A FAST FACE CLUSTERING METHOD FOR INDEXING APPLICATIONS ON MOBILE PHONES

Sudha Velusamy, Pratibha Moogi

Samsung Research & Development Institute, Bangalore, India

ABSTRACT

Tagging of faces present in a photo or video at shot level has multiple applications related to indexing and retrieval. Face clustering, which aims to group similar faces corresponding to an individual, is a fundamental step of face tagging. We present a progressive method of applying easy-to-hard grouping technique that applies increasingly sophisticated feature descriptors and classifiers on reducing number of faces from each of the iteratively generated clusters. Our primary goal is to design a cost effective solution for deploying it on low-power devices like mobile phones. First, the method initiates the clustering process by applying K-Means technique with relatively large K value on simple LBP features to generate the first set of high precision clusters. Multiple clusters generated for each individual (low recall) are then progressively merged by applying linear and non-linear subspace modelling strategies on custom selected sophisticated features like Gabor filter, Gabor Jets, and Spin LGBP (Local Gabor Binary Patterns) with spatially spinning bin support for histogram computation. Our experiments on the standard face databases like YouTube Faces, YouTube Celebrities, Indian Movie Face database, eNTERFACE, Multi-Pie, CK+, MindReading and internally collected mobile phone samples demonstrate the effectiveness of proposed approach as compared to state-of-the-art methods and a commercial solution on a mobile phone.

Index Terms— Face Tagging, Feature Description, Subspace Modeling, Cost Effective, Mobile Applications.

1. INTRODUCTION

With the revolutions on camera phones and internet, a large amount of video and image data is being captured and stored. The problem of indexing such a huge data has been one of the important challenges for decades. Among the various attributes of media content, human face is one of the primary factors for indexing. Automatic tagging of human faces also find other utility applications like person based video summarization, prime casts detection in broadcast video, etc. Reliable grouping of similar faces from a randomly collected pool of images and videos is highly challenging task due to the variations in photometric and person specific factors like lighting conditions, viewing angles, occlusions, wide expressions, hair style, make-up, etc. Factors like processing speed, memory utilization, and power consumption also brings in other dimensions of challenges, when the solution needs to be deployed in resource constrained devices like mobile phones.

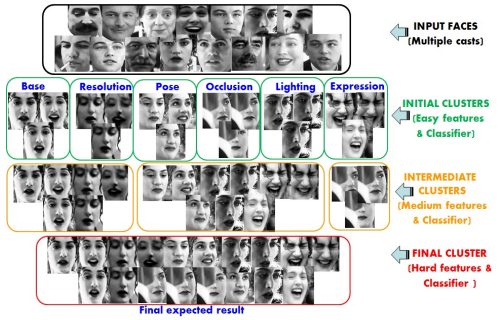


Fig. 1. *An example of progressively generated face clusters*

2. RELATION TO PRIOR WORK

There are many interesting face clustering techniques and methods present in the literature [1, 2, 3]. Recent works in this field focus on one or more of the following face clustering components to achieve high performance solutions: (i) Discriminative feature descriptors that can handle intra-person variabilities and capture inter-personal dissimilarities [4]; (ii) Suitable classifiers and distance metrics [5]; (iii) Effective thresholding/merging criteria [6]; and (iv) Exploiting auxiliary information of human faces like clothing information [7], background matching [3], people co-occurrence [1], etc. Among the above listed points, appropriate selection of a suitable feature descriptor and learning methodology have proved to be the fundamental needs for handling most of the data related challenges [2, 8, 9]. For example, Zisserman et al [2] present a face grouping method that uses kernel PCA to efficiently classify a large corpus of data. Other successful methods include using feature descriptors like LBP, SIFT, HOG [5] and clustering algorithms like K-Means [10], Hierarchical [11] and Spectral clustering [10], etc. While many of the above listed conventional methods are seen to provide high precision rates by using simple to compute descriptors and traditional classifiers, they often suffer from poor recall rates on real-world data.

To achieve robustness in clustering, researchers have proposed methods that apply advanced features and classifiers. Examples include methods using deep learning of 3D modelled faces [12], multi-scale features computed around fiducial points [13], heterogeneous set of features [5, 8], elastic bunch graphs [14], pose-specific techniques [15], and more recently approaches using linear/affine subspaces (assumes each image set spans a linear or affine subspace) [16, 17]. For example, Kim et al [17] proposed an efficient clustering technique for faces with wide photometric variations. The method uses Discriminant Canonical Correlations (DCC) to represent images as linear subspace and compute the principal angles of between the subspaces as similarity metric. There are also methods that perform explicit image level normalization of variations like facial pose or expressions to boost the clustering accuracy [18]. However, there is a trade-off between their ability in handling various real-world challenges and computation cost of the applied features or classifiers. It becomes increasingly more challenging to deploy the above said sophisticated techniques on the cost conscious devices.

In this paper, we present a progressive method of applying easy-to-hard grouping technique at a reduced computation cost. Fig. 1 shows the progressive nature of grouping an example set of faces with wide range of data variations by the proposed method. The method applies simple to compute features and classifiers at the initial stages of clustering, and uses increasingly sophisticated and suitable features, and models at the higher stages of clustering to handle the increased intra-person variabilities. We consider highly subsampled number of faces (non-redundant) from each of the higher stage clusters for further grouping. This makes the proposed method highly compute efficient and suitable for mobile devices. Novelties of the proposed method include:

1. Strategy of applying simple-to-hard feature descriptors and classifiers that are custom selected to achieve a fast face clustering on mobile phone devices;
2. Application of features like LGBP with spin support for handling highly challenging data variations like facial expression, make-up, pose, etc., at a reduced cost;
3. Sub-sampling of clusters based on the amount of intra-face variations and noise pruning at each stages;

3. THE PROPOSED METHOD

Figure 2 shows the complete system flow of the proposed progressive face clustering method. It is particularly designed to suit face tagging application in mobile phone galleries, where the sets may contain a wide range of images/videos including personal and professional contents. The present system include; i) a pre-processing unit that perform frame quality check and face normalization; followed by ii) the progressive clustering of faces from the pre-processed frames.

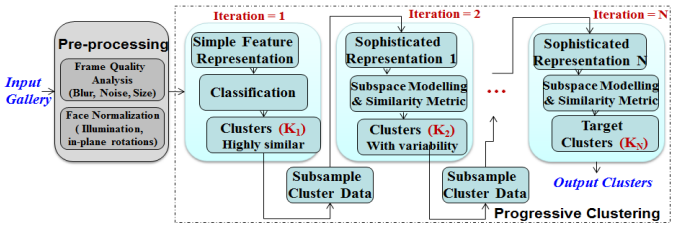


Fig. 2. *The proposed progressive face clustering systems*

3.1. Pre-processing

3.1.1. Input Frame Quality Analysis

Given that the input gallery of videos and photos may include frames with heavy motion blur, compression noises, low contrast, tiny face regions, large occlusions, etc., it is important to select quality frames for further processing. We apply an image analysis module that measure the above mentioned parameters [19] for every input frame and filter out noisy frames based on experimentally arrived thresholds, and allow only good quality samples for further clustering.

3.1.2. Face Normalization

On each of the selected frames, we apply the in-built, low-cost face detector in Samsung Galaxy $S7$ to detect one or more face regions present. The detected faces are then further aligned to be near frontal with the help of automatically detected pupil co-ordinates on the faces. The aligned faces are then cropped to a standard face template size of 96×96 pixels, and passed to the next step of progressive clustering.

3.2. Progressive Face Clustering

3.2.1. Problem Formulation

Given a gallery of face images, $\mathbf{X} = \{x_{11}, ...x_{PN_P} | \mathbf{x_{pn_p}} \in \mathbf{R^d}\}$, where $\{p = 1, ..., P\}$ indices the number of people, and $\{n_p = n_1, ..., N_P\}$ indices the number of samples for each p, our goal is to precisely group them at a low computation cost.

3.2.2. Simple Descriptor & K-Means Clustering

Given a large value of N_p for every p, it is important to use a simple to compute clustering technique at the initial stage. The method extracts face descriptor called Local Binary Patterns from each of the pre-processed faces, and apply K-Means clustering with relatively large K to create the initial clusters. The K-Means technique partition the $P \times N_P$ input faces into $K (< P \times N_P)$ sets $(C = \{C_1, C_2, , C_K\})$ so as to minimize the cost function shown in Equ. 1.

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \quad (1)$$

where μ_i is the mean of points in C_i. The initial set of clusters is allowed to include small or singleton face clusters as well to maintain higher precision. This step of simple clustering is measured to have relatively less computation, but at the cost of low recall rate (faces of a single person falls into multiple clusters). To improve the recall rate, the initial clusters that are likely to refer to the same person are then further processed at the higher stages, as explained below.

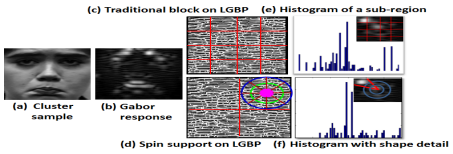


Fig. 3. *SPIN based LGBP feature computation*

3.2.3. Discriminative Descriptor & Linear Subspace Model

As the negligible intra-face variations at the initial clusters may increase to challenging level at the higher stages, it is important to consider the following factors for better recall: (i) Efficient feature descriptors that encode intra-class variations at an optimal cost; (ii) Suitable learning methodologies for better clustering performance; (iii) Noise pruning strategies to handle outliers and cluster sampling to control the cost.

Considering the above points, we derive texture discriminating feature descriptors using Gabor filters and apply linear sub-space modelling that treat each of the (high precision) cluster as an "image-set" corresponding to an individual, to further group them. It is important to note here, we perform sub-sampling of each image-set to select faces that are critical representatives of them, and sampling is done by selecting samples that are farthest from its corresponding centroids. On the selected samples of each set, we apply a bank of 40 Gabor filters (Refer Equ. 2) with 5 scales and 8 orientations to extract the features to encodes appearance variations due to facial expressions and illumination changes.

$$g(x,y) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\phi\right) \qquad (2)$$
$$x' = x\cos\theta + y\sin\theta, \qquad y' = -x\sin\theta + y\cos\theta$$

where, λ, θ, ϕ, $\gamma < 1$ are filters' wavelength, orientation, phase and aspect-ratio, respectively. The constructed Gabor feature vector of each face is represented as a point in a linear feature space and each image-set is characterized as a convex geometric region spanned by its feature points [20]. The distances measured between such sub-spaces are used to merge or assign a new class labels. For example, given image sets $\mathbf{C}$ and $\mathbf{C}^{'}$, the distance between them is the infinum of the distance between any points in $\mathbf{C}$ and any points in $\mathbf{C}^{'}$, where, the linear sub-spaces are approximated by an Equ. 4.

$$D(C, C^{'}) = \min_{\mathbf{x}\in\mathbf{C},\mathbf{y}\in\mathbf{C}'} \|\mathbf{x}-\mathbf{y}\| \qquad (3)$$

$$C_{k1}^{aff} = \{\mathbf{x} = \mu_{\mathbf{k1}} + \mathbf{U_{k1}}\mathbf{v_{k1}} | \mathbf{v_{k1}} \in R^d\} \qquad (4)$$

where $\mu_{\mathbf{k1}}$ is mean of all the samples in the image-set C_{k1}, $\mathbf{U_{k1}}$ is an orthonormal basis for the directions spanned by the linear subspace, and $\mathbf{v_{k1}}$ is a point within a sub-space (expressed with respect to the basis $\mathbf{U_{k1}}$). Plots in Fig. 4 shows the distances computed between the subspaces at two different iterations. Here, we consider 1 out of total N clusters as a test cluster, and compute its distance with remaining clusters to detect its matching clusters.

With the proposed method of modelling, the clusters similar to the selected test cluster result in distinctly low distance values and hence make the further merging easier. The threshold to decide on the clusters close to the test cluster is dynamically decided based on the *moving average of the absolute distances* computed on $D(C, C^{'})$ values(Refer the Fig. 4). We also apply a filtering mechanism for pruning noisy samples that may result in deviations to the modeled subspaces, based on the geometric shape of the subspace. This procedure is repeated until the distances measured across all the image-sets exceeds the set threshold, indicating no further cluster merging is possible with the current strategy. At this stage, we observe that the faces with wide expressions and large head pose changes are still not merged with their individuals and hence propose the below steps to achieve the target clusters.

3.2.4. Complex Descriptor & Non-linear Subspace Model

To handle the highly challenging intra-class variations, we represent the above generated intermediate clusters with a combination of highly sophisticated appearance based and geometric features, and use non-linear affine hulls [9] to find the similar clusters and merge them together. We propose to use spin based LGBP features to encode appearance details and Gabor Jets [21] based features computed around the 40 facial anchor points (tracked using Constrained Local Models (CLM) [21]) to capture the geometric information. Our design of spin LGBP [22] is to provide discriminative feature representation, while handling wide expression and head pose variations. Spin LGBP encodes changes in facial features like brows, eyes, lips, etc., with the relative angle and position information as shown in Fig. 3. This helps to match the individuals with wide variations in these features, and hence improve clustering accuracy. In our experiments, we use spin support with 8 orientations and 3 radial scales on 4 equally divided sub-regions of a face template. The non-linear subspace modelled image-sets are then iteratively merged based on geometric dissimilarity using L2 norm [9].

To reduce the computation cost of using above described complex descriptors and modelling, we again consider highly sub-sampled faces with only high variabilities as input at this stage. Here we use cosine similarity metric shown in Equ. 5 to find the point wise similarity across the faces.

$$\mathbf{e} = \cos^{-1}(f_x(f_y)^T), \qquad (5)$$

where f_x, f_y are features of faces $\mathbf{x}$ and $\mathbf{y}$, respectively. Experiments and results to prove the high clustering performance and computation efficiency of the proposed face clustering are present in the following section.

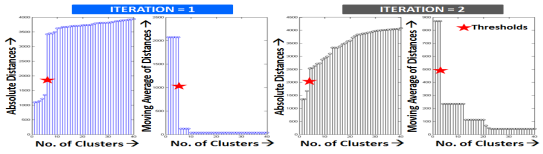


Fig. 4. *Euclidean distances between the affine subspaces*

Table 1. Performance Comparison - Face Clustering of VIDEOS and IMAGES

METHOD	PRECISION (%)								AVG	RECALL (%)								AVG	FPS
DB	**YTF**	**YTC**	**eNF**	**MR**	**Pie**	**CK**	**IM**	**SFD**		**YTF**	**YTC**	**eNF**	**MR**	**Pie**	**CK**	**IM**	**SFD**		
Kmean [10]	95	97	96	99	93	99	92	97	**96.0**	53	69	58	63	54	61	49	51	57.3	**18**
Hierar [11]	93	98	93	100	92	97	89	96	94.8	65	72	72	71	60	68	69	58	66.9	09
Hakan [9]	90	100	90	97	92	100	87	96	94.0	80	74	77	76	81	81	77	76	77.8	06
Seque [23]	89	94	90	92	-	-	-	93	91.3	77	70	69	79	-	-	-	69	72.8	05
Wen [24]	96	98	95	97	94	99	94	95	**96.0**	82	78	83	91	80	86	83	80	**82.8s**	04
S7	96	97	92	94	90	98	87	94	93.5	79	75	74	80	80	75	76	74	76.8	06
Ours	95	97	94	100	93	99	93	94	**95.7**	91	84	92	97	89	92	86	87	**89.8**	**14**

4. EXPERIMENTATION AND RESULTS

4.1. Databases and Evaluation Metrics

The proposed face clustering mechanism is evaluated across a large corpus of images and videos from standard databases like YouTube Faces(YTF), YouTube Celebrities(YTC), eNTERFACE(eNF), Mind Reading(MR), Multi-Pie(Pie), Cohn Kanade Plus(CK), Indian Movie Face database(IM), and internal Samsung database(SFD) of movies, news, and personal galleries from mobile phones. Table 2 gives the details of all the databases, and few example samples of SFD are given at Fig. 5(a), and The samples are carefully chosen to cover at least one primary challenges among illumination, expressions, resolution, occlusions, make-ups, etc. We employ *precision* and *recall* as metrics to measure the clustering accuracy and *frames-per-second (fps)* to measure the processing speed.

Table 2. Database Details [I-Image, V-Video]

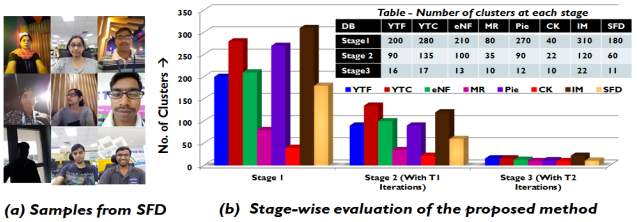
	YTF	YTC	eNF	MR	Pie	CK	IM	Ours
A	V	V	V	V	I	I	I	I/V
B	500	1910	1293	50	750K	2500	34K	8K/200
C	200	47	42	15	337	130	100	35
D	All	MUp	Illu	Exp	Pose	Exp	All	All

(Note: A:Data Type, B:#Samples, C:#Subjects, D:Variations, MUp:Make-up, Illu:Illumination, Exp:Expression)

4.2. Results and Comparison

To evaluate the performance of the proposed solution, we compare the results with several state-of-the-art methods including a conventional method of K-Means [10], Hierarchical clustering [11], linear sub-space clustering [9], Sequential clustering [23], [24] and a commercial face tagging solution present in Samsung Galaxy $S7$ phone. For K-Means and Hierarchical clustering, we preset optimal K value to retain high precisions, and measure recall. In case of sequential data clustering, we experiment only with video samples as proposed [23]. The results present in Table 1 shows the precision and recall rates of the proposed solution in comparison with competitor solutions. The comparison also include processing time taken on $IntelCore-I3$ CPU by each of these methods to cluster same size of data. It is very evident from the results that the proposed method is faster than most of the compared techniques, while achieving comparable precisions and better recall rates.

To further prove the effectiveness of the proposed method, we evaluated the clustering performance at each stages of the system. For this experiment, we selected highly challenging samples of 10 individuals from each database and evaluated the final number of clusters created for them. Fig. 5(b) shows that the proposed method is better in handling most of the data challenges, except a slightly declined result for faces with heavy facial make-ups from YTC. We also measured the proposed solution on Samsung Galaxy $S7$, and found that the solution has an better processing speed of $14fps$ as compare to $8fps$ of the existing solution on the device, and that makes it as a faster solution.



(a) Samples from SFD *(b) Stage-wise evaluation of the proposed method*

Fig. 5. *Stage-wise evaluation of the proposed method*

5. CONCLUSIONS

The proposed face clustering method of applying easy-to-hard grouping techniques is highly suitable for deploying on low-power devices like mobile phones. The concept of approximating the high precision initial clusters as image-sets using increasingly sophisticated features and classifiers aids to achieve improved recall rates, while maintaining high precision. The system complexity is also reduced by processing only non-redundant representative samples of clusters at each stage. The proposed solution is experimented on a large corpus of benchmark databases with wide range of real-world variations, and compared with many state-of-the-art and commercially solution to demonstrate its effectiveness.

6. REFERENCES

[1] P Lei and W.N Chong, "Unsupervised celebrity face naming in web videos," in *Trans. on Multimedia*. IEEE, 2015, pp. 854–866.

[2] O. Arandjelovic and A Zisserman, "Automatic face recognition for film character retrieval in feature-length films," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2005, pp. 860–867.

[3] L. Wolf, T. Hassner, and I Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2011.

[4] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2014.

[5] Z. Cui, W. Li, D. Xu, S. Shan, and X Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2013.

[6] E. Elhamifar and R Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," in *Pat. Analysis and Machine Intell.(PAMI)*. IEEE Trans., 2013, vol. 35(11), pp. 2765–2781.

[7] L. Zhang, V.K. Dmitri, and M Sharad, "A unified framework for context assisted face clustering," in *Proc. of the Intl. Conf. on Multi. Retrieval*. ACM, 2013, pp. 9–16.

[8] L. Wolf, T. Hassner, and Y Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," in *Trans. on Pat. Analysis and Machine Intell.(PAMI)*. IEEE, 2014.

[9] H. Cevikalp and B Triggs, "Face recognition based on image sets," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2010, pp. 2567–2573.

[10] C.C. Aggarwal and C.K Reddy, "Data clustering: algorithms and applications," in *CRC Press*, 2013.

[11] G. Karypis, E. Han, and V Kumar, "Hierarchical clustering using dynamic modeling," in *Computer*, 1999.

[12] S. Florian, K. Dmitry, and P James, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2015, pp. 815–823.

[13] S. Biswas, G. Aggarwal, P.J. Flynn, and K.W Bowyer, "Pose-robust recognition of low-resolution face images," in *Trans. on Pat. Analysis and Machine Intell.(PAMI)*. IEEE, 2013, vol. 35(12).

[14] C. Xianming, Hattiesburg, and Z Chaoyang, "Improve recognition performance by hybridizing principal component analysis and elastic bunch graph matching," in *Comp. Intell. for Multimedia*. IEEE, 2014.

[15] Y. Dong, L. Zhen, and S.Z Li, "Towards pose robust face recognition," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2013, vol. 35(12).

[16] L Chen, "Dual linear regression based classification for face cluster recognition," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2014.

[17] T.K. Kim, J. Kittler, and R Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations.," in *Pat. Analysis and Machine Intell.(TPAMI)*. IEEE Trans., 2007.

[18] Z. Xiangyu, L. Zhen, Y. Junjie, Yi. Dong, and Z.L Stan, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2015.

[19] W. Yongkang, S. Chen, S. Mau, C. Sanderson, and B.C Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPRW)*. IEEE, 2011, pp. 74–81.

[20] K. Fukui and Y Osamu, "Face recognition using multiviewpoint patterns for robot vision," in *The 11th Intl. Symposium on Robotics Research*. Springer Berlin Heidelberg, 2005, pp. 192–201.

[21] F. Jiao, S. Li, H.Y. Shum, and D Schuurmans, "Face alignment using statistical models and wavelet features," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2003.

[22] V. Sudha, G. Viswanath, A. Balasubramanian, M. Pratibha, and K.P. Basant, "Improved feature representation for robust facial action unit detection," in *Proc. of the Conf. on Computer Comm. and Network (CCNC)*. IEEE, 2013.

[23] T. Stephen, G. Junbin, and G Yi, "Subspace clustering for sequential data," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2014.

[24] W. Wen, W. Ruiping, H. Zhiwu, S. Shiguang, and C Xilin, "Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets," in *Proc. of the Conf. on Computer Vision and Pat. Recogn(CVPR)*. IEEE, 2015.