

Image Similarity based on Hierarchies of ICA Mixtures

Arturo Serrano¹, Addisson Salazar¹, Jorge Igual¹, Luis Vergara¹

¹ Universidad Politécnica de Valencia, Departamento de Comunicaciones, Camino de Vera
s/n, 46022 Valencia, Spain
asalazar@dcom.upv.es, jigual@dcom.upv.es

Abstract. This paper presents a novel algorithm to build hierarchies from independent component analyzer mixtures and its application to image similarity measure. The hierarchy algorithm composes an agglomerative (bottom-up) clustering from the estimated parameters (basis vectors and bias terms) of the ICA mixture. Merging at different levels of the hierarchy is made using the Kullback-Leibler distance between clusters. The procedure is applied to merge similar patches on a natural image, to group different images of an object, and to create hierarchical levels of clustering from images of different objects. Results show suitable image hierarchies obtained by clustering from basis functions to higher-level structures.

1 Introduction

Independent component analyzers (ICA) mixture models were introduced in [1] considering a source model switching between Laplacian and bimodal densities. Recently this model has been relaxed using generalised exponential sources [2], self-similar areas as a mixture of Gaussians sub-features [3], and sources with non-Gaussian structures recovered by a learning algorithm using Beta divergence [4]. Real applications of those works span: separation of eye-movement artefacts from EEG recordings, separating ‘back-ground’ brain tissue, fluids and tumours in fMRI images, and the separation of voices and background music in conversations.

It is well known that local edge detectors can be extracted from natural scenes by standard ICA algorithms as Infomax [5], or fastICA [6] or new approaches as Linear Multilayer ICA [7]. In addition there is neurophysiological evidence that suggest relation of primary visual cortex activities with the detection of edges, and some theoretical dynamic models of abstraction process from visual cortex to higher-level abstraction has been proposed [8].

The contribution of this paper is to provide a new algorithm to process the parameters of ICA mixtures in order to obtain hierarchical structures from the basis function level (edges) to higher levels of clustering. Particularly the algorithm is applied to image analysis obtaining promising results in discerning object similarity and suitable levels of hierarchies by processing image patches. This kind of feedforward process would suggest some relation with abstraction. The algorithm is agglomerative and uses the symmetric Kullback-Leibler distance [9] to select the grouping of the clusters at each level.

2 Hierarchy of ICA Mixtures

2.1 Estimation of the ICA mixture parameters

In the ICA mixture model, the observation vectors \mathbf{x} are modelled as the result of applying a linear transformation \mathbf{A}_k ($\mathbf{W}_k = \mathbf{A}_k^{-1}$ is the filter matrix) to a vector \mathbf{s}_k (sources), whose elements are independent random variables, plus a bias vector \mathbf{b}_k , for all the classes $C_k, (k = 1 \dots K \text{ number of ICAs})$. The probability of every available observation vector can be separated into the contributions due to every class.

An iterative learning algorithm based on maximum-likelihood estimation (MLE) is used to adapt the parameters of the ICA mixtures, i.e., the basis functions and the bias terms of each class, using gradient ascent [1]. To estimate the probability density function of the sources different priors could be used as Laplacian [1] or non-parametric densities [10].

2.2 Agglomerative clustering

From the estimated ICA mixture parameters, a procedure that follows a bottom-up agglomerative scheme for merging the mixtures was developed.

The conditional probability density of \mathbf{x} for cluster $C_k^h, k = 1, 2, \dots, K - h + 1$ in level $h = 1, 2, \dots, K$ is $p(\mathbf{x} / C_k^h)$. At the first level, $h = 1$, it is modelled by K ICA mixtures, i.e., $p(\mathbf{x} / C_k^1)$ is:

$$p(\mathbf{x} / C_k^1) = |\det \mathbf{A}_k^{-1}| p(\mathbf{s}_k), \mathbf{s}_k = \mathbf{A}_k^{-1}(\mathbf{x} - \mathbf{b}_k) \quad (1)$$

At each consecutive level, two clusters are merged according to some minimum distance measure until we reach at level $h = K$ only one cluster.

As distance measure we use the symmetric Kullback-Leibler distance between the ICA mixtures. It is defined for the clusters u, v by:

$$D_{\text{KL}}(C_u^h, C_v^h) = \int p(\mathbf{x} / C_u^h) \log \frac{p(\mathbf{x} / C_u^h)}{p(\mathbf{x} / C_v^h)} d\mathbf{x} + \int p(\mathbf{x} / C_v^h) \log \frac{p(\mathbf{x} / C_v^h)}{p(\mathbf{x} / C_u^h)} d\mathbf{x} \quad (2)$$

For level $h = 1$, from (2), we can obtain (we write $p_{\mathbf{x}_u}(\mathbf{x}) = p(\mathbf{x} / C_u^1)$ and omit the superscript $h = 1$ for brevity):

$$D_{\text{KL}}(C_u, C_v) = D_{\text{KL}}(p_{\mathbf{x}_u}(\mathbf{x}) // p_{\mathbf{x}_v}(\mathbf{x})) = \int p_{\mathbf{x}_u}(\mathbf{x}) \log \frac{p_{\mathbf{x}_u}(\mathbf{x})}{p_{\mathbf{x}_v}(\mathbf{x})} d\mathbf{x} + \int p_{\mathbf{x}_v}(\mathbf{x}) \log \frac{p_{\mathbf{x}_v}(\mathbf{x})}{p_{\mathbf{x}_u}(\mathbf{x})} d\mathbf{x} \quad (3)$$

where, imposing the independence hypothesis and supposing that both clusters have the same number of sources M for simplicity (assuming that sources follow the same model and the data they draw are on the same space):

$$\begin{aligned}
p_{\mathbf{x}_u}(\mathbf{x}) &= \frac{\prod_{i=1}^M p_{s_{u_i}}(s_{u_i})}{|\det \mathbf{A}_u|}, \quad s_{u_i} = \mathbf{A}_{u_i}^{-1}(\mathbf{x} - \mathbf{b}_{u_i}) \\
p_{\mathbf{x}_v}(\mathbf{x}) &= \frac{\prod_{j=1}^M p_{s_{v_j}}(s_{v_j})}{|\det \mathbf{A}_v|}, \quad s_{v_j} = \mathbf{A}_{v_j}^{-1}(\mathbf{x} - \mathbf{b}_{v_j})
\end{aligned} \tag{4}$$

The pdf of the sources is approximated by a non-parametric kernel-based density for both clusters:

$$p_{s_{u_i}}(s_{u_i}) = \sum_{n=1}^N a e^{-\frac{1}{2} \left(\frac{s_{u_i} - s_{u_i}(n)}{h} \right)^2}, \quad p_{s_{v_j}}(s_{v_j}) = \sum_{n=1}^N a e^{-\frac{1}{2} \left(\frac{s_{v_j} - s_{v_j}(n)}{h} \right)^2} \tag{5}$$

where again for simplicity we have assumed the same kernel function for all the clusters, with the parameters a, h and number of samples N adapted to each cluster. Note that this corresponds to a Gaussian mixture model where the number of Gaussians is maximum (one for every observation) and the weights are equal. Reducing to standard mixture of Gaussians does not help in order to compute the Kullback-Leibler distance because there is not analytical solution to it. Therefore, we prefer to maintain the non parametric approximation of the pdf in order to model more complex distributions than a mixture of a small finite number of Gaussians.

The symmetric Kullback-Leibler distance between the clusters u, v can be expressed such as:

$$D_{\text{KL}}(p_{\mathbf{x}_u}(\mathbf{x}) // p_{\mathbf{x}_v}(\mathbf{x})) = -H(\mathbf{x}_u) - H(\mathbf{x}_v) - \int p_{\mathbf{x}_u}(\mathbf{x}) \log p_{\mathbf{x}_v}(\mathbf{x}) d\mathbf{x} - \int p_{\mathbf{x}_v}(\mathbf{x}) \log p_{\mathbf{x}_u}(\mathbf{x}) d\mathbf{x} \tag{6}$$

where $H(\mathbf{x})$ is the entropy, defined as $H(\mathbf{x}) = -E[\log p_{\mathbf{x}}(\mathbf{x})]$. To obtain the distance, we have to calculate the entropy for both clusters and the cross-entropy terms $E_{\mathbf{x}_v}[\log p_{\mathbf{x}_u}(\mathbf{x})]$, $E_{\mathbf{x}_u}[\log p_{\mathbf{x}_v}(\mathbf{x})]$.

The entropy for the cluster u can be calculated through the entropy of the sources of that cluster considering the linear transformation of the random variables and their independence (4):

$$H(\mathbf{x}_u) = \sum_{i=1}^M H(s_{u_i}) + \log |\det \mathbf{A}_u| \tag{7}$$

The entropy of the sources can not be analytically calculated. Instead, we can obtain a sample estimate $\hat{H}(s_{u_i})$ using the training data. Denote the i -th source obtained for the cluster u by $\{s_{u_i}(1), s_{u_i}(2), \dots, s_{u_i}(Q_i)\}$. The entropy can be approximated as follows:

$$\begin{aligned}\hat{H}(s_{u_i}) &= -\hat{E}[\log p_{s_{u_i}}(s_{u_i})] = -\frac{1}{Q_i} \sum_{n=1}^{Q_i} \log p_{s_{u_i}}(s_{u_i}(n)), \\ p_{s_{u_i}}(s_{u_i}(n)) &= \sum_{l=1}^N a e^{-\frac{1}{2} \left(\frac{s_{u_i}(n) - s_{u_i}(l)}{h} \right)^2}\end{aligned}\quad (8)$$

The entropy of $H(\mathbf{x}_v)$ is obtained analogously:

$$\begin{aligned}H(\mathbf{x}_v) &= \sum_{i=1}^M H(s_{v_i}) + \log |\det \mathbf{A}_v| \approx \sum_{i=1}^M \hat{H}(s_{v_i}) + \log |\det \mathbf{A}_v| \\ \hat{H}(s_{v_i}) &= -\frac{1}{Q_i} \sum_{n=1}^{Q_i} \log p_{s_{v_i}}(s_{v_i}(n)), p_{s_{v_i}}(s_{v_i}(n)) = \sum_{l=1}^N a e^{-\frac{1}{2} \left(\frac{s_{v_i}(n) - s_{v_i}(l)}{h} \right)^2}\end{aligned}\quad (9)$$

with $\hat{H}(s_{v_i})$ defined analogously to (8). Following the same procedure for j -th source we can estimate $\hat{H}(s_{v_j})$.

Once the entropy is computed, we have to obtain the cross-entropy terms. After some operations and considering the relationships $\mathbf{x} = \mathbf{A}_u \mathbf{s}_u + \mathbf{b}_u$, $\mathbf{x} = \mathbf{A}_v \mathbf{s}_v + \mathbf{b}_v$ and thus $\mathbf{s}_v = \mathbf{A}_v^{-1} (\mathbf{A}_u \mathbf{s}_u + \mathbf{b}_u - \mathbf{b}_v)$, the independence of the sources, and that the samples for clusters u, v follow the corresponding distribution $\{s_{u_i}(1), s_{u_i}(2), \dots, s_{u_i}(Q_i)\}$, $i = 1, \dots, M$, $\{s_{v_j}(1), s_{v_j}(2), \dots, s_{v_j}(Q_j)\}$, $j = 1, \dots, M$; we can estimate,

$$\hat{H}(\mathbf{s}_v, s_{u_i}) = \frac{1}{\prod_{i=1}^M Q_i} \cdot \sum_{s_{v_1}=1}^{Q_M} \dots \sum_{s_{v_M}=1}^{Q_1} \log \sum_{n=1}^N a e^{-\frac{1}{2} \left(\frac{[\mathbf{A}_u^{-1} (\mathbf{A}_v \mathbf{s}_v + \mathbf{b}_v - \mathbf{b}_u)]_i - s_{u_i}(n)}{h} \right)^2} \quad (10)$$

and $\hat{H}(\mathbf{s}_u, s_{v_j})$ defined analogously to (10), with $\mathbf{s}_u = \mathbf{A}_u^{-1} (\mathbf{A}_v \mathbf{s}_v + \mathbf{b}_v - \mathbf{b}_u)$.

Using the terms obtained above, we can estimate the symmetric Kullback-Leibler distance between the clusters u, v :

$$D_{\text{KL}}(p_{\mathbf{x}_u} // p_{\mathbf{x}_v}(\mathbf{x})) = -\sum_{i=1}^M \hat{H}(s_{u_i}) - \sum_{j=1}^M \hat{H}(s_{v_j}) - \sum_{i=1}^M \hat{H}(\mathbf{s}_v, s_{u_i}) - \sum_{j=1}^M \hat{H}(\mathbf{s}_u, s_{v_j}) \quad (11)$$

As we can observe, the similarity between clusters depends not only on the similarity between the bias term, but the similarity between the distributions and the mixing matrices.

Once the distances are obtained for all the clusters, the two clusters with minimum distance are merged in level $h = 2$. This is repeated in every step of the hierarchy until we reach one cluster in the level $h = K$. To merge cluster in level h we can calculate the distances from the distances of level $h-1$. Suppose that from level $h-1$ to h the clusters C_u^{h-1}, C_v^{h-1} are merged in cluster C_w^h . Then, the density for the merged cluster at level h is:

$$p_h(\mathbf{x}/C_w^h) = \frac{p_{h-1}(C_u^{h-1})p_{h-1}(\mathbf{x}/C_u^{h-1}) + p_{h-1}(C_v^{h-1})p_{h-1}(\mathbf{x}/C_v^{h-1})}{p_{h-1}(C_u^{h-1}) + p_{h-1}(C_v^{h-1})} \quad (12)$$

where $p_{h-1}(C_u^{h-1})$, $p_{h-1}(C_v^{h-1})$ are the priors or proportions of the clusters u, v at level $h-1$. The rest of terms are the same in the mixture model at level h that at level $h-1$. The only difference from one level to the next one in the hierarchy is that there is one cluster less and the prior for the new cluster is the sum of the priors of its components and the density the weighted average of the densities that are merged to form it. Therefore, the estimation of the distance at level h can be done easily starting from the distances at level $h-1$ and so on until level $h=1$. Consequently, we can calculate the distances at level h from a cluster C_z^h to a merged cluster C_w^h obtained by the agglomeration of clusters C_u^{h-1} , C_v^{h-1} at level $h-1$ as the distance to its components weighted by the mixing proportions:

$$D_h(p_h(\mathbf{x}/C_w^h) // p_h(\mathbf{x}/C_z^h)) = \frac{p_{h-1}(C_u^{h-1}) \cdot D_{h-1}(p_{h-1}(\mathbf{x}/C_u^{h-1}) // p_{h-1}(\mathbf{x}/C_z^{h-1}))}{p_{h-1}(C_u^{h-1}) + p_{h-1}(C_v^{h-1})} + \frac{p_{h-1}(C_v^{h-1}) \cdot D_{h-1}(p_{h-1}(\mathbf{x}/C_v^{h-1}) // p_{h-1}(\mathbf{x}/C_z^{h-1}))}{p_{h-1}(C_u^{h-1}) + p_{h-1}(C_v^{h-1})}. \quad (13)$$

3 Application on image data

ICA can be used to analyze image patches as a linear superposition of basis functions. Those vectors have been related with the detection of borders in natural images [5]. Therefore basis functions have a physical relation with objects and they can be used to measure the similarity between objects based on ICA decomposition. In image patches decomposition, the set of independent components is larger than what can be estimated at one time, and what we get at one time is an arbitrarily chosen subset [11]. Nevertheless ICA has been applied successfully in several image applications [1].

3.1 Object similarity

For the hierarchical classification of images of objects, the COIL-100 database was used [12]. The database consists of different views of objects over a dark background. The method applied to preprocess the images was this. The images were converted to greyscale, and grouped in different views in order to obtain several images to train up to three classes per object. From each image, patches of 8 by 8 pixels were randomly taken to estimate the basis function previous a whitening process, with a reduction to 40 components. A total of 1000 patches per object were extracted [5].

The basis functions of each class were then calculated with the ICA mixtures algorithm, considering supervision, and using the Laplacian prior to estimate the source pdfs. Fig. 1 shows the 40 basis functions of six classes corresponding to different views of two objects. The basis functions of Fig. 1a correspond to a box with a label inscribed whereas Fig. 1b corresponds to an apple. We can observe the similarity be-

tween the functions of each object and differences, for instance, the lower frequency in the pattern corresponding to a natural object versus the frequency in the pattern of a more artificial object.

The same data were used to measure the distance between classes estimating the symmetric Kullback-Leibler distance from the mixture matrices calculated previously, as we explain in Section 2. Distances reveal that basis functions allow finding the similarity (short distances) between classes corresponding to the same object (intra-object), whereas distances are much longer between classes of different objects (inter-object), see Table 1.

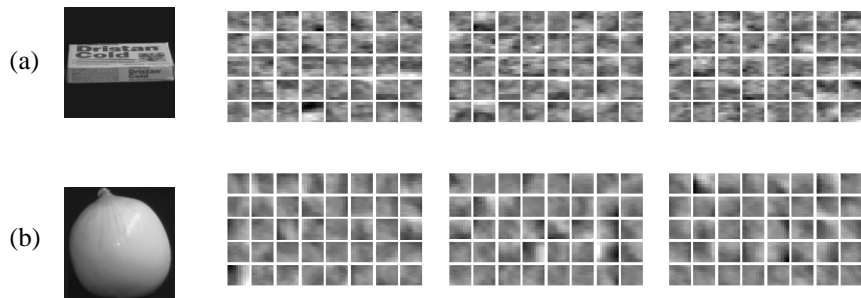


Fig. 1. Two groups of basis functions corresponding to two different objects. Basis functions at top are from a little box and basis functions at bottom are from an apple.

Table 1. Mean distances inter-object and intra-object of Fig. 1.

Object	box (a)	apple (b)
box (a)	12.89	114.90
apple (b)	114.90	13.81

Additionally, experiments in order to create a hierarchical classification of objects were performed. Thus, patches were sampled from a large number of objects, some of them very similar among themselves. A hierarchical representation was then created applying the agglomerative clustering algorithm. Fig. 2 shows an example of classification of eight objects, with three main kinds of objects. The tree outlined by the dendrogram positively shows grouping of objects based on similarity content, and suitable similarities between ‘families’ of objects, e.g., cars were more alike with cans than with apples.

3.2 Natural images

The proposed algorithm was applied to natural images in order to obtain a bottom-up structure merging several zones of an image. Fig. 3 shows an image with 9 zones, some of them clearly different and others more or less similar each other. Dendrogram of Fig. 3 shows how the zones are merged from the patches. It shows two broad kinds

of basis functions that correspond to the part of the image that mainly contains portions of sky, and those zones that correspond to patches where there is a predominant portion of stairs (high frequency).

The dendrogram also shows the distances at which the clusters are merged, it can be used as a similarity measure of the zones of the image. The bottom zones are merged at low distances due to the high similarity in borders.

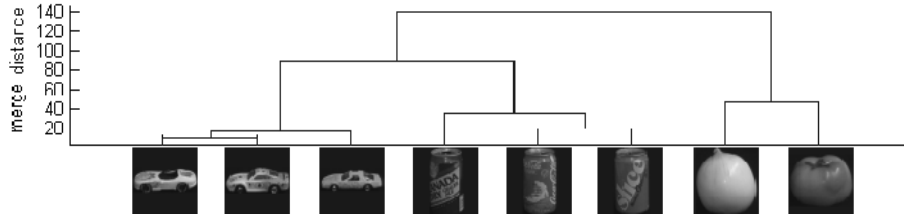


Fig. 2. Hierarchical representation of object agglomerative clustering. Three kinds of object ‘families’ are obtained.

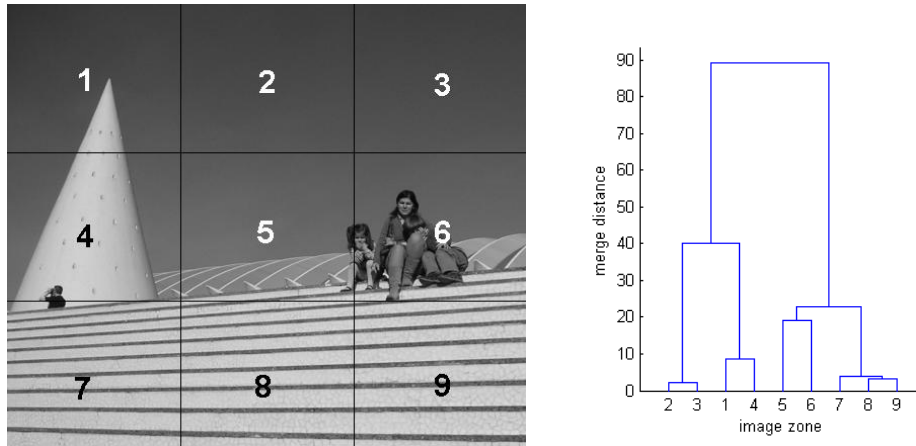


Fig. 3. (Left) Image divided in nine zones. (Right) Hierarchical representation of the zones of the image based on basis functions similarity. It shows two broad groups of zones.

4 Conclusions

The new algorithm for hierarchical ICA mixtures uses the mixture matrices to calculate distances between the distributions of the independent sources based on a symmetric Kullback-Leibler distance. The estimation of the source pdfs is made using a non-parametric kernel-based approach allowing adaptation to several kinds of densities. Clusters are merged using a bottom-up strategy defining hierarchical levels creating higher-level structures.

Results of the hierarchical algorithm application demonstrated its suitability to process image data. Image content similarity between objects based on ICA basis functions allows learning an organization of objects in higher-levels of abstraction where the more separated hierarchical levels more different the objects. Experiments with natural images showed application to image segmentation based on similarity of the different zones. The application of the procedure could be extended to unsupervised or semi-supervised classification of images in order to discover meaningful hierarchical levels.

Many potential applications of the procedure could be approached as defect classification in non-destructive testing. Hierarchical levels would represent concepts as material condition, kind of defect, defect orientation, or defect dimension [13].

Acknowledgements

This work was supported by Spanish Administration under grant TEC 2005-01820.

References

1. Lee T.W., Lewicki M.S. and Sejnowski T.J.: ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation, *IEEE Trans. on Patt. Analysis and Machine Intelligence*, v. 22, n. 10, pp. 1078-1089, 2000.
2. Penny W.D. and Roberts S.: Mixtures of independent component analyzers, *Proc. ICANN2001*, pp. 527-534, Vienna, 2001.
3. Choudrey R. and Roberts S.: Variational Mixture of Bayesian Independent Component Analysers, *Neural Computation*, v. 15, n. 1, pp. 213-252, 2003.
4. Mollah N.H., Minami M. and Eguchi S.: Exploring Latent Structure of Mixture ICA Models by the Minimum β -Divergence Method, *Neural Computation*, v. 18, n. 1, pp. 166-190, 2006.
5. Bell A.J. and Sejnowski T.J.: The 'Independent Components' of natural scenes are edge filters, *Vision Research*, v. 37, n. 23, pp. 3327-3338, 1997.
6. Van Hateren J.H. and van der Shaaf A.: Independent component filters of natural images compared with simple cells in primary visual cortex, *Proceedings of Royal Society of London: B*, v. 265, pp. 359-366, 1998.
7. Matsuda Y. and Yamaguchi K.: Linear multilayer ICA generating hierarchical edge detectors, *Neural Computation*, v. 19, n. 1, pp. 218-230, 2007.
8. Lee T.S. and Mumford D.: Hierarchical Bayesian inference in the visual cortex, *Journal of the Optical Society of America A*, v. 20, n. 7, pp. 1434-1448, 2003.
9. Mackay, D. J.: *Information theory, inference, and learning algorithms*. Cambridge University Press, 2004.
10. Vergara L., Salazar A., Igual J. and Serrano A.: Data Clustering Methods Based on Mixture of Independent Component Analyzers, *Proc. of ICA Research Network International Workshop, ICARN*, pp. 127-130, Liverpool, 2006.
11. Hyvarinen A., Hoyer P. O., and Inki M.: Topographic independent component analysis, *Neural Computation*, v. 13, n. 7, pp. 1527-1558, 2001.
12. Nene S. A., Nayar S. K. and Murase H.: *Columbia Object Image Library (COIL-100)*, Technical Report CUCS-006-96, February 1996.
13. Salazar A., Uni6 J.M., Serrano A., and Gosalbez J.: Neural networks for defect detection in non-destructive evaluation by sonic signals, *LNCS*, v. 4507, pp. 631-638, 2007.