MEAN SHIFT CLUSTERING FOR PERSONAL PHOTO ALBUM ORGANIZATION

E. Ardizzone, M. La Cascia

Universita degli Studi di Palermo
Dipartimento di Ingegegneria Informatica
Viale delle Scienze, 90128 Palermo, ITALY

F. Vella

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni
Viale delle Scienze, 90128 Palermo, ITALY

ABSTRACT

In this paper we propose a probabilistic approach for the automatic organization of pictures in personal photo album. Images are analyzed in term of faces and low-level visual features of the background. The description of the background is based on RGB color histogram and on Gabor filter energy accounting for texture information. The face descriptor is obtained by projection of detected and rectified faces on a common low dimensional eigenspace. Vectors representing faces and background are clustered in an unsupervised fashion exploiting a mean shift clustering technique. We observed that, given the peculiarity of the domain of personal photo libraries where most of the pictures contain faces of a relatively small number of different individuals, clusters tend to be not only visually but also semantically significant. Experimental results are reported

Index Terms— Image databases, Image analysis, Image representations, Clustering methods, Computer applications

1. INTRODUCTION

The popularity of personal imaging systems like digital cameras and camera-equipped cell phones has lead to an outstanding increase in the amount of digital content acquired for personal use. The main risk is to end up with tens of thousand of pictures stored on PCs but with scarce access to them. Currently, the main way to search digital photo libraries is by mean of time of shooting and/or keywords given by the user. This modality of access to the library is definitely unsatisfactory, as it requires the user associate keywords to pictures and as keywords themselves often tend to be ambiguous. Time of shooting is a much more reliable cue and it is available for free as digital cameras attach a timestamp to each pictures. Some sort of content-based organization of the pictures is then needed to improve the efficacy of the image retrieval process, ideally requiring no intervention from the user during the acquisition process. In the last years several approaches have been proposed to attack the particular problem of content-based access and browsing of personal photo libraries. In fact, personal photo libraries sports peculiar characteristics compared to general image collection, namely the presence of people in most of the images and a relatively small number of different individuals across the whole library. These characteristics should be exploited to develop automatic or semi-automatic approaches. In our approach image data is analyzed in two domains of interest. On one side faces are extracted from the images and referred to person identity; on the other side the remaning part is considered as the image context. The main idea is that a large number of faces can be automatically detected, rectified, resampled, cropped [1] and finally projected in a common low dimensional *face space*. A few coefficients of the projection can then be used as face descriptor. The remaining part of the images (the background) can be characterized by mean of low-level features based on color and texture that are useful in discriminating between different contexts (*where*). To automatically organize image data based on faces and background descriptors we use a mean shift based approach. Organization of data does not need any human intervention as parameters of the clustering method are chosen automatically according to a proposed figure of merit. The time when the picture was shot and optical metadata attached to the picture appears to be very useful information [2, 3, 4] but we are not currently using it focusing only on visual features.

2. RELATED WORKS

One of the first personal photo collection browser has been reported by Kang and Shneiderman[5] but it is mainly a very powerful user interface with very limited CBIR capabilities. As in personal photos the objects of interest are often people Zhang et al.[6] addressed the problem of automated annotation of human faces in family album integrating CBIR techniques and face recognition in a probabilistic framework. Another photo management application leveraging face recognition technology has also been proposed by Girgensohn et al.[7]. The authors implemented a user interface that greatly helps the users in face labelling. Other semi-automatic annotation techniques for personal photo libraries have also been proposed recently[8, 9, 10]. Other researcher address the problem of personal photo album management in an image clustering framework. For example hierarchi-

cal clustering enable the users to navigate the levels to find images. Clusters prototypes are a compact representation of classes of similar images and then can be used in browsing or searching the library. The efficacy of the clustering approach, as well as any CBIR system, is obviously affected by the goodness of the features used to describe the images and of the similarity metrics. As similarity metrics may not reflect semantic similarity between images, clusters are not guaranteed to be semantically homogeneous. Several techniques have been proposed also for the clustering of images based on image metadata [2, 4]. In [11] a semi automatic photo annotation system based on enhanced spectral clustering is proposed. They use time, global color correlogram for location/event clustering and local facial features and color correlogram from human body area for face clustering. The presence of faces has also been exploited in an attempt to bridge the gap between visual and semantic content [1, 12].

3. PERSONAL PHOTO ALBUM INDEXING

In the proposed approach, each image in the collection is represented by the presence of faces and by visual background features [13]. A data oriented clustering allows to generate aggregation structures driven by the regularities in the represented data. Faces are preprocessed to reduce the variation of the appearance and are mapped in an auto emerging space employing eigenfaces. The information from background is managed representing the image part not associated to a face as a vector of low-level visual features. Finding faces in general images is a very challenging task due to variations in pose and illumination. Berg et al.[1] analyzed hundred of thousands of images taken from the Internet to detect faces *in the wild*. In a similar way in our approach each image to be archived in the system is searched for faces. Detected faces are then validated and rectified to a canonical pose and size. The face detector we adopted[14] is usually successful in detecting faces in a quite large range of pose, expression and illumination conditions. In order to represent these faces in a meaningful way, useful for subsequent retrieval, some processing is needed. In particular, as suggested by Berg et al.[1], we try to detect five features per face (external corners of left and right eyes, corners of the mouth an tip of the nose) and, if detection is successful, we estimate an affine transformation to rescale and align the face to canonical position. A final crop to 100×100 pixels brings each face to a common reference system. Faces where the feature detector failed to work with an high degree of confidence were rejected. Once a face has been detected and successfully rectified and cropped, a reduced dimension face descriptor is computed. The descriptor is a vector $\mathbf{w}$ containing the projection of the rectified and cropped face in a subspace of the global face space. In practice the average face Ψ is subtracted from the 100×100 cropped and rectified face Γ_i and the obtained image Φ is then projected on the eigenspace to obtain $w_i = \mathbf{e}_i^T \Phi$. The face space, as well as the average face, is learned off-line on a significant subset of the image collection and it is not updated. In our experiments we learned the face space with about 200 images.

Once faces have been detected the remaining part of the image is processed as background and represented with composition of color and texture features. Features are globally evaluated and a single vector for each image is produced. Color information is captured through histograms in the RGB color space. The 60-dimensional global descriptor is computed as the concatenation of the 20-bin histograms of the R, G and B channels. Texture is evaluated through Gabor filters considering 3 orientation and 2 scales. Image is subdivided in 16×16 blocks and the energy is computed in each block for all the filters to obtain a 15-dimensional vector. The sum of the vectors of all the blocks is the global texture feature. Since mean values are less indicative of feature distribution in the data space image features are filtered with a sigmoid to stretch the values towards low or high values.

4. IMAGE CLUSTERING

4.1. The mean shift algorithm

Mean shift is a technique for kernel density estimation that applies gradient climbing to probability distribution[15]. Given n data points $\mathbf{x}_i, i = 1, 2, ..., n$ in the d-dimensional space R^d, a multivariate kernel density estimator $\hat{f}(\mathbf{x})$ is calculated as

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{x_i}}{h}) \tag{1}$$

where h is the bandwidth and the kernel $K(.)$ is the Epanechnikov kernel. Using a differentiable kernel, the estimate of the gradient density can be written as the gradient of the kernel density estimate(1):

$$\hat{\nabla} f(\mathbf{x}) \equiv \nabla \hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} \nabla K(\frac{\mathbf{x} - \mathbf{x}_i}{h}) \tag{2}$$

For the Epanechnikov kernel the density gradient estimate is:

$$\hat{\nabla} f(\mathbf{x}) = \frac{n_c}{nV_d} \frac{d+2}{h^d} \left(\frac{1}{n_c} \sum_{\mathbf{x} \in S(x)} (\mathbf{x}_c - \mathbf{x}) \right) \tag{3}$$

where $S(\mathbf{x})$ is the hyper-sphere of radius h, having volume $h^d V_d$, centered in $\mathbf{x}$ and containing L_c data points. The quantity $M_h(\mathbf{x})$ defined as

$$M_h(\mathbf{x}) \equiv \frac{1}{n_c} \sum_{\mathbf{x} \in S(\mathbf{x})} (\mathbf{x}_c - \mathbf{x}) \tag{4}$$

is called Mean Shift Vector that can be expressed as:

$$M_h(\mathbf{x}) = \frac{h^d}{d+2}\frac{\hat{\nabla} f(\mathbf{x})}{\hat{f}(\mathbf{x})} \tag{5}$$

The Mean Shift Vector at location $\mathbf{x}$ is aligned with the local density gradient estimate and is oriented towards the direction of maximum increase in density. For each point the Mean Shift Vector defines a path leading from the fixed point to a stationary point of estimated density where gradient is equal to zero.

4.2. Mean Shift Clustering for Personal Album

Given a generic point in the feature space (*faces, backgrounds*), the Mean Shift Vector in equation (4) describes a trajectory in the density space converging to points where the density is maximum. The set of all points converging to a local maximum is the *basin of attraction* for the found maximum density point. The procedure for the detection of modes in the data distribution is based on running mean shift to find stationary points for $\hat{f}(\mathbf{x})$, pruning the found points retaining only the local maximum points and unifying adjacent clusters.

4.3. Entropy based Clustering Measure

A number of evaluation indexes have been proposed to evaluate clustering methods(see for example [16]). We evaluated our clustering using a number of images hand labelled both in face space and in the backround space. With this information we define two indexes able to capture the clustering capability. The *Intra-Cluster Entropy* and the *Intra-Label Entropy*. The *Intra-Cluster Entropy* is defined as:

$$E_c = -\frac{1}{N_C * log(N_L)} \sum_{i=1}^{N_L}\sum_{j=1}^{N_C} \frac{u_{ij}}{T_j} log \frac{u_{ij}}{T_j} \tag{6}$$

where N_C is the number of clusters, N_L is the number of labels, u_{ij} is the number of times the $i-th$ label is present in the $j-th$ cluster and T_j is the number of labelled samples in the $j-th$ cluster. This function gives a measure of the "disorder" inside clusters. If many labels are present in a cluster the value u_{ij}/T_j is near the average and the *Intra-Cluster Entropy* is high.

The *Intra-Label Entropy* is defined as:

$$E_l = -\frac{1}{N_L * log(N_C)} \sum_{i=1}^{N_L}\sum_{j=1}^{N_C} \frac{u_{ij}}{S_i} log \frac{u_{ij}}{S_i} \tag{7}$$

where N_C is the number of clusters, N_L is the number of labels, u_{ij} is the number of times the $i-th$ label is present in the $j-th$ cluster and S_i is the number of occurrence of the $i-th$ label. This function gives a measure of the distribution of a label across clusters. If a label is always present in a cluster, or in the opposite way always absent, the ratio u_{ij}/S_i is near 1, or near 0, and the entropy has a low value.

We have a global measure we also defined the *Global Clustering Entropy* as $E_G = \zeta \cdot E_c + (1-\zeta)\cdot E_l$.

4.3.1. Mean Shift Clustering for composite data

The clusterization of data through the mode seeking assumes the possibility to estimate distribution density with a single kernel being the data characterized by the same density distribution in all the space. Assuming for both domains (faces and background) to use the Euclidean norm as metric, a multivariate kernel is defined as product of two radially symmetric kernels:

$$K_{h_f,h_b}(\mathbf{x}) = \frac{C}{h_f^M h_b^P} k\left(\left\|\frac{\mathbf{x}^f}{h_f}\right\|^2\right) k\left(\left\|\frac{\mathbf{x}^b}{h_b}\right\|^2\right) \tag{8}$$

where $\mathbf{x}^f$ is the data in the first domain, $\mathbf{x}^b$ is the data referred to the second domain, h_f and h_b are the corresponding kernel bandwidths, C is the normalization constant. In this case the first part regards faces information having a dimensionality f corresponding to the number of eigenfaces used in face representation. The second part regards background information with a dimensionality equal to b. Instead of evaluating empirically the performance of multiple values of the bandwidth, the *Global Clustering Entropy* is used as performance measure. According to clustering results, the bandwidth value is automatically chosen. The process is run for both the domains, and ideally can be applied to any number of set of orthogonal feature representing input data, then the merging of clusters among multiple domains is performed similarly to [15].

5. EXPERIMENTAL RESULTS

To evaluate the performances of the proposed system we ran a set of experiments on a real photo collection. The digital album used is a subset of a real personal collection of 1008 images taken in the last three years. The presented process for face detection and rectification brought to the extraction of 331 images of rectified faces. The experiments have been aimed to the evaluation of the retrieval capability of the proposed system in terms of faces and background labeling. An entropy based analysis of the clustering process has also been performed. To evaluate our approach with respect to the semantic meaning of clusters we divided all the images of the test collection in six categories representing six typical contexts mainly present in the collection and assigned an identifier to the four people present in most of the photos. Data are clustered using a single domain information. For each domain the optimal point according to the Global Clustering Entropy is chosen. The results for the clustering of background are shown in the table 1 (clusters with a single element are discarded):

	beach	indoor	nature	public garden	snow	urban
Cl 1	11%	32%	2%	40%		15%
Cl 2		89%		11%		
Cl 3		96%		4%		
Cl 4		100%				
Cl 5		13%		63%		25%
Cl 6	6%	42%		52%		
Cl 7		100%				
Cl 8		100%				
Cl 9				67%		33%

Table 1. Percentage of labels in generated clusters

	Id 1	Id 2	Id 3	Id 4	Other
Cl 1		77%	9%	5%	9%
Cl 2	3%	22%	53%	6%	17%
Cl 3	22%	33%	33%		11%
Cl 4	11%	11%	44%		33%
Cl 5			100%		
Cl 6			100%		

Table 2. Percentage of identities in generated clusters

For faces data, the optimal point according the entropy measure is found for a bandwidth of 5756 and a dimension of the eigenspace equal to 131. The distribution of the 6 cluster with more than one element is shown in table 2.

An evaluation of the clusterization using information from both domains is achieved calculating the Global Clusterization Entropy using labels given by couples (*identity*, *context label*). In the table 3 values of Global Clustering Entropy versus face and background bandwidths are shown.

	3.0	3.5	4.0	4.5	5.0
4756	2.67	2.70	2.78	2.87	3.10
5256	2.70	2.72	2.82	2.92	3.11
5756	3.03	3.10	3.20	3.22	3.19
6256	4.31	4.87	5.89	5.71	5.45
6756	6.41	8.31	10.09	9.78	8.07

Table 3. Value of global entropy for clusterization

6. CONCLUSIONS

A novel approach to cluster composite data driven by a entropy-based measure has been presented. The approach has been demonstrated on the very interesting problem of automatic organization of photos in personal album. In our experiments data are represented in two spaces representing people in the photos and low-level visual characterization of the background. Results of experiments on a real set of 1000 pictures are very promising.

7. REFERENCES

[1] Tamara L. Berg et al., "Names and faces in the news," in *Proc. of IEEE CVPR*, 1994.

[2] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd, "Time as essence for photo browsing through personal digital libraries," in *Proc. of ACM JCDL*, 2002.

[3] Pinaki Sinha and Ramesh Jain, "Concept annotation and search space decrement of digital photos using optical context information," in *SPIE Vol. 6820*, 2008.

[4] Bo Gong and Ramesh Jain, "Hierarchical photo stream segmentation using context," in *SPIE Vol. 6820*, 2008.

[5] H. Kang and B. Shneiderman, "Visualization methods for personal photo collections: Browsing and searching in the photofinder," in *Proc. of IEEE ICME*, 2000.

[6] L. Zhang, L. Chen, M. Li, and H. Zhang, "Automated annotation of human faces in family albums," in *Proc. of ACM International Conference on Multimedia*, 2003.

[7] Andreas Girgensohn, John Adcock, and Lynn Wilcox, "Leveraging face recognition technology to find and organize photos," in *Proc. of ACM MIR*, 2004.

[8] Mor Naaman, Ron B. Yeh, Hector Garcia-Molina, and Andreas Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. of ACM JCDL*, 2005.

[9] Benjamin N. Lee, Wen-Yen Chen, and Edward Y. Chang, "A scalable service for photo annotation, sharing and search," in *Proc. of ACM International Conference on Multimedia*, 2006.

[10] Jingyu Cui, Fang Wen, Rong Xiao, Yuandong Tian, and Xiaoou Tang, "Easyalbum: An interactive photo annotation system based on face clustering and re-ranking," in *Proc. of ACM CHI*, 2007.

[11] J. Cuiy, F. Wenz, R. Xiaoz, Y. Tianx, and X. Tang, "Easyalbum: An interactive photo annotation system based on face clustering and re-ranking," in *Proc. of ACM CHI*, 2007.

[12] Y. Song and T. Leung, "Context-aided human recognition clustering," in *Proc. of ECCV*, 2006, vol. 3.

[13] E. Ardizzone, M. La Cascia, and F. Vella, "A novel approach to personal photo album representation and management," in *SPIE Vol. 6820*, 2008.

[14] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE CVPR*, 2001.

[15] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transaction on PAMI*, pp. 603–619, May 2002.

[16] K.L. Wu and M.S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognition Letters*, pp. 1275–1291, 2005.