Celebrity Face-Name Association in Web Videos using Unsupervised Approach

Shweta Tadge[1], Prof. Ranjana Dahake[2]

[1] PG Student, Department of Computer Engineering, MET BKC, Nashik.
[2] Professor, Department of Computer Engineering, MET BKC, Nashik.

ABSTRACT

This paper explores the problem of missing name and missing faces in unconstrained videos with user provided metadata. Rather than depending upon supervised learning, a better relationship built from the content of a video, those relationship includes the arrival of faces in different spatio-temporal contexts and visual similarities between faces. The knowledge base consists of tagged images along with a set of names and celebrity social networks. Celebrity social network is built based on the co-occurrence statistics of celebrities in video metadata. Merging of relationship along with knowledge base is carried out via conditional random field. Two types of face-name association are investigated: within video face labeling and between video face labeling. The within video labeling takes care of noisy as well as incomplete labels in metadata, in which null assignment for the labels is permitted. Furthermore Between video face labeling addresses the flaws within metadata, particularly to correct incorrect names and label faces having no available names in metadata of a video. To do so it considers a gathering of socially associated videos for combined name inference. The experimental result analysis on web video dataset shows that proposed approach is very much powerful for handling the issue of missing names and incorrect names in face labeling problem than existing approaches.

Keyword: - *Celebrity face labeling, social networks, unsupervised learning, unconstrained videos.*

1. Introduction

Individuals do upload large number of videos, in which 80\% are related to people. In those videos 75\% are related to celebrities. The tremendous growth of video on the internet and the rising insufficiency in metadata associated with video forces us to look at the content from the video content for search retrieval and browsing based exposure. A large chunk of users' browsing patterns are centered around people present in the video. With the massive growth of digital videos in the Internet, recognizing and understanding the visual content is becoming an increasingly important problem. These days identification of characters in web videos is a challenging task because of huge deviation in the approach of person or celebrities within web videos. Naming celebrities in the user-generated videos provides great help for both indexing and browsing. In all the top video search engines like YouTube, indexing of these videos is based on user-provided text data like title of videos or description, which found to be noisy and incomplete most of the times. Often a mentioned celebrity may not exist within video, and a celebrity which actually exist within a video is not mentioned in user-provided text. There is no correlation present between video content and metadata associated with video due to these issues, people-related video search results into non satisfactory retrieval of videos. Identifying the direct relation between faces and names can help in rectifying the potential errors in user provided metadata hence it is provided as an initial processing step for indexing of videos. Rich context information cannot be applied directly for face naming in unrestricted videos due to lack of prior knowledge and the context cues.

Fig -1 explores problem using a web video example. In this video nineteen faces are detected in which only four having their names mentioned in metadata. Also, among the five mentioned celebrities in metadata, only four of them are actually present in the video. It concludes that faces are missing within videos and names are missing in text provided with video. Furthermore, faces can appear wildly different because of motion blurriness, resolution and lighting changes. In other words, the problem of face name association can be specific to inadequate metadata, false metadata and visual appearance suggestion. That means it consider the reality that the text provided with video

is not that much accurate. Therefore it is good to perform the video content analysis for labeling the faces from video instead of depending solely to the metadata.

Fig -1: A web video example exploring the problem of relating the names(red) present in metadata with the identified faces (enclosing squares) within video.

The main contribution behind this work is the extension of name-face association to domain unrestricted user created videos for celebrity face naming. Instead of depend on supervised learning which takes the accurate face labels from user side for supervised learning, a rich set of relationships automatically derived from video content and knowledge from image domain and social cues is leveraged for unsupervised face labeling. The relationships refer to the appearances of faces under different spatio-temporal contexts and their visual similarities. The knowledge includes Web images weakly tagged with celebrity names and the celebrity social networks. The face models of celebrities are automatically learned from the web images. The names of celebrities are mined from the web text by name entity detection. A highly accurate and efficient face detection and tracking algorithm is applied to extract faces.

2. Related Work

Existing researches about celebrity face naming are applied in the area of web based images and constrained videos like TV serials, movies as well as news related videos. These works were based on rich set of time coded information, where the emphasis is highly based upon the rich text which assumes the text having no errors and matches completely to the detected faces. In literature celebrity naming problem is about name-face association, in which the aim is to coordinate the detected faces with available names. Zhi-Neng Chen and et al. [9] proposed that Name-face association, which utilizes surrounding metadata to assist recognition, is generally regarded as a feasible methodology for face recognition. A common approach is to weakly associate every name found in metadata with every face detected from images or videos, and then the refinement based on visual similarity and contextual clues is conducted to remove false matches. Existing studies in name-face association mostly defined in the way of how refinement is formulated based on that the categorization is happened into three methods namely classification-based, clustering-based and knowledge-based methods based on how the refinement is happening.

The classification-based methods learn discriminative models to predict the correction of the weakly associated name-face instances. For example, in [2], instead of learning models for each person, a unified SVM classifier was trained to determine the correction of each name-face instance based on multiple modalities extracted from the transcripts, optical character recognition (OCR) result and speech track of news videos. Due to the need of labeling a large number of name-face pairs for learning, the work was later extended to partial learning under multiple instance setting, i.e., MIL[7], where only partial label information is required for model training. Observing that celebrities or major characters usually appear recurrently in news videos, TV series and movies, the clustering based methods investigate name-face association by focusing on mining visual similarities between faces (face tracks) and contextual information derived from video structure and prior knowledge. Satoh et al. [3] describes the first proposal on face-name association in news videos based on the co-occurrence between the detected faces and names extracted from the video transcript. A face is labeled with the name which frequently co-occurs with it. Similarly in [4], global name-face matching in movie domain is proposed to match affinity graphs of faces and names through aligning movie scripts and subtitles.

M. Guillaumin et al. [5] implements influential work is the graph-based clustering methods in Web image domain, where the faces detected from an image collection are jointly modeled as a graph and the relationship between faces is determined based upon the similarity of facial features. With the assumption that the faces of a person should exhibit higher visual similarity and reside in a dense sub-graph, the problem was converted to identifying densely connected sub-graphs corresponding to the names. In contrast, the knowledge-based methods tackle the problem of weak contextual clue by leveraging online information sources for learning face models [6] and identifying social networks [8].

Existing work are divided into three categories model-based face labeling, search based face labeling and constrained clustering based face labeling. In model based approach the mechanism is based upon learning classifier for face identification. This approach requires labeled samples which can be used as training samples for each and every face model. In this approach scaling was not possible as number of names get increased. Number of efforts has been done in order to learn efficient classifier from small sized training samples. In [11], Transductive Kernel Fisher Discriminant (TKFD) algorithm is used, which employs the kernel deformation techniques to exploit both labeled and unlabeled data effectively for annotation tasks. TFKD approach is found to be effective when there are only a small number of labeled data.

In search-based approach there is no need of training samples hence no need of training classifier separately feel to be equivalent to the query faces. Search based strategy extracts names from the fetched examples. Noise in the labels is the most important concern with this approach. Unsupervised label refinement [12] approach is used in the absence of supervisory information in order to clarify labels associated with web images based upon strategies associated with machine learning. Effective optimization algorithm is developed in order to effectively solve the large scale learning task. The name mining problem is solved by majority voting scheme between top n retrieved images. In [14], weak labels are enhanced using LCC (local coordinate coding) during the minimization of effect of incorrect labels at the time of voting for top n images. In [15] the problem is modeled as measurement of weights for the votes which are given by images based upon learning distance functions using multimodal features and optimized combination of these functions. Training examples are used for learning of distance functions and fusion weights based upon offline manner. In the retrieval process a vote is used from top n images to the name of candidate depend upon the multimodal similarities to the query, computed using distance metrics which is learnt and optimal fusion weights associated with them.

Clustering based approach is highly related to this work. This approach gives better performance when there are limited number names present for a face. It is assumed that faces corresponding to a person can be chunk in dense manner and therefore get used in the process of face naming. Existing researches consist of three main approaches GC (graph based clustering), CGMM (constrained gaussian mixture models) and FACD (face name association by commute distance). CGMM uses expectation-maximization algorithm in order to learn Gaussian mixture model for every name. The process of learning includes assignment of faces to best possible model and updation of parameters of model. Graph representation is used by GC [5] and FACD [13] in order to model the density of faces. Firstly GC retrieves images annotated with the names present in metadata. Then graph is constructed online having faces in those images as vertices and similarities between faces as edges. From that graph densest sub-graphs are extracted each corresponds to name to formulate the name assignment problem.

This proposed work focuses on expansion of name face association to unrestricted web based videos for the task of celebrity face labeling [1]. The proposed system is based upon Image Matching technique in which face photos of popular celebrities can be easily searched from the Web and stored, this method matches a face to the Web images of celebrities. The KNN classifier is used for name-face association. Null assignment is activated if the similarity of a face to its nearest neighbor is below an empirically set threshold [9]. Furthermore, the work here concentrates on three important relationships to solve the problem of missing names and missing faces occurring in web videos. This work considers CRF for its capability in consolidating different arrangements of connections and powerful algorithms for label interpretation [10].

3. Proposed Approach

The proposed approach uses rich relationships instead of rich texts in the domain of web video. The strategy is based on three important relationships as below:

- Face-to-name affinity (F2N): it models likelihood of a face assignment to the name, using external knowledge in the domain of image.
- Face-to-face coercion (F2F): it deals with factors like back-ground context, temporal disconnectivity, spatial overlap, and observable affinity to relate the faces of various frames and videos.

- Name-to-name relationship (N2N): Named as social relationship, deals with collective arrival of celebrities by using social network prepared by considering the co-occurrence statistics between celebrities.

First and second relationship is used in order to label faces within a video, which termed as within-video face labeling. The job here is to assign the names present in metadata to the faces identified within video, while considering the problem of missing faces and names such that uncertainty in labeling is permitted. Using social network labeling related to single video is extended to between video, by considering group of videos for labeling of faces where celebrities should be present in the same social network. In between-video naming, the relationships formed between videos permit the correction of names which are tagged incorrectly and the labeling of names which are missing in metadata.

Two types of face annotation task are taken into account:

- Within-video face labeling
- Between-video face labeling

Within-video Face Labeling: Consider a web video V1. For solving the problem of missing names and missing faces this approach construct a graph by considering the faces and names present in video as graph vertices. Furthermore, edges are formed between vertices with the help of F2N and F2F relationships. Name inference is carried out using CRF with the consideration of uncertainty in labeling using null assignment. The inference of face labels can be influenced by situation like there are names present in metadata but respective faces are not appearing in video as well as faces appear in video but names are not mentioned.

Between-video Face Labeling: This approach considers name-to-name relationship which extends the graph of within video face labeling. Between video labeling is done on collection of related videos where celebrities are present under same social network. By associating v1 with a social network collect related videos such as v2 and v3 and generate a bigger graph having faces and names from multiple relevant videos. The augmented graph has advantage such as missing names from video v1 can be produced through the help of relevant videos v2 and v3 as well as faces which are wrongly labeled can be corrected with the help of name-to-name relationship.

4. System Architecture

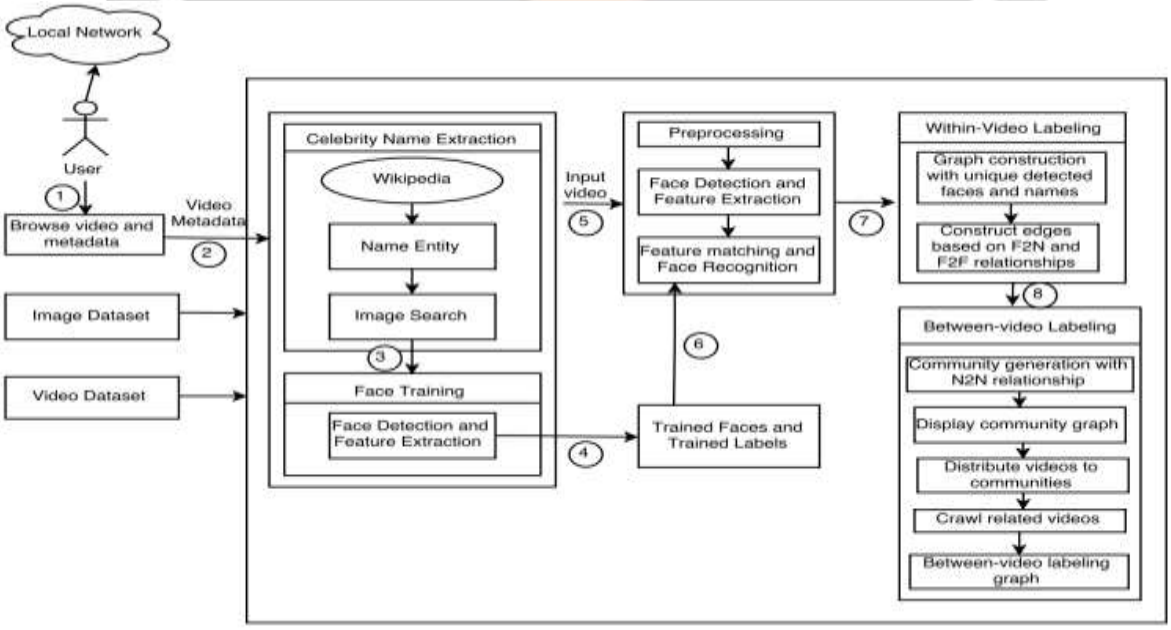

Fig -1: System Architecture

The architectural diagram of proposed system is shown in Fig. 2. User browse a input video with its metadata. Metadata of videos is taken as a input for celebrity name extraction process. To extract the celebrity names the

successive words from metadata are sent to wikipedia. Wikipedia find out the name entity. Identified names are get searched into the image dataset. If the search result has found, then the system will train itself for that image. Face detection and facial feature extraction are executed on each frames of video. If the extracted features are matched with the feature of trained faces then the trained label for that face is displayed. If the detected face having no name present in metadata then "null" label is get assigned to that face. In between-video naming process the communities needs to be generated. The communities are generated with the help of name-to-name relationship i.e, co-occurrence statistics of celebrity names in metadata. The videos from dataset are distributed to each community if the names present in video metadata are same as the celebrity names from communities. In between-video naming related videos are get crawled to solve the problem of celebrity face naming of within-video naming. The expanded graph of between-video labeling has the advantages that the any missing name can be propagated from the list of related videos.

4.1 Stepwise flow of system

1. Name extraction from metadata of video.
2. Training of faces.
3. Unique face detection and face recognition from video.
4. Construct a graph having detected faces and names are modeled as graph vertices.
5. Establish edges of graph based on F2F and F2N relationship between faces and between faces and names.
6. Estimate the conditional probability for each label assignment and select solution that maximizes the probability.
7. Establish social network based upon co-occurrence statistics among celebrities.
8. Constructing a social graph depicting the relationship between celebrities.
9. Using Walktrap algorithm the graph is partitioned into sub graphs i.e. communities corresponding to social network.
10. Distribute each video from dataset to one or more network based on the names mentioned in a video and name of celebrities in community.
11. Crawl related videos having the same social network as this web video.
12. Construct a new graph to create edges among the related videos based on N2N relationship.

5. ALGORITHMS

5.1 Algorithm I - Within Video Face Assignment

Input: Detected faces and extracted names from video.
Processing:
Step 1 = Construct a graph through the modeling of unary potential for each and every detected face by which edges are established between faces and names.
Step 2 = Establishment of edges for any pair of faces in graph that fulfills the condition of spatial, temporal and visual relationships.
Step 3 = Perform loopy belief propagation on this graph for face naming.
Output: Face labels that maximizes conditional probability.

5.2 Algorithm II - Between Video Face Assignment

Input: Video vd*i* and its metadata
Processing:
Step 1 = Build a graph by joining the detected faces and names from video vd*i* with the help of unary potential, spatial, temporal and visual relationships.
Step 2 = Perform community generation using Name-to-Name relationship.
Step 3 = Crawls the related videos having same shared social network as video vd*i.*
Step 4 = Train the system for all related videos and construct a new graph after training has done.
Step 5 = Execute loopy belief propagation on this graph for face naming.
Output: Face Labels.

6. EXPERIMENT

6.1 System requirement specification

The system is run under the Asp.Net framework in Windows 7 OS. Visual studio 2010 community edition is used as the integrated development environment (IDE). C# is the programming language from .Net framework which is used for the implementation of this work. Intel core i3 1.70 GHz processor with 4GB RAM and 1 TB HDD is used as host configuration for experiment.

6.2 Dataset

The proposed system is works on video dataset having social videos of celebrities and image dataset having celebrity images tagged with celebrity name for knowledge base. Proposed system is tested upon 8 videos downloaded from you tube having large and diverse appearance of faces and wide range of celebrities with different professions.

6.3 Analysis of Results

This section elaborates the results of proposed system. Within-video labeling and Between-video labeling are the two major modules tested on various input videos. The results of these two modules are shown in table I and II respectively. Total number of faces present in video are taken from each video and from those faces how many number of faces having name assigned and number of faces having null assigned is calculated. The resulting graph for both the processes of Within-video and Between-video labeling are shown in Chart 1 and 2 respectively. The proposed system is tested with frame rate management and without frame rate management. With frame management proposed system considers nth frame in frame sequence where n can be 3, 5, 7 etc depending on the size of video to handle speed efficiency and minimize time complexity. The resulting graph of time taken by system for face detection, face recognition and graph formation with and without consideration of frame rate management is shown in Chart 3.

Table -I: Result table for Within-Video labeling process

Sr. No.	Metadata of Video	Total faces in video	Total faces having name assigned	Total faces having null assigned
1	Tonight show about hillary clinton	2	0	2
2	Thank you notes with barack Obama	2	1	1
3	Hillary clinton concession speech	3	1	2
4	Emma watson talks about beauty and beast	2	1	1
5	Donald Trump talks muslims, president obama and hillary clinton	2	1	1
6	Barack obama calls out hillary clinton on ambiguity	6	2	4
7	Barack obama demolishes donald trump endorses hillary clinton	2	2	1
8	Narendra Modi talks about hillary clinton to Network 18	2	1	1

Table -II: Result table for Between-Video labeling process

Sr. No.	Metadata of Video	Total faces in video	Total faces having name assigned	Total faces having null assigned
1	Tonight show about hillary clinton	2	2	0
2	Thank you notes with barack obama	2	2	0
3	Hillary clinton concession speech	3	2	1
4	Emma watson talks about beauty and beast	2	2	0
5	Donald Trump talks muslims, president obama and hillary clinton	2	2	0
6	Barack obama calls out hillary clinton on ambiguity	6	2	4
7	Barack obama demolishes donald trump endorses hillary clinton	2	2	1
8	Narendra Modi talks about hillary clinton to Network 18	2	1	1

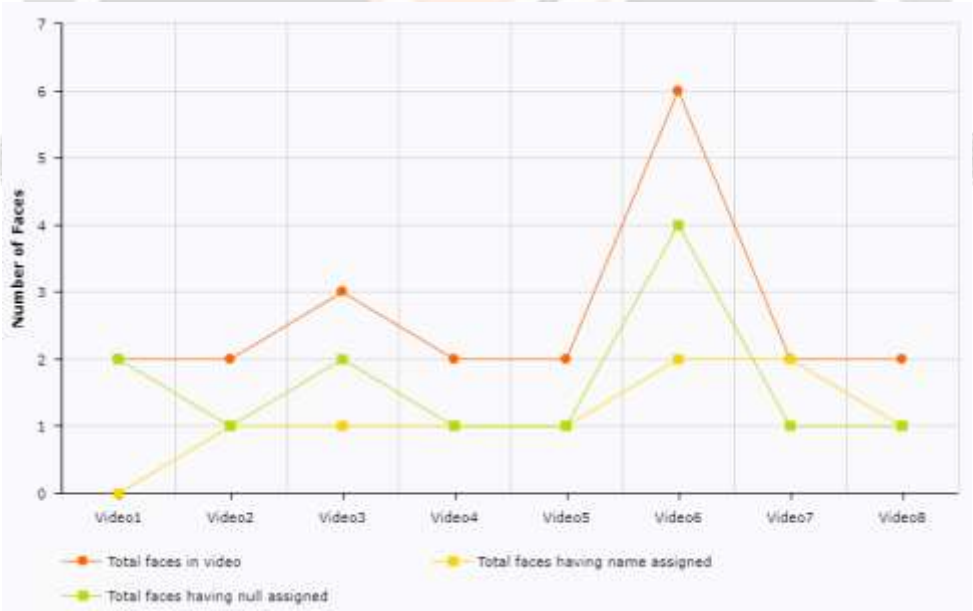

Chart -1: Result of Within-Video labeling process

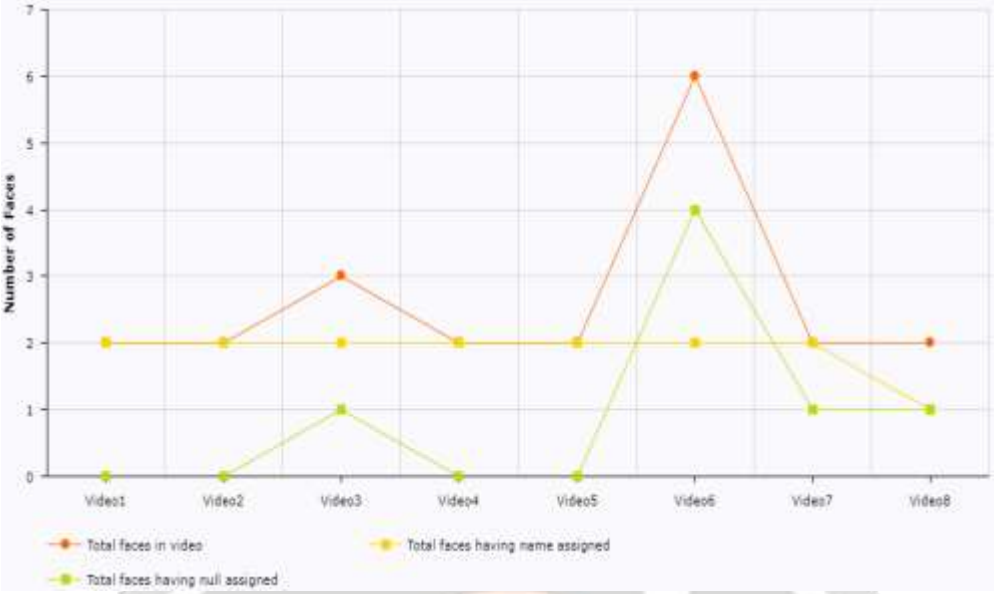

Chart -2: Result of Between-Video labeling process

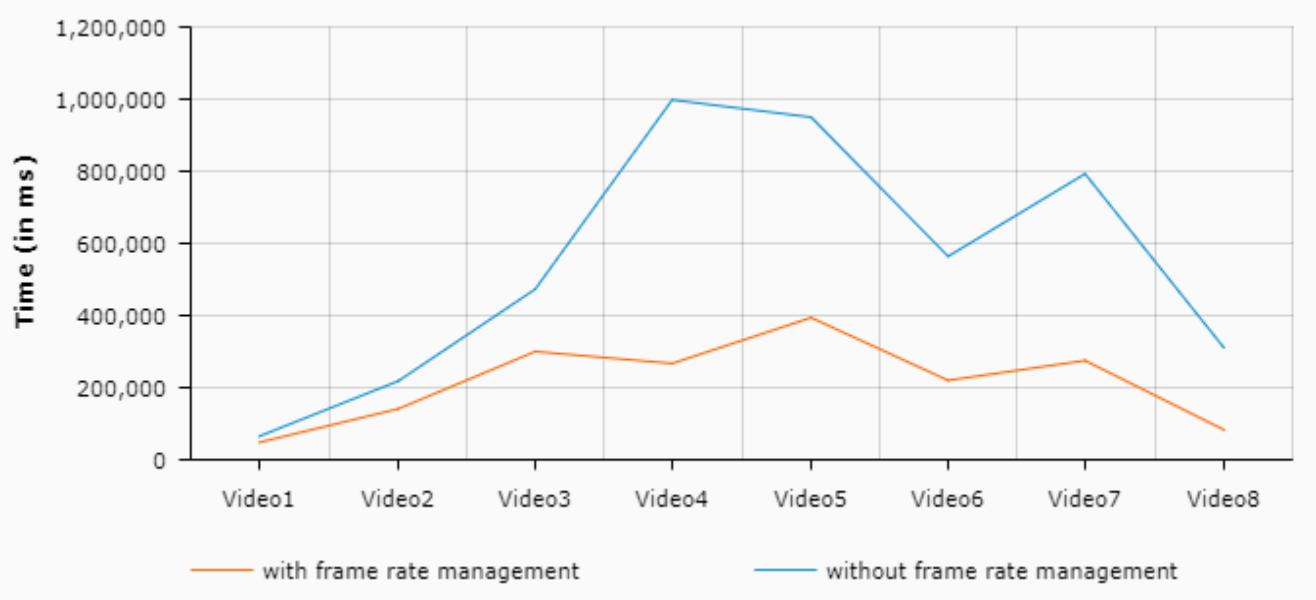

Chart -3: Face detection, face recognition and graph formation time with the consideration of frame rate management and without frame management.

7. CONCLUSIONS

The proposed system solves the problem of celebrity face naming in unconstrained videos. The faces from video are associated with the names from metadata for face labeling. Face name association approach of proposed system is based upon rich relationships rather than rich text approach. The proposed system considers the problem of metadata that metadata provided with video is incomplete as well as inaccurate because their is no proper association between faces and names from video and metadata respectively. CRF does the smooth encoding of Face

to Face and Face to Name relationships, allowing null assignment to the faces which considers the uncertainty within labeling to deal with the incomplete and noisy metadata. Furthermore, system addresses the errors present in metadata, in order to rectify false names and clarify faces having no names present in the metadata of a video, by using socially related web videos. Between-video relationships helps in boosting the performance, having the potential of correcting the errors happened because of missing names and persons. But due to the between video labeling the processing time gets increased as we consider the group of related videos. It will deliver as a initial processing step of video indexing which gives better video retrieval performance. The proper indexing of videos gives the proper searching result for web videos.

8. REFERENCES

[1]. Lei Pang and Chong-Wah Ngo, "Unsupervised Celebrity Face Naming in Web Videos", in IEEE Transactions on Multimedia, vol.17, no.6, june 2015.
[2]. J. Yang and A. G. Hauptmann, "Naming every individual in news video monologues", in Proc. ACM Int. Conf. Multimedia}, 2004, pp. 580–587.
[3]. S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos", in IEEE Multimedia, vol. 6, no. 1, pp. 22-35, Jan.–Mar. 1999.
[4]. Y. F. Zhang, C. S. Xu, H. Q. Lu, and Y. M. Huang, "Character identification in feature-length films using global face-name matching", in IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1276–1288, Nov. 2009.
[5]. M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision", Proc. IEEE Comput. Vis. Pattern Recog., Jun. 2008, pp. 1-8.
[6]. M. Zhao, J. Yagnik, H. Adam, and D. Bau, "Large scale learning and recognition of faces in web videos", in Proc. Int. Conf. Automat. Face Gesture Recog, 2008, pp. 1–7.
[7]. J. Yang, R. Yan, and A. G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video", in Proc. ACM Int. Conf. Multimedia, 2005, pp. 31-40.
[8]. Z. Stone, T. Zickler, and T. Darrell, "Toward large-scale face recognition using social network context", in Proc. IEEE, vol. 98, no. 8, pp. 1408–1415, Aug. 2010.
[9]. Zhi-Neng Chen, Chong-Wah Ngo, Wei Zhang, Juan Cao, and Yu-Gang Jiang, "Name-Face Association in Web Videos: A Large-Scale Dataset, Baselines, and Open Issues", in JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, Sept. 2014, pp. 785-798.
[10]. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, in "Conditional random fields: probabilistic models for segmenting and labeling sequence data", Proc. Int. Conf. Mach. Learn. 2001, pp. 282-289..
[11]. J. K. Zhu, S. C. H. Hoi, and M. R. Lyu, "Face annotation using transductive kernel fisher discriminant", in IEEE Trans. Multimedia, vol. 10, no. 1, pp. 86-96, Jan. 2008.
[12]. D. Y. Wang, S.Hoi, Y.He, and J.K.Zhu, "Mining weakly labeled web facial images for search-based face annotation", in IEEE Trans. Knowl. Data Eng, vol.26, no.1, pp.166-179, Jan.2014.
[13]. J. Bu et al., "Unsupervised face-name association via commute distance", in Proc. ACM Int. Conf. Multimedia, 2012, pp.219-228.
[14]. D. Y. Wang, S. C. Hoi, Y. He, J. K. Zhu, T. Mei, and J. B. Luo, "Retrieval-based face annotation by weak label regularized local coordinate coding", in IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 550-563, Mar. 2014.
[15]. D.Y.Wang, S.C.Hoi, P.C.Wu, J.K.Zhu, Y.He, and C.Y.Miao, "Learning to Name Faces: A Multimodal Learning Scheme for Search- Based Face Annotation", in Proc. ACM Conf. Res. Develop. Inf. Retrieval, 2013, pp. 443-452.