Image Database Browsing

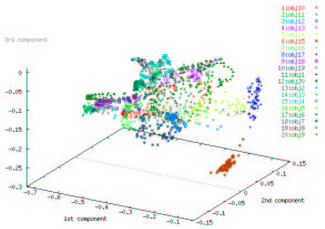
Bertrand Le Saux & Nozha Boujemaa

The aim of browsing is to give the best overview of the database.
The problem is to estimate the distributions of image categories, and to find the best representatives for each category.

1. Nature of data

Images are described by various features :
- color : weighted histogram
  (except black&white database : gray levels histogram).
- texture : Fourier spectrum.
- shape & structure : edge orientations histogram.

3 problems occur :
- Data's natural categories have various shapes and densities, and overlap each other.
- Features space is high dimensional.
- Number of categories is unknown.



Distribution of gray levels histograms of Columbia database according to 3 first principal components.

2. Clustering by Self-Organization of Oscillators Network

SOON algorithm [2] can be summarized as follows :
Each image is represented by one Integrate-and-Fire oscillator.
When an oscillator reaches the threshold, similar oscillators are excited and dissimilar ones inhibited.
Oscillators which reach the threshold too are synchronized with this one.

Similarity between 2 oscillators yj and yk is measured by Jaccard distance : $d^2(\mathbf{y}_j, \mathbf{y}_k) = \frac{\sum_{l=1}^{p} \min(\mathbf{y}_j^l, \mathbf{y}_k^l)}{\sum_{l=1}^{p} \max(\mathbf{y}_j^l, \mathbf{y}_k^l)}$

=> problem : the noise cluster is too large, too much images are not categorized.

3. Robust Image Database Categorization

1st step : **Reduction of dimensionality** by principal components analysis

2nd step : **Robust unsupervised clustering**

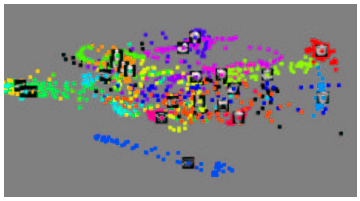
As we don't know the number of clusters, we minimize iteratively the CA objective function [1]:

$$J(\mathbf{B}, \mathbf{U}, X) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^2 d^2(x_j, \beta_i) - \alpha \sum_{i=1}^{C} [\sum_{j=1}^{N} u_{ij}]^2$$

we introduce a **noise cluster** to eliminate ambiguous images : $d^2(x_j, \beta_i) = \delta^2$

good clusters use Mahalanobis distance : $d^2(x_j, \beta_i) = (x_j - \beta_i)^T \mathbf{A}_i (x_j - \beta_i)$

where $\mathbf{A}_i$ is cluster covariance matrix.



Result of robust categorization on Columbia database.

4. Results

1. Choice of prototypes

1. We compute for each image, the mean value of each feature (e.g. mean color for a color feature).
2. Then for all images in the cluster, we compute the mean of features' mean values.
3. The prototype is the nearest image to this mean.

2. Comparison of both methods

Prototypes of clusters obtained with SOON on Columbia database.

Prototypes of clusters obtained with our method on Columbia database.

=> We have less redundancy in prototypes.
=> When a natural category is found, our method retrieves more images from it.

5. Conclusion

We studied the problem of image database categorization through two methods.
Because SOON method misses too much images, we have proposed a new robust clustering algorithm adapted to image databases.
Our method succeeds in presenting a maximum number of images in each category and giving a good overview of the database.

References :

[1] H. Frigui and R. Krishnapuram, Clustering by Competitive Agglomeration, Pattern Recognition, 30(7), 1997.
[2] H. Frigui, N. Boujemaa and S.-A. Lim, Unsupervised Clustering and Feature Discrimination with Application to Image Database Categorization, NAFIPS 2001.
[3] M. Rhouma and H. Frigui, Self-Organization of a population of coupled oscillators with application to clustering, PAMI, 23(2), 2001