Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval

Markus Koskela, Jorma Laaksonen, and Erkki Oja
Laboratory of Computer and Information Science, Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, FINLAND

Abstract

Content-based image retrieval (CBIR) addresses the problem of finding images relevant to the users' information needs, based principally on low-level visual features for which automatic extraction methods are available. For the development of CBIR applications, an important issue is to have efficient and objective performance assessment methods for different features and techniques. In this paper, we study the efficiency of clustering methods for image indexing with entropy-based measures. Furthermore, the Self-Organizing Map (SOM) as an indexing method is discussed further and an analysis method which takes into account also the spatial configuration of the data on the SOM is presented. The proposed methods enable computationally light measurement of indexing and retrieval performance for individual image features.

1. Introduction

Indexing image databases is a different and in many ways more complex problem than indexing traditional databases. The main difficulties arise from the high dimensionality of the used feature vectors, the large sizes of the image databases, and that many feature spaces may have to be used simultaneously. Due to these factors, using basic linear search easily leads to poor performance and specialized techniques are needed so that the most similar images can be determined quickly enough.

In general, there are two broad categories of index structures for high-dimensional spaces. First, one can transform the original feature space into a new space where the needed operations are less demanding. This usually means reducing the dimensionality of the feature space. Alternatively, one can apply a divide-and-conquer type strategy: the data or the feature space is divided into clusters or subspaces with the intention that only one or a few of these have to be processed in one given query. After clustering, each cluster is represented by its centroid or representative data item.

The Self-Organizing Map (SOM) [2] can be considered a method for both clustering and dimensionality reduction. The mapping of feature vectors and their associated images to their best-matching units (BMUs) can be interpreted as clustering. This, however, ignores the topology of the SOM, so a portion of the provided data organization is dismissed.

In this paper, we study how clustered image class distributions can be interpreted in terms of probability densities and how the effectiveness of a clustering method can be assessed with entropy-based methods. Indexing feature vectors with the SOM is then discussed further and a method which takes into account also the spatial configuration of the data on the SOM surfaces is presented.

2. Feature-wise evaluation

Objective performance measures of CBIR are needed for further development in the research field. This includes benchmarking entire CBIR systems with simulated retrieval tasks designed to resemble the actual usage of the system as well as possible. Usually, these evaluations are conducted by defining a number of example queries and corresponding sets or *classes* of relevant images $\mathcal{C}$, and then measuring the ability of the system to retrieve the images belonging to $\mathcal{C}$.

In addition, individual features and indexing methods need to be studied separately. Evaluations of this type are essential in feature extraction method development. In these evaluations, it is often unnecessary to do a time-consuming simulation of the actual retrieval system with an extensive set-up. Instead, a direct measure based on the ability of the feature extraction to discriminate images belonging to a certain set of semantic similarity or relevancy may suffice.

3. Class distributions

The shape of the distribution of a set of high-dimensional feature vectors mapped on a set of k clusters depends on several factors including: (a) The distribution of the *original data* in the very-high-dimensional pattern space, which

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

is generally given and cannot be controlled. (b) The *feature extraction* technique in use affects the formation and thus the distribution of all the generated feature vectors. (c) The *overall shape* of the training set, after it has been mapped from the original data space to the feature vector space, determines the overall organization of the clustering. (d) The *class distribution* of the studied object set or class, relative to the overall shape of the feature vector distribution, specifies the configuration of the class within the clustering.

In the very-high-dimensional pattern space the distribution of any non-trivial object class is most certainly sparse. As a consequence, in most cases it is meaningless to talk about the uni- or multimodality of class distributions in the pattern space. On the other hand, if the feature extraction stage is working properly, semantically similar patterns will be mapped in the feature space nearer to each other than semantically dissimilar ones. In the most advantageous situation, the pattern classes might even match clusters in the feature space, ie. there would exist a one-to-one correspondence between feature vector clusters and pattern classes.

4. Cluster entropy and perplexity

Given a set of cluster centroids, one can in theory calculate the *a priori* probability of each cluster for being the best-matching one for any vector $\mathbf{x}$ of the feature space. This is possible if the *probability density function* (pdf) $p(\mathbf{x})$ is known. When the cluster is denoted by i and its surrounding *Voronoi region* by $\mathcal{V}_i$, one may calculate the unit's *a priori* probability P_i as

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \int_{\mathcal{V}_i} p(\mathbf{x})\, \mathbf{dx} \quad . \tag{1}$$

With discrete data, one needs to replace the continuous pdf with a discrete probability *histogram*. Without danger of confusion, the probability can still be denoted as P_i:

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i \,\}}{N} \quad , \tag{2}$$

where $\#\{\cdot\}$ stands for the cardinality of a set, and N is the size of the training data set, whose members are $\mathbf{x}_j$, $j = 0, 1, \ldots, N-1$. Considering only an image class $\mathcal{C}$ instead of all images, the probability histogram will be

$$P_i^{\mathcal{C}} = P(\mathbf{x} \in \mathcal{V}_i \mid \mathbf{x} \in \mathcal{C}) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i, \mathbf{x}_j \in \mathcal{C} \,\}}{N^{\mathcal{C}}} \quad . \tag{3}$$

A simple and commonly used measure for the randomness of a symbol distribution is its *entropy*. In our case, the cluster indices for the vectors of the training set play the role of symbols. The entropy H of a distribution $P = (P_0, P_1, \ldots, P_{k-1})$ is calculated as:

$$H(P) = -\sum_{i=0}^{k-1} P_i \log P_i \quad , \tag{4}$$

where k is the number symbols in the alphabet of the stochastic information source. P_i is the probability of cluster i being the correct one for an input vector, as defined before. Usually logarithm base of two is used.

If one assumes that every cluster is equally probable as the correct one for an input vector, one can easily calculate a theoretical maximum for the entropy of the clustering:

$$H_{\max} = \max_{\{P_i\}} \left\{ -\sum_{i=0}^{k-1} P_i \log P_i \right\} = \log k \quad . \tag{5}$$

In the discrete case, the above definition for $H_{\max}$ to hold exactly assumes that N is divisible by k. In general this is not the case but the produced error is insignificant with sufficient amount of data, ie. if $N \gg k$. This can generally be assumed when studying the whole database since the overall aim of the clustering is to reduce computational requirements of the retrieval algorithm.

Instead of using entropy directly, perhaps a more illustrative measure is *perplexity* $PPL = 2^H$, which is commonly utilized in text-based information processing, especially speech recognition. Perplexity can be considered as the weighted number of equal choices for a random variable; ie. in this setting, the average number of equivalent clusters that have to be considered. Thus, if entropy has the maximum value, perplexity of a clustering equals the total number of clusters, $PPL_{\max} = k$. A suitable performance measure for feature extraction and the associated clustering methods can be formed by the ratio of perplexity and the total number of clusters, denoted here as *normalized perplexity* $\overline{PPL} = 2^H/k$, which is non-negative and ≤ 1 in all cases. The normalized perplexity of an image class $\mathcal{C}$ can simply be calculated by replacing P_is in Eq. (4) with $P_i^{\mathcal{C}}$s.

In general it can be assumed that the clustering distributes the input vectors roughly evenly to all clusters and the normalized perplexity of the whole data should thus be near unity. On the other hand, images with semantic similarity should be mapped to a small cluster subset, provided that the feature extraction and clustering methods have been favorable to that specific class. In this case, normalized perplexity should be $\ll 1$. However, it should be noted that with small image classes and large values of k, $\overline{PPL}$ will be biased toward smaller values since the perplexity value cannot exceed the size of the class. If this is the case, $N^{\mathcal{C}}$ instead of k can be a more suitable scaling factor for $\overline{PPL}$.

5. SOM entropy

The above measure is general and suited for any clustering method. The SOM algorithm is, however, by nature a trade-off between clustering and topological ordering. This trade-off depends on the size of the SOM; the clustering

property is dominant with relatively small SOMs ($k \ll N$) whereas the topology of the map becomes more significant as the size of the SOM is increased. With larger SOMs, the measure is thus less informative as the number of images sharing a BMU becomes overly small and the perplexity value mostly reflects just the size of the image class. In this setting, the spatial configuration of the data on the SOM grid should be taken into account. In the extreme case of k and N being of the same order of magnitude, the unit-wise clustering performed by the SOM is negligible and the organization of the data lies in the topology of the map.

In devising a topology-supporting entropy measure, we begin by recognizing the analogy of a greyscale image and the SOM. Map units of the SOM correspond to image pixels when the intensity values of map units are determined by the data histogram of the SOM. Let X_s denote the grey-level value of pixel s in an image with M pixels. In this setting, entropy of an image pixel can be written as

$$H(X_s) = -\sum_{i=0}^{G-1} P(i) \log P(i) \ , \tag{6}$$

where G is the number of distinct grey levels and $P(i)$ is the probability of the ith grey level, usually estimated by the ith value on the normalized histogram of the image. With the logarithm base of 2, $H(X_s)$ can be interpreted as the minimum number of bits needed to code the grey-level value of a pixel if we know the histogram of the image.

The entropy of Eq. (6) is clearly insufficient as a measure of a pixel's uncertainty on natural images as it completely neglects any spatial properties. To incorporate the local neighborhood context of a pixel, a common method is to apply Markov Random Fields (MRFs) [1]. Following MRF terminology, let $\mathcal{N}_s$ denote a *neighborhood* of pixel (site) s. $\mathcal{N}_s$ consists of pixels j so that $P(X_s \mid X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_M)$ depends on X_j. However, it is often assumed that $\mathcal{N}_s$ only consists of pixels spatially close to s. In addition, the neighborhood $\mathcal{N}_s$ is in a certain configuration $N_s = \{X_t \mid t \in \mathcal{N}_s\}$. The *spatial entropy* of a pixel can now be defined as

$$H_{\mathrm{sp}}(X_s) = -\sum_{N_s} \sum_{i=0}^{G-1} P(i, N_s) \log P(i \mid N_s) \ . \tag{7}$$

Spatial entropy has been used eg. as a measure of satellite image redundancy [8] and in texture discrimination [7].

Due to the large number of possible configurations N_s, considering all of them requires a lot of data. The number of configurations can be reduced eg. by considering only the number of pixels with the same grey-level value as X_s [8]. Here, a natural choice is to examine only the number of data points I_s that have been mapped to the neighborhood of s. X_s thus now equals the number of data points mapped to s and $I_s = \sum_{t \in \mathcal{N}_s} X_t$. Conditional entropy of map unit s given that the neighborhood contains m data points is then

$$H_{\mathrm{sp}}(X_s \mid I_s{=}m) = -\sum_{i=0}^{G-1} P(i \mid I_s{=}m) \log P(i \mid I_s{=}m) \tag{8}$$

and the spatial entropy is now given by

$$H_{\mathrm{sp}}(X_s) = -\sum_{m} \sum_{i=0}^{G-1} P(i, I_s{=}m) \log P(i | I_s{=}m) \ . \tag{9}$$

A normalized performance measure can now be obtained by

$$\rho = 1 - \frac{H_{\mathrm{sp}}(X_s)}{H(X_s)} \ . \tag{10}$$

The ρ measure is zero for a completely random distribution since the neighborhood does not provide any information about the number of data points mapped to a map unit. Respectively, ρ is near one for a highly localized distribution.

6. Experiments

In the following experiments, we study two clustering methods, viz. k-means and the SOM. As image data we use a Corel database of 59 995 images and as features three MPEG-7 [6] descriptors and a *keyword* (KW) feature computed from related textual data [4]. The used MPEG-7 descriptors were *Color Structure* (CS), *Edge Histogram* (EH), and *Homogenous Texture* (HT). As semantic image classes, we used three manually picked subsets of the images: faces (1115 images), cars (864 images), and sunsets (663 images). A separate set of SOMs was trained for each feature by using the Tree Structured SOM algorithm [3]. The sizes of the SOM layers were $4 \times 4 = 16$, $16 \times 16 = 256$, $64 \times 64 = 4096$, and $256 \times 256 = 65536$ map units. For k-means, the studied values of k were 16, 256, and 4096.

Table 1 shows the resulting perplexity values for both methods. It can be observed that k-means performs better than the SOM as a clustering method, which was to be expected due to the SOM's aforementioned tradeoff between clustering and preserving topology. This result was affirmed also with actual retrieval experiments in [5]. Due to the semantic gap between low-level features and semantic image classes, the results with the visual features remain quite modest. With sunsets, the perplexity values of visual features are the lowest. Correspondingly, it has been previously determined that of these classes, sunsets is the "easiest" one for retrieval [4]. With the *keyword* feature, all perplexities are the lowest, indicating that the feature is able to cluster all three classes. Again, this agrees with previous experiments, in which the superior retrieval performance of *keyword*s was perceived [4]. It can also be noted that with

Table 1. Perplexities and normalized perplexities (in parentheses) of k-means and SOM clustering.

	k	faces (1115)	cars (864)	sunsets (663)	SOM size	faces (1115)	cars (864)	sunsets (663)
CS	16	10.7 (0.67)	13.5 (0.85)	6.66 (0.42)	4×4	11.5 (0.72)	14.3 (0.90)	5.44 (0.34)
	256	109 (0.43)	157 (0.61)	43.4 (0.17)	16×16	129 (0.51)	172 (0.67)	51.3 (0.20)
	4096	553 (0.14)	569 (0.14)	188 (0.046)	64×64	633 (0.15)	673 (0.16)	253 (0.062)
					256×256	1050 (0.016)	840 (0.013)	569 (0.0087)
EH	16	11.2 (0.70)	11.9 (0.75)	4.46 (0.28)	4×4	11.7 (0.73)	13.0 (0.81)	4.71 (0.29)
	256	99.3 (0.39)	98.8 (0.39)	42.6 (0.17)	16×16	107 (0.42)	107 (0.42)	47.0 (0.18)
	4096	505 (0.12)	503 (0.12)	222 (0.054)	64×64	554 (0.14)	523 (0.13)	301 (0.073)
					256×256	1010 (0.015)	818 (0.012)	567 (0.0087)
HT	16	10.1 (0.63)	11.6 (0.73)	9.62 (0.60)	4×4	12.6 (0.79)	13.9 (0.87)	7.18 (0.45)
	256	122 (0.48)	146 (0.57)	86.9 (0.34)	16×16	120 (0.47)	161 (0.63)	85.1 (0.33)
	4096	609 (0.15)	627 (0.15)	352 (0.086)	64×64	674 (0.16)	659 (0.16)	399 (0.097)
					256×256	1040 (0.016)	838 (0.013)	623 (0.0095)
KW	16	4.62 (0.29)	3.91 (0.24)	3.82 (0.24)	4×4	3.04 (0.19)	4.08 (0.26)	5.30 (0.33)
	256	18.5 (0.072)	6.04 (0.023)	18.6 (0.073)	16×16	15.9 (0.062)	8.41 (0.033)	31.3 (0.12)
	4096	139 (0.034)	62.8 (0.015)	74.9 (0.018)	64×64	71.3 (0.017)	44.3 (0.011)	91.6 (0.022)
					256×256	253 (0.0039)	157 (0.0024)	252 (0.0038)

the largest SOMs, the perplexity values of the visual features are close to image class sizes. Most of the map units that contain images in the class thus contain only one of them. This is understandable as the SOM actually has more map units (65536) than there are images in the database.

In order to properly investigate the organization provided by the largest SOMs, we must take the spatial configuration into account. Table 2 shows values of the ρ measure (10) for the largest SOMs. The used neighborhood consists of map units within an L_1 distance of 8 units from the unit in question. In Table 2, these values are compared to the so called τ measure [5] which is an averaged rank-based measure of retrieval performance (larger is better with ρ, smaller is better with τ). The comparison shows that similar results are obtained with both measures, ie. a feature-class pair that works well according to one measure, does so also according to the other. Calculating the τ measure requires N^C full queries with the actual retrieval system, so it is computationally considerably more demanding as the ρ measure which can be obtained directly from the SOM index.

Table 2. Values of ρ (left) and τ (right) for different features and image classes.

	faces	cars	sunsets	faces	cars	sunsets
CS	0.12	0.029	0.27	0.35	0.40	0.25
EH	0.17	0.14	0.24	0.27	0.31	0.075
HT	0.11	0.044	0.17	0.24	0.38	0.24
KW	0.38	0.47	0.32	0.17	0.070	0.084

7. Conclusions

In this paper, we examined how distributions of image feature vectors can be studied with a clustering or on a SOM surface. Entropy and perplexity of the distribution characterize quantitatively the compactness of the class, which is in turn an indicator of the success of the feature extraction and indexing methods for that particular class. Also, more informative results for the SOM were obtained with a proposed entropy measure taking the map topology into account. The feature-wise assessments agreed well with previous experiments obtained with an actual retrieval system.

References

[1] R. Chellappa and A. Jain, editors. *Markov Random Fields: Theory and Application*. Academic Press, 1993.

[2] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 2001.

[3] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. International Joint Conference on Neural Networks*, volume II, pages 279–284, San Diego, CA, 1990.

[4] M. Koskela and J. Laaksonen. Using long-term learning to improve efficiency of content-based image retrieval. In *Proc. Third International Workshop on Pattern Recognition in Information Systems*, pages 72–79, Angers, France, April 2003.

[5] M. Koskela, J. Laaksonen, and E. Oja. Comparison of techniques for content-based image retrieval. In *Proc. 12th Scandinavian Conference on Image Analysis*, pages 579–586, Bergen, Norway, June 2001.

[6] MPEG-7 Overview (version 9), March 2003. ISO/IEC JTC1/SC29/WG11 N5525.

[7] F. Tupin, M. Sigelle, and H. Maître. Definition of a spatial entropy and its use for texture discrimination. In *Proc. IEEE International Conference on Image Processing*, volume 1, pages 725–728, Vancouver, Canada, September 2000.

[8] E. Volden, G. Giraudon, and M. Berthod. Modelling image redundancy. Technical Report 2440, INRIA Sophia-Antipolis, Cedex, France, December 1994.