Content-based image collection summarization and comparison using self-organizing maps

Da Deng *
Department of Information Science, University of Otago
P.O. Box 56, Dunedin, New Zealand

December 5, 2005

Abstract

Progresses made on content-based image retrieval has reactivated the research on image analysis and similarity-based approaches have been investigated to assess the similarity between images. In this paper, the content-based approach is extended towards the problem of image collection summarization and comparison. For these purposes we propose to carry out clustering analysis on visual features using self-organizing maps, and then evaluate their similarity using a few dissimilarity measures implemented on the feature maps. The effectiveness of these dissimilarity measures is then examined with an empirical study.

1 Introduction

Over the last decade, content-based image retrieval (CBIR) has become a popular research topic, much owing to the ever increasing use of multimedia on the Internet, as well as in personal entertainment, education, and mobile communications. Aimed at effective multimedia asset management and efficient information retrieval, a typical CBIR system (e.g. [1], [2]) works basically on low-level visual features such as color, texture, shape or regions. Despite some breakthroughs made in the field, it is generally understood that the problem is still far from being solved [3]. Due to obstacles in object recognition and image understanding in general, it is difficult to overcome the so-called 'semantic gap' and establish the linkage from low-level visual features to high-level concepts that correspond to objects and their semantic content within the image. Consequently, none of commercial image search engines has provided CBIR utilities so far.

Nevertheless, interesting progresses have been achieved with the research on CBIR and relevant fields. For instance, the validation of various visual

*The author thanks the support of School of Business Grant, University of Otago.

schemes carried out in the MPEG-7 core experiments has provided a group of robust visual features that are promising for semantic matching [4][5]. On the other hand, some effective mechanisms have been explored to compare the similarity of these visual features. In [6], a number of tests have been proposed for histogram analysis and examined in comparison along with L1 and Euclidean norms. More dissimilarity measures have been introduced for the comparison of feature distributions of color images [7][8]. Other techniques such as relevance feedback and the use of joint features have been worked out to complement the weakness of the CBIR approach. Even though similarity metrics defined on these low level visual features may not directly reflect the high-level semantic similarity between images and objects, CBIR still provides a valuable interface for multimedia asset management and semantic analysis.

In CBIR applications, usually one is interested in searching out interested patterns or images. There are however circumstances where there is a need to compare different image collections - e.g. images stored in folders named 'family photos' under different paths of a file system, or two collections of images featuring architectures in two countries. In a wider context, people may wish to be able to summarize their MP3 collections stored in a computer network and compare with other people's collections of similar or different music taste. Hence it is desirable to have a multimedia collection management tool that is equipped with such kind of capabilities.

In particular we would like to consider the following two questions in this paper:

- Can we generate some visual summary or *profiles* for image collections?
- Can these profiles be quantitatively compared? In other words, can we define some dissimilarity measures on them to assess the dissimilarity of the original collections?

We propose to extend the CBIR approach in order to tackle the problems of collection summarization and comparison outlined as above. A neural network model is employed for self-organized summarization of image collections. The neural structure is used as a profile of the image collection and provides a graphical interface for collection navigation and visual assessment of dissimilarity between image collections. To enable quantitative assessment, a number of distance measures are implemented and are found to perform effectively in an empirical study.

The rest of this paper is organized as follows. In Section 2 we present an introduction to the computational model, including the summarization of collections via self-organizing feature maps, and different dissimilarity measures defined on the feature maps. In Section 3 empirical results obtained from four image collections and some mixed sets are presented to verify the computational model, and the performance of different dissimilarity measures are discussed. Finally the paper is concluded along with a discussion on some future works in Section 4.

2 The computational model

2.1 Self-organizing maps

Our approach of image collection summarization is based on extracting representative prototypes from the CBIR feature space generated from the image collections. Therefore, a number of clustering or vector quantization algorithms can be employed for this purpose. Among numerous clustering algorithms, one neural network model of particular interest is Kohonen's Self-Organized Map (SOM) [9]. It has been applied widely in a number of information retrieval systems, such as in SOMLib [10] and PicSOM [11].

The SOM features in carrying out vector quantization and multi-dimensional scaling at the same time. The map, usually set in 2-D or 3-D topology, consists of a regular lattice of neurons set in hexagonal or rectangular topology. Each neuron is associated with a weight vector. The map attempts to perform localized clustering on these node vectors, while in the meantime the ordering of nodes on the lattice works to allow similar inputs to be matched to the same node or nodes close to each other, and dissimilar inputs onto nodes far from each other. The nodes are sometimes also called 'units', and unit vectors 'prototypes'.

Assume we have a N-prototype SOM to train. Denote $\mathbf{w}_i(t)$ as the weight vector associated with node i. Given an input $\mathbf{x}(t)$, the algorithm first finds the best-matching unit (BMU) $\mathbf{w}_b$ among all prototypes, i.e.

$$b = \arg_{\min} \|\mathbf{x}(t) - \mathbf{w}_i(t)\|, \ i = 1, ..., N. \tag{1}$$

The weight vectors are then updated according to the following learning rule:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \gamma(t) h_{b,i}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)]. \tag{2}$$

where $h_{b,i}$ is a neighborhood function centered at BMU and shrinking over time, and $\gamma(t)$ the learning rate decreasing over time. There have been quite some variants proposed to this original learning rule, but generally it has been shown that these learning rules lead to the convergence of weight vectors of very good quantization quality.

2.2 Collection summarization by SOMs

The SOM has a number of traits that make it a suitable choice for collection summarization. With the map nodes located on a low-dimensional lattice, it is easy to be visualized and interpreted. SOM also displays good topology preserving capability. Similar inputs are mapped onto the same node or nodes in a neighborhood on the map. This means that similar images can be closely mapped onto the grid, obviously an advantage for the visualization of the collection contents. Hierarchical design of maps can be used to leverage navigation. Another additional feature of the SOM goes to density matching ([12], page 460-461). It represents a cluster of more frequently occurring input stimuli by a

larger area in the feature map. If we denote the number of nodes in a small volume $d\mathbf{x}$ over the input space X as $m(\mathbf{x})$, it is proved that the one-dimensional SOM achieves $m(\mathbf{x}) \propto p_{\mathbf{x}}^{2/3}(\mathbf{x})$, where $p_{\mathbf{x}}(\mathbf{x})$ is the probability density function of the input $\mathbf{x}$ [13]. It also has been shown in [14] that the SOM can be regarded as a simplified Gaussian mixture estimator using a homoscedastic Gaussian mixture model. Trained with a large amount of input data, the feature map will form micro-clusters that can be treated as multivariate normal distributions. This gives the plausibility of defining appropriate dissimilarity measure between protocol vectors.

To extend the use of the SOM for summarization, a straightforward approach is to train a feature map on visual features so as to construct a profile of the whole collection. Further operations such as browsing, search, and comparison can be carried out efficiently using this profile. For a user to quickly assess and compare profiles of different image collections, it is preferable to adopt flat structure for the feature maps. By matching feature vectors of high dimensionality onto prototypes organized on a low-dimensional grid, SOM is by itself a multi-dimension scaling method. However node distance on the grid does not reflect faithfully the distance between prototypes in the high-dimensional space. It is then required to project high-dimensional prototypes onto optimal positions on the low-dimensional, usually 2-D, display space, using linear transform such as principal component analysis (PCA), or nonlinear projection algorithm such as Sammon's mapping [15].

Before considering the comparison of these feature maps or profiles, one has to ensure that the self-organizing maps generated for this purpose are robust and stable. The outcome of the SOM algorithm itself is subject to variation on random initialization, different parameter settings of the learning rate and map size etc., and the presentation order of data samples during training. Moreover, multi-dimension scaling techniques such as Sammon's mapping makes use of random initialization and gradient descent techniques that give no unique result. All these loom an undesirable effect for image collection profiling, as a user would expect a stable profile each time when visualizing an image collection. A remedy for this is to use linear initialization of the network weights using two-dimensional PCA. This not only stabilizes the feature maps generated, but also leads to maps of better quality, with little folding effect. Sammon's mapping can also be stabilized by using PCA-based initialization.

2.3 Dissimilarity measures between feature maps

Although the SOM algorithm has been used widely for data analysis in many disciplines, the problem of comparing two different feature maps has received little treatment in the literature. We could adopt a simple visual approach, projecting all feature maps in a graph of low dimensionality, for instance using PCA. While this may facilitate the visual exploration of multimedia collections, it gives little quantitative information about their dissimilarity.

In [16] a dissimilarity measure is proposed based on the evaluation of the *goodness* of the feature maps by comparing the shortest path on the maps when

matching a given pair of data samples. To calculate the distance measure all pairs of data samples need to be matched onto the feature maps in comparison, which can be rather time-consuming for large data sets. This method was used for comparison of word category maps generated by the SOM in [17]. When dealing with high dimensional feature maps generated from a large volume of multimedia collections, the efficiency of such an approach will however be in question, as the plausibility of retaining the large training data set for map assessment can hardly be assumed. We suggest, given that those feature maps have formed good representation of the original data collections, a comparison process directly based on the map prototypes rather than on the original data sets would be much more efficient.

If we regard the SOM simply as a clustering algorithm that extracts a few prototypes from the overall feature data set, it is straightforward to employ a few distance metrics defined on point sets. By considering its modeling capability, more dissimilarity measures for probabilistic distributions can be explored. We next present a few distance measures extended for the use on self-organizing maps.

2.3.1 Hausdorff distance

Given two point sets X and Y, the Hausdorff distance from X to Y is defined by

$$h(X,Y) = \sup_{x\in X}\inf_{y\in Y} d(x,y), \tag{3}$$

where d is a L_p metric, where usually the Euclidean distance is used. The Hausdorff metric is define as

$$\mathrm{HD}(X,Y) = \max\{h(X,Y), h(Y,X)\}. \tag{4}$$

It is found that the Hausdorff distance satisfies triangular inequality but is very sensitive to outliers in the point sets. Some modification can be done, for example, by generalizing the maximum operator with a quantile or a median. The Hausdorff distance has been applied in fractal image compression, shape matching and object detection etc.

2.3.2 Earth Mover's Distance

The EMD [7] is defined over weighted point sets. Suppose each point set is configured by a normalized weight set. We denote a point set as $A = \{a_1, a_2, ..., a_m\}$, with $a_i = \{(x_i, w_i)\}$, $x_i \in R^k$, and $w_i \in R^+ \cup \{0\}$. The EMD calculates the minimum amount of work needed to transform one configuration to another by moving weight under constraints. Denote the set of all feasible flows as $F = \{f_{ij}\}$, where i is a point label for set A, and j for B. These flows are subject to the following constraints:

1. $f_{ij} \geq 0, i = 1, ..., m, j = 1, ..., n$
2. $\sum_{j=1}^{n} f_{ij} \leq w_i, i = 1, ..., m$

3. $\sum_{i=1}^{m} f_{ij} \leq u_j, j = 1, ..., n$

4. $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(W, U)$

Here W and U are the total weights of A and B respectively. These constraints ensure, for instance, each flow of weight is non-negative; a point at the 'sender' can not send more weight than it holds, and a point at the 'receiver' does not receive more weight than it needs.

The EMD between the two point sets can then be define as

$$\text{EMD}(A, B) = \min_{f \in F} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}. \tag{5}$$

EMD has been applied in image retrieval for similarity comparison of global color features, texture features, and shapes.

To make EMD eligible for SOM comparison, weights need to be assigned to the prototype nodes. An easy solution is to map the original data set onto the trained SOM and assign the probability of each node being selected as the BMU onto the node as the weight. Due to the probability density matching characteristic, there is however a tendency for the nodes to share a flat firing rate distribution. Also to acquire the firing rate over the entire data population can be rather time-consuming, even if on-line resource information is tracked as in some variants of the algorithm (e.g., [18]). Therefore in practice we find it more efficient to assign a uniform weight to all nodes in a map.

2.3.3 Sum of Minimum Distances

In [19] the sum of minimum distances (SMD) as a similarity measure was discussed. It is defined as:

$$\text{SMD}(X, Y) = \frac{1}{2}(\sum_{x \in X} \min_y d(x, y) + \sum_{y \in Y} \min_x d(x, y)). \tag{6}$$

The calculation of SMD between two feature maps is straightforward and is of the same complexity compared with HD. However, like all dissimilarity measures presented so far, it ignores the established neural structure formed among the prototype nodes.

2.3.4 Sum of Average Neighbor Distance (SAND)

As stated earlier a trained feature map not only quantizes its high dimensional features, but also self-organizes into a low dimensional grid structure that reflects the probability density of feature data. The neighborhood topology within a feature map can be examined to further characterize the original feature space. Taking this into account, we propose a modified sum of minimum distance, so called *sum of average neighbor distances* (SAND).

The same as in the calculation process of HD and SMD, for a prototype $x \in X$, its BMU on the peer map Y, i.e., the prototype $y_b \in Y$ with the

minimum distance to x, is found. To calculate SAND, this minimum distance is averaged with the distances between x and y_b's neighbors, before it is summed across all population of X. The same process is then repeated on Map Y.

The calculation process can be summarized in the following steps:

1. Find the BMU $y_b \in Y$ for any $x \in X$, with

$$b = \arg_{y \in Y} \min(\|x - y\|). \tag{7}$$

2. Find out all best-matching pairs (α, β) between the neighborhood of x and y_b, and calculate the averaging distance:

$$d_n(x) = E\{\|\alpha - \beta\|\}, \forall \alpha \in \Omega(x), \beta \in \Omega(y_b) \tag{8}$$

   Here $\Omega(.)$ denotes the neighborhood of a map node.

3. Sum up the individual measures:

$$\text{SAND}(X, Y) = \frac{1}{2}(\sum_{x \in X} d_n(x) + \sum_{y \in Y} d_n(y)). \tag{9}$$

The rationale behind this scheme is the probability density matching ability of the SOM. Examining the matching among a map neighborhood can tell the difference between maps of similar range of spatial span yet originated from different probability distributions. As density differing in the original feature space will result in, on the low dimensional map, either dense grids or sparse grids, the difference can be better reflected by SAND than a plain point-to-point measure.

2.4 Simplified Kullback-Leibler divergence (SKLD)

If we regard a SOM as an approximation of a Gaussian mixture, then we can extend the point-to-point distance measure between prototype vectors onto a dissimilarity measure between two Gaussian distributions. The Kullback-Leibler divergence between two density functions f and g is defined as

$$\text{KL}(f; g) = \int f \log \frac{f}{g} \tag{10}$$

If we assume that the N dimensions of the data are independent and Gaussian distributed, a simplified version of Kullback-Leiberler divergence can be worked out in close form, as presented in [8] for two models p and q:

$$\text{KL}(q; p) = \frac{1}{2} \sum_{j=1}^{N} \left(\log \left(\frac{\sigma_j^{(p)}}{\sigma_j^{(q)}} \right)^2 + \left(\frac{\sigma_j^{(q)} - \sigma_j^{(p)}}{\sigma_j^{(p)}} \right)^2 + \left(\frac{\sigma_j^{(q)}}{\sigma_j^{(p)}} \right)^2 - 1 \right) \tag{11}$$

As the SOM tends to produce prototype vectors of similar variance because of its density matching capability, we simply estimate the deviation $\sigma^{(p)}$ at

each prototype as the average deviation of all map prototypes. Having this simplified measure between two prototypes, we can then define a simplified Kullback-Leibler divergence (SKLD) of two feature maps as the averaged KL measure between best-matching prototypes of the two maps:

$$\mathrm{SKLD}(X,Y) = \frac{1}{2}(E_x\{\mathrm{KL}(x,y_b)\}, E_y\{\mathrm{KL}(y,x_b)\}) \tag{12}$$

2.5 Coupling index

The above dissimilarity measures are all based on the distance between a pair of map nodes and lead to a scale value as the overall dissimilarity assessment between two maps. On the other hand, when two maps representing the high-dimensional manifolds formed by similar visual features are generated, they may overlap or fold into each other. To assess the overlap of feature maps it may be interesting to see how well a feature map compete to represent the others collectively. Hence we define a *coupling index* between a pair of feature maps as a cross referencing ratio to nearest neighbors. Here a cross referencing occurs when the nearest neighbor of a map prototype is located on the peer map rather than its own. We can therefore consider coupling index as a similarity measure, as the more similar the maps are, the more cross referencing occurs, and the bigger the coupling index is.

Given two maps X and Y, the coupling index can be defined for a node $x \in X$:

$$\eta(x) = \begin{cases} 0, & \text{if } \min_{x'}(d(x,x')) < \min_y(d(x,y)), \forall x' \in X, x' \neq x; y \in Y \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

The overall coupling index is then defined as the average of the coupling indexes on all the nodes in X and Y:

$$\mathrm{CI} = \frac{1}{2}(\sum_{x \in X} \eta(x) + \sum_{y \in Y} \eta(y)). \tag{14}$$

3 Empirical study

3.1 Data sets

To verify the image collection profiling and comparison approach as presented above, a series of experiments have been conducted. We use color pictures downloaded from the SUNET FTP site. There are four categories of images used, namely *views*, *sports*, *animals*, and *vehicles*. Each category holds images differing in background and objects in focus. There are 413 images in 'animals', 170 in 'views', 391 in 'sports' and 356 in 'vehicles'. There are 1380 images used in total. Selected thumbnails of each category are shown in Fig.1(a)-(d).

Even though images of each category are pinned under the same subject, we can see from these thumbnails that their content is rather heterogeneous,

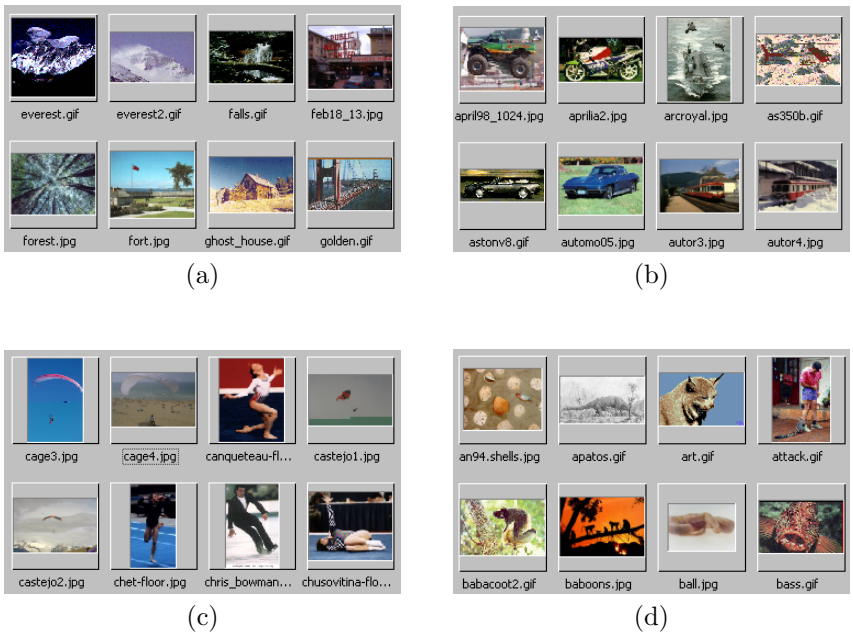

Figure 1: Thumbnails from the collections: (a) 'views', (b) 'vehicles', (c) 'sports', and (d) 'animals'.

presenting different objects or scenes with big variance in both background and foreground. For instance, sky background is present in all four sets; 'vehicles' includes vehicle on the road, in the sky, as well as on water; and 'sports' features outdoor scenes as well as indoor scenes with or without players.

3.2 Training of SOMs

The first feature scheme we explore is regional average colors (RAC) in five non-overlapping zones, the same as in PicSOM [11]. It is actually a simplified version of the color layout descriptor (CLD) as defined in [4], using 5 overlapping zones instead of 8×8 blocks as in CLD. To assess the similarity of texture components of the image collections, a set of Gabor filters in 4 frequency level and 8 orientations is used [20]. Texture energy on these filters gives a 32-dimension feature vector for each image. We denote this feature scheme as GFH.

For sake of simplicity, feature sets of all four categories are clustered on 8×8 SOMs. The SOM_PAK toolbox [21] was used to generate these feature maps under hexagonal topology and Gaussian neighborhood. Two passes of training as suggested in [21] were conducted, with the initial learning rates of 0.05 and 0.03 respectively. Initial neighborhood size were set as 6 and 3 respectively. The

run length was set to be 6400.

The RAC profiles generated as the feature map obtained from the 'views' and the 'vehicles' collection respectively are shown in Fig.2 and Fig.3. These maps are visualized by Sammon's mapping, with some nodes labeled with images that carry the best matching feature to the node vectors.

Figure 2: The profile of image collection 'views'.

3.3 Visual assessment of collection profiles

For visual assessment on the dissimilarity of the four profiles generated on the RAC feature scheme, they are all projected onto a 2-D plot generated by projecting their prototypes over the first two eigenvectors of the overall prototype set. This is shown in Figure 4. Comparing the node 'clouds' shown in Figure 4, we may have some clues of the dissimilarity of the collection profiles. It is hard however to assess the coupling effect from this projection. To get more accurate assessment of their dissimilarity, quantitative measures need to be worked out.

3.4 Quantitative assessment of collection profiles

The mutual dissimilarity measures between the 'sports' profile and the other three, assessed on the RAC and GFH features, are listed in Table 1 and Table 2 respectively.

In Table 1 all distance measures calculated between the 'sports' collection and the other three using the RAC feature are listed together with their CPU time consumed. It is noted that all the dissimilarity measures report similar

Table 1: Distance calculated between the RAC profiles of 'sports' and other three collections.

Measures	Collections			CPU Time
	animals	views	vehicles	
HD	125.4	153.3	135.1	0.01
EMD	73.4	101.0	102.9	0.10
SMD	55.4	66.1	56.7	0.01
SAND	66.7	76.1	68.5	0.02
SKLD	1.187	1.7	3.76	0.01

Table 2: Distance calculated between the GFH profiles.

Measures	Collections			CPU Time
	animals	views	vehicles	
HD	0.54	0.64	0.59	0.01
EMD	0.23	0.29	0.59	0.12
SMD	0.33	0.38	0.26	0.02
SAND	0.35	0.39	0.31	0.04
SKLD	15.36	14.72	27.98	0.01

Table 3: Coupling index matrix of the four profiles.

Collections	animals	vehicles	views
vehicles	17.2%		
views	14.8%	14.1%	
sports	11.7%	16.4%	4.7%

Figure 3: The profile of image collection 'vehicles'.

ranking of the similarity between image collections, with 'animals' on the near side, and 'views' on the far side. Interestingly EMD reports rather close distances to both the 'views' and 'vehicles' profiles. On the other hand, while most measures are efficient to calculate, EMD requires the longest time, requiring more than 0.10 seconds to complete while others need only about 0.01 seconds. This result agrees with other empirical findings on using EMD for image dissimilarity computation [22]. The CPU time here were collected from a Linux 2.2 system running on a Pentium-II PC. Although the difference is not significant here, for applications with much larger collections to compare it may become an important factor to consider.

The coupling index matrix of these collection profiles are also calculated, as shown in Table 3. While these coupling indexes indicate significant overlap of 'animals' - 'views', and 'sports' - 'vehicles' pairs, it is also revealed that little coupling exists between 'sports' and 'views', which can be hardly found out from visual assessment using profiles projection. This may suggest that these two profiles have a large dissimilarity measure, which we shall further verify with another feature scheme.

Table 2 gives the results with profiles generated by the GFH feature. Since the feature codes are in a different range the dissimilarity values are also quite different when compared with those in Table 1. A similar dissimilarity ranking of the collections is obtained, even though in general dissimilarity assessed from different maps trained on different features may differ. It is however noted that the SKLD measure managed to single out the 'views' collection mapped with the texture feature. To obtained an overall dissimilarity ranking, results from

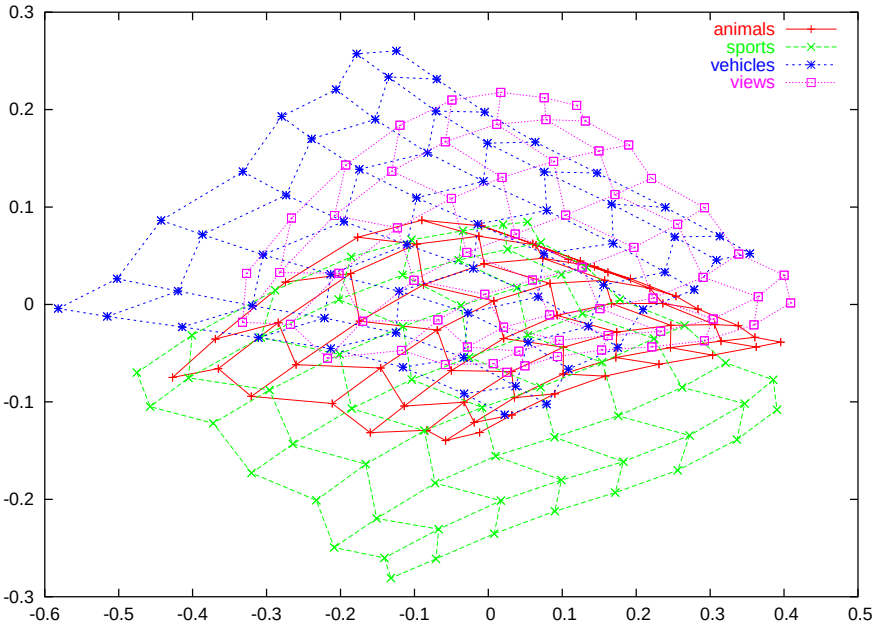

Figure 4: Four RAC-feature SOMs visualized by PCA.

different features can be normalized and then multiplied together, in this case resulting in the normalized joint dissimilarity measures as shown in Fig. 5.

All these dissimilarity measures suggest almost consistently that 'views' and 'sports' have the largest dissimilarity between each other. This is most likely due to the fact that these two collections share the biggest difference in overall visual contents. The 'views' collection mainly features outdoor scenes, while the 'sports' collection has a large portion of indoor scenes featuring close-ups humans in various outfits. On the other hand, 'animals' and 'vehicles' mostly have outdoor objects or man-made objects and therefore should be closer to 'sports'. The dissimilarity between the 'views'-'sports' pair is ranked the second on EMD, originated from the difference measured on the texture feature. This difference however can be balanced if using a smaller weight on texture dissimilarity when producing the joint measures.

3.5 Robustness of the dissimilarity measures

To test the robustness of the dissimilarity measures, we constructed a few mixed data sets from the feature sets generated from the four collections. Table 4 gives the normalized dissimilarity measures of two mixed sets compared with the original sets. For Mixed Set 1, generated by half of 'sports' and half of 'views', measures were normalized with the dissimilarity measure between 'sports' and 'views'. Likewise the measures on Mixed Set 2 (generated with half of 'animals'

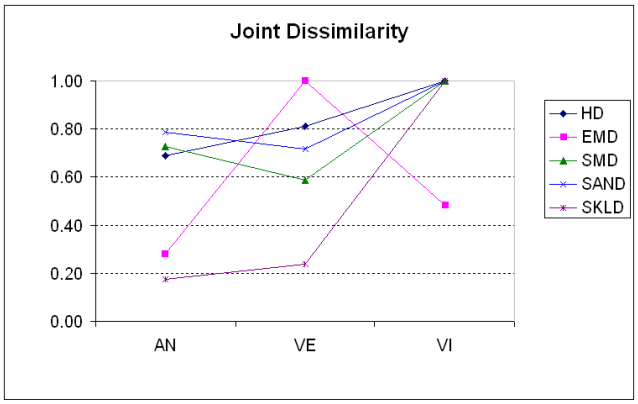

Figure 5: Joint dissimilarity measures between 'sports' and AN: 'animals', VE: 'vehicles', and VI: 'views'.

Table 4: Mixed sets and their normalized dissimilarity measures

Dissimilarity	HD	EMD	SMD	SAND	SKLD	CI
Mixed Set 1						
to 'sports'	0.425	0.444	0.417	0.478	0.238	74.20%
to 'views'	0.788	0.700	0.706	0.783	0.689	31.20%
Mixed Set 2						
to 'animals'	0.76	0.60	0.61	0.68	0.61	50.70%
to 'vehicles'	0.52	0.50	0.55	0.63	0.48	63.30%

and half of 'vehicles') were also normalized. We can see that all the dissimilarities measured between the mixed set and the original sets are smaller than the dissimilarity between the original sets. Also, the dissimilarity order is consistent with the coupling index. The general tendency is that, the smaller the coupling index is, the bigger the dissimilarity will be. Other mixed sets generated also give similar outcome, suggesting that the dissimilarity measures as well as the coupling index are all robust to assess the dissimilarity of mixed collections.

On the other hand, when the size of a feature map is changed, its quantization ability, location and modeling of its prototypes will be directly impacted. It will be interesting to see how dissimilar the profiles of the same collection can be if being generated in different sizes. The following experiment was conducted to test the robustness of the dissimilarity measures calculated for feature maps of different sizes. Different map sizes, from 4×4, to 16×16, were used and maps were generated from the 'sports' RAC feature. We then used the original 8×8 'sports' map as a reference map, calculated its dissimilarity measures to these new maps. We also normalized these values by the smallest inter-collection measure (between 'sports' and 'animals', as given in Table 1) to see how the scaling on the maps will affect the dissimilarity assessment between collections.

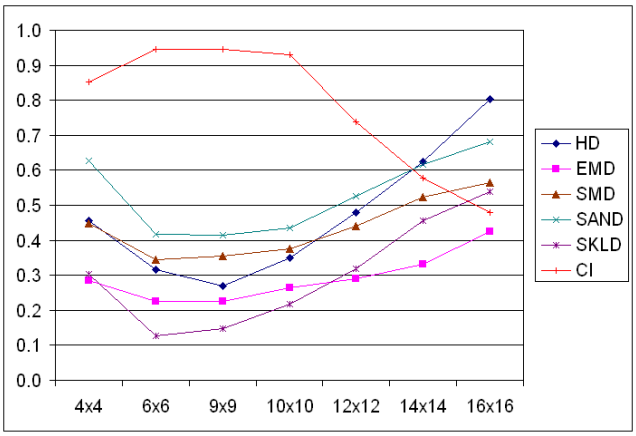

Figure 6: Normalized dissimilarity measures and CI for maps of different sizes compared with `Sports_8x8_RAC`.

The results are presented in Fig. 6. As we can see all dissimilarity measures will more or less be affected by map scaling. Among them, SKLD gives consistently the best performance especially around the reference size (6×6, $9times9$, and 10×10), only starting to be become less robust than EMD after the size of 12×12. The CI curve also indicates that from then on the coupling of the bigger map with the reference map drops significantly. In this case, local Gaussian models should undergo significant changes, making a dissimilarity measure such as SKLD defined on the Gaussian mixture less invalid. It is noted that EMD as a point set approach suffers less from the further scaling-up of the feature maps, but a 40% dissimilarity compared with the reference is still reached when the number of map nodes approaches four times the original size.

These results indicate that SKLD and EMD are robust to limited change on the feature map size. To mitigate the effect of map size variation so as to avoid possible failures in assessing map similarity, we can impose some control on the map size, possibly by maintaining the quantization error on a certain level.

4 Conclusion

We proposed to use self-organizing maps trained on low level CBIR features extracted from image collections for content-based summarization and comparison. A number of different dissimilarity measures are examined in a four collection problem with more than 1000 images in total. Especially we have found that a simplified Kullback-Leibler dissimilarity measure outperforms the Earth Mover's Distance in terms of efficiency, ranking accuracy, and robustness to small map size changes. The coupling index, introduced as a simple concept, also gives consistent result in assessing the overlap of collection profiles. We would like to test our approach on more image collections of larger scales,

assessing feature maps generated on more feature schemes used such as in [4].

While the SOM may seem a natural and effective approach for document and multimedia data analysis, a few drawbacks exist. There is no efficient mechanism for a feature map, once trained, to adapt its own size when there is a need to allocate new resource for novel inputs. The lack of incremental learning ability in the SOM also makes on-line adapting of the network implausible. Considering the use of summarization for multimedia collections under constant variation, it is desirable to adapt the existing profiles with new data without the retraining of the whole model. This is hard to achieve with the SOM.

Other self-organizing neural networks may be considered in this respect, including online incremental clustering models and hierarchical models [23][24]. A hierarchical online learning process will improve the efficiency for profiling large image collections. Better probabilistic modeling ability may also improve the robustness of dissimilarity measures defined on the profiles. Also, by exploring more powerful feature descriptors, it is promising for this approach to be extended for content-based summarization and comparison of audio and video data collections.

References

[1] C. Carson, M. Thomas, S. Belongie, et al., Blobworld: A system for region-based image indexing and retrieval, in: Proc. Int. Conf. Visual Inf. Sys., 1999, pp. 509–516.

[2] J. Smith, S. Chang, Visualseek: a fully automated content-based image query system, in: Proc. of ACM Multimedia 96, 1996, pp. 87–98.

[3] A. Smeulders, M. Worring, S. Santini, A. Gupta, J. R., Content-based image retrieval at the end of the early years, IEEE Transaction on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.

[4] B. Manjunath, J. Ohm, V. Vinod, A. Yamada, Color and texture descriptors, IEEE Trans. Circuits and Systems for Video Technology Special Issue on MPEG-7.

[5] M. Bober, Mpeg-7 visual shape descriptors, IEEE Trans. on Circuits and Systems for Video Technology 11.

[6] R. Brunelli, O. Mich, Histograms analysis for image retrieval, Pattern Recognition 34 (2001) 1625–1637.

[7] Y. Rubner, C. Tomasi, L. Guibas, A metric for distributions with applications to image databases, in: Proc. of IEEE ICCV, 1998, pp. 59–66.

[8] J. R. Mathiassen, A. Skavhaug, K. Bø, Texture similarity measure using kullback-leibler divergence between gamma distributions, in: ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part III, Springer-Verlag, London, UK, 2002, pp. 133–147.

[9] T. Kohonen, Self-organizing Maps, 2nd Edition, Springer-Verlag, 1997.

[10] A. Rauber, D. Merkl, The `somlib` digital library system, in: Proc. of European Conference on Digital Libraries, 1999, pp. 323–342.

[11] J. Laaksonen, M. Koskela, E. Oja, Content-based image retrieval using self-organizing maps, in: Visual Information and Information Systems, 1999, pp. 541–548.

[12] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd Edition, Prentice Hall, 1999.

[13] H. Ritter, Asymptotic level density for a class of vector quantization processes, IEEE Trans. Neural Networks 2 (1991) 173–175.

[14] H. Yin, N. Allison, Self-organizing mixture networks for probability density estimation, IEEE Trans. on Neural Networks 12 (2) (2001) 405–411.

[15] W. Sammon, A nonlinear mapping for data analysis, IEEE Trans. on Computers 5 (1969) 401409.

[16] S. Kaski, K. Lagus, Comparing self-organizing maps, in: J. Vorbruggen, B. Sendhoff (Eds.), Proceedings of ICANN96 International Conference on Artificial Neural Networks, Vol. 1112 of Lecture Notes in Computer Science, Springer, Berlin, 1996, pp. 809 – 814.

[17] T. Honkela, Comparisons of self-organized word category maps, in: Proceedings of WSOM97, Workshop on Self-Organizing Maps, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997, pp. 298–303.

[18] B. Fritzke, A growing neural gas network learns topologies, in: Proc. of NIPS, 1994.

[19] T. Eiter, H. Mannila, Distance measures for point sets and their computation, Acta Informica 34 (1997) 109–133.

[20] B. S. Manjunath, W. Ma, Texture features for browsing and retrieval of image data, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI - Special issue on Digital Libraries) 18 (8) (1996) 837–42.
URL `http://vision.ece.ucsb.edu/publications/96PAMITrans.pdf`

[21] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, Som_pak: The self-organizing map program package (1996).

[22] J. Puzicha, Y. Rubner, C. Tomasi, J. Buhmann, Empirical evaluation of dissimilarity measures for color and texture, in: Proceedings the IEEE International Conference on Computer Vision(ICCV-1999), 1999, pp. 1165–1173.

[23] D. Deng, N. Kasabov, On-line pattern analysis by evolving self-organizing maps, Neurocomputing 51 (2003) 87–103.

[24] P. Tino, I. Nabney, Hierarchical GTM: Constructing localized nonlinear projection manifolds in a principled way, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 639–656.