A Review on Automatic Face-Name Association for Web Videos

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

[1]*PG Student, Department of Computer Engineering, MET's Institute of Engineering, Maharashtra, India*
[2]*Professor, Department of Computer Engineering, MET's Institute of Engineering, Maharashtra, India*

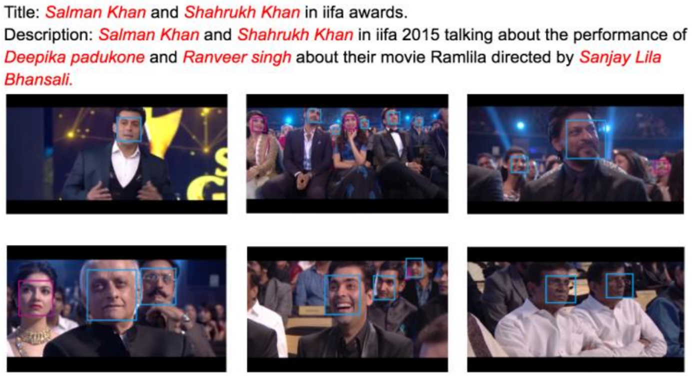
---***---

Abstract - *Celebrity face labeling in web videos is huge and challenging task because of large deviation in the appearance of person or celebrity in the web videos. This work explores the problem of missing name and missing faces in unconstrained videos with user created metadata. Instead of depending upon supervised learning, a better relationship built from content of video, those relationship includes arrival of faces in different spatio-temporal contexts and visual similarities between faces. The knowledge base consist of weakly tagged images along with set of names and celebrity social networks. 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, where null assignments for the names is permitted. Furthermore Between video face labeling addresses the flaws in metadata, particularly to correct incorrect names and label faces with missing names in the metadata of a video, by considering a gathering of socially associated videos for joint name inference.*

Key Words: **celebrity face labeling, social networks, unconstrained videos, supervised learning.**

1. INTRODUCTION

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. Individuals do upload large number of videos, in which 80% are related to people. In those videos 75% are related to celebrities. 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, 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 and thus provided as a pre-processing stage for the process of video indexing. Rich context information cannot be applied directly for face naming in unrestricted videos due to lack of prior knowledge and the context cues.



Title: Salman Khan and Shahrukh Khan in iifa awards.
Description: Salman Khan and Shahrukh Khan in iifa 2015 talking about the performance of Deepika padukone and Ranveer singh about their movie Ramlila directed by Sanjay Lila Bhansali.

Fig -1: A web video example exploring the key challenge of associating the names(red) in metadata with the identified faces (enclosing boxes) within video.

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. Which concludes that faces are missing in videos and names are missing in text. 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 text data, strident text and visual deficiency. This work reviews the face name association based upon the rich relationship rather than rich context.

2. EXISTING APPROACH

In literature celebrity naming problem is about name-face association, in that the goal is to coordinate the detected faces with a given set of names [2]. 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 is free of noise and matches completely to the detected faces. Existing researches 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 [5], 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 Multiple instance learning approach [6] no complete information on data labels is present. Each and every face-name association is treated as a data instance with a two correct/incorrect labels, and supervised method is used for the training of classifier using labeled data to predict the probability of association between name and face. In [7], images are crawled from web which are weakly labeled and used for learning of face models. Consistency learning approach is a bootstrap strategy used to minimize labeling efforts which filter out the false samples from web based images which are weakly labeled. Deepface [10] is one of the latest approach able to achieve success using techniques based on deep learning. One of the deficiency in Deepface was the requirement of large training samples to cover the variations in visual appearance.

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 [8] approach is used in the absence of supervisory information in order to clarify the labels of web facial images based upon techniques 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 numbers of retrieved images. In [11], weak labels are enhanced using LCC(local coordinate coding) during the minimization of impact of noisy labels while voting of top n number of images. In [12] the problem is modeled as measurement of weights for the votes which are casted 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 of name candidates to be considered for a face. It is assumed that faces corresponding to a person can be clustered in dense manner and hence get exploited 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 [3] and FACD [9] 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. The process of graph creation is much faster in FACD by using face pair indexing into inverted index. FACD has two stages named off-line index stage and on-line index stage. In first stage the Name-Face index structure is created to efficiently retrieve the faces. In second stage independent processing of every request image-caption item is carried out. Particularly FACD creates a unified graph for every item and calculate the face-name relationship using commute distance on graph.

This work focuses on expansion of name face association to unrestricted web based videos for the task of celebrity face labeling [1]. Furthermore, this work concentrates on three important relationships to solve the problem of missing names and missing faces occurring in weakly-tagged web videos. This work considers CRF(Conditional Random Field) for its capability in combining diverse sets of relationships and algorithms for label inference.

3. FACE NAME ASSOCIATION APPROACH

This approach uses rich relationships instead of rich texts in the domain of web video [1]. Conditional random field [4] based algorithm used to deal with the problem of face labeling. Conditional random fields is a popular probabilistic method for structured prediction. Many tasks involve predicting a large number of variables that depend on each other as well as on other observed variables. Structured prediction methods are essentially a combination of classification and graphical modeling. They combine the ability of graphical models to compactly model multivariate data with the ability of classification methods to perform prediction using large sets of input features.

This strategy is based on three important relationships as below:

1) Face-to-name affinity: it models likelihood of a face assignment to the name, using external knowledge in the domain of image.
2) Face-to-face coercion: it deals with factors like background context, temporal disconnectivity, spatial overlap, and visual similarity for relating faces from various frames and videos.
3) Name-to-name relationship: Named as social relation, deals with the joint 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 naming within a 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. Fig -2 shows the detailed approach of face labeling. Two types of Face annotation task are taken into account: within-video face labeling and between-video face labeling.

Within-video Face Labeling:

Consider a web video. For solving the problem of missing names and missing faces this approach construct a graph by considering the names and faces in the video as vertices. Furthermore, with the help of face-to-name and face-to-face relationships, edges are formed between the vertices for inference of face labels. Inference is carried out using CRF with the consideration of uncertainty in labeling (i.e. null assignment) is allowed. The inference of face labels can be influenced by situation like there are names mentioned in metadata but respective faces are not appearing in video as well as faces appear in video but names are not mentioned.

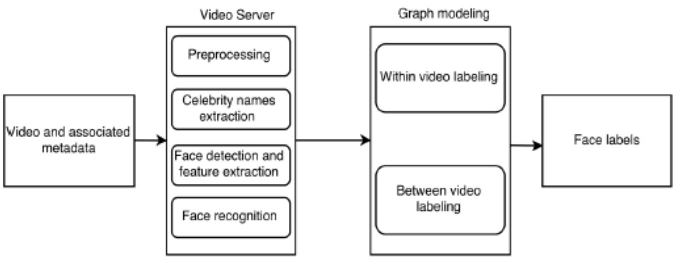


Fig -2: Approach of face labeling

Between-video Face Labeling:

This approach considers social relationship (i.e. name-to-name relationship) which extends the graph of within video face labeling. Between video labeling is done on group of videos whose celebrities fall under same social network. Consider a video having a name $V1$. By associating $V1$ with a social network collect relevant videos such as $V2$ and $V3$ and generate a larger 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. RELATIONSHIP REPRESENTATION

Consider a video in which there are two inputs to the problem of face-name association. First input is a set of detected faces from a video and second is celebrity names present in metadata. Celebrity faces are displayed as a sequence $F = \{ f1, f2, f3,..., fN \}$ and celebrity names from metadata are displayed as a set $N = \{ n1, n2, n3,..., nM \}$ where N and M represent the count of faces and names respectively. The problem of celebrity face naming can be represented as assigning at most one ni in N to fi in F, such that every face is annotated with a single name otherwise no name(i.e. assign null). Output for this problem is a label sequence displayed as $X = (x1, x2,..., xN)$, where each particular element xi is an indexed variable which signify that the fi face in the sequence F is annotated with a name ni in set N or annotate as a null.

Conditional random field is used to model the graph with face and label sequences for the purpose of name inference. The graph is denoted as G = (V, E) having vertices V= {F, X} denotes set of faces and label sequences and edges E shows the relationship exist between faces and between faces and names. Example graph is shown in Fig -3 having 11 faces in detected sequence and there is index variable, yi , representing label for each and every faces [1]. Basically here the problem is to take account of all possible label assignments and then pick out the best one assignment as the solution that maximizing the assignments probability.

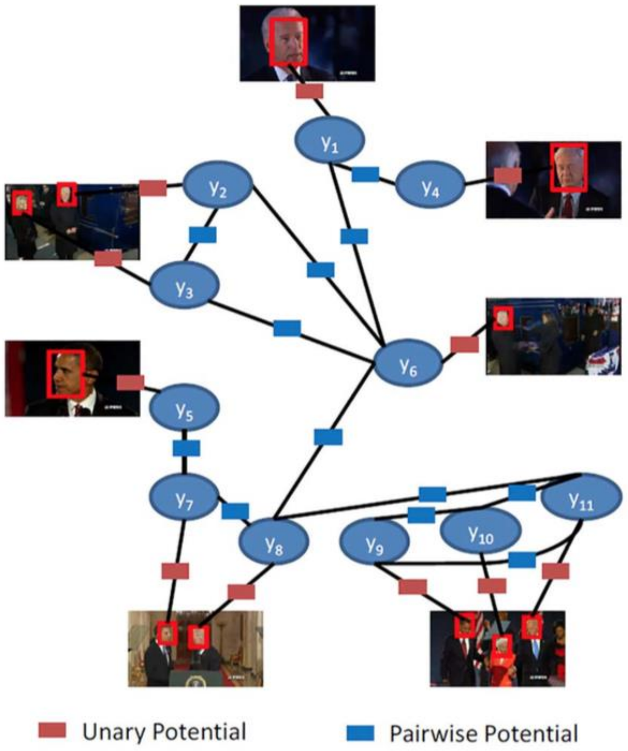


Fig -3 Example of graph describing modeling of relationships for label inference.

In graph modeling example, according to CRF properties if there is no edge i.e. relationship between two indexed variables then these two variables are considered as an independent one. Therefore, variable y1 and y2 are independent of each other where y1 and y4 dependent variable.

In view of modelling the relationship consider two types of potentials named as unary and pairwise potential. These two potentials models face-to-name and face-to-face relationship respectively.

- **Unary Potential:**
  The unary potential identify the possibility of a face being labelled with a particular name or a null. This potential create the edge between face and a label. To do this, perform modelling of each and every name as Multivariate Gaussian Distribution of faces. Uncertainty in labeling means assigning null category to a face as a label due to the uncertainty of assigning the face to available label. If the probability distribution for labeling a face with particular name is uniformly distributed then uncertainty is higher, otherwise when the probability of labelling face with a name is very high then uncertainty becomes lower.

- **Pair-Wise Potential:**
  Pairwise potential identify the relationship between two faces. Pairwise potential consist of three relationship, namely spatial, temporal and visual relationship which describes the pairwise potential. The pairwise represents noticeable relationship between two faces. In spatial relationship, to find out the relationship between two faces some approaches providing clues such as two frames of different shots, the spatial locations of faces, as well as their overlapping area. Wherein temporal relationship, appearance of face at different instant provides clues whether the names assigned to the faces should be same or unique. Label inference based on dissimilarity between faces is wrong due to conditions like lighting changes as well as changes in viewpoint. Furthermore two mostly similar faces is not even so a necessary clue as an evidence for name identity. Based upon these conditions visual relationship models Background changes and color differences.

5. CONCLUSIONS

This paper reviews the method which is based upon rich relationships rather than rich texts for face name association. CRF does the smooth encoding of Face to Face and Face to Name relationships, allowing null assignment which considers the uncertainty within labelling to deal with the incomplete and noisy metadata. Furthermore, it addresses the errors present in metadata, in order to correct false names and clarify faces with missing names in the metadata of a video, by using socially related web videos. Between-video relationships helps in boosting the performance, which is having the capability of correcting the errors due to missing names and persons. It will serve as preprocessing step for video indexing which gives better video retrieval performance.

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] S. Satoh, Y. Nakamura, and T. Kanade, in "Name-It: Naming and detecting faces in news videos", IEEE Multimedia, vol. 6, no. 1, pp.2235, Jan.Mar. 1999.
[3] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, in "Automatic face naming with caption-based supervision", Proc. IEEE Comput. Vis. Pattern Recog., Jun. 2008, pp. 18.
[4] 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.
[5] J. K. Zhu, S. C. H. Hoi, and M. R. Lyu, "Face annotation using transductive kernel sher discriminant", in IEEE Trans. Multimedia, vol. 10, no. 1, pp. 8696, Jan. 2008.
[6] 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. 3140.
[7] 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. 17.
[8] 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.166179, Jan.2014.
[9] J. Bu et al., "Unsupervised face-name association via commute distance", in Proc. ACM Int. Conf. Multimedia, 2012, pp.219-228.
[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification", in Proc. IEEE Comput. Vis. Pattern Recog., Jun. 2014, pp. 1701-1708.
[11] 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. 550563, Mar. 2014.
[12] 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. 443452.