Semi-supervised Cast Indexing for Feature-length Films

Wei Fan[1], Tao Wang[2], JeanYves Bouguet[2], Wei Hu[2], Yimin Zhang[2], and Dit-Yan Yeung[1]

[1] Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong
{fwkevin, dyyeung}@cse.ust.hk
[2] Intel China Research Center, Beijing, P.R. China, 100080
{tao.wang, Jean-yves.bouguet, wei.hu, yimin.zhang}@intel.com

Abstract. Cast indexing is a very important application for content-based video browsing and retrieval, since the characters in feature-length films and TV series are always the major focus of interest to the audience. By cast indexing, we can discover the main cast list from long videos and further retrieve the characters of interest and their relevant shots for efficient browsing. This paper proposes a novel cast indexing approach based on hierarchical clustering, semi-supervised learning and linear discriminant analysis of the facial images appearing in the video sequence. The method first extracts local SIFT features from detected frontal faces of each shot, and then utilizes hierarchical clustering and Relevant Component Analysis (RCA) to discover main cast. Furthermore, according to the user's feedback, we project all the face images to a set of the most discriminant axes learned by Linear Discriminant Analysis (LDA) to facilitate the retrieval of relevant shots of specified person. Extensive experimental results on movie and TV series demonstrate that the proposed approach can efficiently discover the main characters in such videos and retrieve their associated shots.

1 Introduction

The ongoing expansion of multimedia information in the world wide web and the entertainment industry has generated increasing requirements for semantic based video mining techniques, such as news/sports summarization, film/TV abstraction and home video retrieval. Among various contents in these video data, characters are always the major focus of interest to the audience. In this paper, we utilize one of the most important visual cues, human face, to discover active characters who frequently appear in the feature-length films and retrieve their associated shots for efficient browsing.

Over the past few decades, there has been a good deal of investigation into automatic face detection and recognition techniques in the field of computer vision and pattern recognition [9]. However, due to the large variation of pose, expression and illumination conditions, robust face recognition is still a challenging goal to achieve, especially for the scenario of still images. Recently, a

significant trend in performing video-based face analysis has emerged, which aims to overcome the above limitations by utilizing visual dynamics or temporal consistence to enhance the recognition performance. In [6] Arandjelovic and Zisserman apply affine warping to mitigate the effect of various poses. However, it is unable to deal with the out-of-plan face rotation problem. The person spotting system [4] associates multiple exemplars of each person in the shot as a compact face-track to cover a person's range and expression changes. The approach constructs multiple patterns to improve the performance, but may fail in some shots with insufficient exemplars, which is often the case in movies and TV series. The multi-view 3D face model is described in [3] to enhance the video-based face recognition performance. However, it is very difficult to accurately recover the head pose parameters by the state-of-art registration techniques, and therefore not practical for real-world applications.

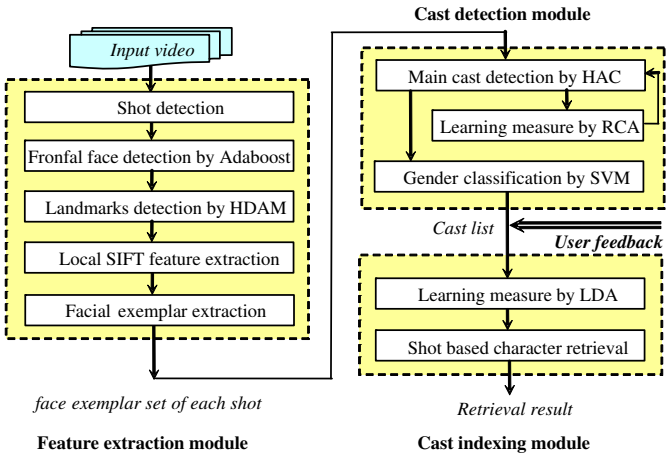

Fig. 1. Framework of the cast indexing system

As mentioned above, feature-length films contain multiple instances of each person's face that can be associated by visual tracking, speech identification and user feedback. Thus it is possible to improve the cast indexing performance by utilizing the complementary facial information under different pose, illumination and expression conditions. Motivated by this idea, we propose a novel semi-supervised cast indexing approach for feature-length films by hierarchical clustering, relevant component analysis and linear discriminant analysis. The framework consists of three modules as shown in Figure 1. In the feature-extraction module, near frontal faces are sequentially detected from sampling frames of the whole video, and then multiple facial exemplars in each shot are extracted by clustering and connected by tracking. We calculate the SIFT features in 5 local facial regions to jointly describe the face image. In the cast detection module, main characters are detected by partial Hierarchical Agglomerative Clustering (HAC) [8] and a semi-supervised learning algorithm – Relevant Component Anal-

ysis (RCA) [10] iteratively. These face clusters are sorted by detected gender and appearing frequency (corresponding to the cluster size). Since faces of the same person with significant pose or expression variations may be unavoidably classified into a few separate clusters, it is necessary to utilize user feedback to further merge these duplicate clusters. Finally, the cast indexing module applies RCA and Linear Discriminant Analysis (LDA) [12] to learn a discriminative distance measure from the HAC output and the refined cast list, and then, in this discriminative feature space, retrieves associated shots for the characters of interest for the users.

The rest of this paper is organized as follows. In section 2, we describe the proposed method in detail, including feature extraction, main cast detection, and main cast retrieval. To evaluate the performance of this approach, extensive experiments are reported in section 3, followed by some concluding remarks in section 4.

2 Method details

2.1 Shot detection

Similar to document mining by parsing the textual content in the form of words, sentences, paragraphs and the whole document, video mining can be analyzed in four hierarchical levels – frame, shot, scene and the whole sequence. To well characterize the video content, shot detection is a prerequisite step and the basic processing unit of most video mining systems.

A shot is a set of video frames captured by a single camera in one consecutive recording action. According to whether the transition between shots is abrupt or not, the shot boundaries are categorized to two types, namely, Cut Transition (CT) and Gradual Transition (GT). In our work, we use a shot detection algorithm from Tsinghua University which achieved the best result in TRECVID 2004 and 2005 [5]. Its CT detector uses the 2nd order derivatives of color histogram, a flash light detector and a GT filter. Its GT detector uses motion vectors and the feature outputs from the CT detector.

2.2 Facial feature extraction

After shot detection, we use Viola and Jones' *'AdaBoost + Cascade'* face detector [7] to extract near frontal faces from temporal sampling frames in each shot. By automatic localization of four facial landmarks (centers of two eyes, nose and mouth) [2], each face is geometrically aligned into the standard normalized form to remove the variation in transition, scale, in-plane rotation and slight out-of-plane rotation. Then facial features are extracted from the normalized gray face images.

It is demonstrated that local features outperform global ones in most recognition and verification tasks, since they are more robust to partial occlusions, pose and illumination variations [1]. In our approach, we first apply Hierarchical

Direct Appearance Model (HDAM) [2] to detect facial landmark points and then extract the SIFT features [1] in five local rectangular regions, covering two eyes, central region of two eyes, nose, and forehead, as shown in Figure 2.

(a) (b) (c)

Fig. 2. Local SIFT feature extraction. (a) The original face image; (b) Detected 30 facial landmarks using HDAM; (c)Five local face regions for SIFT feature extraction.

As the basic processing unit in videos, a shot may contain NULL, one or more actors' faces. The faces of the same person in one shot can be easily detected by tracking the continuous positions of facial landmark points. To effectively characterize the variation of different poses, expressions and illumination conditions, we perform the *basic leader-follower clustering* algorithm [8] to generate multiple face exemplars for the same person in each shot. Thus a person appearing in one shot is represented by a representative face-exemplar set. The face-set distance measure between two shots S_i and S_j is defined by the shortest element-pair distance between the two sets as Eq(1):

$$d(S_i, S_j) = \min_{m,n} |x_{i,n} - x_{j,n}|/dim \tag{1}$$

where $x_{i,m} \in S_i, x_{j,n} \in S_j$ are the concatenated local SIFT feature vectors, $|.|$ is the L_1 distance and $dim = 5 \times 128$ is the dimension of the feature vector. The *basic leader-follower clustering* algorithm is described as following:

Algorithm (Basic leader-follower clustering)
1 begin initialization $\theta = threshold$
2 $C_1 = \{x\}, N = 1$
3 **do** accept new x
4 $j = \arg\min_i \|x - C_i\|$ $(i = 1, \ldots, N)$ //find the nearest cluster C_j
5 **if** $distance(x, C_j) < \theta$ //belong the same person
6 $C_j = C_j + \{x\}$
7 **else** create new cluster $C_{N+1} = \{x\}, N = N + 1$
8 **until** no more samples x
9 **return** $C_1, C_2, \ldots, C_N$
10 **end**

2.3 Main cast detection using HAC

In most feature-length films, the main characters are the persons who frequently appeare in different shots, resulting in large numbers of similar face images, e.g.

frontal faces. Based on this observation, the main characters can be discovered by clustering all the selected shots using the distance measure proposed in Eq(1).

It is well known that facial features, represented as high-dimensional pixel arrays, often belong to a nonlinear manifold of intrinsically low dimensionality [11]. The variations between the facial features of the same person under different pose, illumination and expression are almost always larger than the variations due to changes in face identity. Therefore, in the clustering process, we do not partition all the shots by "flat" algorithms (e.g. K-means or spectral clustering) which will unavoidably group different persons into the same cluster. Instead, we perform Hierarchical Agglomerative Clustering (HAC) [8] to merge similar face shots whose distances are below a strict threshold, i.e. the clustering process will terminate once the merging face-set distance exceeds a pre-selected threshold. The threshold is set low enough to make sure that the two merged clusters are from the same person. As illustrated in Figure 3, the dendrogram shows how the shots are grouped by HAC, which well reflects the similarity relationship among different characters.

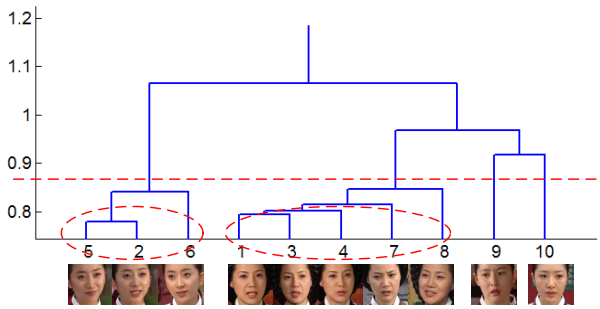

Fig. 3. Main cast detection by HAC on 10 shots. The HAC stops clustering when the face-set distance between shot 9 and shot 10 exceeds the threshold $\theta = 0.87$.

Algorithm (Agglomerative hierarchical clustering)
1 **begin initialization** $S_1, S_2, \ldots, S_n, \theta = threshold, N_{end}, F$
2 $N = n, C_i = \{S_i\}, i = 1, 2, \ldots, N$
3 **do** $N = N - 1$
4 Find nearest clusters, say C_i and C_j
5 **if** $\|C_i - C_j\| < \theta$ //make sure to be the same person by θ
6 merge C_i and C_j
7 **else** break
8 **until** $N = N_{end}$
9 **return** *sorted cluster with cluster size* $> F$ *(shots)*
10 **end**

After HAC procedure, the output clusters are sorted according to their sizes. Only clusters which contain more than F shots (i.e. the frontal face appears at least in F shots) are selected as the main characters. Furthermore, genders

of the main cast are detected by an SVM classifier using the local SIFT facial features (Figure 5). In our work, RBF kernel based SVM classifier is trained on a dataset of 2000 labeled samples and performs well for most of the videos with an averaged precision of 90%. According to the user's preference, the cast list can be also organized by their ages, poses or expressions for convenient browsing.

The main cast detection process is fully automatic. Although exemplars of each cluster belong to the same person, it is unavoidable that a person may appear in a few clusters due to the large variation of poses and expressions etc. The accuracy can be further refined by semi-supervised learning in section 2.4 and user's feedback.

2.4 Refine main cast detection using RCA

For many clustering and classification algorithms, such as K-means, SVM, and K nearest neighbor (KNN) etc., learning a good distance metric from training examples is the key to their success. Since exemplars of each cluster belong to the same person, each cluster is a *chunklet* [10]. We define "chunklet" as a subset of data points that are known to belong to the same although unknown class. From this kind of side-information in the form of *equivalence relations*, we learn a better distance metric in a semi-supervised manner and further perform the main cast detection using HAC.

In our approach, we employ Relevant Component Analysis (RCA) [10] to improve the feature space of HAC. The RCA algorithm has been theoretically shown to be an optimal semi-supervised learning procedure from the information theoretic perspective. By learning a Mahalanobis metric from chunklets, RCA transforms the original feature x into a new representation y, which assigns large weights to "relevant dimensions" and low weights to "irrelevant dimensions". Thus in the new feature space, the inherent structure of the data can be more easily unraveled for clustering.The RCA algorithm is described as following:

Algorithm (Relevant Component Analysis)
1 **Begin initialization** k chunklets $\{x_{ji}\}_{i=1}^{n_j}$ with means $m_j, j = 1, \ldots, k$
2 Compute the scatter matrix
- $C = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - m_j)(x_{ji} - m_j)^T$
3 Compute the whitening transformation matrix by SVD
- $W = C^{-1/2}$
4 Transform the original feature x to the new feature $y = W \cdot x$
5 **end**

In the case of singular matrix C of high dimensional features, SVD is applied to calculate the transformation matrix W. Figure 4 (a) simulates the manifolds of facial features of two persons, where two chunklets are marked as red circles and blue circles respectively. Figure 4 (b) is the transformed features using RCA. It can be seen that transformed manifold becomes more separate. A constrained k-means clustering over the original feature space gives poor result with an accuracy of 60% (Figure 4 (c)). However, through the RCA transformation, the

constrained K-means achieves significant improved performance with an accuracy of 96% (Figure 4 (d)).

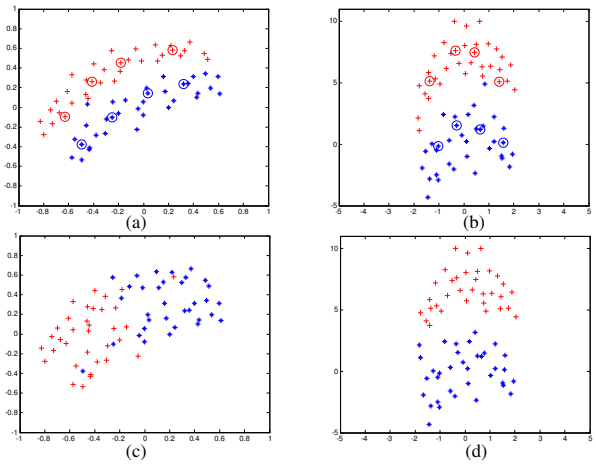

Fig. 4. (a) A 2-class clustering problem where each cluster has 4 labeled feedback samples as chunklets; (b) Data set after RCA transformation; (c) Constrained k-means clustering over the original space; (d) Constrained k-means clustering over the transformed RCA feature space.

2.5 Main cast retrieval using LDA

By main cast detection of section 2.3 and 2.4, we discovered main characters and most of their multi-view facial exemplars in the video. Since faces of the same person may be classified into a few different clusters, it is necessary to utilize the user's feedback to refine the finial cast list by indicating which clusters belong to the same person. To retrieval relevant shots of these main characters for efficient browsing, we apply a nearest neighbor matching in the Linear Discriminant Analysis (LDA) [12] subspace of the above feature space.

LDA is a well-known technique for dealing with the class separability problem and determining the set of the most discriminant projection axes. The most widely used LDA approach seeks an optimal projection from the input space onto a lower-dimensional discriminating feature space as Eq(2).

$$W_{opt} = \arg\max_{W} \frac{|W^T S_b W|}{|W^T S_w W|} \tag{2}$$

with the within class scatter matrix $S_w = \sum_{i=1}^{L} \sum_{x_k \in X_i} (x_k - m_i)(x_k - m_i)^T$ and the between class scatter matrix $S_b = \sum_{i=1}^{L} n_i (m_i - m)(m_i - m)^T$. Here m_i is the mean face of class X_i, m is the mean face of all classes, and n_i is the number of samples in class X_i. The optimal projection matrix W_{opt} can be constructed by the eigenvectors of $S_w^{-1} S_b$. To avoid degeneration of S_w, we

first reduce the feature dimensionality by PCA, and then perform discriminant analysis in the reduced PCA subspace. By applying this method, we find the projection directions that maximize the Euclidean distance between the face images of different classes and minimize the distance between the face images of the same class. An example of main cast retrieval is illustrated in Figure 6.

3 Experiment

To demonstrate the performance of the proposed cast indexing approach, extensive experiments were conducted on a story TV series of "Da ChangJin" and an action movie of "007 Die Another Day", totaling up to 3 hours of videos. "DaChangjin" is a hot Korea TV series with 594 shots and 67006 frames (45min). The main characters are Chang Jin, Jin Ying, Cui ShangGong, Shang Shan, Min ZhengHao etc. "007 die another day" is a famous action movie with 1652 shots and 237600 frames (132min). The main cast includes James Bond, Jinx Johnson, Gustav Graves, Miranda Frost, Zao etc.

In the experiments, we temporally sample each shot by 5 frames per second to reduce the duplicated images and computational burden. The detected main cast of "Da ChangJin" and "007 Die Another Day" are shown in Figure 5, which are organized according to their gender for convenient browsing. It can be observed that there are some duplicate faces which correspond to large pose, illumination and expression variations of the same character. The gender is detected by RBF kernel based SVM classifier on local SIFT features. Table 1 illustrates the gender classification performance with F-score above 93%. The F-score $= 2 \times Pr \times Re/(Pr + Re)$ evaluates the comprehensive performance.

Table 1. Performance of Gender classification by SVM.

Gender	Precision (%)	Recall (%)	F-score (%)
female	97	90	93
male	95	98	97

Fig. 5. Automatically detected main cast of "Da Changjin" and "007 Die Another Day" by HAC and RCA.

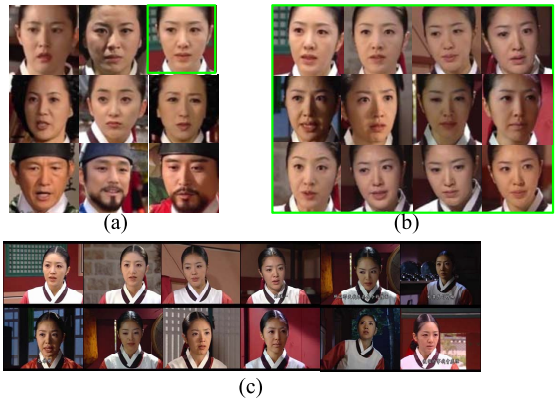

Fig. 6. An example of main cast retrieval. (a) The main cast list. (b) The face-exemplar set of one actress "Jin Ying". (c) Key frames of the retrieved shots for the query person "Jin Ying".

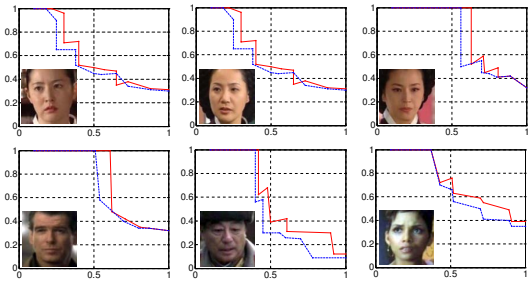

Fig. 7. The precision vs. recall curves of main cast retrieval of "Da ChangJin" and "007 Die Another Day". The red solid curves are the RCA based retrieval result. The blue dashed curves are the retrieval results without RCA.

According to user's feedback, we manually merge clusters of the same person to refine the finial cast list and get the multiple exemplars of each character. By these exemplars, LDA learns the discriminative transform W to retrieve relevant shots of the query person. Figure 6 illustrates one retrieval procedure of a main actress "JinYing" in the TV series "Da ChangJin". The curves of six main actors in "Da ChangJin" and "007 Die Another Day" videos are shown in Figure 7 and Table 2. It can be observed that LDA significantly improves the shot retrieval performance and achieves good cast retrieval result.

4 Conclusion

In this paper, we proposed a novel semi-supervised cast indexing approach using HAC, RCA and LDA. The method first detects near frontal faces from temporal sampling frames of each shot and then adopts partial hierarchical agglomerative clustering (HAC) and semi-supervised learning algorithm RCA to discover the

Table 2. Performance of the main cast retrieval of "DaChangJin" using RCA and LDA.

Character	Precision (%)	Recall (%)	F-score (%)
CuiShangGong	85.7	93.1	89.25
HanShangGong	78.1	100	87.70
JinYing	85.2	100	92.01
ChangJin	95	57.1	72.69
LingLu	100	55.6	71.47
HuangShang	100	100	100
ShangShan	100	54.5	70.55

main cast. To refine the accuracy of automatic main cast detection, user's feedback is employed by indicating which clusters belong to the same person. Then by these multiple exemplars of main characters, Linear Discriminant Analysis (LDA) algorithm learns a discriminative distance measure to retrieve relevant shots of the query person in the whole video. Extensive experimental results on movies and TV series demonstrate the effectiveness of the approach. In future work, we'll take advantage of multiple cues such as speech, music, clothing, close caption, and tracking etc. to improve the cast indexing performance and further retrieve the highlight scenes of main characters.

References

1. D. Lowe: Distinctive image features from scale-invariant keypoints. IJCV. **60** (2004) 315–333
2. G. Song, H. Ai, G. Xu: Hierarchical direct appearance model for elastic labeled graph localization. Proc of SPIE (2003) 139–144
3. J. Kittler, A. Hilton, M. Hamouz, J. Illingworth: 3D assisted face recognition: a survey of 3D imaging, modelling and recognition approaches. Proc. of IEEE CVPR (2005) 144–144
4. J. Sivic, M. Everingham, and A. Zisserman: Person spotting: video shot retrieval for face sets. Proc. of IEEE CIVR (2005) 226-236
5. J.H. Yuan, W.J. Zheng, L. Chen, etc.: Tsinghua University a TRECVID 2004: shot boundary detection and high-level feature extraction. NIST workshop of TRECVID. (2004)
6. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell: Face recognition with image sets using manifold density divergence. Proc. of IEEE CVPR (2005) 581- 588
7. P. Viola, M. Jones: Rapid object detection using a boosted cascade of simple features. Proc. of IEEE CIVR (2001) 511–518
8. R. Duda, P. Hart, D. Stork: Pattern Classification. Wiley (2000)
9. W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld: Face recognition: a literature survey. ACM Comput. Surv. **35** (2003) 399-458
10. BarHillel, T. Hertz, M. Shental, D. Weinshall: Learning distance fucntions using equivalence relations. Proc. of ICML (2003)
11. S. Roweis and L. Saul: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000)
12. P. Belhumeur, J. Hespanha, D. Kriegman: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. on PAMI. **19** (1997) 711-720