Probability interpretation of distributions on SOM surfaces

Jorma Laaksonen, Markus Koskela and Erkki Oja
Helsinki University of Technology
P.O.BOX 5400, 02015 HUT, Finland
tel.+358-9-4513269, fax.+358-9-4513277,
{*jorma.laaksonen,markus.koskela,erkki.oja*}*@hut.fi*

Keywords: feature extraction, entropy, mutual information, content-based image retrieval

Abstract— In this paper, the distributions of training vectors on SOM surfaces are studied in terms of probabilities. The entropy of a single feature's distribution and the mutual information of two features' distributions are used to describe the form of the distributions quantitatively. Qualitatively different distributions can be obtained from the same data by using different feature extraction techniques and by studying different semantically related object subsets. In addition, the effect of low-pass filtering the SOM surfaces prior to the calculation of the entropy is studied. This technique facilitates analyzing the compactness and internal structure of an object class after mapping on the two-dimensional SOM surface. Illustrations and examples come from content-based image retrieval (CBIR), especially our PicSOM CBIR system.

1 Introduction

The Self-Organizing Map (SOM) [2] is a powerful tool for exploring huge amounts of high-dimensional data. Many studies have been made on the clustering and visualization capabilities of the SOM. In the case of labeled training samples, some people have used SOMs also for classification even though the technique is not intended for such use. For labeled data, the SOM is best used as a visualization tool for the class distributions.

In this paper, we study how object class distributions on SOMs can be given interpretations in terms of probability densities and information theoretic measures such as entropy and mutual information. The latter measures arise when the distributions of the same objects after two different feature extraction stages are compared.

The discussions in the paper are meant to be general in their nature. However, the specific application used in illustrations is *content-based image retrieval* (CBIR) [1, 10], namely our PicSOM CBIR system [6, 7]. PicSOM uses SOMs in implementing *relevance feedback* and *query by example* (QBE) paradigms [8] in interactive and iterative information retrieval from unannotated databases.

In the illustrations, we use MPEG-7 [9] and keyword features extracted from a Corel Gallery database of 59 995 images. We have trained a separate SOM for each feature type by using the Tree Structured SOM (TS-SOM) algorithm [4, 3]. The sizes of the TS-SOM layers have been 4×4, 16×16, 64×64, and 256×256. This results in approximately 3750, 234, 15, and 0.92 training set vectors on the average per SOM unit, respectively.

2 Class Distributions

The shape of the distribution of a set of high-dimensional vectors mapped on the surface of a SOM depends on a multitude of factors including:

- The distribution of the *original data* in the very-high-dimensional pattern space is generally given and cannot be controlled.
- The *feature extraction* technique in use affects the formation and thus the distribution of all the generated feature vectors.
- 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 SOM.
- The *class distribution* of the studied object set or class, relative to the overall shape of the feature vector distribution, specifies the location of the class on the formed SOM.

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 nearer to each other in the feature space than semantically dissimilar ones. In the most advantageous situation, the pattern classes match clusters in the feature space, ie. there exists a one-to-one correspondence between feature vector clusters and pattern classes. In actual circumstances this situation is however rare.

In feature extraction, some pattern space directions are retained better than others. This is known as a particular type of *feature invariance*. Depending on the application, different types of invariances are needed.

The relative distances between the feature vectors of a class compared to the overall distribution of the feature space data determine how well the class is concentrated in nearby SOM units. If the class is truly multimodal with wide relative variance, one cannot in general avoid its splitting in non-contiguous map regions.

Figure 1 visualizes how the pattern space is projected to feature space, the vectors of which are then used in training the SOM. The areas occupied by objects of a particular class are shown with gray shades.

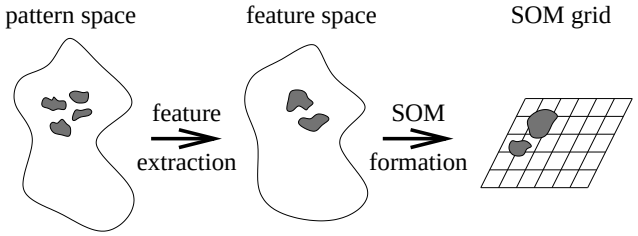

Figure 1: Stages in dimensionality reduction from the very-high-dimensional pattern space through the high-dimensional feature space to the two-dimensional SOM grid.

3 BMU Probabilities

In theory, one can calculate the *a priori* probability of each SOM unit for being the *best-matching unit* (BMU) 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 SOM unit 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

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

Empirical data in engineering applications is, however, generally discrete and finite. One needs therefore to replace the continuous pdf with a discrete probability *histogram*. Without danger of confusion, the probability can still be denoted as P_i:

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

where $\#\{\cdot\}$ stands for the cardinality of a set, and N is the size of the training data set, whose members are $\mathbf{x}_j$, $j = 0, 1, \ldots, N-1$.

In what follows, we will study also the distributions of specific subsets of the training data. We may assume that the members of such a subset fulfill a specific *ground truth* criterion by which each object can be classified as either a member or non-member of the class. The classes will be denoted with superscripts and the probability histogram of class $\mathcal{C}$ will thus be

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

4 BMU Entropy

We will now turn to study the randomness of the distributions of feature vectors' BMUs on the SOM surface. A simple and commonly used measure for the randomness of a symbol distribution is its *entropy*. In our case, the BMU 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:

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

where K is the number symbols in the alphabet of the stochastic information source, in our case thus the number of map units. P_i is the probability of map unit i being the BMU of an input vector, as defined before.

If one assumes that every map unit is equally probable as an input vector's BMU, one can easily calculate a theoretical maximum for the entropy of the BMU distribution

$$H_{\max} = -\sum_{i=0}^{K-1} P_i \log P_i = -K \cdot \frac{1}{K} \log \frac{1}{K} = \log K \quad .$$

Thus, for a map of size $K = 16 \times 16 = 256$ units, $H_{\max} = 8$ when logarithm base two is assumed.

The entropy of BMU histograms can to some extent be made independent of the size of the SOM by dividing the entropy H by its theoretical maximum. The *relative entropy* $\bar{H}$ of the distribution is thus

$$\bar{H} = \frac{H}{H_{\max}} \quad .$$

In general it can be assumed that the training of a SOM distributes the training vectors evenly over the map surface. Therefore, the relative entropy $\bar{H}$ of the training set should be near unity. On the other hand, all object subsets of semantic similarity should be concentrated in few SOM units if the feature extraction and SOM training phases have been favorable to that specific subset. The relative entropy $\bar{H}^{\mathcal{C}}$ of class $\mathcal{C}$ can simply be calculated by replacing P_is with $P_i^{\mathcal{C}}$s.

Table 1 illustrates this with four SOM grids of different sizes (4×4, 16×16, 64×64, and 256×256 map units). The relative entropies of all images in the Corel database and three semantic image classes are shown. The used image classes were faces (1115 images, *a priori* probability 1.85%), cars (864 images, 1.44%), and sunsets (663 images, 1.11%). Four feature extraction

Table 1: Relative entropies of different SOM sizes with the *Color Structure* (CS), *Scalable Color* (SC), *Edge Histogram* (EH), *Homogenous Texture* (HT), and *keyword* (KW) descriptors, respectively.

	SOM size	all	faces	cars	sunsets
CS	4×4	0.990	0.882	0.961	0.611
	16×16	0.995	0.877	0.928	0.710
	64×64	0.994	0.776	0.783	0.665
	256×256	0.947	0.627	0.607	0.572
SC	4×4	0.979	0.877	0.966	0.749
	16×16	0.995	0.885	0.927	0.775
	64×64	0.995	0.788	0.777	0.697
	256×256	0.943	0.629	0.606	0.575
EH	4×4	0.975	0.888	0.925	0.559
	16×16	0.996	0.843	0.843	0.694
	64×64	0.995	0.759	0.752	0.686
	256×256	0.941	0.624	0.605	0.572
HT	4×4	0.989	0.914	0.949	0.711
	16×16	0.997	0.864	0.916	0.801
	64×64	0.995	0.783	0.780	0.720
	256×256	0.948	0.627	0.607	0.580
KW	4×4	0.865	0.401	0.507	0.601
	16×16	0.922	0.499	0.384	0.621
	64×64	0.905	0.513	0.456	0.543
	256×256	0.844	0.499	0.456	0.499

methods defined in MPEG-7 were experimented with, viz. the *Color Structure* (dimensionality 256), *Scalable Color* (256), *Edge Histogram* (80), and *Homogenous Texture* (62) descriptors. In addition, a *keyword* feature obtained from the original keyword annotations for the images was used. The keyword feature was composed of 4538 keywords and then reduced to 150 dimensions with latent semantic indexing (LSI) (see [5] for more details).

First of all, it can be observed that the relative entropy results with the whole database are close to one and clearly higher than the ones computed with only one image class. Relative entropies of semantic image classes express distinct differences, providing estimates about the discriminating ability of those particular feature extraction methods with the used object class.

In previous experiments (eg. [7]), it has been determined that of the three studied classes, sunsets is the "easiest" one with the low-level visual features, ie. it yields by far the best retrieval results. The two other classes exhibit more similar retrieval behavior, with the class of cars resulting in slightly worse retrieval precision. These findings are validated also by the relative entropies of the classes. With the *keyword* feature, all the class-wise relevant entropies are smallest, indicating that the feature is able to cluster all three semantic classes. Again, this agrees with previous experiments [5], in which the transcendent retrieval performance of the *keyword* feature was perceived.

An overall trend seems to be that the relative entropy decreases as the size of the SOM increases. This is especially true for the distributions of the semantic classes. Similar results are obtained with all feature extraction methods. An explanation for this behavior can be found from the fact that the maximum entropy $H_{\max}$ increases faster than the actual entropy H when the size of the SOM is increased. This in turn is a consequence of the clustering property of the SOM: the training vectors are not distributed strictly evenly, but mutually similar vectors form clusters that are separated by gaps of empty map regions.

5 SOM Surface Convolutions

It should be noted that the above described calculation of entropies does not take into account the spatial topology of the SOM units. This is a direct consequence of the fact that the SOM indices are regarded as discrete symbols without any connection to the SOM grid. One may, however, obtain a different kind of interpretation from the same data. If the BMU of each training vector is given a small positive "hit" value equal to the inverse of the number of training vectors, the SOM surface can be turned to a discrete value field. When the hit values are summed in the SOM units, the overall sum of all values in the map units equals to one and the field can be regarded as a discrete probability distribution. The entropy of this distribution is naturally the same as that of the BMU index distribution.

One may then force the neighboring SOM units to interact by *low-pass filtering* or *convolving* the hit distributions on the SOM surface. When the surface is convolved, the one-to-one relationship between input vectors' SOM indices and "hits" on the SOM surface is broken. Instead, each hit results in a spread point response around the BMU. This can be seen as a form of Parzen or Gaussian mixture estimation applied with the distribution centers located on the SOM grid.

We will first study the effects the convolution has on the surface entropy with a series of simple artificial examples. The left column of images in Figure 2 shows a series of SOM surface value fields prior to convolution. The right image column displays the same surfaces after a convolution with a mask

$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$

.

The left and right value columns of Figure 2 show the corresponding entropy values. One can see how the entropy measure is unable to make a difference between the three non-convolved two-unit distributions. On the other hand, it is evident from the entropy values obtained after the convolution has been performed

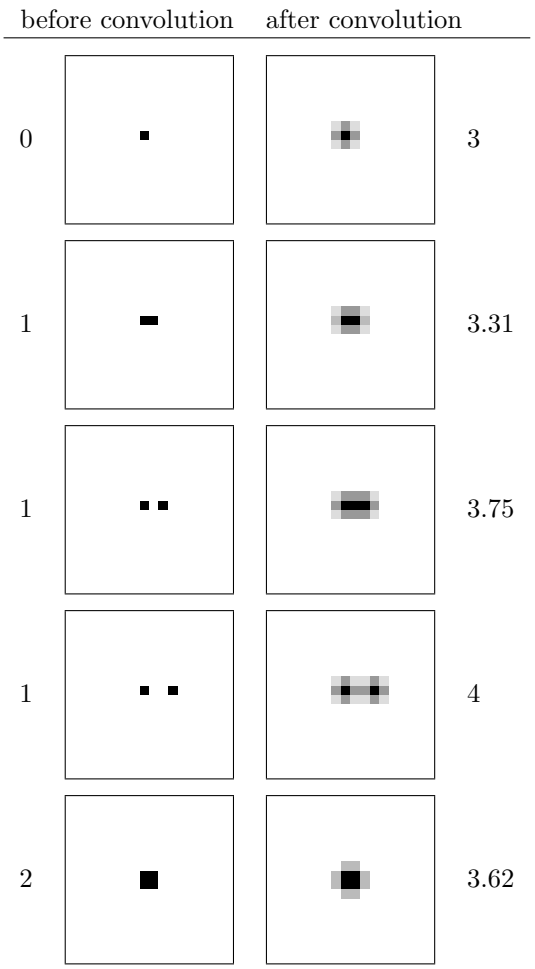

Figure 2: Entropies of value fields before and after a convolution. The gray shades have been scaled in each image separately so that the darkest shade corresponds to the largest value in that particular image.

that the nearer the two peaks are to each other, the more the distribution resembles the one-peak distribution of the first row. Even for the distribution on the last row, where the non-convolved entropy is the largest, the convolved entropy is smaller than that of any of the two-peak cases.

An example illustration with real data is again obtained from the CBIR application, using the sunsets image class and a 64×64-sized SOM trained with the *Scalable Color* descriptor. Figure 3 visualizes a real distribution of "hits" before convolution and Figure 4 after it. It can be seen in both images that the sunsets class is concentrated in a certain area of the SOM, but, on the other hand, is split in two separate regions. The observation is much clearer in the convolved surface image.

Figure 5 displays two histograms of SOM surface values obtained again with the sunsets class and the 64×64-sized SOM of the *Scalable Color* descriptor as in Figures 3 and 4. The first histogram is calculated

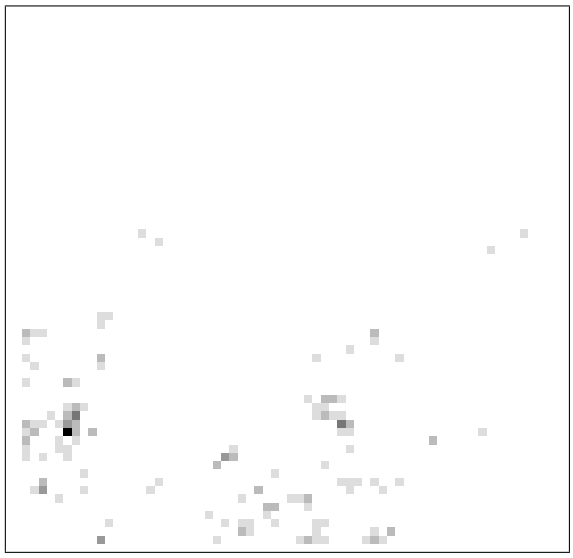

Figure 3: Distribution of the sunsets class on a 64×64-sized SOM trained with the *Scalable Color* descriptor *before* convolution.

from the non-convolved "raw" value field and it peaks strongly in two locations, corresponding to zero and one images being mapped to a particular BMU. The first peak (zero images) actually rises well above the scale of the figure, to the value 0.9. There are also weaker but still visible peaks for two, three, and four image cases. The second histogram has been obtained after the convolution and it can be seen to be much smoother, as it is the result of a low-pass filtering operation on the discrete value field. Especially, there is a notable amount of values that are slightly greater than zero.

The relative entropies of the distributions of objects can be calculated also after the convolution has been performed. Similarly as in Table 1, Table 2 shows the relative entropies of all images and the three image classes with the five feature extraction methods and four SOM grids, this time calculated after convolving the discrete value fields with triangular masks. The size of the used mask depends on the size of the SOM: symmetric triangular masks with a radius of 2, 4, 6, and 8 map units were used for the 4×4, 16×16, 64×64, and 256×256 sized SOMs, respectively.

From the resulting relative entropies in Table 2, it can be seen that the convolution further increases the entropies of the distributions, so that when using all images the entropy approaches the theoretical maximum. For the image classes, the relative entropy values continue to express the discriminating abilities of the features with the used image classes. For sunsets images, being the easiest class for the features, the relative entropies are the lowest. In this setting, the relative entropies on SOMs of different sizes are rather similar and the inverse proportionality with re-

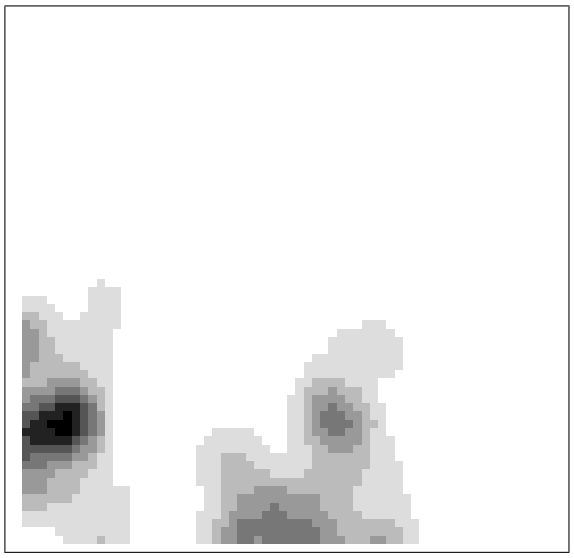

Figure 4: Distribution of the sunsets class on a 64×64-sized SOM trained with the *Scalable Color* descriptor *after* convolution with a triangular mask of size 6 map units.

spect to SOM size present in Table 1 is not observed here. These findings can be explained as resulting from the fact that the convolution spreads the dense clusters and partially fills the empty gaps between them, causing thus increased entropy.

6 Multiple Feature Extractions

In some application areas it is possible to use more than one feature extraction method in parallel. Our example case, content-based image retrieval, is such an area. In CBIR, three different feature categories are generally recognized: color, texture, and shape features. Each of them is useful in CBIR by its own right, and it is wise not to combine them all in one descriptor.

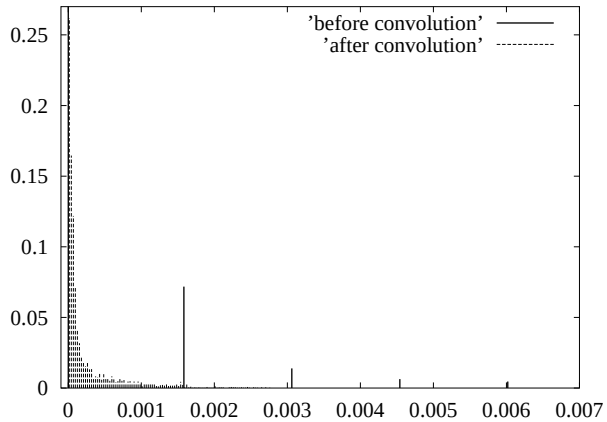

Figure 5: Histogram of SOM surface values before and after the convolution with the sunsets class distributed on the 64×64-sized SOM trained with the *Scalable Color* descriptor.

Table 2: Relative entropies of convolved SOMs with the *Color Structure* (CS), *Scalable Color* (SC), *Edge Histogram* (EH), *Homogenous Texture* (HT), and *keyword* (KW) descriptors, respectively.

	SOM size	all	faces	cars	sunsets
CS	4×4	0.997	0.936	0.982	0.780
	16×16	1.000	0.949	0.984	0.837
	64×64	1.000	0.946	0.977	0.848
	256×256	1.000	0.926	0.954	0.835
SC	4×4	1.000	0.945	0.997	0.861
	16×16	1.000	0.950	0.993	0.894
	64×64	1.000	0.944	0.980	0.889
	256×256	1.000	0.930	0.949	0.867
EH	4×4	0.994	0.947	0.986	0.734
	16×16	0.999	0.937	0.950	0.826
	64×64	0.999	0.927	0.931	0.835
	256×256	0.999	0.903	0.908	0.842
HT	4×4	0.995	0.955	0.972	0.823
	16×16	0.999	0.949	0.979	0.879
	64×64	0.999	0.938	0.971	0.889
	256×256	1.000	0.923	0.946	0.880
KW	4×4	0.996	0.797	0.913	0.826
	16×16	0.999	0.851	0.853	0.844
	64×64	0.996	0.803	0.764	0.822
	256×256	0.992	0.772	0.713	0.783

Let us denote by $P = (P_0, P_1, \ldots, P_{K-1})$ and $Q = (Q_0, Q_1, \ldots, Q_{K-1})$ the probability distributions on two SOM surfaces, as explained in Section 4, obtained from the same data with two different feature extractions. Then the question of the independence of the features arises. As entropies $H(P)$ and $H(Q)$ measure the distributions of the single feature vectors, *mutual information* $I(P, Q)$ can be used for studying the interplay between them:

$$I(P,Q) = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} R_{ij} \log \frac{R_{ij}}{P_i Q_j}$$

where R_{ij} is the joint probability distribution.

The dependency of the mutual information on the SOM size can be canceled by dividing it with the smaller of the two entropies:

$$\bar{I}(P,Q) = \frac{I(P,Q)}{\min\{H(P), H(Q)\}}$$

Table 3 illustrates the relative mutual information. The pairwise relative mutual information of the four studied MPEG-7 descriptors and the *keyword* feature are shown in two SOM sizes (4×4 and 16×16 map

Table 3: Relative mutual informations of 4×4 and 16×16 sized SOMs in the upper and lower triangles of the table, respectively.

	CS	SC	EH	HT	nHT	KW
CS	–	0.16	0.077	0.057	0.057	0.032
SC	0.33	–	0.016	0.019	0.018	0.020
EH	0.17	0.13	–	0.14	0.14	0.031
HT	0.19	0.15	0.21	–	0.64	0.013
nHT	0.18	0.14	0.21	0.62	–	0.017
KW	0.17	0.17	0.16	0.15	0.15	–

units). In addition, a normalized version (normalized to zero mean, unit variance) of the *Homogenous Texture* descriptor was used to train separate SOMs. With the larger SOMs, the measure is less informative as the number of images sharing a common BMU becomes too small. When using larger SOMs with respect to the number of available data items, information about the spatial configuration of the data on the SOMs should be taken into account.

The results in Table 3 show that the SOMs trained with the unnormalized and normalized versions of the *Homogenous Texture* feature (HT and nHT) have by far the largest values for mutual information as can be well expected. Of the separate features, the two color-based features, *Color Structure* (CS) and *Scalable Color* (SC) have the largest value on both SOMs, which again was to be expected. Additionally, the mutual information of *Edge Histogram* (EH) and *Homogenous Texture* (HT) is high on the smaller SOM, but not so much on the larger SOM with more resolution. Of the visual features, the pair of the smallest relative mutual information is *Edge Histogram* and *Scalable Color*, which could thus be used together to produce the least correlated joint information.

7 Conclusions

In this paper, we have shown how distributions of feature vectors calculated from objects of mutual semantic similarity can be studied on the SOM surfaces. We demonstrated that the entropy of the distribution characterizes quantitatively the compactness of an object class. More informative results were obtained if the SOM surface was low-pass filtered prior to the calculation of the entropy.

When studying two different SOMs created with different feature extraction methods, we showed that the mutual information of the distributions could be used to identify both the most similar and the most uncorrelated pair of features.

The described techniques can be utilized in selecting an effective set of features in various application areas.

Acknowledgements

This work was supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *New information processing principles*, the latter being part of the Finnish Centre of Excellence Programme.

References

[1] Alberto Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, Inc., 1999.

[2] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, third edition, 2001.

[3] Pasi Koikkalainen. Progress with the tree-structured self-organizing map. In A. G. Cohn, editor, *11th European Conference on Artificial Intelligence*. European Committee for Artificial Intelligence (ECCAI), John Wiley & Sons, Ltd., August 1994.

[4] Pasi Koikkalainen and Erkki Oja. Self-organizing hierarchical feature maps. In *Proceedings of International Joint Conference on Neural Networks*, volume II, pages 279–284, San Diego, CA, USA, 1990.

[5] Markus Koskela and Jorma Laaksonen. Using long-term learning to improve efficiency of content-based image retrieval. In *Proceedings of Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003)*, pages 72–79, Angers, France, April 2003.

[6] Jorma Laaksonen, Markus Koskela, Sami Laakso, and Erkki Oja. Self-organizing maps as a relevance feedback technique in content-based image retrieval. *Pattern Analysis & Applications*, 4(2+3):140–152, June 2001.

[7] Jorma Laaksonen, Markus Koskela, and Erkki Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, July 2002.

[8] Michael S. Lew, editor. *Principles of Visual Information Retrieval*. Springer-Verlag, 2001.

[9] MPEG-7 Overview (version 8.0), July 2002. ISO/IEC JTC1/SC29/WG11 N4980.

[10] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, March 1999.