Face Clustering in Videos with Proportion Prior

Zhiqiang Tang[1], **Yifan Zhang**[1*], **Zechao Li**[2], **Hanqing Lu**[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology
zqtang2013@gmail.com, {yfzhang, luhq}@nlpr.ia.ac.cn, zechao.li@njust.edu.cn

Abstract

In this paper, we investigate the problem of face clustering in real-world videos. In many cases, the distribution of the face data is unbalanced. In movies or TV series videos, the leading casts appear quite often and the others appear much less. However, many clustering algorithms cannot well handle such severe unbalance between the data distribution, resulting in that the large class is split apart, and the small class is merged into the large ones and thus missing. On the other hand, the data distribution proportion information may be known beforehand. For example, we can obtain such information by counting the spoken lines of the characters in the script text. Hence, we propose to make use of the proportion prior to regularize the clustering. A Hidden Conditional Random Field(HCRF) model is presented to incorporate the proportion prior. In experiments on a public data set from real-world videos, we observe improvements on clustering performance against state-of-the-art methods.

1 Introduction

We investigate the problem of clustering faces in real-world videos. It is an important problem in computer vision [Cinbis *et al.*, 2011; Wu *et al.*, 2013; Xiao *et al.*, 2014], which can be applied to many fields, including automatically determining the cast of a feature-length film, content based video retrieval, rapid browsing and organization of video collections, automatic collection of large-scale face data set, etc. However, this task is challenging. In real-world videos, lighting conditions, facial expressions and head poses may drastically change the appearance of faces.

Face clustering is traditionally viewed as an unsupervised process. However, due to the difficulties mentioned above, many efforts have been devoted to seeking for extra knowledge beyond the data instances themselves to obtain weak supervision for clustering. In videos, the most often used knowledge is the relationships within and between the face tracks (where each face track is a sequence of faces) [Wu *et al.*, 2013; Xiao *et al.*, 2014; Zhang *et al.*, 2009], which form two types of constraints during clustering: 1) the must-link constraint: the faces in the same face track must belong to the same person, no matter how different the appearances of the faces look like; 2) the cannot-link constraint: if two face tracks overlap in some frames, then faces in these two tracks must belong to different persons, no matter how similar they look like. Such kind of knowledge can be considered as the "instance-level prior". They cannot regularize the clustering process globally.

Usually, most of the existing clustering algorithms cannot well cluster the data with unbalanced distributions. However, the unbalanced data are common in real-world applications. In movies or TV series videos, the leading casts appear quite often and the rest appear much less. However, many clustering algorithms cannot well handle such severe unbalance between the data distribution, resulting in that the large class is split apart, and the small class is merged into the large ones and thus missing. To deal with this problem, we propose to make use of the data distribution proportion prior during clustering to protect the large class from being split and the small class from being merged. Such knowledge can be easily obtained by exploiting the external sources such as the script text. One can count the spoken lines of each character in the script text to get the proportion information. Such kind of knowledge can be considered as the "cluster-level prior".

In this paper, we propose a probabilistic clustering model named as Hidden Conditional Random Field (HCRF). This model can incorporate both the instance-level prior and the cluster-level prior in a unified framework. The must-link and cannot-link constraints are naturally embedded in the pairwise potentials of the HCRF model. The proportion prior is set as a regularizer when we optimize the posterior probability of the model given the data.

Our HCRF model is different from the traditional CRF model as the observation model of the traditional CRF is learned in a supervised way. In our model, as there is no class label information available for the hidden nodes during model training, the parameters of the observation model are optimized in an Expectation-Maximization (EM) fashion. Our model is also different from the Hidden Markov Random Fields (HMRF)[Koller and Friedman, 2009] which is used in [Wu *et al.*, 2013]. HMRF is a generative model which focuses on describing how the labels can probabilistically "generate" observations, whereas HCRF is a discrim-

*Corresponding author.

inative model which does not expend modeling efforts on the observations. It directly describes how to take observations and assign them labels [Sutton and McCallum, 2012; Lafferty *et al.*, 2001].

In summary, the contributions of our work include:

1. We propose to use the new proportion prior knowledge generated from external textual sources in the face clustering.
2. As there is no class label known during clustering, it is difficult to determine which cluster best represents which class. Hence, we design a regularizer using the lower bound of KL divergence between the prior proportions and the clustering proportions to minimize the proportion loss.
3. By incorporating both the instance-level and cluster-level knowledge on the video face data, a HCRF model is developed for face clustering, whose effectiveness is demonstrated in experiments on a public data set with six real-world videos.

The rest of the paper is organized as follows. After reviewing relevant previous works in Section 2, we describe the HCRF model and the use of proportion prior in Section 3, and provide the optimization in Section 4. Experimental evaluations and comparisons of our method are reported in Section 5 and Section 6 concludes the paper.

2 Related work

Face clustering in videos has become a hot topic in recent years. The related works can be grouped into two categories: purely data-driven methods and clustering with prior knowledge. Most of the unsupervised methods[Fitzgibbon and Zisserman, 2002; 2003; R.Wang *et al.*, 2008; Hu *et al.*, 2011; Arandjelovic and Cipolla, 2006] focused on obtaining a good distance measure or mapping raw data to a new space for better representing the structure of the inter-personal dissimilarities from the unlabeled faces.

Above data-driven methods exploit the information inside of the data instances without using any external knowledge. In the videos, the easily available must-link and cannot-link constraints generated from the relationships within and between the face tracks have been explored in the face clustering. In [Vretos *et al.*, 2011], the constraints are exploited to modify the distance matrix and to guide the clustering. However, the method is very computationally expensive. As reported in [Vretos *et al.*, 2011], it takes about 6 days on a data set of 10000 faces. Cinbis et al. [Cinbis *et al.*, 2011] proposed an unsupervised logistic discriminative metric learning (ULDML) method. A metric is learned such that must-linked faces are close, while cannot-linked faces are far from each other. More recently, Wu et al. [Wu *et al.*, 2013] proposed a probabilistic constrained clustering method based on the Hidden Markov Random Fields (HMRF) model for face clustering in videos. The latest work on face clustering with priors was presented by Xiao et al. [Xiao *et al.*, 2014]. They learn a weighted block-sparse low rank representation (WBSLRR). A weighted block-sparse regularizer on the data representation is designed to incorporate the available constraints, so that the resultant data representation is more discriminative.

Face clustering using prior knowledge can be treated as a constrained clustering problem, which has been studied in the works such as COP-KMEANS [Wagstaff *et al.*, 2001], constrained EM [Shental *et al.*, 2004], HMRF-KMeans [Basu *et al.*, 2006] and Penalized Probabilistic Clustering(PPC) [Lu and Todd, 2007]. However, most of the prior knowledge used is on the instance level. Few works exploit the data distribution proportion prior in clustering, which has effects on the cluster level. This is because during the clustering, there is no class label known. It is difficult to determine which cluster best represents which class. Hence, a few works use the proportion prior on classification problems. Yu et al. [Yu *et al.*, 2013] proposed a method called proportion-SVM, which explicitly models the latent unknown instance labels together with the known group label proportions in a large-margin framework. Lefort et al. [Lefort *et al.*, 2011] addressed the inference of probabilistic classification models using the proportion prior.

3 Our method

The faces in videos are usually collected in the form of face tracks. A video usually contains many face tracks appearing along the timeline in the video. During the face clustering, faces in a face track should belong to one cluster while the faces from temporally overlapped face tracks should be assigned into different clusters. Therefore, the problem of face clustering in videos with proportion prior can be defined as follows:

3.1 Problem formulation

Given unlabeled data $X = \{x_1, x_2, \cdots, x_n | x_i \in \mathbb{R}^d\}$, our goal is to partition X into K (predefined) disjointed clusters. The latent categorical class label set is denoted as $Y = \{y_1, y_2, \cdots, y_n\}, y \in \{1, 2, \cdots, K\}$. Further provided is the proportion prior for each class in the whole data set, denoted as $\Pi = \{\pi_1, \cdots, \pi_K\}$, where $\pi_i \in [0, 1]$ is the proportion of class i; and the must-link and cannot-link constraints, denoted as $C = \{C_{ml}, C_{cl}\}$, where the must-link constraint $C_{ml}(x_i, x_j)$ indicates that x_i and x_j should belong to the same cluster while the cannot-link constraint $C_{cl}(x_i, x_j)$ indicates that x_i and x_j should be assigned into different clusters. Based on this information, we want to find a model $f : X \to Y$ which can best predict $y \in Y$ for observations $x \in X$, pairwise constraints C and proportion prior Π.

3.2 Hidden Conditional Random Field

We introduce a novel model by generalizing the traditional CRF, named as Hidden Conditional Random Field (HCRF), which is shown in Fig. 1. The label variables Y are unobserved, and conditioning on the observations X. Y constitute a Markov random field. The general formulation is given as follows:

$$P(Y|X; \Theta, \beta) = \frac{1}{Z(X)} \prod_{i}^{n} \psi_u(y_i|X; \Theta) \psi_p(y_i, y_{N_i}|X; \beta), \tag{1}$$

where ψ_u is the unary potential function, ψ_p is the pairwise potential function, $Z(X) =$

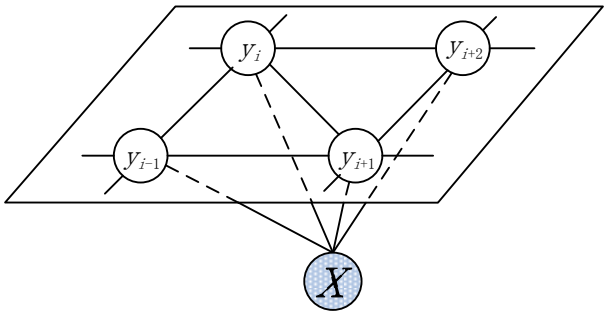

Figure 1: Hidden Conditional Random Field model

$\sum_y \prod_i^n (\psi_u(y_i|X;\Theta)\psi_p(y_i, y_{N_i}|X;\beta))$ is the partition function, y_{N_i} is the neighborhood set of y_i, Θ and β are the parameters.

A model with the same name "HCRF" was also given in [Quattoni *et al.*, 2007], which is different from ours. They assume there is an additional latent layer between the observations X and the observed label variables Y. That HCRF is a three-layer model, and used in supervised learning, while our HCRF is a two-layer model, and used in unsupervised learning.

Unary potential function

The unary potential models the relations between observations X and label variables Y. Different from the traditional CRF in which the unary potential function is formulated as the multi-class logistic regression, we adopt the simple exponential function as follows:

$$\psi_u(y_i|x_i;\Theta) = \prod_{k=1}^{K} (exp(\theta_k^T x_i))^{\mathbb{I}(y_i=k)}, \tag{2}$$

where $\Theta = \{\theta_1, \cdots, \theta_K\}$. $\mathbb{I}(\cdot)$ denotes the indicator function: if $a = 1$, then $\mathbb{I}(a) = 1$, otherwise $\mathbb{I}(a) = 0$.

Pairwise potential function

The pairwise potential embeds the correlations between label variables Y. Since the initial pairwise constraint matrix W that represents Constraints C is usually sparse, it provides limited information to guide clustering. Constraint propagation based on the local smoothness assumption can produce many soft constraints. We use the constraint propagation in [Lu and Ip, 2010] to compute our neighborhood system V as follows:

$$V = (1-\alpha)^2(I - \alpha L)^{-1} W (I - \alpha L)^{-1}, \tag{3}$$

where $\alpha \in (0, 1)$ is the constraint propagation degree and L is the laplacian matrix. V is a dense constraint matrix where $-1 \leq v_{ij} \leq 1$. $v_{ij} > 0$ means the positive correlation, i.e., y_i and y_j should be same; $v_{ij} < 0$ means the negative correlation, i.e., y_i and y_j should be different; $v_{ij} = 0$ means no correlation. $|v_{ij}|$ indicates the constraint confidence.

Based on neighborhood system V, the pairwise potential between y_i and y_j is defined as follows:

$$\begin{aligned} \psi_p(y_i, y_j|X;\beta) &= \exp(\beta\phi(y_i, y_j)) \\ &= \exp(\beta[P(y_i = y_j) - \mathbb{I}(v_{ij} > 0)]v_{ij}), \end{aligned} \tag{4}$$

where $\phi(y_i, y_j)$ denotes the pairwise energy function, $P(y_i = y_j)$ is the probability that $y_i = y_j$, β is the trade-off parameter between the unary and pairwise potentials, and it will be learned. $\mathbb{I}(\cdot)$ denotes the indicator function.

3.3 Cluster proportion loss

In order to deal with the unbalanced distribution of cluster proportions, we use the KL divergence between prior proportions and the actual proportions as the proportion loss. The difficulty to use the proportion prior is that the class label of each cluster is unknown in the clustering. Thus, we cannot determine which prior proportion corresponds to which cluster. There are $K!$ possible correspondences between the prior proportions and the clusters. It is computationally infeasible and unnecessary to enumerate all the correspondences to find the most likely one that minimizes the KL divergence between the two proportion distributions. Assume the actual cluster proportions are denoted as $B = \{b_1, \cdots, b_K\}$, we use the KL divergence $D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B}))$ as the proportion loss, where $\mathbf{sort}(\cdot)$ is the vector in which the elements are sorted in the ascending(or descending) order.

Theorem 1. $D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B}))$ is the lower bound of $D_{KL}(\Pi||B)$.

Proof. Given any bijection correspondence between sets Π and B in which any element of set Π corresponds to exactly one element of set B, we sort elements of set Π in the ascending(or descending) order to get vector $\mathbf{sort}(\mathbf{\Pi}) = \pi_1 \cdots \pi_K$ while maintain the bijection relation to get another vector $\mathbf{B_1} = b_1 \cdots b_K$, where π_i corresponds to b_i and $1 \leq i \leq K$. Then we choose any two elements b_i and b_j, subject to that $i < j$ and $(\pi_i - \pi_j) \ln \frac{b_i}{b_j} \leq 0$, i.e., $(\pi_i - \pi_j)(b_i - b_j) \leq 0$, and swap them until vector $\mathbf{B_1}$ becomes vector $\mathbf{sort}(\mathbf{B})$. The change of the KL divergence for the h^{th} swapping can be computed as follows:

$$\begin{aligned} \nabla_h &= D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{B_{h+1}}) - D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{B_h}) \\ &= \pi_i \ln \frac{\pi_i}{b_j} + \pi_j \ln \frac{\pi_j}{b_i} - \pi_i \ln \frac{\pi_i}{b_i} - \pi_j \ln \frac{\pi_j}{b_j} \\ &= \pi_i \ln \frac{b_i}{b_j} + \pi_j \ln \frac{b_j}{b_i} \\ &= (\pi_i - \pi_j) \ln \frac{b_i}{b_j} \\ &\leq 0 \end{aligned}$$

where vector $\mathbf{B_h}$ becomes new vector $\mathbf{B_{h+1}}$ after the h^{th} swapping. After all the swappings, the total change of the KL divergence is $\nabla = D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B})) - D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{B_1}) = \sum_h \nabla_h \leq 0$. Therefore, $D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B}))$ is the lower bound of $D_{KL}(\Pi||B)$.

4 Objective function and optimization

4.1 Objective function

The goal is to find the optimal labels Y^* and parameters Θ^* and β^* that can best explain observations X, pairwise constraints C and proportion prior Π. Since computing the partition function Z in Eq. (1) is intractable, we use the pseudo-

Algorithm 1 EM for Hidden Conditional Random Field

Input: data X, cluster number K, neighborhood system V and proportion prior Π

Output: optimal parameters Θ^*, β^* and optimal labels Y^*

1: Initialize $\Theta^{(0)}$ and $P^{(0)}(Y|X)$ by Kmeans;
2: **while** not converge **do**
3: t=t+1;
4: learn $\Theta^{(t)}$ based on $P^{(t-1)}(Y|X)$, $\beta^{(t-1)}$ and Π;
5: learn $\beta^{(t)}$ based on $P^{(t-1)}(Y|X)$, $\Theta^{(t)}$ and Π;
6: get $P^{(t)}(Y|X)$ from $P^{(t-1)}(Y|X)$, $\Theta^{(t)}$ and $\beta^{(t)}$;
7: **end while**
8: **return** $\Theta^* = \Theta^{(t)}$, $\beta^* = \beta^{(t)}$ and $Y^* = \arg\max_Y P^{(t)}(Y|X)$

likelihood to approximate it. In this way, the posterior probabilities of data X can be factorized as follows:

$$P(Y|X;\Theta,\beta) = \prod_i^n P(y_i|x_i, y_{N_i};\Theta,\beta), \tag{5}$$

with

$$P(y_i|x_i, y_{N_i};\Theta,\beta) = \frac{\psi_u(y_i|x_i;\Theta)\psi_p(y_i, y_{N_i}|X;\beta)}{z_i}, \tag{6}$$

$$\psi_p(y_i, y_{N_i}|X;\beta) = \prod_{j\in N_i} \psi_p(y_i, y_j|X;\beta), \tag{7}$$

where $z_i = \sum_{y_i} \psi_u(y_i|x_i;\Theta)\psi_p(y_i, y_{N_i}|X;\beta)$ is the local partition function.

In order to avoid the extremely small values caused by the production of many posterior probabilities, we turn to the log likelihood. Therefore, the objective function based on the log pseudo-likelihood can be written as follows:

$$\begin{aligned} Y^*, \Theta^*, \beta^* &= \arg\max_{\Theta,\beta}(\log P(Y|X;\Theta,\beta) \\ &\quad - \mu D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B}))) \\ &= \arg\max_{\Theta,\beta}(\sum_{i=1}^n \log P(y_i|x_i, y_{N_i};\Theta,\beta) \\ &\quad - \mu D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B}))). \end{aligned} \tag{8}$$

4.2 Optimization

Our HCRF is first initialized via Kmeans and then optimized in an expectation-maximization(EM) framework. The posterior probabilities of each $x_i \in X$ in K different states are expected and then parameters Θ and β are learned by maximizing the expectations. At last, the optimal labels Y^* can be inferred with optimal parameters Θ^* and β^*. The main framework is given in Algorithm 1.

E step

We use the mean field theory to approximate the Markov dependencies between label variables Y. In the t^{th} EM iteration, when estimating the probability of label variable y_i in state k, i.e., $P^{(t)}(y_i = k|x_i, y_{N_i};\Theta^{(t)},\beta^{(t)})$, we consider the probabilities of its neighborhood label variables y_{N_i} in that state, i.e., $\{P^{(t-1)}(y_j = k|x_j, y_{N_i};\Theta^{(t-1)},\beta^{(t-1)})\}_{y_j\in y_{N_i}}$ to compute the pairwise potentials of y_i. For clarity, we abbreviate $P^{(t)}(y_i|x_i, y_{N_i};\Theta^{(t)},\beta^{(t)})$ as $P^{(t)}(y_i|x_i)$.

M step

In the t^{th} EM iteration, given the posterior probability $P^{(t-1)}(y_i|x_i)$, we update the parameters by maximizing the expectation of the log pseudo-likelihood in Eq. (8). Parameters Θ and β are learned once in a sequential manner. First, parameter Θ can be optimized by maximizing the following objective function.

$$\begin{aligned} \Theta^{(t)} = \arg\max_{\Theta}(&\sum_{i=1}^n \sum_{y_i} P^{(t-1)}(y_i|x_i)\log P(y_i|x_i, y_{N_i};\Theta, \\ &\beta^{(t-1)}) - \mu D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B})) + \lambda_1\|\Theta\|^2), \end{aligned} \tag{9}$$

where the regularization term $\lambda_1\|\Theta\|^2$ is to prevent the overfitting of Θ. We set $\lambda_1 = 10^{-4}$ in the experiments. Vector $\mathbf{sort}(\mathbf{\Pi}) = \pi_1 \cdots \pi_K$ and vector $\mathbf{sort}(\mathbf{B}) = b_1 \cdots b_K$ have the same sorted order. Since $\Theta = \{\theta_1, \cdots, \theta_K\}$, we need to compute K gradients. The gradient of Eq. (9) with respect to $\theta_j (1 \le j \le K)$ is

$$\begin{aligned} \nabla\theta_j = &\sum_{i=1}^n x_i(P^{(t-1)}(y_i = j|x_i) - P(y_i = j|x_i, y_{N_i};\Theta, \\ &\beta^{(t-1)})) - \mu\sum_k \pi_k(\sum_k \frac{\partial b_k}{\partial\theta_j}/\sum_k b_k - \frac{\partial b_k}{\partial\theta_j}/b_k) \\ &+ 2\lambda_1\|\theta_j\|, \end{aligned} \tag{10}$$

with

$$\begin{aligned} \frac{\partial b_k}{\partial\theta_j} = &\sum_{i=1}^n x_i P(y_i = k|x_i, y_{N_i};\Theta,\beta^{(t-1)})(\mathbb{I}(j = k) - \\ &P(y_i = j|x_i, y_{N_i};\Theta,\beta^{(t-1)})), \end{aligned} \tag{11}$$

$$b_k = \sum_{i=1}^n P(y_i = k|x_i, y_{N_i};\Theta,\beta^{(t-1)}). \tag{12}$$

Then parameter β can be learned through maximizing the following objective function.

$$\begin{aligned} \beta^{(t)} = \arg\max_{\beta}(&\sum_{i=1}^n \sum_{y_i} P^{(t-1)}(y_i|x_i)\log P(y_i|x_i, y_{N_i}; \\ &\Theta^{(t)},\beta) - \mu D_{KL}(\mathbf{sort}(\mathbf{\Pi})||\mathbf{sort}(\mathbf{B})) + \lambda_2\|\beta\|^2), \end{aligned} \tag{13}$$

where $\lambda_2\|\beta\|^2$ it to avoid the overfitting of β. λ_2 is set to 10^{-4} in the experiments. The gradient of Eq. (13) with regard to β is

$$\begin{aligned} \nabla\beta = &\sum_{i=1}^n \sum_{y_i} (P^{(t-1)}(y_i|x_i) - P(y_i|x_i, y_{N_i};\Theta^{(t)},\beta))\phi(y_i, \\ &y_{N_i}) - \mu\sum_k \pi_k(\sum_k \frac{\partial b_k}{\partial\beta}/\sum_k b_k - \frac{\partial b_k}{\partial\beta}/b_k) \\ &+ 2\lambda_2\|\beta\|, \end{aligned} \tag{14}$$

with

$$\begin{aligned} \frac{\partial b_k}{\partial\beta} = &\sum_{i=1}^n P(y_i = k|x_i, y_{N_i};\Theta^{(t)},\beta)(\phi(y_i = k, y_{N_i}) - \\ &\sum_k P(y_i = k|x_i, y_{N_i};\Theta^{(t)},\beta)\phi(y_i = k, y_{N_i})), \end{aligned} \tag{15}$$

$$b_k = \sum_{i=1}^{n} P(y_i = k | x_i, y_{N_i}; \Theta^{(t)}, \beta), \tag{16}$$

where $\phi(y_i = k, y_{N_i}) = \sum_{j \in N_i} \phi(y_i = k, y_j)$ is the neighborhood energy of y_i in state k.

5 Experiments

5.1 Experimental settings

We evaluate the performance of our method in the public face data set Big Bang Theory(BBT) given in [Bauml *et al.*, 2013]. It contains episodes 1-6 in the first season of BBT. Each episode has about 20 minutes with 4-6 characters. We get their proportion prior from the script by counting their spoken lines. The character proportions from both the groundtruth and script in the 6 episodes are presented in Fig. 3. It can be observed that the proportion distribution is unbalanced as the biggest proportion is usually 3-10 times the smallest proportion. We can also find that the prior proportion distributions from the script are similar to the groundtruth proportion distributions. This verifies the rationality of our motivation to use the script proportion prior to facilitate the face clustering in the video.

We directly utilize the extracted face data in [Bauml *et al.*, 2013]. Each episode consists of a list of face tracks and each face track has a sequence of faces. The feature of each face is represented by a 240 dimensional Discrete Cosine Transform (DCT) vector. In the experiments, we use the same data preprocessing and result statistics for all the methods. Considering the huge amount of face data in the videos and that the nearby faces in a face track are very similar, we sample the faces to reduce the data volume. First, we uniformly sample 3 faces from each face track. For all the sampled faces in one video, we compute a laplacian matrix in the which the k-nearest neighbor graph is used and $k = 10$. Then the Laplacian Eigenmaps reduces the feature dimension from 240 to the cluster number. After we get the posterior probability of each sampled face belonging to each cluster, we compute the posterior probabilities of each track. The probability of track t being labeled with k is calculated as $p(t, k) = \frac{1}{n} \sum_{i=1}^{n} P(y_i = k | x_i)$, where n is the sampled face number and x_i is the i^{th} sampled face in a track. Then we can get the track label by $y(t) = \arg\max_k p(t, k)$. At last, the face track labels are used to evaluate the clustering performance based on the measures of both accuracy and normalized mutual information(NMI).

5.2 Competing methods

We compare our method to [Wu *et al.*, 2013] and [Lu and Ip, 2010]. They both utilize the pairwise constraints in the clustering and separately achieve state-of-the-art performance when in comparison with different former methods. [Wu *et al.*, 2013] is based on the Hidden Markov Random Fields(HMRF), while [Lu and Ip, 2010] is a constrained spectral clustering called E^2CP. Although WBSLRR[Xiao *et al.*, 2014] is more recent and achieves better results, it mainly seeks for the better data representation to improve the clustering performance. Since our method focuses on the clustering rather than the data representation, the comparison with it is not provided. The Kmeans is used as the baseline.

Besides, since the proportion prior in the video face clustering is first investigated in this work, we also design another strategy to utilize it for the comparison. It is straightforward to consider the proportion prior as the cluster prior and use it in the Bayes' rule. Thus, we use it as the third potential function in our model. By incorporating the third potential function, Eq. (6) becomes

$$P(y_i | x_i, y_{N_i}; \Theta, \beta) = \frac{\psi_u(y_i | x_i; \Theta) \psi_p(y_i, y_{N_i} | X; \beta) \pi_{y_i}}{z_i}. \tag{17}$$

In order to use the proportion prior in this way, we need to first find the correspondences between the clusters and the prior proportions. In every iteration of the EM algorithm, we use the cluster sizes as the cues to assign the priors to clusters. We call this strategy as "Bayes prior" in later comparisons.

5.3 Experimental results and analysis

Since we use the EM algorithm to solve our HCRF, we want to investigate how the face clustering performance changes along with the iterations. Fig. 2 shows the average face clustering performance curves of baseline HCRF and its extension with the KL divergence regularizer in 6 videos of BBT. It can be observed that our HCRF largely improves the initial Kmeans clustering accuracy and NMI by about 13% and 10% and it converges quickly in only about 5 iterations. Adding the KL divergence regularizer delays the the convergence but brings obvious improvements over baseline HCRF by about 5% accuracy and about %6 NMI.

Tables 1 and 2 give the comparison of different methods measured by the clustering accuracy and NMI. We can find that our HCRF largely outperforms HMRF and E^2CP by more than 7% in both accuracy and NMI. HMRF assumes that the observation node given the hidden node follows the Gaussian distribution. The last step of E^2CP is the Kmeans clustering which also favors the Gaussian distribution assumption. If the Gaussian assumption better satisfies, then the two approaches may perform better. However, in real applications, it is known that the face data usually locate on a manifold, thus do not follow the Gaussian distribution. In contrast, our HCRF uses an exponential function to represent the unary potential, which does not expend modeling efforts on the distribution over the observations.

Based on Table 1 and Table 2, our two strategies of using the script proportion prior can both promote the face clustering performance. For the average improvements of accuracy and NMI on the basis of baseline HCRF, the Bayes prior brings about 3.6% and 3.5%, while the KL divergence regularizer contributes about 5.6% and 5.9%. Obviously, the KL divergence regularizer performs better than the Bayes prior. The reason can be explained as follows. For the Bayes prior strategy, the prior proportions are assigned to the clusters once in each EM iteration and then the correspondences between them are fixed in the iteration. This is a greedy assignment strategy. If the assignment is wrong, it cannot be corrected in the current iteration and the wrongness may be magnified in the later iterations. In contrast, the KL diver-

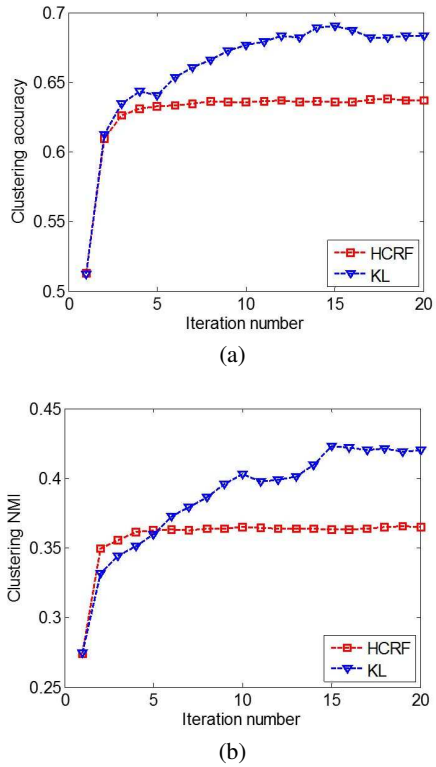

Figure 2: Clustering performance in EM iterations. (a) and (b) show clustering accuracy and NMI curves with the EM iteration number. KL denotes HCRF adding the KL divergence regularizer.

gence regularizer is minimized in the optimization of parameters. During the optimization, the correspondences between the prior proportions and the clusters are not fixed. In each iteration of the gradient descent algorithm, the prior proportions are reassigned to the clusters to compute the KL divergence. By using some techniques to avoid many local minimas, the gradient descent algorithm can take more advantage of the KL divergence regularizer to find better clustering results.

Table 1: Clustering accuracies. "Bayes" and "KL" denote the two strategies of using the proportion prior based on HCRF.

	BBT1	BBT2	BBT3	BBT4	BBT5	BBT6	**Avg.**
Kmeans	0.588	0.523	0.587	0.476	0.433	0.459	0.512
E^2CP	0.648	0.548	0.615	0.513	0.495	0.510	0.556
HMRF	0.668	0.564	0.649	0.543	0.488	0.523	0.574
HCRF	0.737	0.667	0.650	0.600	0.588	0.603	0.643
Bayes	0.784	0.699	0.659	0.612	0.624	0.637	0.679
KL	**0.787**	**0.736**	**0.700**	**0.658**	**0.629**	**0.671**	**0.699**

6 Conclusion

In this paper, we have presented the studying on data clustering with cluster proportion prior. We have shown how to exploit the readily available proportion prior to guide the face clustering. A HCRF model has been proposed to predict the

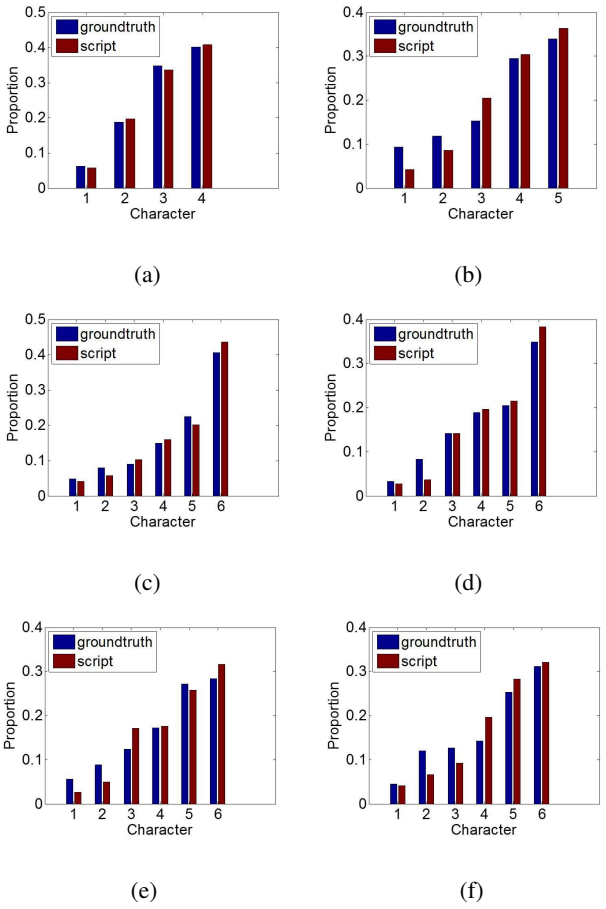

Figure 3: Character proportion distributions. (a)-(f) are the proportion distributions of characters in the 6 BBT episodes. The blue histogram is the groundtruth distribution of the video faces, while the red histogram represents the name distribution from the script.

Table 2: Clustering NMIs. "Bayes" and "KL" denote the two strategies of using the proportion prior based on HCRF.

	BBT1	BBT2	BBT3	BBT4	BBT5	BBT6	**Avg.**
Kmeans	0.314	0.256	0.311	0.256	0.266	0.240	0.274
E^2CP	0.352	0.280	0.344	0.291	0.290	0.277	0.305
HMRF	0.376	0.274	0.346	0.276	0.272	0.262	0.302
HCRF	0.426	0.371	0.351	0.332	0.381	0.379	0.375
Bayes	0.485	0.402	0.392	0.359	0.399	0.411	0.410
KL	**0.491**	**0.454**	**0.400**	**0.401**	**0.403**	**0.440**	**0.434**

class labels of observations. The parameters of the model are learned with the regularization of the proportion loss term in a unified iterative optimization framework. The proposed method has been verified on one public face data set with six real-world TV series episodes. The experiments demonstrate that our HCRF largely outperforms previous state-of-the-art methods and the proposed proportion prior can further improve the face clustering performance.

Acknowledgments

This work was supported in part by the 863 Program (2014AA015104), and the National Natural Science Foundation of China (61332016, 61202325, 61402228).

References

[Arandjelovic and Cipolla, 2006] O. Arandjelovic and R. Cipolla. Automatic cast listing in featurelength films with anisotropic manifold space. In *CVPR*, pages 1513–1520, 2006.

[Basu *et al.*, 2006] S. Basu, M. Bilenko, A. Banerjee, and R. J. Mooney. Probabilistic semi-supervised clustering with constraints. *Semi-Supervised Learning*, chapter 5:71–98, 2006.

[Bauml *et al.*, 2013] Martin Bauml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3602–3609. IEEE, 2013.

[Cinbis *et al.*, 2011] R. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *Proceedings of International Conference on Computer Vision*, 2011.

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

[Fitzgibbon and Zisserman, 2003] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *CVPR*, 2003.

[Hu *et al.*, 2011] Y. Hu, A.S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[Lefort *et al.*, 2011] R Lefort, R Fablet, and JM Boucher. Object recognition using proportion-based prior information: Application to fisheries acoustics. *Pattern Recognition Letters*, 2011.

[Lu and Ip, 2010] Zhiwu Lu and Horace H. S. Ip. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *Proceedings of European Conference on Computer Vision*, pages 1–14, 2010.

[Lu and Todd, 2007] Z. Lu and K. Todd. Penalized probabilistic clustering. *Neural Computation*, 19(6):1528–1567, 2007.

[Quattoni *et al.*, 2007] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1848, 2007.

[R.Wang *et al.*, 2008] R.Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008.

[Shental *et al.*, 2004] N. Shental, A. Bar-Hillel, T. Hertz, and D.Weinshall. Computing gaussian mixture models with em using equivalence constraints. In *Advances in Neural Information Processing Systems*, 2004.

[Sutton and McCallum, 2012] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[Vretos *et al.*, 2011] N. Vretos, V. Solachidis, and I. Pitas. A mutual information based face clustering algorithm for movie content analysis. *Image and Vision Computing*, 2011.

[Wagstaff *et al.*, 2001] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrdl. Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, pages 577–584, 2001.

[Wu *et al.*, 2013] Baoyuan Wu, Yifan Zhang, Bao-Gang Hu, and Qiang Ji. Constrained clustering and its application to face clustering videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[Xiao *et al.*, 2014] Shijie Xiao, Mingkui Tan, and Dong Xu. Weighted block-sparse low rank representation for face clustering in videos. In *Proceedings of European Conference on Computer Vision*, 2014.

[Yu *et al.*, 2013] Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. Proportion-svm for learning with label proportions. In *International Conference on Machine Learning*, pages 504–512, 2013.

[Zhang *et al.*, 2009] Yi-Fan Zhang, Changsheng Xu, Hanqing Lu, and Yeh-Min Huang. Character identification in feature-length films using global face-name matching. *IEEE Transactions on Multimedia*, 11(7):1276–1288, 2009.