Modelling image complexity by independent component analysis, with application to content-based image retrieval

Jukka Perkiö[1,2] and Aapo Hyvärinen[1,2,3]

[1] Helsinki Institute for Information Technology,
[2] Department of Computer Science,
[3] Department of Mathematics and Statistics
P.O. Box 68, FI-00014, University of Helsinki, Finland
{jperkio,ahyvarin}@cs.helsinki.fi
http://www.hiit.fi

Abstract. Estimating the degree of similarity between images is a challenging task as the similarity always depends on the context. Because of this context dependency, it seems quite impossible to create a universal metric for the task. The number of low-level features on which the judgement of similarity is based may be rather low, however. One approach to quantifying the similarity of images is to estimate the (joint) complexity of images based on these features. We present a novel method to estimate the complexity of images, based on ICA. We further use this to model joint complexity of images, which gives distances that can be used in content-based retrieval. We compare this new method to two other methods, namely estimating mutual information of images using marginal Kullback-Leibler divergence and approximating the Kolmogorov complexity of images using Normalized Compression Distance.

Key words: Image complexity; ICA; NCD; Kolmogorov complexity

1 Introduction

Measuring image similarity is not a simple task. Similarity is always defined at two levels: The semantics and the syntax of an image. Two images containing cars may be judged similar based on the fact that there are cars in both of the images but on the other hand they may be judged dissimilar based on the make of the car. This is an example of the semantic level. Similarly two versions of the same image may be judged similar or dissimilar based on – for example – different colorspaces, which is an example of the syntactic level.

The semantics of an image are dependent on the context. When one decides whether the images containing cars are similar, it is the context that defines whether similarity is dependent on the bare fact that there are cars in the image or whether the make of the cars is also important. The less context-dependently one defines similarity, the simpler the interpretation of semantics is and the more general the similarity measure is.

There certainly exist features which give a lot of information on the similarity of images. The problem is that sometimes one simply does not know what the discriminating features are and sometimes there are no clear dominating features. In general, manually selecting one or a few simple low level features works only for specific tasks, whereas using a large number of low level features raises the complexity of estimation process to impractical level.

The complexity of images is a universal property which is related to similarity. Intuitively it may be easy to decide between two images which one is more complex, but one can also imagine situation when semantically completely different images may appear equally complex. This is not a desirable result, hence complexity alone may not be very good measure of similarity or distance between images. If one is mostly interested in pair-wise distances, one can try remedy this by looking at the joint complexity of images versus the complexity of images separately [7]. The difference between complexity of a single image and the joint complexity of two images is more descriptive than arbitrary complexity values of arbitrary images alone. Of course – depending on the method used – these values have to be normalized appropriately.

Whether the difference between joint complexity and complexity of single image is good enough measure of similarity depends on the task in hand. As in all data-analysis, results depend a lot on the preprocessing and especially feature extraction. For example, measuring general image similarity may not require any specific feature extraction (pixel level intensity and color are the lowest level features and directly available) but if one wants to perform more specific tasks, the importance of features used grows. For specific tasks, there may be well established working methods and complexity-based measures of similarity may not be very attractive. On the other hand, the attractiveness of using complexity-based similarities is based on its universality, and the fact that in principle one can do this completely model-free—although the results will depend on the complexity measure chosen.

Two options for estimating the complexity of images are Shannon's classical information theory and algorithmic information theory. Although fundamentally different in some basic concepts, the two theories are connected [3]. Classical information theory have been utilized extensively in data analysis for clustering, feature selection, blind signal separation, etc. These methods maximize or minimize certain information theoretic measures. Kolmogorov complexity based similarity measures have been studied and used for different data [7, 2]. In those papers the authors develop and use data compression based techniques to approximate the Kolmogorov complexity. They call the distance measure *normalized compression distance* [7].

Complexity-based methods have been applied to image analysis. In [1] these methods are applied to earth observation imagery and in [8] approximation of Kolmogorov complexity is applied to image classification. Both of the above papers use normalized compression distance as the measure of difference, hence they belong to the methods based on algorithmic information theory.

In this paper we present a new method based on a model that approximates the complexity of the data. The model that we use is *independent component analysis* (ICA) [5]. We first build the ICA model and then estimate the image complexity from the properties of the model. Our method can be justified from the information-theoretic framework, and it incorporates the sparsity of data in the complexity measure. Sparsity is a prominent statistical property of images which may not be well-captured by other methods.

The rest of this paper is organized as follows: In Section 2 we present our method and discuss it in the context of other complexity measures, namely measuring complexity by marginal Kullback-Leibler divergence and approximating Kolmogorov complexity. In Section 3 we present experiments using natural images and in Section 4 we present our conclusions.

2   Estimating image complexity

Given a general complexity measure $\mathcal{C}(x)$ for an image x one can try to estimate similarities between images. A naive assumption would be that the difference $|\mathcal{C}(x_0) - \mathcal{C}(x_1)|$ tells the similarity between images x_0 and x_1. Unfortunately such a general complexity measure does not exist. The closest thing that exists is the Kolmogorov complexity or algorithmic entropy $K(x)$ of the image (or any string) x. Kolmogorov complexity is not computable, however.

Even if the complexity measure $\mathcal{C}(x)$ existed or Kolmogorov complexity were computable, their value as measures of similarity would be questionable. Intuitively, the similarity between images does not always equal to the difference in complexity. This is because the context plays an important role even at the syntactic level, although not as much as in the semantic level.

An obvious way of introducing the context in the picture is to estimate the joint complexity of images. This is still at a very low level but estimating the complexity in the context of other image versus the complexity of single image is more informative than arbitrary complexity values alone. Hence we are interested in the distance that is defined as

$$D(x_0, x_1) = \mathcal{C}(x_0|x_1) - \min\{\mathcal{C}(x_0), \mathcal{C}(x_1)\}, \tag{1}$$

assuming that the joint complexity is symmetric, i.e. $\mathcal{C}(x_0|x_1) = \mathcal{C}(x_1|x_0)$. Also one wants to ensure that the distance is normalized appropriately.

As it was noted above the ideal complexity measure does not exist and Kolmogorov complexity is not computable. One can approximate the ideal complexity measure in different manners, however. Shannon's information theory introduced the concept of entropy, which is easily estimated from data. Entropy can be seen also as a statistical measure of complexity. Even though Kolmogorov complexity is not computable it can be approximated using compression based methods. Complexity can also be estimated from a model that approximates the log-pdf of data as we do in this paper.

2.1 Relative entropy as distance measure

Given a discrete probability distribution P Shannon's entropy $H(x)$ is defined as

$$H(x) = -\sum_x P(x) \log P(x). \tag{2}$$

Entropy is a natural measure of complexity, since it estimates the degree of uncertainty with random variables. Intuitively it is appealing: The more uncertain we are about an outcome of an event, the more complex the phenomenon (data, image, etc.) is.

Given another distribution Q, the Kullback-Leibler divergence is defined as

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{3}$$

KL-divergence is also called relative entropy and it can be interpreted as the amount of extra bits that is needed to code samples from P using code from Q. If the distributions are the same, the need for extra information is zero and the divergence is zero as well. KL-divergence is nonnegative but not symmetric and as such it can not be used directly as a measure of distance or dissimilarity between distributions. The symmetry is easy to obtain, however, just by calculating and summing the KL-divergence from Q to P and from P to Q, hence the symmetric[1] version is simply

$$KLS(P, Q) = KL(P||Q) + KL(Q||P). \tag{4}$$

This is not a true metric but it can be used directly as measure of distance or dissimilarity between distributions.

Using the symmetric version of KL-divergence (Eq. 4) as the pair-wise distance between two images is straight forward. It is not quite the ideal distance measure in Eq. 1, but it captures the idea of estimating the complexity in the context of another image.

2.2 Algorithmic complexity

Kolmogorov complexity $K(x)$ of string x is the length of shortest program p using given description language L on a universal Turing machine U that produces the string x.

$$K(x) = \min_p \{|p| : U(p) = x\}, \tag{5}$$

where $|p|$ denotes the length of the program p. Kolmogorov complexity is not computable.

Conditional Kolmogorov complexity $K(x_0|x_1)$ of string x_0 given string x_1 is the length of shortest program that produces output x_0 from input x_1

$$K(x_0|x_1) = \min_p \{|p| : U(p|x_1) = x_0\}. \tag{6}$$

---

[1] Actually this is the original formulation that Kullback and Leibler give [6].

Normalized information distance [7] is based on the Kolmogorov complexity and is defined as

$$NID(x_0, x_1) = \frac{\max\{K(x_0|x_1), K(x_1|x_0)\}}{\max\{K(x_0), K(x_1)\}}. \tag{7}$$

As Kolmogorov complexity is not computable, NID neither is computable. It can be approximated, however, using the normalized compression distance (NCD) [7]. NCD approximates NID by using a real world compressor C and it is defined as

$$NCD(x_0, x_1) = \frac{C(x_0, x_1) - \min\{C(x_0), C(x_1)\}}{\max\{C(x_0), C(x_1)\}}. \tag{8}$$

To use the NCD for measuring pair-wise distances between images one just compresses images separately and concatenated and observes the difference between the compression results.

2.3   Using ICA as an approximation for entropy

A practical approximation of entropy can be attained by fixing some model which approximates the log-pdf. We propose here to use this approach, in connection with the model of independent component analysis (ICA), or equivalently sparse coding [4]. These models are widely used in statistical image modelling. In ICA, the pdf is approximated as

$$\log p(\mathbf{x}; \mathbf{W}) = \sum_i G(\mathbf{w}_i^T \mathbf{x}) + \log|\det \mathbf{W}| \tag{9}$$

where n is the dimension of the space, the $\mathbf{w}_i$ are linear features, collected together in the matrix $\mathbf{W}$. The function G is a non-quadratic function which measures the sparsity of the features; typically $G(u) = -|u|$ or $G(u) = -\log\cosh(u)$ are used. The latter can be considered as a smooth approximation of the former, which improves the convergence of the algorithm. A number of algorithms have been developed for estimation of the ICA model, in particular the matrix of features $\mathbf{W}$ [5].

After the model has been estimated, we can then approximate the complexity of $\mathbf{x}$ as

$$-E\{\log p(\mathbf{x}; \mathbf{W})\} = E\{-\sum_i G(\mathbf{w}_i^T \mathbf{x}) - \log|\det \mathbf{W}|\} \tag{10}$$

where the expectation is taken, in practice, over the sample.

An intuitive interpretation of the ensuing complexity measure is also possible. First, note that in ICA, the variance of the $\mathbf{w}_i^T \mathbf{x}$ is fixed to one. The first term on the right-hand-side in (10) can thus be considered as a measure of sparsity. In other words, it measures the non-Gaussian aspect of the components, completely neglecting the variance-covariance structure of the data. In fact, this term is minimized by sparse components. What is interesting is that the second term

does measure the covariance structure. In fact, we have in ICA the well-known identity

$$2|\det \mathbf{W}| = |\det \mathbf{W}\mathbf{W}^T| = |\det C(\mathbf{x})|^{-1} \tag{11}$$

where $C(\mathbf{x})$ is the covariance matrix of the data. This formula shows that the second term in (10) is a simple function of the data covariance matrix. In fact, $\log|\det \mathbf{W}|$ is maximum if the data covariance has a minimum determinant. A minimum determinant for a covariance matrix is obtained if the variances are small in general, or, what is more interesting for our purposes, if some of the projections of the data have a very small variances. Since in ICA, we constrain the variances of the components to be equal to one, only the latter case is possible. Thus, our entropy measure becomes small if the data is concentrated in a subspace of a limited dimension.

Thus, this measure of entropy (complexity) is small if the components are very sparse, or if the data is concentrated in a subspace of limited dimension, both of which are in line with our intuition of structure of multivariate data.

Practicalities Remembering the ideal complexity distance in Eq. 1 we present some remarks about the use of ICA model.

- Assuming that we want to estimate the distance between two images, we estimate the ICA model from both images separately and combined.
- The complexity value that we get using Eq. 10 is normalized in similar manner as the NCD in Eq. 8.
- In practice the ICA model for images is estimated from data that contains a large number of randomly sampled image patches.

3   Experiments

We wanted to evaluate how our method relates to other complexity based methods. For that we performed experiments using a subset of images in the University of Washington content-based image retrieval database[2].

We estimated the pair-wise distances between the subset of images using ICA, marginal KL-divergence and NCD. All the images were in RGB colorspace. The experiments were conducted as follows:

- The ICA models were estimated from data that contained 10,000 16×16 randomly sampled patches for each image. The data was normalized to be of zero-mean and of unit variance as is customary.
- Marginal KL-divergences were estimated from RGB intensity histograms.
- NCDs were estimated from RGB image matrices using zlib[3], which uses the DEFLATE algorithm for compression.

---

[2] http://www.cs.washington.edu/research/imagedatabase/groundtruth/
[3] http://www.zlib.net/

All the experiments were implemented in Python[4]. KL-divergence and NCD experiments were done for comparison. At this point we are not interested in image classification or clustering: We want to inspect the results visually and using some quantitative measure.

For the quantitative evaluation we turned the distances into rankings. This was done relative to every image in the data set. Rankings capture quite nicely the essential differences between the methods. For the rankings we calculated the Spearman rank correlation in order to understand the differences. Figure 1 shows for each image the rank correlation between all the methods we tried.
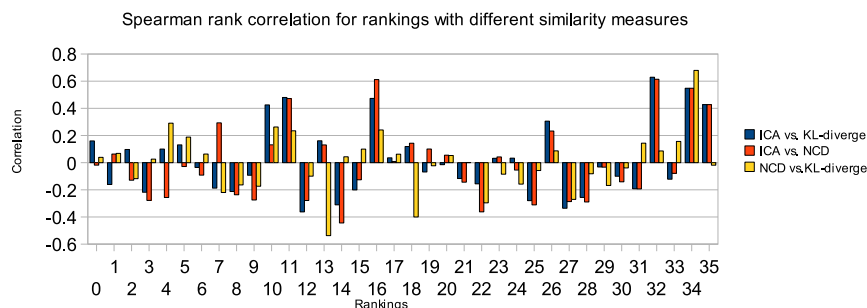
Spearman rank correlation for rankings with different similarity measures

Fig. 1. The Spearman rank correlation between the different methods is showed when the test images are ranked relative to every other image shown. Within each experiment and ranking, the significance level $\alpha = 0.05$ is attained by an absolute value 0.26 or higher of correlation.

First, we observe that the correlations between rankings differ significantly depending on the image the ranking is relative to. This is actually somewhat surprising. Second, we notice that for most statistically significant correlations our method agrees more with both the KL-divergence- and the NCD-based methods, whereas the KL-divergence and NCD rankings are less correlated. This may suggest that our method captures more general features than the other two. Whether this works in real world applications is not sure though. Lastly we also observe surprisingly many negative correlations and the average correlation is rather low. This is different though if we only observe the absolute values of the correlation, which is justifiable, since correlation – negative or positive – is interesting, whereas non-correlated data does not tell us much.

Images 2 and 3 show two-dimensional Sammon mappings estimated from the pair-wise distances between images using ICA, KL-divergence and NCD respectively. Image 4 show example rankings for one reference image using all the methods. Visually inspecting it is clear that all the methods produce different results. It is harder to judge one better than the other, however.

---

[4] http://www.python.org

Fig. 2. Two-dimensional Sammon mapping calculated from the pair-wise distances between images, when the distances were estimated using ICA as an approximation for entropy. Even though the Sammon mapping is used to preserve the distances in the two dimensional visualization as well as possible, the individual rankings are not directly comparable to the mapping.

It seems that the ICA method (Fig. 2, Fig. 4 left) is affected mostly by the texture of the images. It is able to nicely group different kinds of trees according to their appearance. The method do not seem to be very specific with regards to the grass appearing in the images.

For the marginal KL-divergence visual experiment (Fig. 3 left, Fig. 4 middle) the first impression is that it seem to be mostly affected by the different intensity in the lighting in the images. That is actually quite natural since the distances were estimated from RGB-intensity histograms. Nevertheless it also produces reasonable results.

The results for NCD visual experiment (Fig. 3 right, Fig. 4 right) are quite intuitive also but it is quite hard to find a common factor on which the grouping is based. NCD seems to be mostly affected by the complexity of rather low level features.

Finally one have to note that at their current state none of the methods presented can compete with more specialized application specific image similarity measures. The similarity that the methods measure is rather generic low level

Fig. 3. Two-dimensional Sammon mapping calculated from the pair-wise distances between images, when the distances were estimated using the KL-divergence (left) and compression-based approximation for Kolmogorov complexity, NCD (right). Even though the Sammon mapping is used to preserve the distances in the two dimensional visualization as well as possible, the individual rankings are not directly comparable to the mapping.

similarity. On the other hand that is exactly what one expects from complexity based similarity measures.

4 Conclusions

We have presented a novel method to estimate image complexity in order to derive a pair-wise similarity measure for natural images. Our method is based on using ICA model to estimate the entropy of images separately and combined. The similarity is derived from the normalized difference between the single image complexity and the pair-wise complexity. This method is comparable but not similar to other complexity based measures such as normalized compression distance and other information theoretic entropy based methods.

Based on quantitative analysis our method seem to be somewhere in between NCD and KL-divergence based distance measures. Visually all the methods tried, produce reasonable results, the ICA method being more responsive to textures.

For future work one has to consider applications of the method for clustering and classification, if not for other reasons than to get more decisive quantitative results than those obtained from the present analysis.

Acknowledgements

Reference image

ICA   KL-divergence   NCD
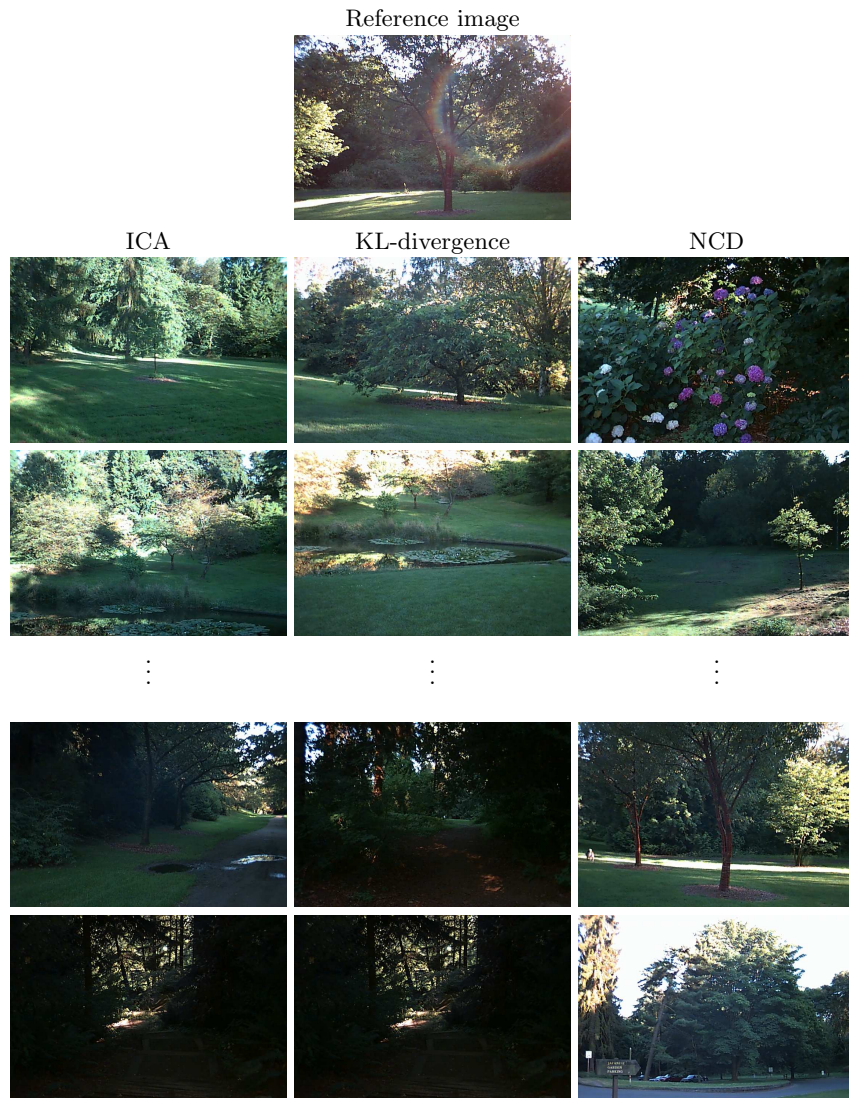

Fig. 4. An example of rankings produced by the three methods. The four rows below the reference image show two most similar and two least similar images to the reference image. The columns are from left to right ICA, KL-divergence and NCD.

References

1. D. Cerra, A. Mallet, L. Gueguen, and Mihai Datcu. Complexity based analysis of earth observation imagery: an assessment. In *ESA-EUSC 2008: Image Information Mining: pursuing automation of geospatial intelligence for environment and security*, March 2008.
2. R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Transactions in Information Theory 51*, pages 1523–1545, 2005.
3. P. Grunwald and P. Vitanyi. Shannon information and kolmogorov complexity. http://arxiv.org/abs/cs.IT/0410002, October 2004.
4. A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics*. Springer-Verlag, 2009.
5. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.
6. S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics 22 (1)*, pages 79–86, 1951.
7. M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *IEEE Transactions in Information Theory 50*, pages 3250–3264, 2004.
8. M. Li and Y. Zhu. Image classification via lz78 based string kernel: A comparative study. In *PAKDD*, pages 704–712, 2006.