LEARNING TOPOGRAPHIC REPRESENTATION FOR MULTI-VIEW IMAGE PATTERNS

Stan Z. Li, XiaoGuang Lv, HongJiang Zhang

Microsoft Research China, Beijing, China
http://research.microsoft.com/ szli

QingDong Fu, Yimin Cheng

Dept of Electronic Science and Technology
Univ of Science and Technology of China

Abstract

In 3D object detection and recognition, the object of interest in an image is subject to changes in view-point as well as illumination. It is benifit for the detection and recognition if a representation can be derived to account for view and illumination changes in an effective and meaningful way. In this paper, we propose a method for learning such a representation from a set of un-labeled images containing the appearances of the object viewed from various poses and in various illuminations. Topographic Independent Component Analysis (TICA) is applied for the unsupervised learning to produce an emergent result, that is a topographic map of basis components. The map is topographic in the following sense: the basis components as the units of the map are ordered in the 2D map such that components of similar viewing angle are group in one axis and changes in illumination are accounted for in the other axis. This provides a meaningful set of basis vectors that may be used to construct view subspaces for appearance based multi-view object detection and recognition.

1. INTRODUCTION

Many image and vision applications have to deal with images containing objects of interest seen from various viewing points and under various illumination conditions. A challenge for such tasks is how to represent the object under the varying conditions.

Appearance based methods [15, 16, 7] avoid difficulties in 3D modeling by using images or appearances of the object viewed from possible viewpoints. The appearance of an object in a 2D image depends on its shape, reflectance property, pose as seen from the viewing point, and the external illumination conditions. The object is modeled by a collection of appearances parameterized by pose and illumination. Object detection and recognition is performed by comparing the appearances of the object in the image and in the model.

In a view-based representation, the pose is quantized into a set of discrete values such as the view angles. A view subspace defines the manifold of possible appearances of the object viewed at a certain angle, subject to illumination. One may use one of the following two methods when constructing subspaces: (1) Quantize the pose into several discrete ranges and decompose the whole data set into corresponding subsets, each composed of appearances viewed from a particular pose; then construct a subspace for each subset. A subspace thus constructed represents the object in that pose and also explains variation in illumination. This is the strategy used in [18]. This method requires that the training data be labeled according to the pose. (2) With training data labeled and sorted according to the pose value (and perhaps also illumination values), one may be able to construct a manifold describing the distribution across views [15, 6, 1]. This requires more detailed labeling of the training data.

In the past two decades, many methods are explored to extract features from training data and these can be useful for image and vision applications. Principal component analysis (PCA) has been a popular tool in data analysis and has been used in image and vision for constructing object subspaces. PCA decorrelates second order moments corresponding to low frequency property. However, interpretation of an image has to deal with important information contained in high-order relationships among three or more image pixels, which has been ignored by PCA.

Independent component analysis (ICA) is a linear non- orthogonal transform which makes unknown linear mixtures of multidimensional random variables as statistically independent as possible. It not only decorrelates the second order statistics but also reduces higher-order statistical dependencies [5]. It extracts independent components even if their magnitudes are small whereas PCA extracts components having largest magnitudes.

Originally for solving the blind source separation problem in signal processing, the ICA mixture model is also applied to unsupervised learning of basis functions for representation of images. When performed on image patches randomly sampled from natural images, ICA produces some interesting results. Olshausen and Field [17] obtain spatially localized, oriented, bandpass basis functions comparable to those in certain wavelet tansforms. Bell and Sejnowski [3] find that independent component of natural scenes are edge-like filters. Lee, Lewicki and Sejnowski [11] derive an ICA model to represent a mixture of several mutually exclusive classes each of which is described as a linear combination of independent non-Gaussian densities. It is found that the two different class of images have different types of basis functions. In image analysis applications, ICA has also been used for face recognition and texture analysis [2, 14, 12, 13], as a hopefully a better method than PCA.

Topographic ICA (TICA) [8] is an extension to ICA, in which the independence assumption is relaxed. Higher-order dependency is allowed within a scope defined by a neighborhood system, as in self-organizing maps [10]. The TICA model thus defined has the following properties: (1) All the components are uncorrelated. (2) Components that are not neighbors to each other are independent, at least approximately. (3) Components that are neighbors tend to be active (nonzero) at the same time, *i.e.* have correlated energies, the energies being high order statistics. In TICA, every neighborhood defines the scope of one feature subspace and corresponds to one complex cell in visual cortex. TICA applied to image patches results in simultaneous emergence of topography and complex cell properties [8]. A linear representation is thus obtained in which the basis vectors and hence the coefficients have a topographic organization that reveals information on the statistical higher-order structure of the data.

In this paper, we propose a method for learning such a representation from a set of un-labeled images containing the appearances of the object viewed from various poses and in various illuminations. Topographic Independent Component Analysis (TICA) [8] is applied for the unsupervised learning to produce an emergent result, that is, a topographic map of basis components. The map is topographic in the following sense: the basis components as the units of the map are ordered in the 2D map such that view-correlated basis components are located nearby in the topographic map; more specifically, components of similar viewing angle are grouped in one axis and changes in illumination are accounted for in the other axis. The emergent topographic map not only reveals relationships between the basis components but also provides a means for describing appearances viewed from various angles. View subspaces can be constructed based on the components for appearance based multi-view object detection and recognition.

The rest of the paper is organized as follows: Section 2 introduces basic concept of ICA in signal processing and its application in image processing. Section 3 presents the use of TICA for unsupervised learning of topographic maps of basis components. Section 4 presents experimental results.

2. LEARNING OF MULTI-BASED TOPOGRAPHIC MAP

TICA is used to learn, in an unsupervised way, a topographic map of basis components from a set of un-labeled training data, such that the components are ordered according to the view.

2.1. Data Description

The training set is composed of training examples. Here in this work, each example is an image patch containing a face viewed at a certain unknown left-right rotation angle between -90° and 90° (from left side view to right side view). Every patch is normalized to the size of 20×20 pixels. Without loss of generality, left-rotated face patches are mirrored to the right-rotated, and so only images of faces rotated between 0° and 90° are used in the experiments. Fig.1 shows some examples. The main causes of variations in face images include changes in the view of face, in illumination, in facial shape.

Fig. 1. Some multi-view face samples.

Fig. 2. Schematic illustration of ICA representation of images.

2.2. Classic ICA

In ICA based image analysis, a gray-level image $\mathbf{x} = \{x(u,v)\}$, where (u,v) is the pixel location, is represented as a linear combination of m basis functions $\mathbf{b} = \{b_1(u,v), \ldots, b_m(u,v)\}$:

$$\mathbf{x}(u,v) = \sum_{i=1}^{m} b_i(u,v) s_i \tag{1}$$

as illustrated by Fig. 2, where the coefficients $\mathbf{s} = (s_1, \ldots, s_m)$ are different for each image given b's. We restrict the $b_i(u,v)$ to be an invertible linear system, so that the equation above could be inverted by using the dot-product

$$s_i = < \mathbf{w}_i, \mathbf{x} > = \sum_{u,v} w_i(u,v) x(u,v) \tag{2}$$

where the $\mathbf{w} = \mathbf{b}^{-1}$ is the inverse filter.

The crucial assumption made in ICA is that s_i are non-gaussian, mutually independent random variables. The latter means

$$p_s(\mathbf{s}) = \prod_{i=1}^{m} p_{s,i}(s_i) \tag{3}$$

where p_s denotes the density of the $\mathbf{s}$. This is a factorial coding. The ICA learning problem is to estimate both the basis functions $b_i(u,v)$ and the realizations of the s_i, for all i and (u,v), using a sufficiently large set of the training images $\{\mathbf{x}_k(u,v)\}$; so that for any given sample $\mathbf{x}_k(u,v)$ from the training set, information about one of the s_i gives as little information as possible about the others. In other words, the s_i are as independent as possible.

There are several approaches for formulating independence in the ICA model [9] such as minimum mutual information, maximum neg-entropy; a very popular approach is the maximum likelihood [19, 4]. Let $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ represent an ICA model and the density of $\mathbf{s}$ be given as p_s in Eq.(3). The density of the observation $\mathbf{x}$, or the likelihood of the model, can be formulated as $p(\mathbf{x} \mid \mathbf{w}, p_s) = |\det \mathbf{b}|^{-1} p_s(\mathbf{b}^{-1}\mathbf{s}) = |\det \mathbf{w}| p_s(\mathbf{w}\mathbf{s})$. Given N_T training images, $\mathbf{x} = \{\mathbf{x}_k \mid k = 1, \ldots, N_T\}$, the logarithm likelihood can be derived as

$$\log p(\mathbf{x} \mid \mathbf{w}, p_s) = \sum_{k=1}^{N_T} \sum_{i=1}^{m} \log p_{s,i} + N_T \log |\det \mathbf{w}| \tag{4}$$

where $p_{s,i} = p_{s,i}(s_{i,k}) = p_{s,i}(< \mathbf{w}_i, \mathbf{x}_k >)$ (the forms of $p_{s,i}$ are assumed to be known). Learning an ICA model can be simply achieved by maximizing the likelihood function with respect to $\mathbf{w}$.

2.3. Topographic ICA

In classic ICA, there is no order relationship between independent components (ICs) $\mathbf{s}_i$. This is due to the assumption of complete statistical independence. For this reason. classic ICA is unable to describe relationships between different ICs. In many applications, the independence assumption is not always satisfied, and dependent can be seen between some estimated ICs. This can be the case for images containing appearances of an object viewed from different poses. In this case, it is desirable to make use of such dependencies to reveal relationships between ICs.

Topographic ICA (TICA) [8], an extension of ICA, is a learning method by which the basis components are ordered by some higher-order statistic. In TICA, the independence assumption is

relaxed. Components that are close to each other, *i.e.* those within a neighborhood, are not assumed to be independent; they are allowed to be correlated in their energies. For example, the residual dependency structure of the compoenents, which cannot be canceled by ICA, is allowed to exist. The scope of the dependency is defined by a neighborhood system, as in self-organizing maps [10].

A kind of higher order correlations that can be used to define the topographic ordering is the correlation between the energies of the components [8]

$$\mathrm{cov}(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \tag{5}$$

Intuitively, such a correlation means that the components tend to be active, *i.e.* nonzero, at the same time. The TICA model thus defined has the following properties: (1) All the components are uncorrelated. (2) Components that are not neighbors to each other are independent, at least approximately. (3) Components that are not neighbors tend to be active (nonzero) at the same time, *i.e.* have correlated energies.

TICA can also be defined using a likelihood function by introducing a neighborhood weighting into Eq.(6). Let there be m components and denote by $\mathcal{N}_j$ the set of indices of the components neighboring to component j. The log likelihood can be defined as

$$\log p(\mathbf{x} \mid \mathbf{w}) = \sum_{k=1}^{N_T} \sum_{i=1}^{m} \log p_s(\sum_{j \in \mathcal{N}_i} s_{j,k}^2) + N_T \log |\det \mathbf{w}| \tag{6}$$

where p_s are the known density functions of the norm, and $s_{j,k} =< \mathbf{w}_j, \mathbf{x}_k >$. Learning a TICA model can be simply achieved by maximizing the likelihood function with respect to $\mathbf{w}$ and can be implemented by using a gradient ascent algorithm [8].

2.4. Learning View-Based Topographic Map

The objective here is to use TICA to learn a topographical map of basis components so that the basis components are ordered according to left-right rotation angle. In the resulting TICA map, we can see a gradual change in the pose in one of the two TICA map directions, and changes in the illumination and other factors such as facial shape in the other direction. As such, basis components of similar view are grouped together. They can be used to form the subspace of faces in the corresponding view. The span of these components in a view group defines the manifold of possible appearances of human face viewed at that angle, subject to illumination and so on. In other words, it represents the view-subspace of facial appearances in the whole face space. These may be used for applications in appearance based multi-view object detection and recognition.

3. EXPERIMENTAL RESULTS

The following experiments demonstrate unsupervised TICA learning of view subspaces which produces emergent topographic maps of view-ordered basis components for representing multi-view face images. There are about 30,000 face examples in the database, roughly evenly distributed with respect to the view angle in the range between 0° and 90°.

There are several parameters to choose in a TICA algorithm: the dimensions ($H \times W$) of the map, where H and W are the height and width, and the shape and mode of the neighborhood system. Two neighborhood shapes are used, respectively: (1) 3×3 and (2) $H \times 3$. Two modes are: (1) the "torus" mode which takes into account the dependence around the basis on the edge of the TICA map so that the left- and right- most columns are neighbors, and so are top- and bottom- rows; (2) the "standard" mode in which there are no neighboring relations between left-right most columns and top-bottom most rows. The TICA software downloaded from `www.cis.hut.fi/projects/ica` is used. As a standard practice of ICA, the training data is preprocessed by whitening and then dimensionality reduction using PCA.

Figure 3 shows a 5×18 TICA topographic map obtained using the 3×3 "torus" neighborhood. Initialized at random, the view-angle related topographic ordering emerges as the iteration goes on, and the map always converges orderly whatever the initialization is. From the left to the right of the map, we can see the orderly change from the frontal view to the side view and then to near frontal view again. Such a fashion of view change is due to the constraint that the left- and right- most columns are neighbors. Vertically, we can see variations in other aspects such as illumination and facial shape. In this case, the components in every 3×3 neighborhood form an independent subspace.

Figure 4 shows a 20×9 TICA topographic map obtained using the 20×3 "standard" neighborhood. In this case, the left- and right- most columns are no longer neighbors and are therefore independent. From the left to the right of the map, we can see the orderly change from the side view to the frontal view. Although an ordering always emerge, however in this non-torus neighborhood, the ordering can be reversed (to from frontal view to side view) depending the initialization. Nonetheless, we can set proper initialization to induce desired ordering. Another note is that using the 20×3 neighborhood, there is no ordering between components within a column. Similar to the previous case, we can see vertical variations in illumination and facial shape. In the latter, the components in every 3 columns form an independent subspace.

4. CONCLUSION

The contribution of this paper is the use of the TICA method for learning a view-based representation of a 3D object from its 2D appearances. It is shown that the TICA algorithm is able to form topographic of basis components in which the components are ordered by its view. The ordering provides a natural way for clustering the basis components into view subsets. As such, a view-subspace can be constructed by the span of components of that view. This is the significance of the view-based topographic map.

Potential applications of such a view-subspace representation is 2D appearance based multi-view object detection and recognition. An appearance or image of the object can be represented as its projection point in a view-subspace. The amplitude of the projection coefficients for that view corresponds the activity of the complex cell tuned for the view. This suggests a means of detecting the object in a particular view by thresholding the norm of the coefficient vector in the corresponding view-subspace. Multi-view object recognition may be performed by comparing distances between the projection points of the observed image and the prototype image.

Fig. 3. A topographic map of 5×18 basis components learned using 3×3 "torus" neighborhood.

5. REFERENCES

[1] S. Baker, S. Nayar, and H. Murase. Parametric feature detection. *IJCV*, 27(1):27–50, March 1998.

[2] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski. "Independent component representations for face recognition". In *Proceedings of the SPIE, Conference on Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, 1998.

[3] A. J. Bell and T. J. Sejnowski. "The 'independent components' of natural scenes are edge filters". *Vision Research*, 37:3327–3338, 1997.

[4] J.-F. Cardoso. "Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90(8).

[5] P. Comon. "Independent component analysis - a new concept?". *Signal Processing*, 36:287–314, 1994.

[6] S. Gong, S. McKenna, and J. Collins. "An investigation into face pose distribution". In *Proc. IEEE International Conference on Face and Gesture Recognition*, Vermont, 1996.

[7] J. Hornegger, H. Niemann, and R. Risack. Appearance-based object recognition using optimal feature transforms. *PR*, 33(2):209–224, February 2000.

[8] A. Hyvärinen and P. Hoyer. "Emergence of topography and complex cell properties from natural images using extensions of ica". In *Advances in Neural Information Processing Systems*, volume 12, pages 827–833, 2000.

[9] A. Hyvärinen and E. Oja. "Independent component analysis: Algorithms and applications". *Neural Networks*, 13(4):411–430, 2000.

[10] T. Kohonen. *Self-Organizing Maps*. Information Sciences. Springer, Heidelberg, second edition, 1997.

[11] T. Lee, M. Lewicki, and T. Sejnowski. ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind separation. *PAMI*, 22(10), October 2000.

[12] C. Liu and H. Wechsler. "Comparative assessment of independent component analysis (ica) for face recognition". In *Proc. Second Int'l Conf. on Audio- and Video-based Biometric Person Authentication*, Washington D. C., March 22-24 1999.

[13] R. Manduchi and J. Portilla. "Independent component analysis of textures". In *Proceedings of IEEE International Conference on Computer Vision*, Corfu, Greece, 1999.

[14] B. Moghaddam. "Principal manifolds and bayesian subspaces for visual recognition". In *Proceedings of IEEE International Conference on Computer Vision*, Corfu, Greece, 1999.

[15] H. Murase and S. K. Nayar. "Visual learning and recognition of 3-D objects from appearance". *International Journal of Computer Vision*, 14:5–24, 1995.

[16] S. Nayar, S. Nene, and H. Murase. Subspace methods for robot vision. *RA*, 12(5):750–758, October 1996.

[17] B. A. Olshausen and D. J. Field. "Natural image statistics and efficient coding". *Network*, 7:333-339, 1996.

[18] A. P. Pentland, B. Moghaddam, and T. Starner. "View-based and modular eigenspaces for face recognition". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 84–91, 1994.

[19] D.-T. Pham, P. Garrat, and C. Jutten. "Separation of a mixture of independent sources through a maximum likelihood approach". In *Proc. EUSIPCO*, pages 771–774, 1992.

Fig. 4. A topographic map of 20×9 basis components learned using 20×3 "standard" neighborhood.