing image content.

A. 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 [8], [9]. The relevance of image content in CBIR systems in the Query-by-Example (QBE) paradigm 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 dependent on a sensible feature aggregation scheme as different features are not directly comparable with pure quantity of them as different features describe different aspects of the image content. For instance, a colour feature distance of 0.5 does not convey a message of any equivalent significance of a texture feature distance of the same value in describing image content. A feature aggregation scheme is to effectively and quantitively determine which aspects 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 some schemes. In the context of relevance feedback, a linear combination of various features were used [1], [2]. The Euclidean distance is also proposed [3], [4] to measure the aggregated similarity of various features. Those two schemes treat the feature aggregation problem in the vector space. In [5], [6], 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, [7] introduced the decision fusion formulated based on fuzzy logic to extend AND and OR operations in Boolean logic. In all above schemes, individual features are aggregated in parallel into one overall distance that is used to rank the final retrieved images.

B. Motivation of the Work

The assumption of conventional feature aggregation methods is that normalized feature distances can be comparable to each other so that the image similarity could be obtained through combining different feature distances into one total distance. Generally, this assumption does not carry any intuitive meanings in visual image similarity. The motivation of our work is to propose a new feature aggregation approach that avoids to combine different visual features from visually unrelated spaces.

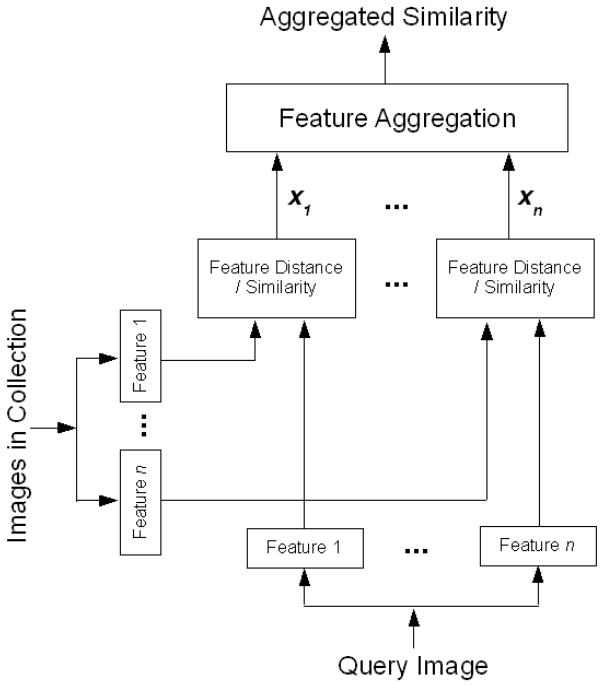


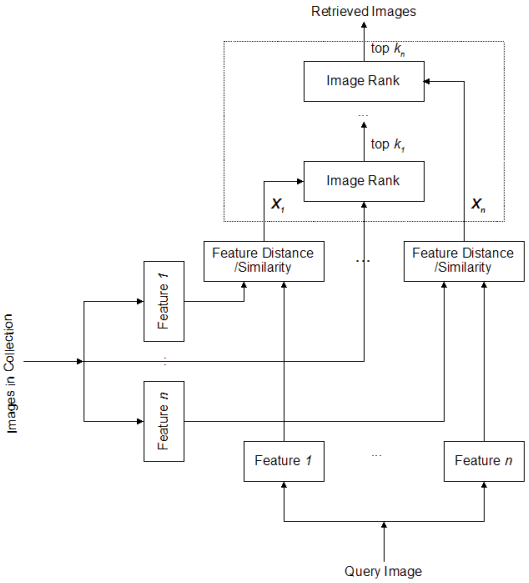
Fig. 1: Feature aggregation in CBIR

We treat the image retrieval problem as a process of selecting relevant images from the image collection based on their relevance to each individual features. Top ranked images using one feature in the collection are selected and form a sub-collection in which images are to be selected using another feature. Effectively, this process filters out irrelevant images using individual features in series stages and the resultant images are relevant to all features. The relevance of images to a feature is measured by the distance in its feature space. In practice, the distance in a feature space is defined to reflect the visual similarity measured by that feature and a shorter distance means, for a good visual feature, more similarity between two images in respect of that feature.

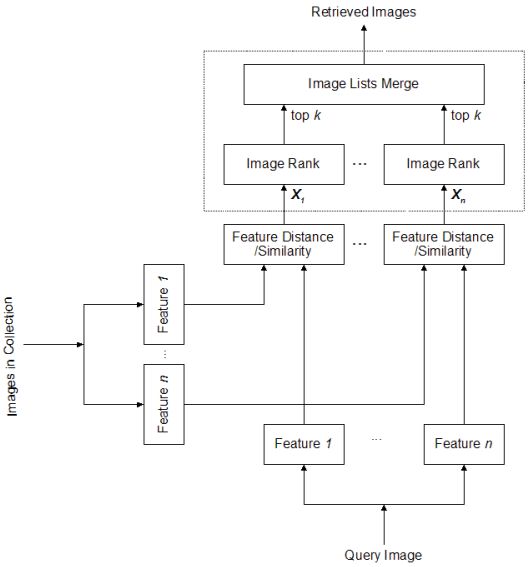
C. Series Feature Aggregation

There are basically two structures in feature aggregation that differ in the way how individual features are used to measure the aggregated image similarity. In accordance of the order of features used to measure the visual similarities, they are series and parallel feature aggregation, as depicted in Fig.2. Parallel feature aggregation has been used in various names such as fusion or merging of multiple streams. Series feature aggregation is a new approach proposed in this paper. Considering that different feature distances can not be directly comparable to each other, SFA does not merge different features or compare distances of different features.

Fig.2(a) depicts the structure of series feature aggregation. The top k_i images ranked by a feature in ith stage form the



(a) Series Feature Aggregation



(b) Parallel Feature Aggregation

Fig. 2: Structures of Feature Aggregation

sub-collection of images for $(i+1)$th stage. The final retrieval result is obtained with n stages where n is the number of features used to describe the image content. There are two key factors in SFA. One is the order of the application of features and the other is the numbers of images, $k_i (i = 1, 2, ..., n)$, retained in each stage. Ideally, the order of features applied for retrieval should correspond to their capabilities to describe the query concept, which varies from query to query. If k_i increases, more images that are less relevant to a specific feature are retained and used as candidates in the next stage, the recall may increase and the precision may decrease and vice versa.

As a comparison, Fig.2(b) depicts the structure of parallel feature aggregation. The final retrieval result is obtained by merging multiple sorted image lists. The top k images ranked by each feature are merged into one list as the retrieval result. Assume that n features are used in the system, there will be n sorted image lists.

In both series and parallel feature aggregation approaches, the operation of feature distances normalization and the operation of feature distance combination are not needed.

III. Experimental Results

In Section II, we proposed a new feature aggregation approach. In this section, we will present experimental results of a comparative study on various feature aggregation schemes.

A. The System

An experiment system is implemented to evaluate the performance of SFA with comparisons to various feature aggregation schemes. For parallel feature aggregation, the following steps are executed in the system.

Parallel Feature Aggregation:

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: Images in collection are ranked according to different feature distances respectively. System returns n image lists, where n equals the number of features applied in system.

Step 4: Top k images in every list will be merged to obtain final retrieval result and display.

The mid-rank strategy is applied for merging top k images, which is to rank images using the sum of their ranks in n lists. If one image does not exist in top k of a special list, its rank in this list will be set to $2.5k$.

For SFA, the Step 1 and Step 2 are the same as above, but Step 3 and Step 4 are different.

Series Feature Aggregation:

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: If the first feature is considered, all images in collection are ranked based on the first feature distance and top k_1 images in the ranked list will be returned. Else if the $(i+1)$th feature is considered, the k_i images returned by last iteration will be ranked according to the $(i+1)$th feature distance and top k_{i+1} images in the ranked list will be returned.

Step 4: If all features have been considered, then system display k_n images. Else, consider the next feature and return to Step 3.

Three standardized MPEG-7 visual descriptors [8] are used in the system including the Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD) and the Homogeneous Texture Descriptor (HTD).

B. The Experiments

The IAPR TC-12 benchmark image collection (ImageCLEF2006) [10] 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.

To evaluate the performance of SFA, parallel feature aggregation and linear combination of feature distances are implemented as reference schemes.

The first set of experiments is designed for SFA. In this scheme, feature order and k_i are key parameters. Experiments for the parallel feature aggregation are designed and the tuned configurations that perform well are found, which are conducted with variable k. k determines how many images in every ranked list are used for the following merging operation. As discussion in Section II-C, the choice of k can affect the precision and recall of the final retrieval results. The linear combination scheme of feature distances [1], [2] is implemented with unbiased weighting on all features.

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

$$precision = \frac{FG(k)}{k}, \tag{1}$$

and

$$recall = \frac{FG(k)}{NG}, \tag{2}$$

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.

C. The Results

Table. I presents the performance of the parallel feature aggregation scheme with different k. The results of eight different k are presented that show the effect of k to the retrieval performance. N is the number of images in the collection, where $N = 20000$ in our experiments (the same in all experiments as presented in this paper). To compare the performances of different schemes, the result of linear scheme is also provided in the table.

The observation of experiments result reveals that the performance of the parallel feature aggregation scheme is not inferior to that of linear combination scheme. The choice of k can slightly affect the performance of this scheme. When $recall < 0.3$, the performances of the parallel feature aggregation scheme with different k are diverse while $recall > 0.3$, they converge. The average performance of the linear combination schemes are about 5 to 10 percent better than the parallel feature aggregation scheme.

TABLE I: The performance of the parallel feature aggregation scheme with different k

Precision	Recall 0.1	Recall 0.2	Recall 0.3	Recall 0.4	Recall 0.5
Linear	0.63	0.45	0.32	0.25	0.19
$k = 0.01N$	0.53	0.33	0.24	0.21	0.16
$k = 0.02N$	0.57	0.33	0.23	0.17	0.14
$k = 0.03N$	0.63	0.37	0.27	0.20	0.14
$k = 0.04N$	0.62	0.41	0.27	0.20	0.15
$k = 0.05N$	0.60	0.42	0.26	0.20	0.15
$k = 0.10N$	0.62	0.42	0.28	0.21	0.14
$k = 0.25N$	0.62	0.41	0.30	0.21	0.15
$k = 0.50N$	0.62	0.41	0.30	0.22	0.16

TABLE II: The optimal retrieval parameters for different queries

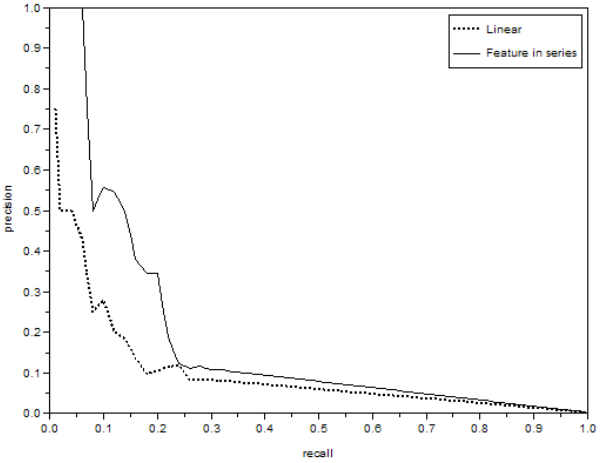
Parameters	Feature order	k_1	k_2	k_3
Query 1	HTD-CLD-EHD	0.050N	0.015N	0.001N
Query 2	CLD-HTD-EHD	0.020N	0.015N	0.001N
Query 3	HTD-EHD-CLD	0.500N	0.100N	0.001N

Experiments show that the orders of individual features in SFA are critical to the performance and different k_i have effects on optimal performance as well. Fig.3 shows examples of the retrieval performances of SFA for three different queries. The parameters for the queries in Fig.3 are listed in Table. II. For comparison, the performances of linear combination scheme are also plotted in the figures. It shows that the SFA can outperform the linear combination scheme. The SFA outperforms the linear combination scheme about 15 to 40 percent when $recall < 0.4$ and the performances converge after $recall > 0.4$. This pattern of performance improvement is significant in applications as more relevant image are highly ranked in SFA that brings better user experience in finding more relevant images quickly.

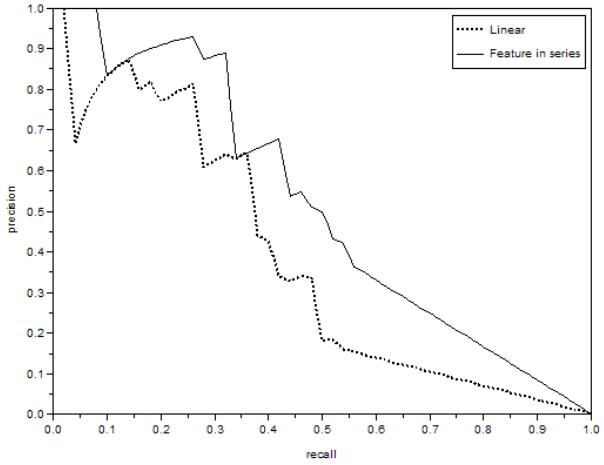
To observe the difference of performances manifested in the ranked retrieval results, we present some image retrieval results. Figs.4 to 6 are 10 top ranked images from SFA and the linear combination schemes for the three queries, named "Group people before mountain", "Scenes of Footballers in Action" and "People on Surfboards" in the IAPR TC-12 benchmark image collection (ImageCLEF2006) [10]. The first image at the top-left in these figures is the query image. In all the results, SFA is able to retrieve more relevant images from the collection. Relevant images are defined in the corresponding ground truth sets.

IV. Conclusions

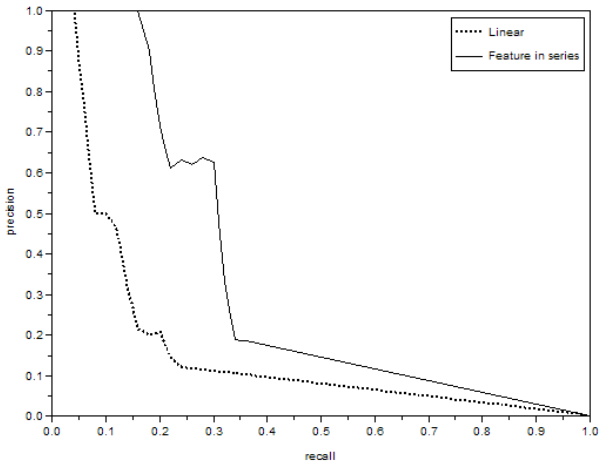
The feature aggregation in content-based image retrieval using multiple visual features is a challenging problem as various feature distances are not directly comparable with



(a) Performance with Query 1:“Group people before mountain”



(b) Performance with Query 2:“Scenes of Footballers in Action”



(c) Performance with Query 3:“People on Surfboards”

Fig. 3: Comparisons between FSA and the linear combination scheme

each other. Previous work treated this problem using either a vector model or a logic model. In this paper, we proposed a new feature aggregation approach, series feature aggregation. The proposed approach does not merge incomparable feature distances in different feature spaces and avoids the problem that conventional feature aggregation methods suffered from. 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. Experiments have shown that SFA can outperform the parallel feature aggregation and linear distance combination schemes. Furthermore, SFA is able to retrieve more relevant images in top ranked outputs that brings better user experience in finding more relevant images quickly. SFA is effectively filtering out irrelevant images using individual features in each stage and the remaining images are images that collectively described by all features.

REFERENCES

[1] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: A power tool for interactive content-based image retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep 1998.

[2] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, “The bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments,” *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan 2000.

[3] J. Shih and L. Chen, “A context-based approach for color image retrieval,” *Int. J. Pattern Recognit. Artif. Intell*, vol. 16, no. 2, pp. 239–255, 2002.

[4] S. Aksoy, R. Haralick, F. Cheikh, and M. Gabbouj, “A weighted distance approach to relevance feedback,” in *Proceedings of 15th International Conference on Pattern Recognition*, vol. 4, 2000, pp. 870–876.

[5] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. Huang, “Supporting ranked boolean similarity queries in mars,” *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 909–925, 1998.

[6] C. Carson, S. Belongie, H. Greenspan, and J. Malik, “Blobworld: image segmentation using expectation-maximization and its applications to image querying,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, 2002.

[7] A. Kushki, P. Androutsos, and K. N. P. A. N. Venetsanopoulos, “Retrieval of images from artistic repositories using a decision fusion framework,” *IEEE Trans. Image Process.*, vol. 13, no. 3, pp. 277–292, Mar 2004.

[8] T. Sikora, “The mpeg-7 visual standard for content description-an overview,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, Jun 2001.

[9] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, June 2001.

[10] M. Grubinger, P. Clough, H. Mller, and T. Deselaers, “The iapr tc-12 benchmark: A new evaluation resource for visual information systems,” in *Proceedings of International Workshop OntoImage2006 Language Resources for Content-Based Image Retrieval, held in conjuction with LREC’06*, Genoa, Italy, 22 May 2006, pp. 13–23.

(a) Linear Combination Scheme

(b) SFA

Fig. 4: Retrieval results for the query “Group people before mountain”

(a) Linear Combination Scheme

(b) SFA

Fig. 5: Retrieval results for the query “Scenes of Footballers in Action”

(a) Linear Combination Scheme

(b) SFA

Fig. 6: Retrieval results for the query “People on Surfboards”