Multi-cue Augmented Face Clustering

Chengju Zhou[1,2,3], Changqing Zhang[1,2], Huazhu Fu[3], Rui Wang[1]*, Xiaochun Cao[1]
[1]State Key Laboratory of Information Security, IIE, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[3]School of Computer Engineering, Nanyang Technological University, Nanyang Avenue 639798, Singapore
chengjuzhou@outlook.com, zhangchangqing@tju.edu.cn, huazhufu@gmail.com,
wangrui@iie.ac.cn, caoxiaochun@iie.ac.cn

ABSTRACT

Face clustering is an important but challenging task since facial images always have huge variation due to change in facial expressions, head poses and partial occlusions, etc. Moreover, face clustering is actually an unsupervised problem which makes it more difficult to reach an accurate result. Fortunately, there are some cues that can be used to improve clustering performance. In this paper, two types of cues are employed. The first one is pairwise constraints: must-link and cannot-link constraints, which can be extracted from the temporal and spatial knowledge of data. The other is that each face is associated with a series of attributes (*i.e*, gender) which can contribute discrimination among faces. To take advantage of the above cues, we propose a new algorithm, Multi-cue Augmented Face Clustering (McAFC), which effectively incorporates the cues via graph-guided sparse subspace clustering technique. Specially, facial images from the same individual are encouraged to be connected while faces from different persons are restrained to be connected. Experiments on three face datasets from real-world videos show the improvements of our algorithm over the state-of-the-art methods.

Keywords

Face Clustering, Graph-guided, Sparse Representation

1. INTRODUCTION

Given a set of facial images, face clustering aims to separate them into different groups according to different individuals. This technique can be used in many fields, such as movie summation, content based image retrieval and automatic collection of large-scale face dataset, etc. General face clustering methods focus on how to separate faces only according to visual information. Most of them try to use the unlabeled facial images to obtain a good similarity representation [8, 9, 10, 6, 2, 15]. In [8], an affine invariant distance metric is proposed which is robust to different face poses and then [9] extends to Joint Manifold Distance (JMD) which represents a set of facial images of the same person detected in consecutive video frames as independent subspace. Hu et al. [10] introduces a between-set distance called Sparse Approximated Nearest Point (SANP) distance, where the dissimilarity of two sets is measured as the distance between their nearest points. In addition to the fully unsupervised clustering, there are also some methods with weak prior information. In [6], the scripts and subtitles are used to obtain cues as to which charactersare present. These weak cues for character presence are then combined with facial similarities to help the clustering. In [2], a multi-view clustering framework, called Diversity-induced Multi-view Subspace Clustering (DiMSC), is proposed to boost clustering performance by exploring the complementary information among multi-view features. Wolf et al. [15] proposes approach called Matched Background Similarity, in which can tell the differences between images with similar background, so that the similarities due to pose, lighting, and viewing conditions can be ignored.

An individual's facial images may have large variation due to change in facial expressions, head poses and partial occlusions, which make face clustering difficult for promising result. In some specific situation, some cues can be employed. For instance, in videos there are some inherent benefits: *faces in the same face track must be the same person while faces in the overlapped tracks can not belong to the same person*. This observation is referred as must-link and cannot-link constraints respectively and have been explored in [5, 16, 17, 19]. Cinbis and Verbeek [5] propose an unsupervised logistic discrinative metric learning (ULDML) method to learn a distance metric. The faces in the same track are pulled closer, while faces with the inter-track relation are pushed away from each other. Based on the Hidden Markov Random Fields (HMRF) model, a probabilistic constrained clustering method called HMRF-com [16] is proposed, in which the pairwise constraints, label-level and constraint-level local smoothness assumptions are incorporated together to guide the clustering process. The work in [19] proposes a video face clustering method which incorporates must-link and cannot-link constraints through constrained sparse representation. [17] develops a Weighted Block-Sparse Low Rank Representation (WBSLRR) to learn a more discriminative representation. However, these methods ignore the higher level attributes such as gender, skin color and hair style, etc. These attributes are consistent and robust in general. Some of these attributes are hard which means that they can not be changed. Therefore, the hard biometric attributes of facial images can be employed to guide the face clustering.

In this paper we explore the effectiveness of high level attributes for face clustering where the hard biometric attribute is employed to improve the clustering. We develop a novel face clustering method, Multi-cue Augmented Face Clustering (McAFC), in which the prior knowledge that can be represented as prior graphs (*i.e.*, pairwise

*Corresponding Author

MM'15, October 26-30, 2015, Brisbane, Australia ©2015 ACM.
ISBN 978-1-4503-3459-4/15/10 . . . $15.00
DOI: http://dx.doi.org/10.1145/2733373.2806290
.

constraints and attribute information) is integrated through graph-guided sparse representation effectively. We compare the proposed method with the state-of-art methods and show its improvements on real-world datasets.

The remaining of this paper is structured as follows: In Section 2, the Multi-cue Augmented Face Clustering is detailed. Section 3 shows the face clustering experiments on three face datasets from real-world videos. Finally, we conclude our paper in Section 4.

2. PROPOSED METHOD

Given a set of facial images $\{f_1, f_2, ..., f_N\}$, where N is the number of faces. Each facial images is represented as a feature vector $\mathbf{x}_i \in \mathbb{R}^D$. Then the whole dataset can be represented as a feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$.

2.1 Attributes and Constraints

There are several attributes that can be integrated into our proposed method, such as gender and skin color. In this paper we focus on using the gender information. We employ [4] to extract the gender information from each facial image. Note that faces in a face track belong to a same person and should have same gender information. Hence, the gender of a track should be determined as the mode value of the gender information in this face track. To utilize this prior, we build up the attribute-link matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, where N is the number of total facial images. The attribute-link matrix is constructed from the attribute information, in which the indices of face pairs that are different in gender are set to a negative value while others are 0. For convenience, we define $\mathcal{A} = \{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{A} | \boldsymbol{A}_{ij} \neq 0\}$ as the set of attribute-link information.

To describe the constraints of the video faces, we also build up two spatial-temporal constraint matrices: must-link matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ and cannot-link matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$. The matrix $\boldsymbol{M}$ represents the must-link constraints, where the indices corresponding to the face pairs in the same track are set to 1 while others are set to 0. The matrix $\boldsymbol{C}$ represents the cannot-link constraints, the elements of which corresponding to the face pairs belonging to the overlapped tracks are set to -1 while others are set to 0. We also define $\mathcal{M} = \{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} | \boldsymbol{M}_{ij} = 1\}$ and $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} | \boldsymbol{C}_{ij} = -1\}$ as the sets of the must-link and cannot-link constraints respectively.

2.2 Multi-cue Augmented Face Clustering

Ideally, the face $\mathbf{x}_i$ can be sparsely represented by small subset of faces from the same person [7] in the dataset. The relationship can be written as:

$$\mathbf{x}_i = \mathbf{X}\boldsymbol{\beta}_i \text{ s.t. } \boldsymbol{\beta}_{ii} = 0, i = 1, \ldots, N \tag{1}$$

where $\boldsymbol{\beta}_i \triangleq [\boldsymbol{\beta}_{1i}, \boldsymbol{\beta}_{2i}, ..., \boldsymbol{\beta}_{Ni}]^\top$ is the sparse representation of face $\mathbf{x}_i$, and the constraint $\boldsymbol{\beta}_{ii} = 0$ eliminates the trivial solution of representing a face as itself. Ideally, the coefficient vector $\boldsymbol{\beta}_i$ should have non-zero entries for these few facial images from the same person while the coefficient corresponding to different persons are zeros. In other words, the matrix $\mathbf{X}$ is a *self-expressive* dictionary in which each face can be rewritten as a linear combination of other faces in $\mathbf{X}$. To find a non-trivial sparse representation of $\mathbf{x}_i$, the tightest convex relaxation of the ℓ_1-norm is often employed, i.e.,

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}_i\right\|^2 + \lambda\left\|\boldsymbol{\beta}_i\right\|_1 \text{ s.t. } \boldsymbol{\beta}_{ji} = 0, i = 1, \ldots, N \tag{2}$$

where λ is regularization parameter which control the sparsity of $\boldsymbol{\beta}$. The optimization problem 2 can be solved efficiently using convex programming tools [1, 11].

2.2.1 *Graph-guided Sparse Representation*

Intuitively, the must-links can be regarded as a kind of positive prior that tell sparse representation which faces it should choose. On the contrary, the cannot-links and attribute-links can be referred as a kind of negative prior that indicates the sparse representation which faces it should not choose as its representation. Given these information, it is reasonable to assume that a sparse representation tends to choose the positive faces and neglect the negative ones. Therefore, we propose a graph-guided sparse representation as follows,

$$\begin{aligned}
&\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}_i\right\|^2 + \lambda\left\|\boldsymbol{\beta}_i\right\|_1 \\
&+ \gamma_M \sum_{(i,j)\in\mathcal{M}} \tau(\boldsymbol{M}_{ij}) \sum_{k=1}^{K} |\boldsymbol{\beta}_{ki} - sign(\boldsymbol{M}_{ij})\boldsymbol{\beta}_{kj}| \\
&+ \gamma_C \sum_{(i,j)\in\mathcal{C}} \tau(\boldsymbol{C}_{ij}) \sum_{k=1}^{K} |\boldsymbol{\beta}_{ki} - sign(\boldsymbol{C}_{ij})\boldsymbol{\beta}_{kj}| \\
&+ \gamma_A \sum_{(i,j)\in\mathcal{A}} \tau(\boldsymbol{A}_{ij}) \sum_{k=1}^{K} |\boldsymbol{\beta}_{ki} - sign(\boldsymbol{A}_{ij})\boldsymbol{\beta}_{kj}| \\
&\text{s.t. } \boldsymbol{\beta}_{ii} = 0, i = 1, \ldots, N
\end{aligned} \tag{3}$$

where $\lambda, \gamma_M, \gamma_C$ and γ_A are regularization parameters that control the complexity of the model. The third term tells which faces should tend to choose as its sparse representation. The last two terms show which faces should be avoided as its sparse presentation. The larger γ_M, γ_C and γ_A lead to a greater graph effect. In this paper, we use $\tau(r) = |r|$ [3]. Actually, any positive monotonically increasing function of the absolute value of correlations can be used. The $\tau(r)$ weights the fusion penalty for each face pairs such that β_{ki} and β_{kj} for highly correlated face pairs with large $|r|$ receive a greater graph effect than other pairs with weaker relationships. The $sign(r)$ indicates that two negatively correlated relationships are encouraged to have the same set of relevant representation coefficients with opposite sign. Note that the similarity coefficient used in spectral clustering should be non-negative. Therefore, we can approximately rewrite the optimization problem 3 as follows,

$$\begin{aligned}
&\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}_i\right\|^2 + \lambda\left\|\boldsymbol{\beta}_i\right\|_1 \\
&+ \gamma \sum_{(i,j)\in\mathcal{Y}} \tau(\boldsymbol{Y}_{ij}) \sum_{k=1}^{K} |\boldsymbol{\beta}_{ki} - sign(\boldsymbol{Y}_{ij})\boldsymbol{\beta}_{kj}| \\
&\text{s.t. } \boldsymbol{\beta}_{ii} = 0 \quad \text{and} \quad \boldsymbol{\beta} \geq 0, i = 1, \ldots, N
\end{aligned} \tag{4}$$

with

$$\boldsymbol{Y} = \gamma_M \boldsymbol{M} + \gamma_C \boldsymbol{C} + \gamma_A \boldsymbol{A}, \quad \mathcal{Y} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{Y} | \boldsymbol{Y}_{ij} \neq 0\}, \tag{5}$$

whereγ_M, γ_C and γ_A are the corresponding weight coefficients which can reveal effectiveness of different prior knowledge. Without loss of generality, we can rewrite the optmization problem 4 for all face $i = 1, \ldots, N$ in matrix form as follows,

$$\begin{aligned}
&\arg\min_{\boldsymbol{B}} \frac{1}{2}\left\|\mathbf{X} - \mathbf{X}\boldsymbol{B}\right\|^2 + \lambda\left\|\boldsymbol{B}\right\|_1 \\
&+ \gamma \sum_{(i,j)\in\mathcal{Y}} \tau(\boldsymbol{Y}_{ij}) \sum_{k=1}^{K} |\boldsymbol{B}_{ki} - sign(\boldsymbol{Y}_{ij})\boldsymbol{B}_{kj}| \\
&\text{s.t. } \boldsymbol{B}_{ii} = 0 \quad \text{and} \quad \boldsymbol{B} \geq 0, i = 1, \ldots, N
\end{aligned} \tag{6}$$

where $\boldsymbol{B} \triangleq \{\boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_N\} \in \mathbb{R}^{N\times N}$ is the sparse representation matrix, and the *i*-th column of which corresponds to the sparse representation of $\mathbf{x}_i$. $\boldsymbol{\beta}_i \in \mathbb{R}^N$ is the vector of elements of $\boldsymbol{B}$. The optimization problem 6 can be solved effectively utilizing [3].

2.2.2 *Constrained Spectral Clustering*

After solving the proposed optimization problem 6,we build a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V}$ denotes the set of *N* nodes in graph $\mathcal{G}$ corresponding to the set of *N* faces, and $\mathcal{E}$ denotes the edges between nodes. $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a symmetric non-negative similarity matrix representing the weights of the edges. Intuitively, an ideal similarity graph $\mathcal{G}$ should have connections corresponding to the same person and have no connections corresponding to different individuals. In the sparse representation solution $\boldsymbol{B}$, nonzero elements can be regarded as a measurement of the relationships between faces. This provides a choice of constructing the similarity matrix [7],

$$\mathbf{W} = \boldsymbol{B} + \boldsymbol{B}^{\mathsf{T}}, \tag{7}$$

where $\boldsymbol{B}$ is normalized as $\boldsymbol{\beta}_i \leftarrow \boldsymbol{\beta}_i / \|\boldsymbol{\beta}_i\|_\infty$ to make sure the weights in similarity graph are of the same scale. In this way, nodes in the graph are connected to each other with a same weight. Besides, the must-link faces should have a much higher weight coefficient than the cannot-links and faces with different hard biometric traits. In spectral clustering, we can also take advantages of the prior knowledge. A straightforward way is [19] as follows,

$$\mathbf{W}^{new} = \mathbf{W} + \eta_M \mathbf{M} + \eta_C \mathbf{C} + \eta_A \mathbf{A}, \tag{8}$$

where η_M, η_C and η_A are the corresponding weights and balance the effect of different kinds of prior knowledge in spectral clustering. In the clustering result, the faces belongs to must-links should be divided into a same cluster. This can be approximately reached by setting η_M with a value that is slighter larger than two which is the maximun value in weight matrix $\mathbf{W}$. The negative prior, attribute-links and cannot-links, means the corresponding face pairs should have a lower similarity coefficient. Therefore, a small value of η_C and η_A are taken for a lower coefficient while maintaining the nature of similarity matrix. Finally, the spectral clustering [13] is conducted on the new similarity matrix $\mathbf{W}^{new}$ to obtain the final clustering result.

3. EXPERIMENTS

3.1 Experimental Setting

Datasets: We conduct our experiments on three real-world datasets: Notting-Hill, TBBTS06E12 and YouTube_6. The Notting-Hill [16, 18] is extracted from the movie "Notting-Hill". Faces of 5 main casts are used, including 4660 faces in 76 tracks. The original dataset consists of the facial images with size of 120×150. The TBBTS06E12 is from the Season 6 Episodes 12 of TV series "The Big Bang Theory" [19]. The detected faces of 9 main casts are employed, including 17168 faces in 385 tracks. The third dataset is YouTube_6, which is a part of YouTube Face Dataset [15]. The facial images are derived from different videos and only face tracks are provided but no frame indices. Accordingly, there are no cannot-link constraints. We select the individuals of whom has at least 6 face tracks. Finally, we get 7266 facial images corresponding to 8 people, each of whom has 6 face tracks. Due to facial images from a face track belong to a same individual, we sample a part of faces from each track instead of using the whole track. The sample number are set 3, 3, 10 in three datasets respectively. We also downsample the original facial images to a corresponding size, which is 40×50, 50×50 and 50×50 for three datasets respectively and then vectorize the gray image as feature. These measurements can significantly reduce the computation complexity. In HMRFs, we follow [16] to use PCA to project the original gray scale feature space to a lower dimensional space which is equal to the number of casts.

Comparisons and Evaluation Criteria: In experiments, we compare our algorithm with some baselines and state-of-the-art methods: K-means [12], ULDML [5], HMRF-com [16], SSC [7], CS-VFC [19] and WBSLRR [17]. We also report the performance of method with only must-link and cannot-link constraints (McAFC$_c$) , only attribute-links (McAFC$_a$) and both of prior knowledge above mentioned (McAFC$_{c\&a}$) respectively. To obtain a comprehensive comparison, two standard measurements are employed to evaluate the clustering result: Accuracy, and Rand Index (RI) [14]. The Accuracy is calculated based on confusion matrix, which is derived from the match between the predicted labels of all faces and the ground-truth labels. The Rand Index is a measure of the similarity between clustering results. It evaluates true positives within clusters and true negatives between clusters. For each of the metrics, the higher it is, the better the performance is.

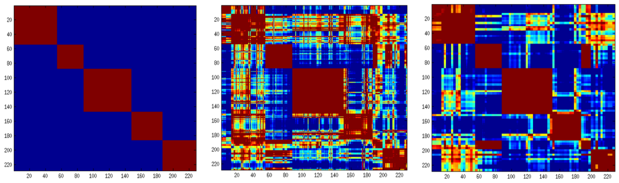
3.2 Quantitative and Qualitative Results

The detailed quantitative results are shown in Table 1. First, our proposed methods (McAFC$_c$, McAFC$_a$, McAFC$_{c\&a}$) achieve much better performances in three datasets. In Notting-Hill, CS-VFC and WBSLRR achieve about 7% improvement over SSC while McAFC$_c$ obtains about a more 2% improvement in terms of accuracy. The McAFC$_a$ also reaches the same performance with CS-VFC. With attribute and pairwise constraint cues, McAFC$_{c\&a}$ reaches about 96%. In dataset TBBTS06E12, our proposed graph-guided approach makes a remarkable improvement. McAFC$_c$ acquires a better performanc than CS-VFC and WBSLRR at least 6% improvement. With pairwise constraints and attributes, McAFC$_{c\&a}$ achieves about 28% improvement than SSC. The dataset YouTube_6 is more challenging due to the faces are from different videos and have a larger variation. The CS-VFC and WBSLRR reach about 20% improvement over SSC while McAFC$_c$ obtains about 23% improvement. McAFC$_a$ also reaches about 45.83% which has about 14% improvement. With two kinds of prior knowledge used, the McAFC$_{c\&a}$ achieves about 58.33% which is a comparable performance considering the challenges in YouTube_6. One can observe that McAFC$_c$ always achieves a better performance than McAFC$_a$ in all cases despite that attribute-links is much larger than constraints in quantity. This is not surprise because McAFC$_a$ just utilize the negative representation cues while McAFC$_c$ not only uses negative cues but also positive ones simultaneously. Our McAFC$_c$ achieves a better result than CS-VFC in three datasets, which shows that our proposed graph-guided approach is more effective than utilizing pairwise constraints with simple manner (*i.e.*, setting indices of must-link and cannot-link constraints to zeros). To sum up, our method outperforms the comparisons thanks to the multiple cues. Moreover, according to the results, we can find that, both the pairwise constraints (*i.e.*, must-links and canot-links) and the high level attributes contribute to improve the clustering performance.

The groundtruth, similarity and confusion matrices of SSC and McAFC are shown in Fig. 1 and Fig. 2. In Fig. 1, with prior knowledge, the similarity matrix of McAFC is more clear than SSC and it reveals the underlying data structure better. This can be further verified in the corresponding confusion matrix in Fig. 2. From the confusion matrices, we can find that 11 instances are wrongly clustered by SSC while McAFC only have 6 ones, so the whole clustering accuracy raises from 85.52% up to 92.10%. This also proves that our proposed method can significantly improve the clustering performance with multiple cues.

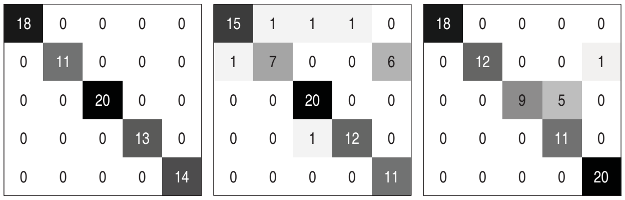
Table 1: Results (Mean ± Standard) on Notting-Hill, TBBTS06E12 and YouTube_6

	Notting-Hill		TBBTS06E12		YouTube_6	
	Accuracy	RI	Accuracy	RI	Accuracy	RI
K-Means [12]	59.21 ± 8.40	77.18 ± 3.49	53.00 ± 4.64	82.38 ± 0.99	40.25 ± 3.83	80.64 ± 3.01
ULDML [5]	55.26 ± 2.78	74.50 ± 2.98	56.73 ± 5.93	82.67 ± 0.34	33.33 ± 3.62	69.29 ± 0.21
HMRF-com [16]	70.21 ± 0.99	83.19 ± 1.53	55.32 ± 0.96	83.23 ± 1.57	40.13 ± 1.01	80.41 ± 0.81
SSC [7]	85.52 ± 0.94	88.52 ± 0.36	52.99 ± 0.23	81.92 ± 0.34	31.63 ± 3.47	63.88 ± 6.79
CS-VFC [19]	92.11 ± 0.89	95.41 ± 0.67	72.47 ± 0.89	87.55 ± 0.56	51.10 ± 3.11	82.39 ± 1.43
WBSLRR [17]	92.11 ± 0.24	95.52 ± 0.27	69.09 ± 2.46	86.58 ± 1.42	52.08 ± 2.09	83.36 ± 1.76
McAFC$_c$	94.73 ± 0.74	**96.59 ± 0.57**	78.70 ± 0.23	90.52 ± 0.83	54.17 ± 3.45	83.64 ± 2.34
McAFC$_a$	92.10 ± 0.68	93.56 ± 0.31	56.36 ± 0.16	82.17 ± 0.23	45.83 ± 0.63	80.12 ± 0.56
McAFC$_{c\&a}$	**96.05 ± 0.39**	96.07 ± 0.28	**80.51 ± 0.32**	**91.44 ± 0.12**	**58.33 ± 0.26**	**84.86 ± 0.19**



(a) Groundtruth (b) SSC (c) McAFC

Figure 1: Visualization of similarity matrices on Notting-Hill.



(a) Groundtruth

18	0	0	0	0
0	11	0	0	0
0	0	20	0	0
0	0	0	13	0
0	0	0	0	14

(b) SSC

15	1	1	1	0
1	7	0	0	6
0	0	20	0	0
0	0	1	12	0
0	0	0	0	11

(c) McAFC

18	0	0	0	0
0	12	0	0	1
0	0	9	5	0
0	0	0	11	0
0	0	0	0	20

Figure 2: Visualization of confusion matrices on Notting-Hill.

4. CONCLUSION

In this paper, we have proposed a face clustering approach, termed as Multi-cue Augmented Face Clustering (McAFC) to effectively take advantages of multiple cues. Specifically, the pairwise constraints, must-link and cannot-link constraints and face attributes knowledge, gender information, are effectively incorporated into the clustering process via graph-guided sparse representation to improve face clustering performance in two steps: sparse representation and spectral clsutering. The proposed approach is flexible to integrate prior knowledge for boosting the clustering performance. We have conducted experiments on three real-world video face datasets which demonstrate the effectiveness of our method. For future work, we hope to develop a framework to take advantage of multiple facial image attributes and explore how to utilize the relative attribute efficiently.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 61422213, 61332012,), 100 Talents Programme of The Chinese Academy of Sciences, and "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010-701).

6. REFERENCES

[1] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.

[3] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, E. P. Xing, et al. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[4] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, pages 3515–3522, 2013.

[5] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *ICCV*, pages 1559–1566, 2011.

[6] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking pictures: Temporal grouping and dialog-supervised person recognition. In *CVPR*, pages 1014–1021, 2010.

[7] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on PAMI*, 35(11):2765–2781, 2013.

[8] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. In *ECCV*, pages 304–320. 2002.

[9] A. W. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *CVPR*, pages 19–26, 2003.

[10] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011.

[11] S. J. Kim, K. Koh, S. Lustig, M. Byod, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized logistic regression. *JMLR*, 8(8):1519–1555, 2007.

[12] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, USA., 1967.

[13] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2:849–856, 2002.

[14] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[15] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[16] B. Y. Wu, Y. F. Zhang, B. G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *CVPR*, pages 3507–3514, 2013.

[17] S. Xiao, M. Tan, and D. Xu. Weighted block-sparse low rank representation for face clustering in videos. In *ECCV*, pages 123–138. 2014.

[18] Y. F. Zhang, C. S. Xu, H. Lu, and Y. Huang. Character identification in feature-length films using global face-name matching. *IEEE Transactions on Multimedia*, 11(7):1276–1288, 2009.

[19] C. Zhou, C. Zhang, X. Li, G. Shi, and X. Cao. Video face clustering via constrained sparse representation. In *ICME*, pages 1–6, 2014.