A Novel Method for Movie Character Identification Based on Graph Matching: A Survey

Mr. B. S. Salve
ME II Computer
VPCOE, Baramati
Pune University (MH), India.
E-mail: salvebs1486@gmail.com

Prof. S. A. Shinde
Assistant Professor
VPCOE, Baramati
Pune University (MH), India
E-mail: meetsan_shinde@yahoo.com

Abstract— Automatic face identification of character in movies received tremendous attention from both video content understanding and video annotation because of their application in movie industry such as video semantic analysis, video summarization, and personalized video retrieval.

Character identification of movie is challenging problem due to huge variation in the appearance of each character and complex background, large motion, non-rigid deformation, occlusion, huge pose, expression, wearing, clothing, even makeup and hairstyle changes and other uncontrolled condition make the result of face detection and face tracking unreliable.

In particular, character identification for movie used video and script. Face tracking and clustering from video and name of person extract from script. Many challenges for face clustering and face-name matching are present. In good situation and clean environment existing methods gives better result, but in a complex movie scene performance is limited because face tracking and clustering process generate a noise.

In this paper we present a comparative study of three methods using textual cues like cast list, script, subtitle and closed caption based on local and global face-name matching.

Keywords*- Character identification; Multimedia databas; Multiple Kernel Learning (MKL); Multimedia Information system; Histogram of Oriented gradient (HOG); Video; Optical character recognition (OCR); Earth mover distance (EMD); Error correcting graph matching (ECGM).*

*****

I. INTRODUCTION

With the big development of movie industry a large amount of movie data is being generated every day. When you watch a movie or TV, mostly you don't know the all characters names in movie video. In movie audience focused on character and real name of character, Sometime people takes third person references for real character name identification.

A large amount of digital video data produces for making TV serial and movie videos, but it required efficient and effective techniques for video content understanding and organization of video data. So, automatic face identification of characters in movies is called video annotation, this technique is to identify the faces of character in the video and label them corresponding names in the cast list. Also some methods are used textual cues like cast list, script, subtitle and closed caption for character identification. The existing method gives promising result in clean environment, but challenging problem due to the huge variation in the appearance of each character.

II. NEED OF AUTOMATIC CHARACTER IDENTIFICATION

A. For movie Index and Retrieval.
B. For scene segmentation.
C. For movie summarization.
D. For Dynamic Captioning Style.

A. for movie Index and Retrieval -

The objective is to label television or movie footage with the names of people in each frame of video. For poor image quality and motion blur of video this techniques is essential to improve the performance of video.

B. For scene segmentation -

This method is used for analysis and alignment of co-occurrences in movie video and script of the videos. Scene is the elemental unit to constitute a sub story in the movie. Accurate scene segmentation not only facilities movie content understanding but also affect sub story detection.

Content based method segments movies scene have low level audio visual feature, which lack of necessary information. Before scene segmentation each movie shot first represented a bag of character, first, for character identification to construct character histogram and identify the leading characters. Then script of movie that record complete scene structure and related character name mapping from script to movies a semantic scene structure can be made. After character identification the movie is converted into a shot sequence.

C. Movie Summarization-

Character based movie summarization is helpful for movie producer to promote the movie as well as audience capture the theme of the movie before watching the whole movie. Mostly movie summarization approaches based on video content only which may not deliver ideal result due to the semantic gap between computers calculated low level feature and human used high level understanding. The purpose of this need is to select portion that most attract audience, attention from the origin movie.

Modification theory “All film are about nothing-nothing but characters,” which reveals character are important for movie summarization. Movie summarization based in character analysis utilizes a character relation to exploit the movie structure including the scene segmentation and sub story discovery.

D. Captioning Style-

For video accessibility enhancement and hearing impairment used dynamic captioning. There are more than 66 million people suffering from hearing impairment and this disability brings them differently in video content understanding due to loss of audio information. So by using script and dynamic captioning are help them in certain degree by synchronously illustrating the script during the playing of video. So that these techniques to help having impaired audience better recognize the speaking character and the hearing impaired audience enjoy video.

III. CHALLENGES

1. Weakly supervised textual cues.
2. Character identification in video is more difficult than in image.
3. The same character appears quit differently during movie.

IV. CLASSIFICATION

The character identification problem is occur relation between videos and associated texts in order to label the faces of character with the name.

A. METHOD I: CAST LIST BASED

In this method only the cast list textual resources are used for face-name relationship in matching.

In this method only the cast list textual resources are used for face-name relationship in matching.

The ‘cast list problem’ discovery problem are founded by Andrew Fitzgibbon and A. Zisserman [03] in clustering and automatic cast list in movies. This system presented a new affine invariant distance metric which efficiently manages prior on the transformation parameter and showing the use of “trust region” and ‘Levenberg- Marquardt’ strategies in nonlinear optimization. The power of this metric for unsupervised clustering has been demonstrated by automatically extracted the principal cast from video.

O.Arandjelovic and cipolla [04**]** uses anisotropic manifold space to determine automatically cast list of feature length film. This is difficult because the cast size not known, with appearance changes of faces caused by extrinsic imaging factor such as illumination, pose expression often greater than due to different identities. This method proposed on algorithm for clustering over face appearance manifold. Also algorithms for exploiting coherence of dissimilarity between manifold.

Ramanana [05] proposed system for labeling character in large archival video collections. In this system first group a frontal face together such that they have same label. Build a color histogram model for face, hair and then it tracks in neighboring frames around this grouped detection and adding dynamic constrains in addition to using aforementioned feature. The remaining group of tracks means character faces grouped based to wearing consistent clothing during a scene and based on body appearance from matching across shot changes, hence temporally disjoint tracks are merged into groups with the same identity. In this method construct labeled dataset by hand labeling a small number of groups which can done quickly because one is labeling clustering rather than individual tracks. So the labeled library is acquired by labeling the track cluster from single episode. In this method implementing a system for 11 year worth of archival footage from television show friends, the 611770 faces dataset is implement order of magnitude.

Mark everingham and Andrew Zisserm [06**]** have addressed the problem of finding a particular character by building a classifier of the character appearance from the cast list.

This method consists of – 1.A 3D model of an individual’s face and head is built .This to be rendered in novel views, giving extrapolation from the few training image provided. 2. The tree structure classifier is trained to detect the individual and estimate the pose over a very wide range of scale and pose. 3. Initial estimate of pose are refined and identify, verified using a generative approach and employing edge feature and matching to give robustness to lighting and expression changes.

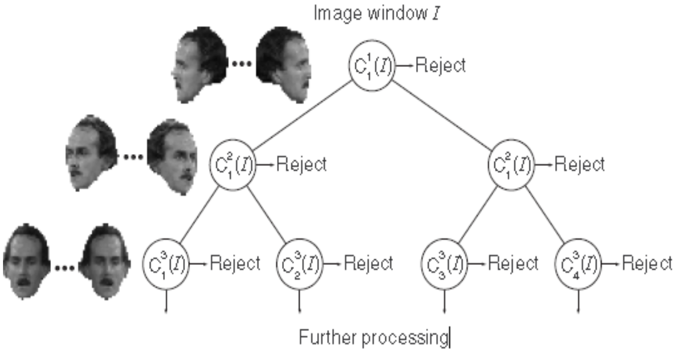

Figure 1- Detection and pose estimation using the tree structured classifier [06].

The classifier at each node detects a range of poses which is subset of the parents.

The detector builds using a tree structure. The space of 3D rotation is divided into successively smaller partition using a tree structure also binary tree is shown for clarity. The root node corresponds to the full ranges of poses from +/- 90 azimuths and +/- 30 elevations and in plane rotation. Tree consist 1024 leaves corresponding to different poses. Each node consist classifier trained to detect images of the head in the corresponding ranges of poses. If the classifier responds to the input image, the children of the node are explored; else the entire branch of the tree is deleted or pruned.

Ramzan G Cinbis, Jokob V and Cardelid schmid [07] introduced face identification in TV video by using unsupervised metric learning. It is to decide whether two faces depict the same person or not. This system automatically label characters in TV series or movies based on subtitle and script, but the problem it is enable effective transfer of the sparse text based supervision to other faces. In [07] addressed this problem, in this system without manually labeling any examples, metric learning can be effectively used.

A.1. METHOD I: ADVANTAGES

- The method in [04] extracting faces appearance manifold and anisotropically growing their class boundaries in the corresponding manifold space. It has been demonstrated to achieve good automatic cast listing in film.
- The process of [05] tracking and clustering is fully automatic. Detecting frontal face, building face and hair model for detection, tracking using the body models, then the cluster the body models from across a video to link up tracks from different shots.
- In [06] detection and pose estimation over a very wide range of poses is computationally efficient because of early pruning of the search. The accuracy of the detector is improved greatly by using a sequence of classifier instead of single classifier.

A.2. METHOD I: DISADVANTAGES

- The current version of [03] given algorithm is poor tolerance to change in expression of character. The results in merging of cluster containing different characters.
- For increasing the clustering robustness it required to employ a more sophisticated way of comparing appearance manifold. In [03], [04] faces are clustered by appearance and faces of a particular characters are expected to be a collected in few pure cluster. Names for the clusters are then manually selected from the cast list.
- The resulting images [05] are based to be response of frontal face detector.

B. METHOD II: SUBTITLE OR CLOSED CAPTION, LOCAL MATCHING BASED

Mark Everingham, Josef Sivic and Andrew Zisserman [08] implement automatic naming of character in TV video series. The appearance of each character represented exemplar based and robustness of pose, lighting and expression variation of the facial appearance is obtained by using parts based descriptor extracted around detected facial feature.

Authors of [08] extended their work in [09], in this method implement the seamless tracking, integration and recognition of profile and frontal detection and a character specific multiple kernel classifier which is able to learn the feature best able to discriminate between characters. To improve coverage i.e. the number of character that can be identified and the number of frames over which they are tracked and to improve accuracy i.e. correctly identify the characters. The frontal views consist of detection, tracking, facial feature, speaker detection. To improve the accuracy author defining a kernel for each descriptor and learning a discriminative classifier using a linear combination of these kernels.

For classification used linear combination of kernels i.e. Multiple Kernel Learning (MKL).It is used to determine the combination of feature used. Optimal combination of feature is learnt and other has considered multiple features. E.g. eye, a spatial region of hair etc. Two face detector used one for approximately frontal faces and other for approximately "3/4 view" to full left profile. Also detector implemented using multiscale sliding window classifiers. Histogram of oriented gradients (HOG) feature extraction is used and linear support vector machine (SVM) for two tracks i and j the composite kernel has the form

$$K(i,j) = \sum b_f k_f(i, j)$$

Where, $k_f(i, j)$ - kernel to feature f between tracks i and j.

b_f- base kernels.

Cour Jordan and Taskar [10] implemented movie/script alignment and parsing of video and text transcription. It contains - 1.Novel probabilistic model and inference procedure from shot treading and scene alignment driven by text. 2. Extraction of verb frames and pronoun resolution from screen play and 3. Retrieval of corresponding action inform by scene structure and character naming.

Following figure 2, Shows alignment between video, screenplay and closed caption.

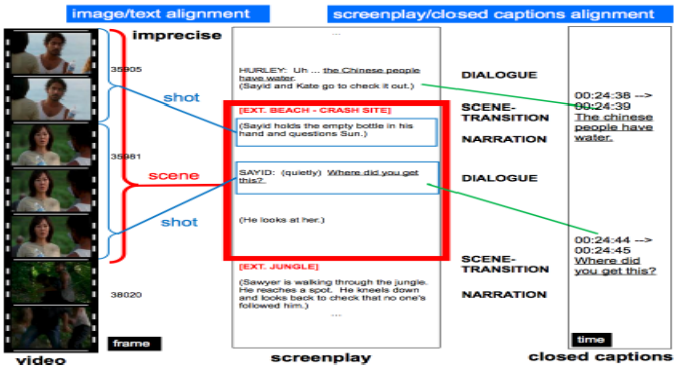
image/text alignment
screenplay/closed captions alignment
imprecise
35905
shot
35981
scene
shot
38020
frame
video
...
HURLEY: Uh ... the Chinese people have water.
(Sayid and Kate go to check it out.)
[EXT. BEACH - CRASH SITE]
(Sayid holds the empty bottle in his hand and questions Sun.)
SAYID: (quietly) Where did you get this?
(He looks at her.)
[EXT. JUNGLE]
(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one's followed him.)
...
screenplay
DIALOGUE
SCENE-TRANSITION
NARRATION
DIALOGUE
SCENE-TRANSITION
NARRATION
00:24:38 --> 00:24:39
The chinese people have water.
00:24:44 --> 00:24:45
Where did you get this?
time
closed captions

Figure: 2 Alignment between video, screenplay and closed caption [10].

This method presents a frame work for automatic parsing of movie or video into hierarchy of shot and scene recovery of shot interconnection recovery. It required input image sequence, closed caption and screen play or movie.

Timothee Cour, Benjamin Sapp, Akash Nagle, Ben Taskar [11] proposes talking picture using temporal grouping and dialog supervised person recognition. This model partitions face tracks across multiple shots while respecting appearance, geometric and film editing cues and constraints. To overcome a problem of character name identification in video, the two phase approach implemented for naming character. In first phase, detect and cluster a sequence of face cluster into small number of groups. Second phase, resolved identities of individual face tracks using first, second and third person references as weak supervision as well as gender cues and grouping cues obtaining in first phase.

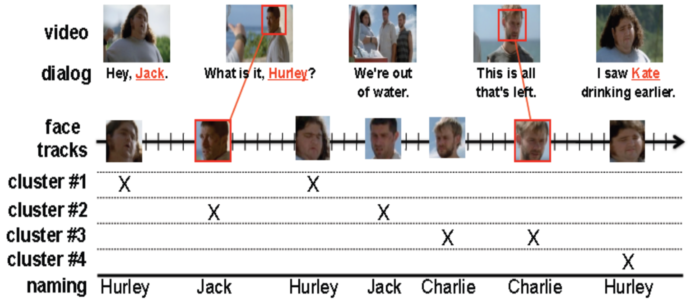

Figure 3: Temporal grouping and dialog supervised person recognition system [11].

Figure 3 shows novel temporal grouping model that groups faces based on not only appearance but also on local film editing cues.

The local matching methods required the time stamped information which is extracted by OCR i.e. subtitle or closed caption.

B.1. METHOD II: ADVANTAGES

- In [08] method, automatic generation of time stamped character annotation by aligning subtitle and transcripts. The supervisory information by identifying when characters are speaking. By using complementary cues of face matching and clothing matching to propose common annotation for face tracks.
- The basic element [10] of movie structure hierarchy of scene and shots and continuity of shot scene. The structure useful for many intelligent movie manipulation task such as by character or object crediting.
- Fully automatic system [11] for character naming in video it uses dialog in the common case when screenplay is not available. A temporal grouping model that may be of independent interest which incorporate arbitrary non pairwise cues including novel film editing cues.

B.2. METHOD II: DISADVANTAGES

- In [08] this method i.e. proportion of video labeled and generalization was limited by a restriction to frontal faces and nearest neighbor classification.
- The system [10] does not provide more fine-grained alignment of movies and screenplay using coarse scene geometry and pose estimation.

C. METHODS – III SCRIPT OR SCREENPLAY, GLOBAL MATCHING BASED

This method required only script and video. The possibility of character identification is without OCR based subtitle or closed caption. Without any local time information the task of character identification in done between face detected from video and the names extracted from the movie script.

Yi-Fan Zhang, Changsheng Xu, Jian Cheng, Hanqing Lu [12] address problem of finding faces in film using video and film script. Here use the global matching name and faces as it is not easy to obtain enough local name cues in the film. In this method cluster the faces into groups corresponding to character and build face network according to face co-occurrences relationship. In the film script a name network is also built according to name co-occurrences relationship. The vertices of two graphs are matched by a hypergraph matching method. For face tracking multi view face track used. The script are obtained from the internet movie script database, for face track clustering define the similarity measurement between two face track, which is represented as,

$$S\ (T_{m}, T_{n}) = \mu,\ \max\ i, j(S(f_{mi}, f_{nj}))$$

Where, T_{m}, $T_{n}$are two tracks.$S(f_{mi}, f_{nj})$ is similarity between two track m and n.µ is normalization.

SIFT descriptor is used for face covering overhead i.e. two eyes, nose and mouth. For clustering constrains K-means clustering is performed. Here number of cluster is set as the number of speaker names. Hypergraph constructed for 'm' face track cluster is G (V, E) for face occurrences matric O_{face}= $[O_{ij}]$ m x n

Where m is number of face , n is number of scenes, $O_{ij}$is matrix of the face count i th character in j th scene.

Also for name occurrences O_{name}= $[O_{ij}]$m x n

m is number of name, n is number of scenes, $O_{ij}$is matrix of the name count i th character in j th scene.

Hypergraph matching, two graphs G_{face}= (V_{f}, E_{f}) and G_{name}= (V_{n}, E_{n}) So, matching between $G_{face}$and $G_{name}$is vertex to vertex m: $V_{f} \rightarrow V_{n}$and edge to edge m: $E_{f} \rightarrow E_{n}$ matching.

Yi-Fan Zhang, Changsheng Xu, Hanqing Lu,and Yeh-Min Huang,[02]implement a character identification in feature length film using global face-name matching. To identified the faces of character in film and label them with their names. To investigate the problem of identifying character in film using video and film script. It implemented for character centered film browsing, which enables users to easily use the name as a query to search related video clips and digest the film content. For measuring face track distance Earth Movers Distance (EMD) is used and multiview face tracker used to detect and track faces on each frame of video. Constrained K-means clustering is performed on group of face track. To reduce the noise in clustering to refine the clustering result by pruning marginal point which have low confidence belonging to the current cluster? For face- name association build a name affinity network and a face affinity network in their domains i.e. script and video. Making the two networks i.e. face and name affinity network. Number of face cluster set a number of speakers

Figure 4 shows flow of character identification using global face-name matching.

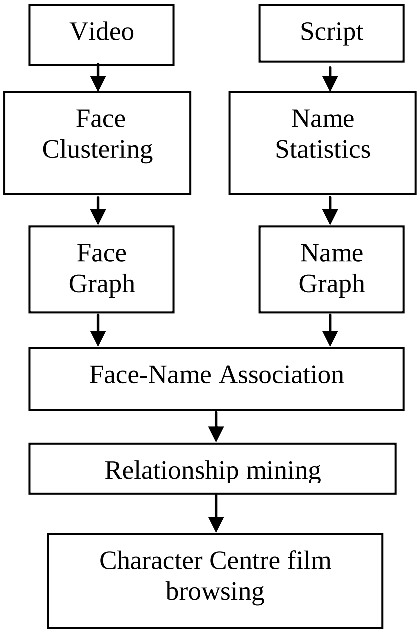

Figure 4: character identification using global face-name matching [02].

Jitao Sang, Changsheng Xu [01] implemented robust face-name graph matching for movie character identification system. Automatic face identification of character in movie or video is challenging problem due to huge variation in appearance and extrinsic parameter such as light, pose, complex background. The exiting methods gives a promising result in clean environment but the performance is limited in complex movie scenes due to the noise generated during the face tracking and face clustering process. The main objective of author is to identify the faces of character in the video and label them with the corresponding names in the cast list. It implemented two schemes, difference in the pre-specification in number of cluster. Also sensitivity analysis by introducing two types simulated noise i.e. coverage and intensity noise. It proposes a global face-name matching based framework for robust movie character identification. Both scheme having input as videos and script of videos.

Scheme I – Face-name graph matching with cluster pre-specified. (Face-name graph matching having the same structural topology.)

By using K-means clustering clustered face tracks. In K-means the number of cluster is set as the number of distinct speakers. Co-occurrences of names in script and face cluster in videos is corresponding face graph and name graph. Authors modify the traditional global matching method framework by using ordinal graph. For representation and introducing an ECGM based graph matching method. For face name graph matching ECGM algorithm is used. In ECGM the difference between two graphs is measured by edit distance which is a sequence of graph edit operation. The optimal match is achieved with least edit distance and to obtain.

Scheme II – Face-name graph matching without cluster pre-specified. (Face- name graph matching does not have same topologists.)

The proposed system for scheme II is shown in figure 6. It has two differences from scheme I, 1. No cluster number is required for the face track clustering process. 2. The face graph and name graph may have different number of vertex, so a partition component is added before ordinal graph representation.

C.1. METHOD –III ADVANTAGES

- In [01] method no cluster is required for the face track clustering step.
- The face graph and name graph in [02] may have different number of vertexes a graph partition component is added before ordinal graph representation.
- The system [02] addresses the problem of people identification in real world video.

C.2. METHOD –III DISADVANTAGES

- In [01] scheme II, the appearances of the same character vary significantly and it is difficult to group them in unique cluster, e.g. the hero and heroine go through a long time period from their childhood, youth, middle age to the old age. The interclass variance is even larger than the interclass variance.
- The main disadvantages [01] of scheme I is here number of character names in script is set a same number of speakers in video, but sometime speaker does not show in video and name of the character is extracted from script. In such situation this method may fail.

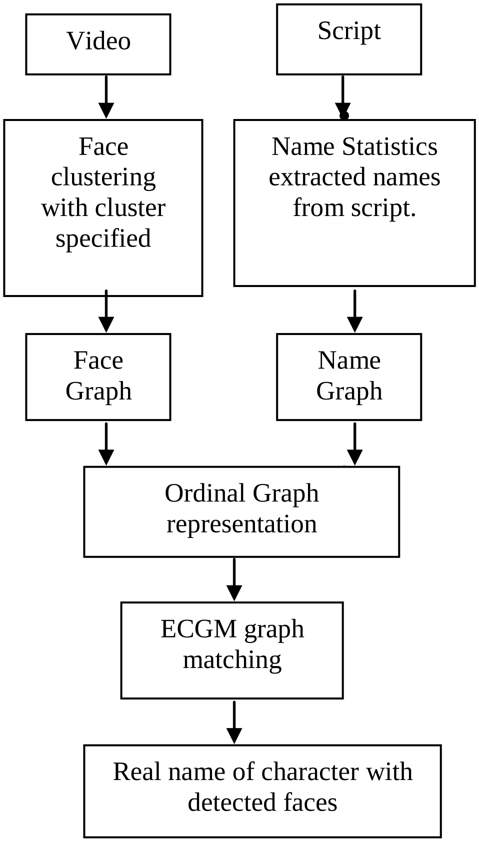
Video
Script
Face clustering with cluster specified
Name Statistics extracted names from script.
Face Graph
Name Graph
Ordinal Graph representation
ECGM graph matching
Real name of character with detected faces

Figure 5: Face-name graph matching with cluster pre-specified [01].

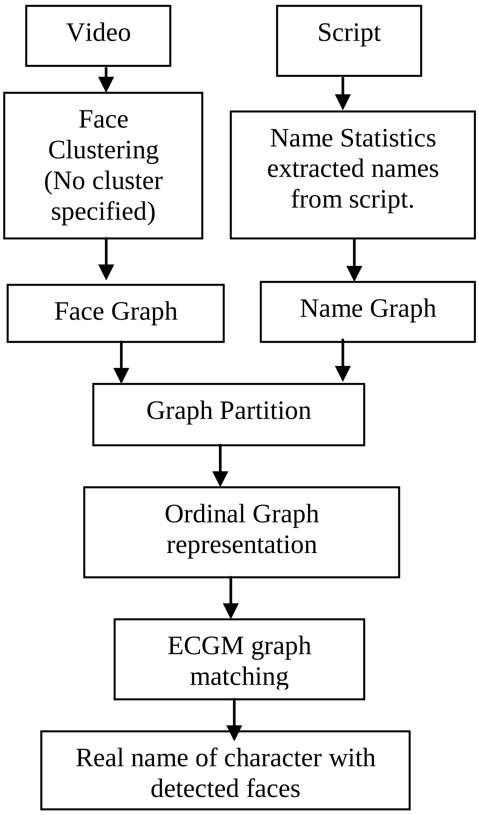
Video
Script
Face Clustering (No cluster specified)
Name Statistics extracted names from script.
Face Graph
Name Graph
Graph Partition
Ordinal Graph representation
ECGM graph matching
Real name of character with detected faces

Figure 6: Face-name graph matching without cluster pre-specified [01].

V. COMPARISION

Among all the methods all are using different approach for face-name detection for character identification. That differs according to the information used for analysis and according to techniques that are employed to face clustering and name clustering. We classify them on the basis of underlying approach they are using.

Cast List Based	Local Matching Based	Global Matching Based
Cast list textual resource used.	Subtitle or Closed caption textual resources used.	Script or Screenplay textual resources used.
No time stamp dialogue is required for alignment.	A time stamped dialogue is required for alignment.	No time information of character is required.
Faces are clustered by appearance of character.	Faces are clustered into face exemplars which extracted from nearest neighbor classifier.	Faces are clustered from extracted from video frames.
Easy to understand and implemented.	Not easy to understand and implemented.	Not easy to understand and implemented.
Names of character extracted from cast list of movie.	Names are extracted from closed caption.	Names are extracted from script of movie.
Need manual labeling for clustering.	Time stamped information is used which extracted by OCR.	Face-name association is used for clustering.
The large intraclass variances, there is no quantative of clustering and classification performance.	It is more sensitive to face detection and tracking noise.	The robustness of the algorithm is good

TAABLE I: COMPARISION OF METHODS

VI. CONCLUSION

In this survey paper, we discussed the three methods for movie character identification based on textual resources i.e. cast list, screenplay, closed caption, script etc. and their advantages and disadvantages. Performance of global matching based method is better than local based and cast list based methods. According to variations in the clustering of faces and names of characters from video and textual resources.

We conclude there is no single method exists for character identification, a lot of technique available, but as per survey global matching based method gives better face tracking and clustering result with minimum noise generation. We have discussed the need and challenges of movie character identification with their application. The comparisons of three methods have been shown in this survey paper. We have tried to present almost all possible techniques of movie character identification and their relationship in movie.

ACKNOWLEDGMENT

I express great many thanks to Prof. Santosh A. Shinde, for his great effort of supervising and leading me, to accomplish this fine work. To college and department staff, they were a great source of support and encouragement. To my friends and family, for their warm, kind encourages and loves. To every person gave us something too light my pathway, I thanks for believing in me.

REFERENCES

[1] Jitao Sang and Changsheng Xu, "Robust Face-Name Graph Matching for Movie Character Identification,"IEEE Transaction on multimedia, Vol. 14, No. 3, June 2012.

[2] Y. Zhang, C. Xu, H. Lu, and Y. Huang, "Character identification in feature-length films using global face-name matching,"IEEE Trans. Multimedia syatem, vol. 11, no.7, pp.1276-1288, Nov. 2009.

[3] A. W. Fitzgibbon and A. Zisserman, "On affine invariant clustering and automatic cast listing in movies,"in Proc. ECCV, 2002, pp. 304-320.

[4] O. Arandjelovic and R. Cipolla,"Automatic cast list-ing in featurelength films with anisotropic manifold space,"in Proc. Comput. Vis. Pattern Recognition, 2006, pp. 1513-1520

[5] D. Ramanan, S. Baker, and S. Kakade,"Leveraging archival video for building face datasets," in Proc. Int. Conf. Comput. Vis., 2007, pp. 1-8.

[6] M. Everingham and A. Zisserman,"Identifying individuals in video by combining generative and discriminative head models," in Proc. Int. Conf. Comput. Vis., 2005, pp. 1103-1110.

[7] R. G. Cinbis, J. Verbeek, and C. Schmid,"Unsupervised metric learning for face identification in TV video,"in Proc. Int. Conf. Comput. Vis., 2011, pp. 1559-1566.

[8] M. Everingham, J. Sivic, and A. Zissserman, "Hello! My name is Buffy automatic naming of characters in TV video," in Proc. BMVC, 2006, pp. 889-908D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[9] J. Sivic, M. Everingham, and A. Zissserman,"Who are you?-Learning person specific classifiers from video," in Proc. Comput. Vis. Pattern Recognition, 2009, pp. 1145-1152.

[10] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar,"Movie/script: Alignment and parsing of video and text transcription," in Proc. ECCV, 2008, pp. 158-171.

[11] T. Cour, B. Sapp, A. Nagle, and B. Taskar,"Talking pictures: Temporal grouping and dialog-supervised person recognition," in Proc. Comput. Vis. Pattern Recognition, 2010, pp. 1014-1021.

[12] Y. Zhang, C. Xu, J. Cheng, and H. Lu, "Naming faces in films using hypergraph matching," in ICME, 2009,pp.278-281.

Authors

Salve Bhausaheb S. received his B.E. degree in Information Technology engineering from the Pune University, Pune, in 2010. He is currently working toward the M.E. degree in Computer engineering from the University of Pune, Pune. He has 04 years of teaching experience at undergraduate level. His research interests lies in Digital Image Processing.

Shinde Santosh A. received his B.E. degree in computer engineering (First Class with Distinction) in the year 2003 from Pune University and M. E. Degree (First Class with Distinction) in Computer Engineering in 2010 from Pune University. He has 11 years of teaching experience at undergraduate and postgraduate level. Currently he is working as Assistant Professor in Department of Computer Engineering of VPCOE, Baramati, Pune University. His autobiography has been published in Marquie’s Who’ Who, an International Magazine of Prominent Personalities of the World in the year 2012. He is also a Life Member of IACSIT and ISTE professional bodies. His research interests are Digital Image processingandwebservices.